Особенности создания поискового индекса к фотографиям в цифровом альбоме.

Андрей Талбонен

Петрозаводский Государственный университет

Особенности проекта

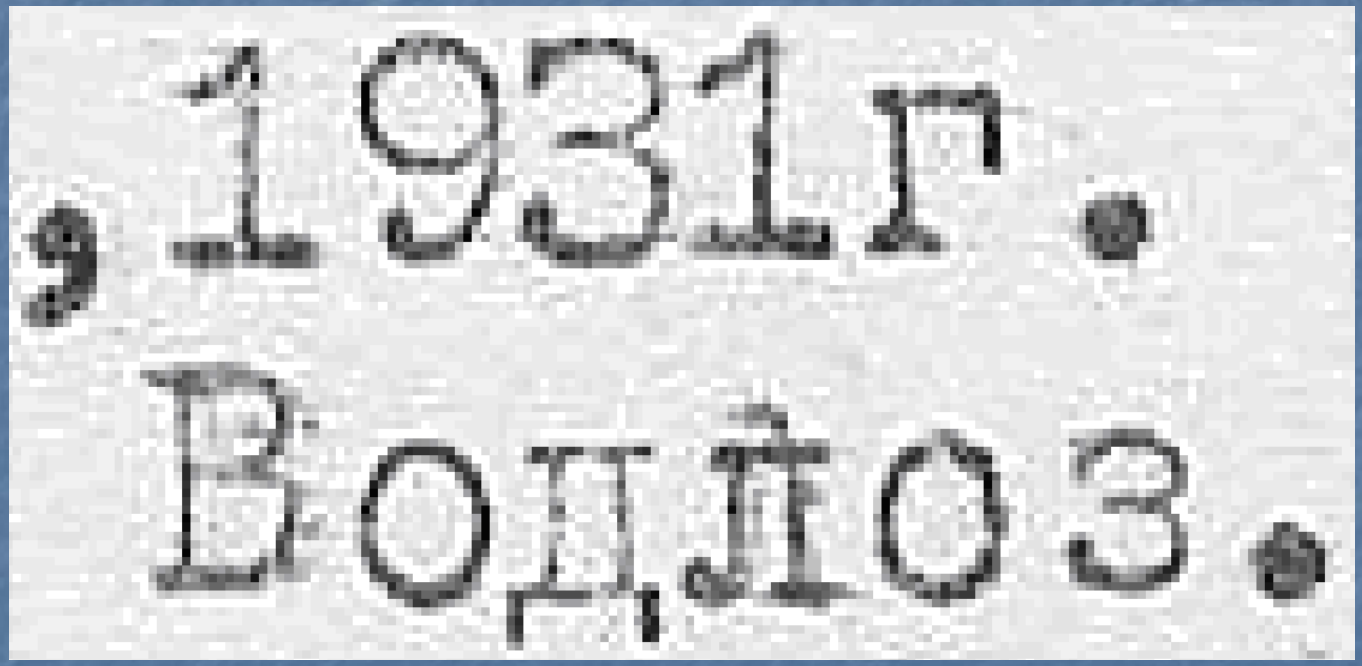
- Низкое качество исходных изображений
- Черно-белые изображения
- Изображения содержат текстовую и графическую информацию

