Розробка комп’ютерної системи зберігання та обробки юридичних документів з розпізнаванням тексту.

Дипломна робота
студент групи 1кс-15сп
Поуданєн Юрій Євгенович

Вінниця 2016

Актуальність

1.Сучасні документи тяжко оформляти за допомогою мобільних пристроїв, тому потрібно спрощувати їх оформлення за допомогою оцифрування зображення.

2.Операційна система Android досить поширена і це є першим важливим критерієм розробки саме під цю платформу.

3.Системи розпізнавання символів дуже стрімко розвиваються і впроваджуються.

4.Проаналізувавши аналоги програмного забезпечення, було виявлено що в аналогах не міститься системи розпізнавання.

Мета та задачі

- Метою дипломної роботи є розробка комп’ютерної системи зберігання та обробки юридичних документів з розпізнаванням тексту для полегшення та підвищення ефективності в роботі та обробці документів за допомогою сучасних технологій.

В дипломній роботі мають бути виконані такі задачі:

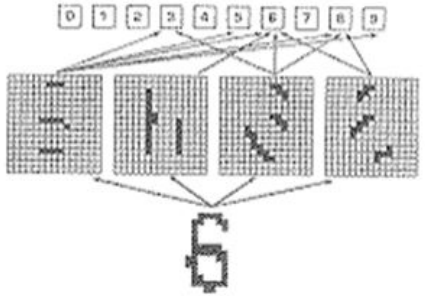
- провести аналіз систем аналогів даній дипломній роботі;
- обрати метод для розпізнавання тексту на зображенні та використати його в дипломній роботі;
- обрати хмарне сховище яке задовольняє усім потребам розроблюваної системи;
- розробити архітектуру програмного додатку для платформи з можливістю розширення та доповнення його в майбутньому при покращенні та внесенні нових функцій;
- розробити програмний додаток згідного технічного завдання та за допомогою останніх сучасних засобів розробки програмного забезпечення;
- розробити можливість попередньої обробки зображення для кращого розпізнавання системою;
- провести аналіз показників якості зображення до та після попередньої обробки;

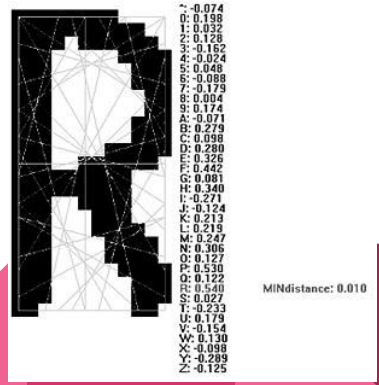
Методи виділення тексту

1.Шаблонний метод (на основі порівняння з готовим зразком).

2.Структурний метод (порівняння на основі топології)

3.Ознаковий метод (на основі порівняння з еталонним набором ознак)





Бібліотека OCR Tesseract

Для дипломної роботи був обраний форк бібліотеки OCR Tesseract під мобільні пристрої.

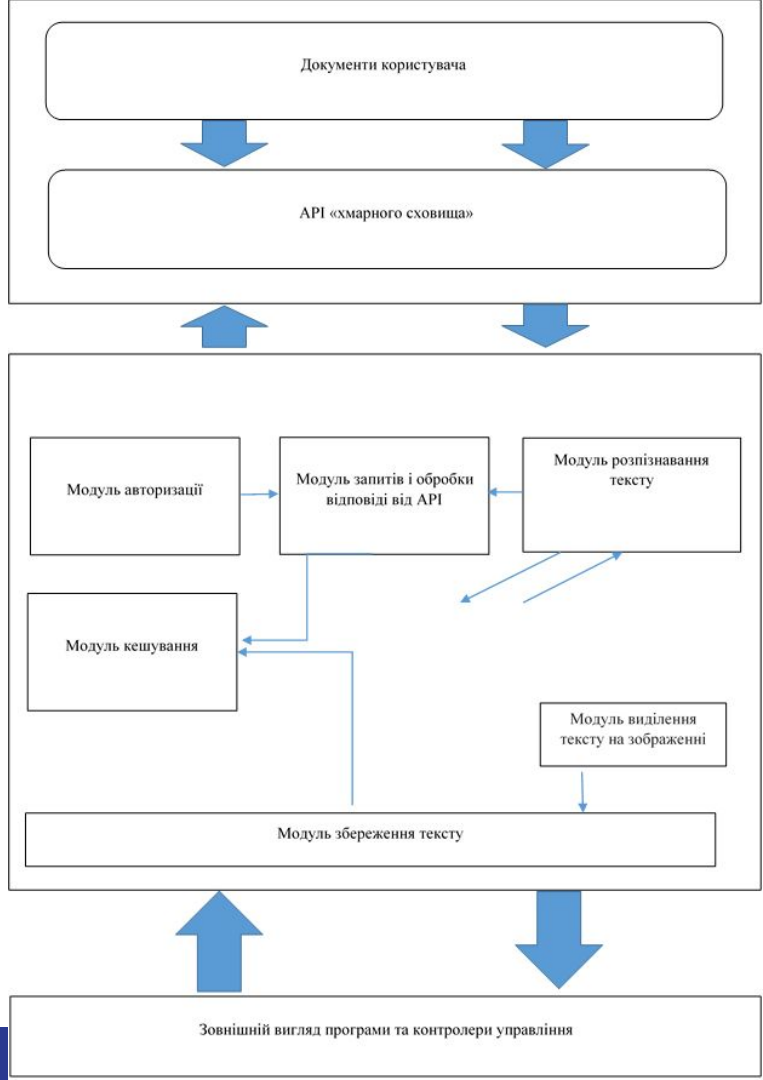
Бібліотека компілюється і використовується як окремий модуль, який потрібно приєднати до основного проекту.

