

Лабораторна робота №4

Тема: Призначення й основні можливості програми сканування та оптичного розпізнавання текстів FineReader.

Мета роботи: вивчення основних можливостей роботи програми FineReader 6.0

Стислі теоретичні відомості

Системи оптичного розпізнавання символів (Optical Character Recognition - OCR) призначені для автоматичного введення друкованих документів у комп'ютер.

FineReader - омніфонтна система оптичного розпізнавання текстів. Це означає, що вона дозволяє розпізнавати тексти, набрані практично будь-якими шрифтами, без попереднього навчання. Особливістю програми FineReader є висока точність розпізнавання і мала чутливість до дефектів друку, що досягається завдяки застосуванню технології "цілісного цілеспрямованого адаптивного розпізнавання".

Процес уведення документа в комп'ютер можна розділити на два етапи:

1. **Сканування.** На першому етапі сканер відіграє роль "ока" Вашого комп'ютера: "переглядає" зображення і передає його комп'ютеру. При цьому отримане зображення є не чим іншим, як набором чорних, білих чи кольорових крапок, картинкою, що неможливо відредагувати в жодному текстовому редакторі.

2. **Розпізнавання.** Обробка зображення OCR-системою.

Обробка зображення системою FineReader містить у собі аналіз графічного зображення, переданого сканером, і розпізнавання кожного символу. Процеси аналізу макета сторінки (визначення областей розпізнавання, таблиць, картинок, виділення в тексті рядків і окремих символів) і розпізнавання зображення тісно зв'язані між собою: алгоритм пошуку блоків використовує інформацію про розпізнаний текст для більш точного аналізу сторінки.

Як уже згадувалося, розпізнавання зображення здійснюється на основі технології "цілісного цілеспрямованого адаптивного розпізнавання".

- **Цілісність** - об'єкт описується як ціле за допомогою значимих елементів і відносин між ними.
- **Цілеспрямованість** - розпізнавання будується як процес висування і цілеспрямованої перевірки гіпотез.
- **Адаптивність** - здатність OCR-системи до самонавчання.

Відповідно цим трьом принципам система спочатку висуває гіпотезу про об'єкт розпізнавання (символі, частині символу чи декількох склеєних символах), а потім підтверджує чи спростовує її, намагаючись послідовно знайти всі структурні елементи і пов'язуючи їх відносини. У кожному структурному елементі виділяються частини, значимі для людського сприйняття: відрізки, дуги, кільця і крапки. Наслідуючи принцип адаптивності,

програма самостійно "настроюється", використовуючи позитивний досвід, отриманий на перших упевнено розпізнаних символах. Цілеспрямований пошук і облік контексту дозволяють розпізнавати розірвані і перекручені зображення, роблячи систему стійкою до можливих дефектів листа.

У результаті роботи у вікні FineReader з'явиться розпізнаний текст, який можна відредагувати і зберегти в найбільш зручному форматі.

FineReader 6.0 окрім 177 звичайних мов, розуміє також основні мови програмування і прості хімічні формули, причому вміє розпізнавати різномовний текст. Має функцію навчання. Успішно інтегрується з Microsoft Office (не тільки з Word, але й з Excel). Крім того, відсканований файл можна відразу відправити електронним листом або завантажити в браузер у вигляді веб-сторінки. FineReader дозволяє відкривати і розпізнавати PDF-файли. PDF - один з найбільш популярних форматів збереження документів в Internet, в архівах і т.д. Відкривши PDF-файл у FineReader, Ви можете розпізнати його, відредагувати і зберегти або в PDF, або в будь-якому іншому підтримуваному форматі збереження.

Головне вікно програми має наступний вигляд (рис.4.1).

