Система розпізнавання та перекладу текстів для мобільних пристроїв

Дипломну роботу виконав: студент гр. КА-34
Баляс Олександр Олегович

Науковий Керівник: к.т.н., доцент Тимощук О.Л.

Актуальність дослідження

- За останні 7-8 років за допомогою систем компанії «Гугл» було оцифровано більше 15 мільйонів книжок
- За останніми дослідженнями середньостатистична компанія в середньому втрачає близько 10 відсотків свого прибутку тільки через ту причину, що хтось із співробітників при ручному введенні документів допускав досить суттєві помилки.
- Іншою шокуючою цифрою є те, що 40 відсотків робочого часу витрачається лише на пошук оригіналу документу.

Мета, об’єкт, предмет

- Мета: розробити програмний продукт, за допомогою якого можна посимвольно розпізнавати текст з декількох різних мов та перекладати отриманий текст на більш зрозумілу мову для користувача.

- Об’єкт: цифрові знімки з наявним текстом

- Предмет: нейронні мережі, розпізнавання символів, переклад тексту.

Поточний стан технологій оптичного розпізнавання тексту

- Точне розпізнавання латинських символів у друкованому тексті зараз можливе тільки, якщо доступні чіткі зображення, такі як друковані документи. Точність при такій постановці задачі перевищує 99%, абсолютна точність може бути досягнута тільки шляхом наступного редагування людиною
- На зображеннях із рукописним «друкованим» текстом без артефактів може бути досягнута точність у 80% — 90%, але з такою точністю зображення буде перетворене в текст із десятками помилок на сторінці

Загальний алгоритм оптичного розпізнавання

- 1. Оцифровування певного тексту
- 2. Сегментація
- 3. Вирівнювання
- 4. Очистка від зайвих елементів
- 5. Бінаризація
- 6. Виокремлення стовпців та рядків з текстом
- 7. Приведення до зручного масштабу
- 8. «Зіставлення з образом» чи «Адаптивне розпізнавання»

Алгоритм перекладу тексту

- 1. Визначення мови вхідного тексту
- 2. Визначення змісту кожного слова із вхідного тексту
- 3. Визначення мови перекладу
- 4. Зіставлення із словник іншої мови смислового значення слова
- 5. Підібрати відповідник до слова, які сходяться за смисловим значенням

Сутність методу «Адаптивного розпізнавання»

- «Адаптивне розпізнавання» - метод, який поєднує одночасно універсальність і технологічність безшрифтового підходу і високу точність розпізнавання шрифтового.
- Як остаточний критерій можна використовувати точний шрифтозалежний алгоритм, база характеристик якого побудована прямо в процесі роботи ("на льоту") за результатами попередніх кроків розпізнавання.
- Метод, що володіє зазначеним вище властивістю, будемо називати адаптивним розпізнаванням, тому що він використовує динамічне налаштування (адаптацію) на конкретні вхідні символи.

Математична модель «Адаптивного розпізнавання»

- P- якість розпізнавання, отримана на етапі первинного розпізнавання.
- σ- міра спотворення символів, яка дає числове вираження кількості випадкових змін в конфігурації пікселів серед примірників символів, що позначають одну і ту ж літеру алфавіту.
- F- фінальна якість розпізнавання досяжна за допомогою шрифтозалежного алгоритму, який адаптований до даної вибірки символів.
- V- надійність розпізнавання символу;
- V = f(x,P), де x-відстань від даного символу до центру кластеру (ідеального символу).

Математична модель «Адаптивного розпізнавання»

- Основне положення моделі полягає в тому, що відстань від символу, який прийшов на розпізнавання до кластера є нормально розподіленою випадковою величиною з щільністю ймовірності.

- Тоді по заданій мінімальній допустимій надійності Vmin обчислимо максимальну відстань Xm на яку символ може відхилитися від кластера і при якому V >Vmin
- $Xm = f^{-1}(Vmin, P)$

- Далі за визначенням функції розподілу отримуємо наступне:
- $F = \int_0^{Xm} p(x)dx$
- Ця рівність дає відповідь на питання якою буде якість розпізнавання при заданих надійності і мірі спотворення символів.

Математична модель «Адаптивного розпізнавання

- Великий практичний інтерес представляє вимірювання величини σ - середньоквадратичного відхилення; тому що вона надає числове вираження важливого поняття - "якість тексту". У цій моделі σ знаходить конкретний фізичний зміст - описує варіації які виникають в конфігурації пікселів, що описують оригінал символу, в процесах друку і сканування.

Математична модель «Адаптивного розпізнавання

Основні аспекти розпізнавання символів, в яких залучена шкала, заснована на засобах розсіювання:

- Верифікація результатів кластеризації. Мається на увазі, що кластер з розсіюванням істотно відмінним від середнього по вибірці повинен викликати підозру і бути кандидатом на додаткову перевірку.
- Динамічне налаштування різних порогових констант, що керують розпізнаванням.
- Екстремальні значення σ можуть вказувати на ситуацію в якій сама адаптація до даної вибірці є не вигідною бо необхідна статистична інформація в ній відсутня.
- Автоматична селекція документів для подальшої обробки.