Бібліотека виділяє текст на основі шаблонного методу.

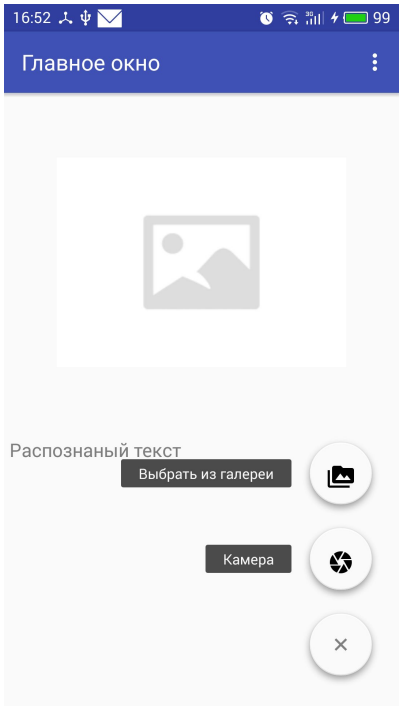
Бібліотека дуже гнучка та зручна в налаштуванні та має великий потенціал.

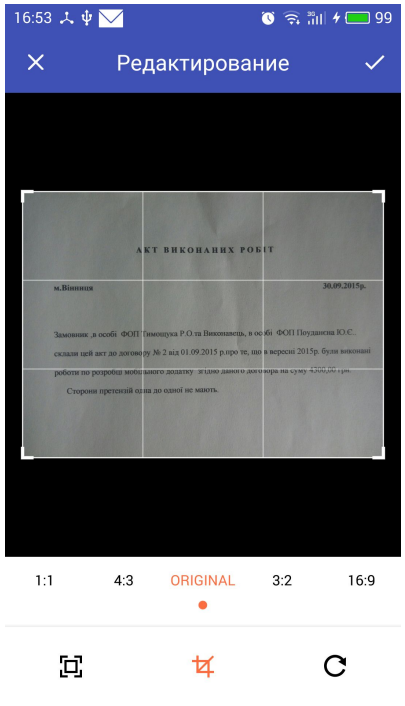
Структура системи

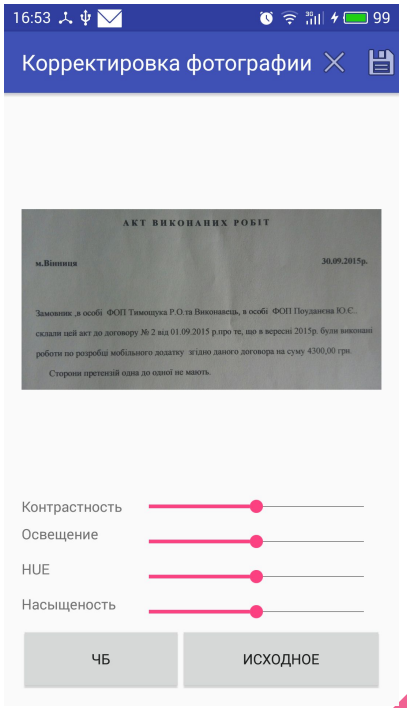
•Структура проста, але завдяки простій архітектурі програми її можна без зайвих зусиль доповнити новим функціоналом та бібліотеками розширення.