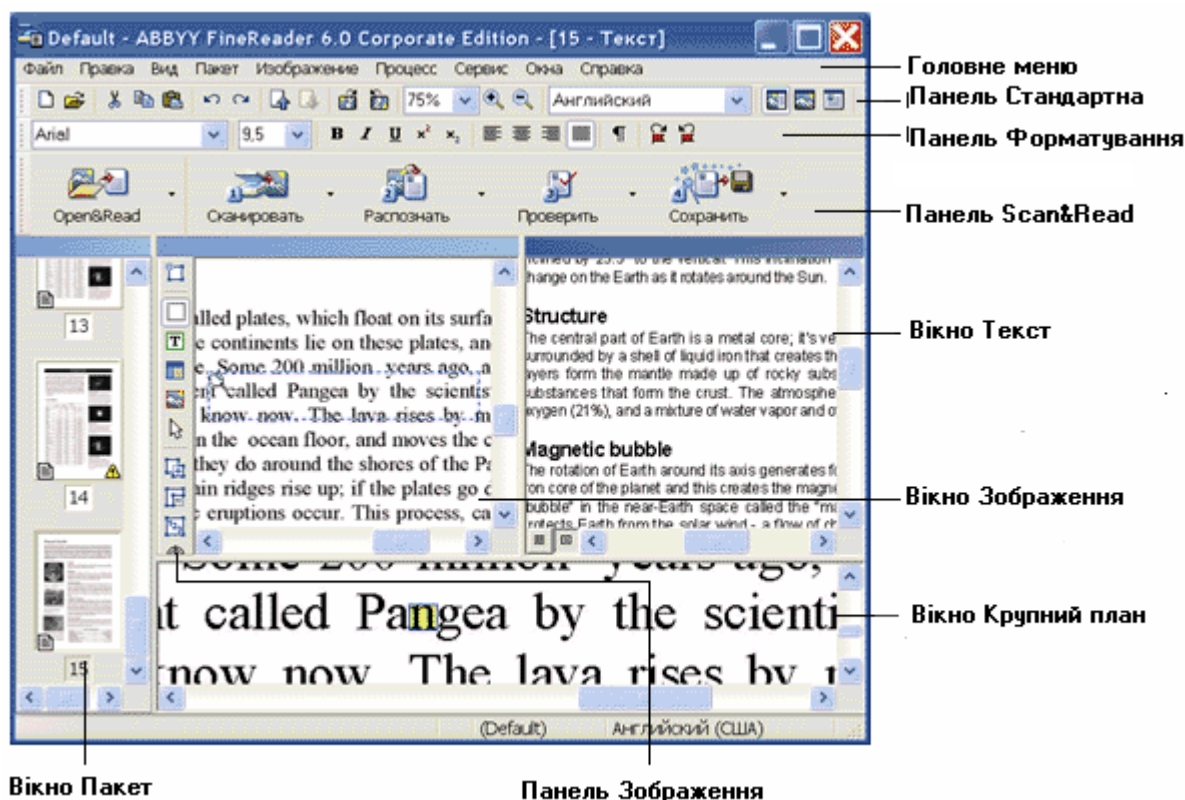


Рис.4.1. Головне вікно програми FineReader 6.0

Угорі Головного вікна FineReader знаходиться меню системи, під ним - інструментальні панелі. У програмі їх чотири: **Стандартна**, **Форматування**, **Зображення** і **Scan&Read**.

Відсканований документ існує в програмі в трьох формах:

- у вигляді значка або ескізу – на лівій панелі (вона називається **Пакет**);

- у вигляді зображення – на середній панелі (**Зображення**);
- у вигляді розпізнаного тексту – справа (панель **Текст**).

На самій нижній горизонтальній панелі міститься збільшене зображення того участка тексту, який переглядається (панель **Крупний план**).

Вікна **Зображення**, **Крупний план** і **Текст** пов'язані між собою: при подвійному кліку на зображенні у вікні **Зображення** курсор у вікнах **Крупний план** і **Текст** (при наявності розпізнаного тексту) переміститься на ту ж позицію, що й у вікні **Зображення**.