Исходные данные

- Пример изображения коллекции

Исходные данные

- Пример изображения коллекции



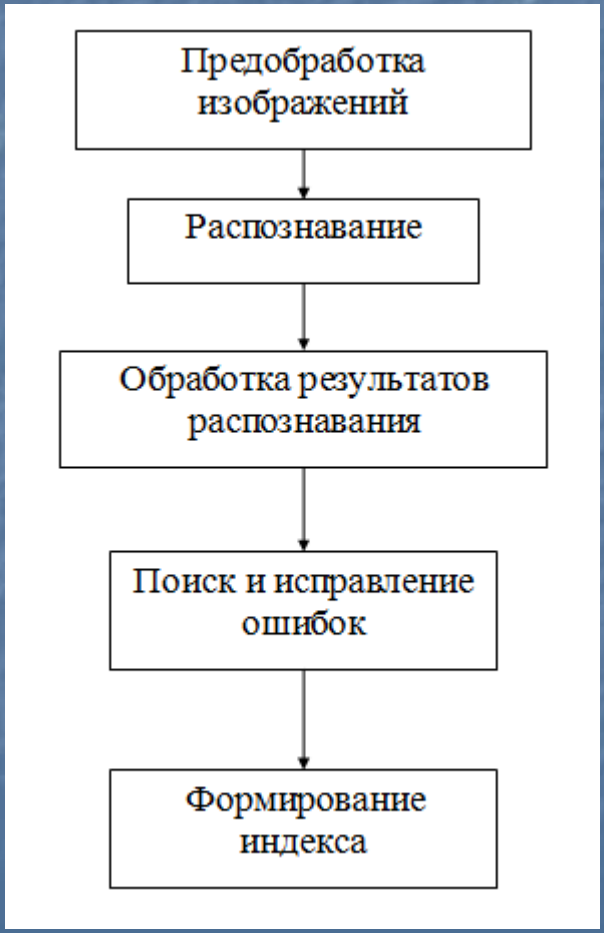
Направления исследований

- Извлечение текста и организация поиска
  - Улучшение качества распознавания текста
  - Формирование тематических словарей и таксономии
  - Повышение полноты и точности поискового запроса
- Поиск графических объектов
  - Поиск лиц
  - Поиск на основе контурных характеристик
  - Поиск на основе текстурных характеристик

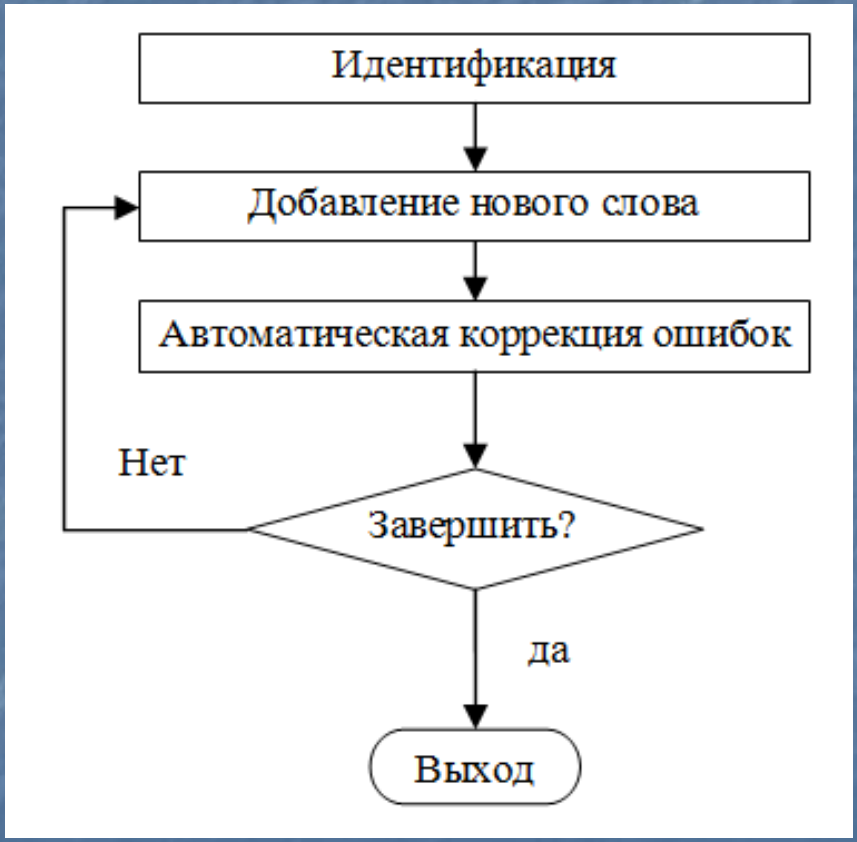
Формирование поискового индекса

- Сравнение и комбинирование результатов распознавания текста после обработки изображений несколькими методами
- Обработка текстовой информации на основе словарей и поиска нечетких слов
- Формирование морфологического словаря за счет морфологических анализаторов и за счет добавленных слов

Формирование индекса



Поиск и исправление ошибок



Результат

- Определен алгоритм формирования текстового индекса
- Удалось повысить качество коллекции с 38% до 66% за счет применения метода сравнения результатов распознавания
- За счет обработки текстовой информации можно повышать качество почти до 100%

Основные проблемы

- Необходимость формирования тематических словарей. Возможные подходы:
  - Формировать за счет добавления слов.
  - Извлекать из онлайновых ресурсов.
- Необходимость определения тематики текста для выполнения автоматической коррекции ошибок
- Необходимость формирования онтологии/тезауруса

Обработка поискового запроса Повышение точности

Основная идея – поиск по лексическому правилу и выполнения преобразований на результатом поиска

Пример

Текст:

шлюз 17

Правило поиска:

const “шлюз”, simple number

Преобразование:

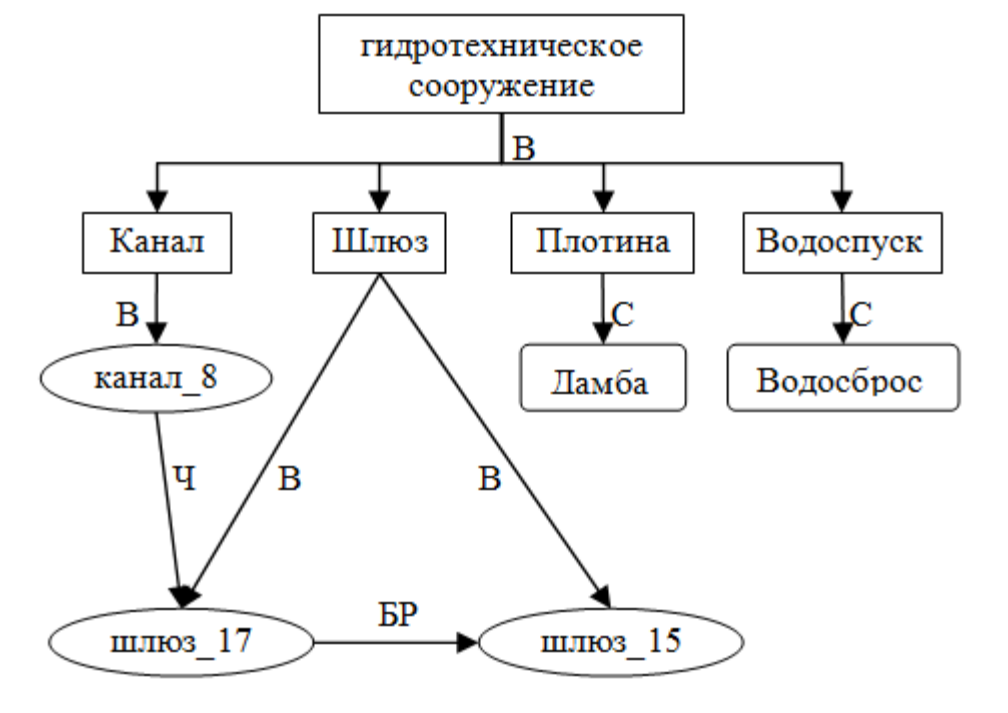
replace from {0}, to {1} with value concat {item {0}, “_”, item {1}}

Результат:

шлюз_17

Обработка поискового запроса Повышение полноты

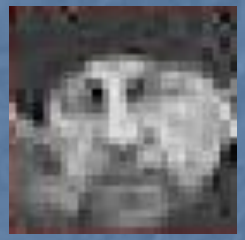
Основная идея – расширение запроса за счет включения узлов онтологии/тезауруса



Поиск объектов на изображениях Текущие исследования:

- Поиск лиц с помощью функции detectMultiScale библиотеки openCV

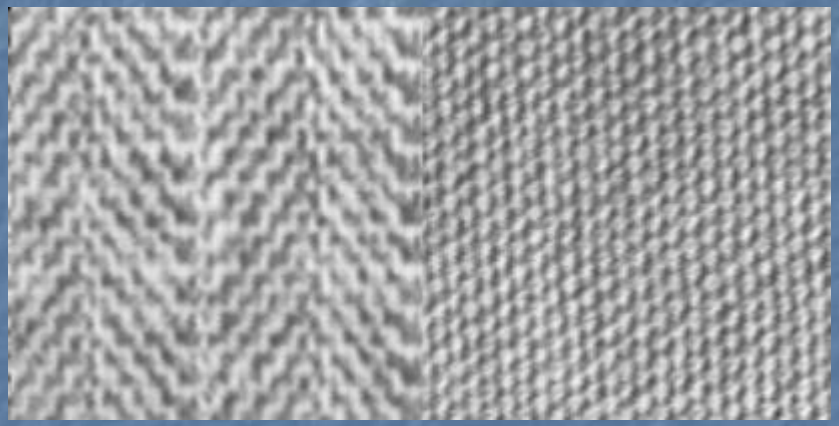
66x66 pix



28x28 pix

Поиск объектов на изображениях Текущие исследования:

- Анализ текстур методом моментов [Tuceryan]



Исходное изображение

Сегментированное изображение

Точность сегментации: 98%

Спасибо за внимание!