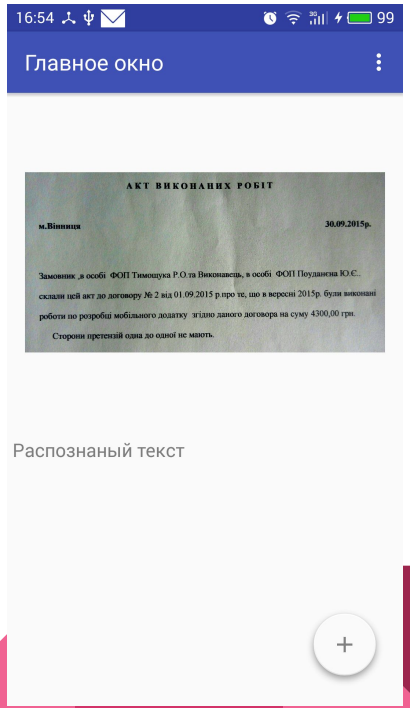


Інтерфейс та функції обробки зображення







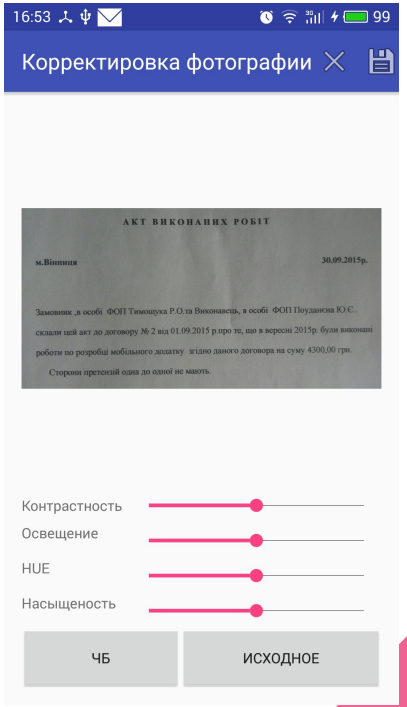


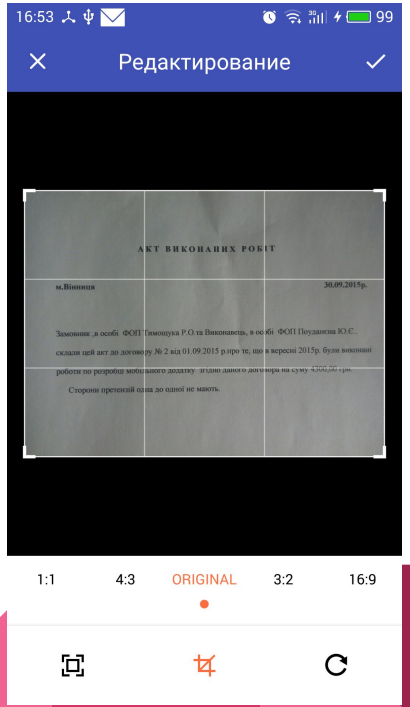
Підготовка зображення перед розпізнаванням

Перша стадія обробки ображення - це редагування її положення та обрізки країв.

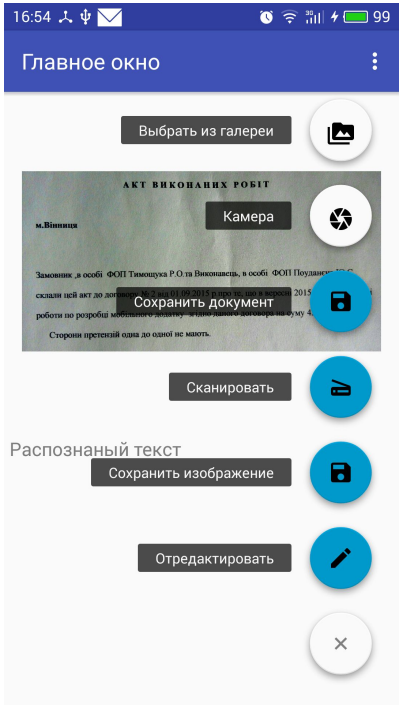
Друга стадія - це редагування за чотирма параметрами:

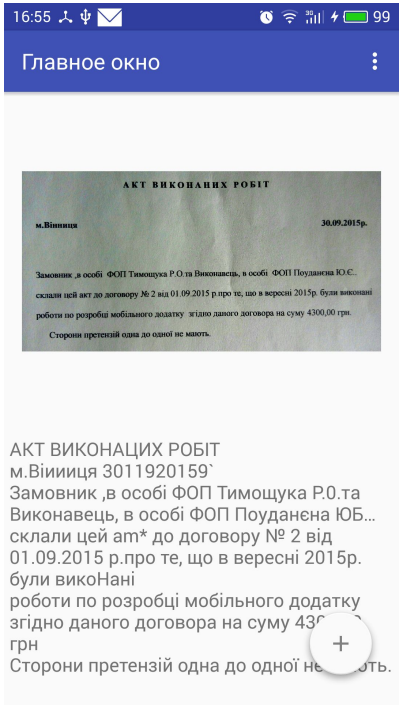
1. Контрастність
2. Посилення відтінку
3. Освітлення
4. Насиченість

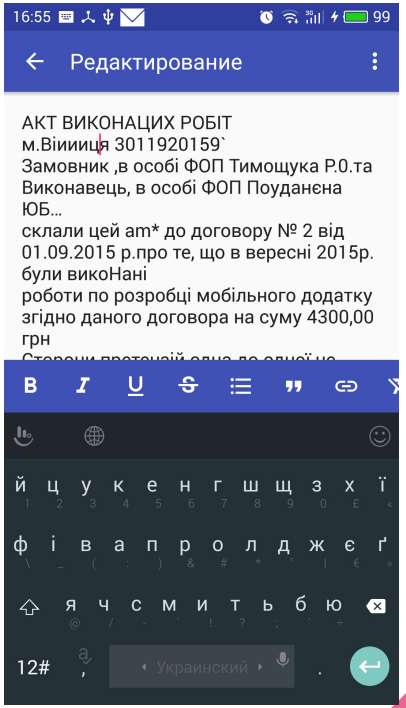


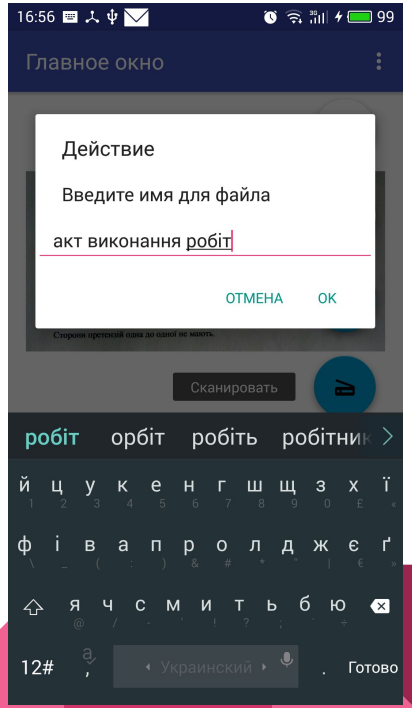


Редагування та збереження текстового файлу









Формули розрахунку показників якості

Точність :

$T = 100 – (X + P) * 100 / W$

де ,

- X – пропущені слова;
- P – нерозпізнані слова;
- W – загальна кількість слів
- T – точність;

Якість :

$S = R * 100 / W$

де ,

- R– розпізнані слова;
- W – загальна кількість слів;
- S – якість.

Достовірність :

$D = 100 – N * 100 / R$

де ,

- R – розпізнані слова;
- N – нерозпізнані слова;
- Д – достовірність.

Аналіз якості без попередньої підготовки зображення

Точність	Якість	Достовірність	Пропущені слова	Нерозпізнані слова	Розпізнані слова
59,71	59,71	75,90	72	40	166
60,07	60,07	54,49	35	76	167
64,75	64,75	87,78	76	22	180
72,30	72,30	80,10	37	40	201
63,31	63,31	59,09	30	72	176
53,96	53,96	56,67	63	65	150

Аналіз якості з попередньою підготовкою зображення

Точність	Якість	Достовірність	Пропущені слова	Нерозпізнані слова	Розпізнані слова
75,54	75,54	89,52	46	22	210
79,14	79,14	89,09	34	24	220
68,35	68,35	92,11	73	15	190
71,58	71,58	84,92	49	30	199
71,22	71,22	86,36	53	27	198
77,70	77,70	80,09	19	43	216

Висновки

- Отже, в ході роботи над дипломною роботою, була створення мінімальна система зберігання документів з можливістю розпізнавання та редагування тексту для подальшого його збереження.
- В ході роботи були досліджені методи покращення
- Описані методи виділення тексту на зображенні.
- Створена архітектура проекту з можливістю її розширення
- Програма розроблена для популярної мобільної платформи Android на мові програмування Java, а також була використана широко відома бібліотека розпізнавання OCR Tesseract.

Дякую за увагу