Основна робота програми FineReader 6.0 ведеться в пакетному режимі, оскільки кожне відскановане зображення записується як окрема сторінка пакету. У пакеті зберігаються як вихідні зображення, так і відповідний їм розпізнаний текст. Більшість установок FineReader зберігаються на пакет (опції сканування, розпізнавання, збереження, а також створені в процесі роботи користувальницькі еталони, мови і групи мов).

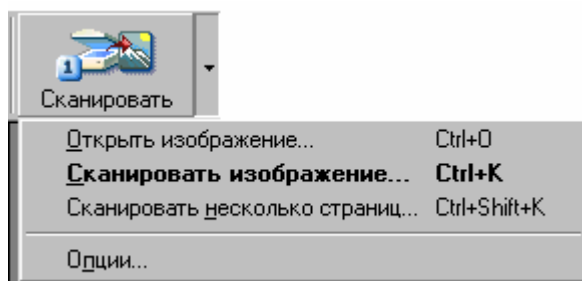
Процес обробки документу програмою FineReader 6.0 складається з наступних етапів:

1. Сканування документу.
2. Розпізнавання документу.
3. Перевірка орфографії.
4. Збереження документу.

Сканування. Кожна модель сканера має свою програму, свої настройки і свої можливості. Але всі такі програми роблять швидке попереднє сканування (**Preview**), після якого можна:

- виділити мишею область сканування;
- обрати режим сканування – кольоровий файл, чорно-білий, з відтінками сірого і т.д.;
- виставити параметри яскравості, контрасту та ін. або обрати автоматичне визначення цих параметрів;
- запустити основне сканування (**Scan**).

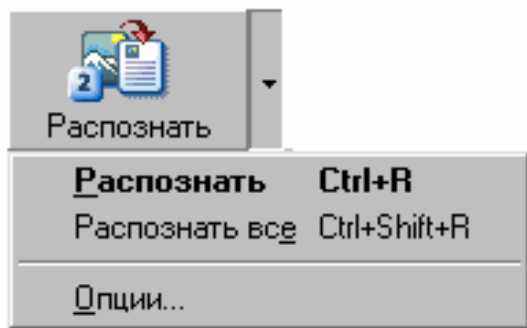
Щоб запустити процес сканування слід натиснути кнопку „Сканировать”



Якщо необхідно відсканувати декілька сторінок, слід в локальному меню обрати пункт „Сканировать несколько страниц”. В результаті

сканування у Головному вікні програми FineReader з'явиться вікно **Зображення** з "фотографією" вставленого листа.

Розпізнавання. Перед початком розпізнавання необхідно перевірити опцію „**язик розпознавання**” а потім натиснути кнопку „**Распознать**”:



Якщо на попередньому етапі було відскановано декілька сторінок, в локальному меню слід обрати пункт „**Распознать все**”.

FineReader не просто розпізнає текст, але й відтворює розмір та накреслення шрифту – підкреслений, напівжирний, курсивний та їх поєднання. Відтворюється й оформлення абзаців – вирівнювання, відступи, маркіровані списки. Те, що FineReader не може розпізнати як текст, він вважає малюнком і вставляє в документ у вигляді графічного фрагменту.

Після завершення розпізнавання результат з'являється у вікні **Текст**. Вікно **Текст** - це вбудований редактор програми FineReader; у ньому можна перевірити результати розпізнавання і відредагувати розпізнаний текст.

Перевірка орфографії. Одна з можливостей текстового редактора FineReader - це вбудована перевірка орфографії.

Система вбудованої перевірки орфографії дозволяє:

- знаходити непевно розпізнані слова (слова, у яких є непевно розпізнані символи);
- знаходити орфографічні помилки (неправильно написані слова);
- додавати невідомі системі FineReader слова в словник для того, щоб вони розпізнавалися впевнено.



Непевно розпізнані символи і слова, яких немає в словнику, виділяються різними кольорами. За замовчуванням для виділення непевно розпізнаних символів використовується блакитний, для несловникових слів - рожевий.