Математична модель «Адаптивного розпізнавання»

- Візьмемо довільну комірку кластера. Нехай p - ймовірність появи тут чорного пікселя при черговому додаванні символу в цей кластер. Очевидно, що ця ймовірність фіксована самою моделлю і залежить тільки від положення комірки всередині сітки. Таким чином процес появи чорних пікселів в даній комірці здійснюється схемою випробувань Бернуллі. У процесі фізичної реалізації попадання символів в кластер в цьому осередку існує ξ - частота потрапляння сюди чорного пікселя. Це випадкова величина, зосереджена близько p і згідно центральною граничною теоремою відхиляється від неї згідно нормального закону розподілу.

Математична модель «Адаптивного розпізнавання»

- Отже,

$$-x_\alpha \le \sqrt{\frac{N}{p(1-p)}}(\xi - p) \le x_\alpha$$

- де
  X α - квантиль рівня α
  α - рівень значимості
  N – кількість символів в кластері

Схема роботи «Адаптивного розпізнавання»

- 1. Кластеризація – розбиття символів на певні групи (за змістом, за типом, за розміром і т.д.)
- 2. Розпізнавання – відновлення цифрового вигляду символів з вхідного тексту за допомогою словників еталонних символів.
- 3. Дорозпізнавання – перевірка правильності попереднього етапу та залучення при цьому певних баз правил

Демонстрація роботи

Основне вікно програми

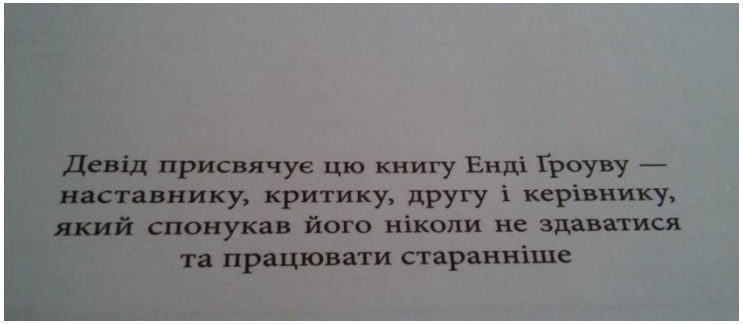

Фото тексту для розпізнавання

Результат роботи програми

Демонстрація роботи

Вибір мови перекладу

Стан програми після здійснення перекладу тексту

Результати роботи. Якість розпізнавання невеликих шматків тексту.

№ експерименту	К-ть символів	К-ть помилок	% розпізнавання
1	107	1	99,07
2	94	6	93,62
3	126	4	96,83
4	119	3	97,48
5	97	5	94,85
6	81	0	100
7	106	5	95,28
8	79	2	97,47
9	110	13	88,18
10	121	5	95,87
11	102	7	93,14
12	95	2	97,89
13	99	6	93,94
14	108	10	90,74
15	123	6	95,12
16	105	8	92,38
17	96	2	97,92
18	105	3	97,14
19	103	7	93,2
20	121	2	98,35

Результати роботи. Якість розпізнавання великих частин тексту.

№ експерименту	К-ть символів	К-ть помилок	% розпізнавання
1	894	12	98,66
2	621	18	97,1
3	1023	40	96,09
4	883	81	90,83
5	901	52	94,23
6	851	38	95,54
7	1015	119	88,28
8	699	97	86,12
9	890	53	94,04
10	1045	165	84,21
11	1207	188	84,42
12	679	51	92,49
13	784	160	79,59
14	825	165	80
15	894	66	92,62
16	901	315	65,04
17	968	182	81,2
18	805	61	92,42
19	710	89	87,46
20	1021	195	80,9

Висновки

- Було розроблено програмний продукт для платформи операційної системи Android, який дозволяє користувачеві розпізнавати посимвольно текст з п’яти різних мов за допомогою бібліотеки Tesseract та з можливістю подальшого перекладу за допомогою безкоштовної бібліотеки Transltr.
- Для фотографій, які містили в близько 100 символів середній результат розпізнавання становив 95 %, при чому, завжди мінімальний показник досягав значення більшого за 85 відсотків.
- У випадку роботи з великими шматками тексту цей показник суттєво впав та робота програми значно вповільнилася. Середній показник на цьому етапі склав 88 відсотків. І у більшості випадків система розпізнає більше, ніж 80% отриманого тексту.
- Подальшим дослідженням для цієї роботи могли б стати впровадження словарних та контекстних розпізнавань, які б покращили якість розпізнавання та майже б нівелювали ймовірність неточності одного-двох символів у слові.

Дякую за увагу!