Щоб перевірити результати розпізнавання слід натиснути кнопку „**Проверить**” на панелі інструментів **Scan&Read**, в результаті відкриється діалог „**Перевірка**” (рис.4.2):



Рис.4.2. Диалогове вікно “Перевірка”

У діалозі **Перевірка** три вікна. Верхнє вікно - аналог вікна "Крупний план" програми FineReader, у ньому показане зображення слова з можливою помилкою. Середнє вікно показує саме слово з можливою помилкою, у рядку над цим вікном виводиться назва типу помилки. У нижньому вікні, **Варіанти**, пропонуються варіанти заміни даного слова (якщо такі існують). Для варіантів використовується словник, зазначений у полі **Мова словника**. Можна використовувати будь-як словник із запропонованого списку.

Для більш швидкої перевірки результатів розпізнавання можна скористатися кнопками  чи  для переміщення до наступного чи, відповідно, що попередньо непевно розпізнаного слова.

Також для переміщення по непевно розпізнаних словах можна використовувати гарячі клавіші: F4 (SHIFT F4).

Збереження результату. Для збереження результату слід скористатись кнопкою „**Сохранить**” на якій є випадаючий список (рис.). FineReader може просто зберігати файли або передавати їх безпосередньо в буфер обміну або одну з перерахованих програм.

Команда „**Сохранить текст в файл**” дозволить записати текст на диск в одному з відомих FineReader’у форматів (рис.4.3).

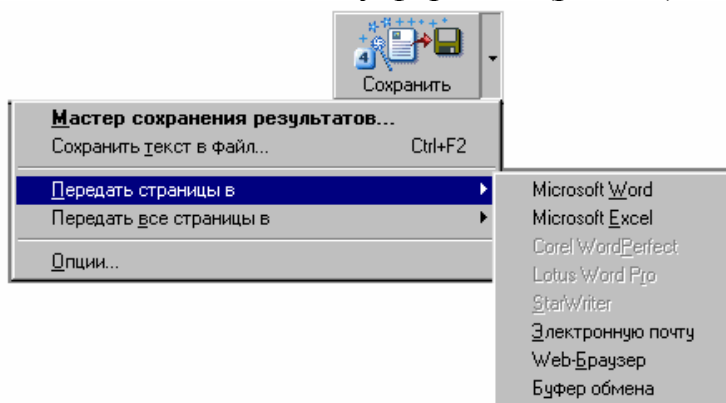


Рис.4.3 Варіанти збереження готового документу.

У стрічці **Опції** можна задати:

- щоб кожна сторінка була записана в окремий файл;
- щоб кожна сторінка була записана в окремий файл, ім'я якого співпадає з іменем вхідного файлу;
- щоб всі сторінки записались в один файл;
- щоб все записалось в один файл, але кожна сторінка – з нового листка.

При прямій передачі файлів у текстовий або табличний редактор ніяких додаткових параметрів вказувати не потрібно.

Зауваження. В процесі розпізнавання FineReader автоматично розмічає скановане зображення блоками трьох видів – текстовими (виділяються рамкою зеленого кольору), табличними (сині) та графічні (червоні). Відповідно програма і відноситься до кожного такого блока: текст розпізнає, картинку не розпізнає – просто вставляє в документ, а в таблиці спочатку шукає стрічки і стовпці, а потім розпізнає – по коміркам.

Іноколи в результаті автоматичного аналізу програма неправильно розмічає сторінку блоками. Тоді необхідно виділяти та редагувати блоки вручну. Границі блоків можна рухати мишею за верхні і бокові сторони та за вузли. Коли нові границі блоку повністю перекривають якийсь зі старих блоків, той за непотрібністю зникає. При значних помилках в розмітці і складній структурі макету сторінки простіше знищити всі блоки клавіатурною комбінацією **Ctrl-Del** і розмітити вручну всю сторінку.

Для цього:

1. встановити курсор миші в кут передбачуваного блоку;
2. натиснути ліву кнопку миші і, не відпускаючи кнопки, потягнути у протилежний по діагоналі кут;
3. відпустити кнопку миші. Виділена частина зображення буде укладена в рамку;
4. привласнити виділеному блоку один з існуючих типів: **Зона розпізнавання, Текст, Таблиця, Картинка** чи **Штрих-код**). Для цього: клацнути на блоці правою кнопкою миші й у локальному меню вибрати **Тип блоку**, а потім - потрібний пункт.

Також для виділення та редагування блоків вручну можливе використання панелі інструментів **Зображення** (рис.4.4).

Панель Зображення містить кнопки, що дозволяють робити аналіз макета сторінки (наприклад, створити і відредагувати блоки), а також кнопки, що дозволяють збільшити/зменшити масштаб зображення, відредагувати зображення (наприклад, стерти непотрібні ділянки зображення, такі, як підписи чи великі ділянки сміття).

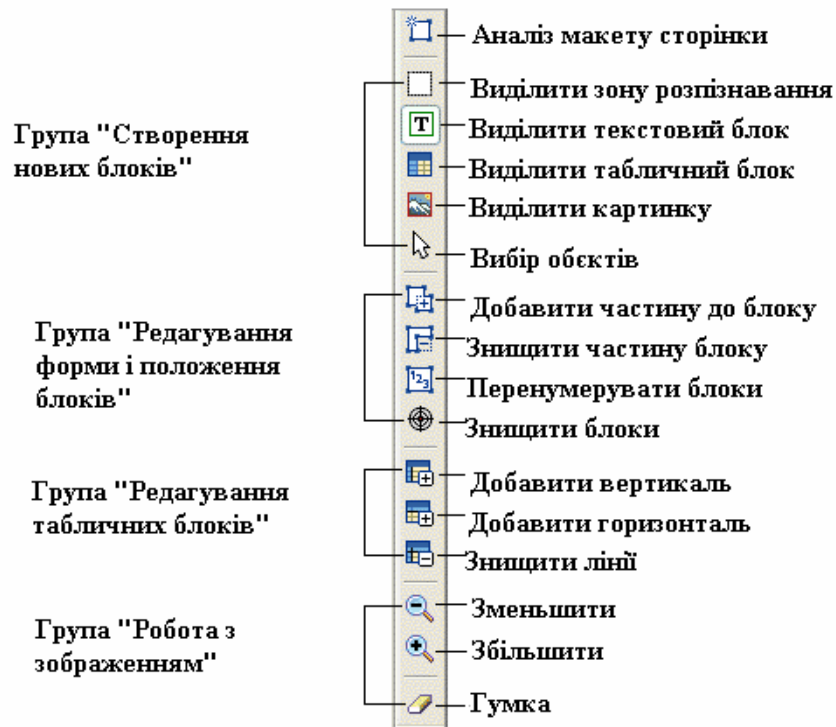


Рис. 4.4. Панель інструментів **Зображення**

Кнопки, що дозволяють створити і відредагувати блоки, можна використовувати не тільки у вікні **Зображення**, але й у вікні **Крупний план**.

Порядок виконання роботи

1. Завантажити програму FineReader. Для цього активізувати Пуск\Програми\Abby FineReader 6.0 Corporate Edition.
2. Відсканувати документ який окрім текстової інформації повинен містити рисунок, таблицю, текст розбитий на декілька колонок.
3. Розпізнати відскановані дані.
4. Перевірити орфографію розпізнаних даних.
5. Зберегти результат у текстовому редакторі Word.

Запитання для самоконтролю

1. В чому полягає процес обробки зображення системою FineReader?
2. Перерахуйте основні елементи Головне вікно програми FineReader 6.0.
3. З яких етапів складається процес обробки документу програмою FineReader 6.0?
4. Які існують типи розпізнаних блоків?
5. Яке призначення панелі інструментів **Зображення**?
6. Які заходи слід застосувати якщо в результаті автоматичного аналізу програма неправильно розмітила сторінку блоками?