Разработка алгоритма детектирования таблиц на изображениях при не фиксированном стиле оформления

Выполнил:

Сафонов Георгий Романович, студент 1 курса Магистратуры Группа 22503

Актуальность

- Цифровизация информации
- Структуризация данных
- Автоматизированное создание отчётной документации

Проблема

Веб-страницы:

- Не стандартное применение тегов языка верстки
- Отсутствие стандартизации названий атрибутов структурных элементов
- Разнообразие стилей оформления страницы
- Динамическая генерация контента на стороне клиента

Документы (отчёты, статьи, презентации):

- Отсутствие исходного кода документа (т.е. имеется только итоговая визуализация)

Проблема

Описание

Мужские кроссовки из натуральной замши - это совершенно другие ощущения и абсолютно другой уровень комфорта. Такая обувь максимально удобно садится на ногу и при этом сочетает в себе еще и стиль! Модель представлена в трех цветах: выбирайте хаки, коричневые или синие кожаные кроссовки. Удобная подошва, крепкая шнуровка, современный дизайн со вставками из текстиля. PATROL - это прекрасная синергия всех параметров в одной модели. Для спорта, прогулок, походов хоть на работу, хоть на неформальную вечеринку. Универсальная обувь для любых ситуаций.

Характеристики	Е₊ Добавить к сравнению		
Материал	Текстиль, Натуральная замша	Российский размер (обуви)	43
Материал стельки	Текстиль	Коллекция	Весна-лето 2022
Материал подошвы	Филон	Материал верха	Замша
Сезон	На любой сезон	Внутренний материал	Текстиль
Пол	Мужской	Вид застёжки	Шнурки
Целевая аудитория	Взрослая	Страна-изготовитель	Китай
Бренд в одежде и обуви	Patrol	Цвет	Коричневый

Рис. 1a: Пример веб-страницы с нестандартизованной вёрсткой (ozon.ru)

Проблема

```
▼ <div class="j2j" data-widget="webCharacteristics">
  ▼ <div id="section-characteristics" class>
    ▼ <div class="sj9"> flex
     ▶ <h2 class="t1i">...</h2>
     <div class="s9i">...</div>
     </div>
     <span class="jj"></span>
    ▼<div class>
     ▼ <div class="tj"> flex
       ▶ <div style="width: calc(50%);">...</div>
       ▶ <div style="width: calc(50%);">...</div>
       </div>
     </div>
    ▶ <small class="tj1">...</small>
   </div>
 </div>
 <div class="ci6" data-widget="separator" style="height: 16px;"></div>
▶ <div class="g1t" data-widget="row">...</div> flex
 <!--->
</div>
```

Рис. 16: Пример HTML-кода веб-страницы с нестандартизованной вёрсткой (ozon.ru)

Предмет

Задача детектирования таблиц на изображениях.

Объект

Автоматизированный сбор информации на основе изображений.

Цель

Разработать наименее ресурсозатратный алгоритм выявления таблиц на изображениях при не фиксированном стиле оформления

Задачи

- 1. Обзор существующих решений
- 2. Сбор данных для анализа (изображений)
- 3. Анализ полученных изображений
- 4. Выявление необходимых метрик и порогов для детектирования таблиц
- 5. Программная реализация алгоритма

Гипотеза

Обнаружение таблиц на веб-странице или документе возможно реализовать на основе информации получаемой из визуального отображения страницы/документа.

Методы

- Анализ (выявление необходимых метрик и пороговых значений путём анализа собранных данных)
- Сравнение (определение отличий в значениях метрик для различных элементов веб-страницы/документа).

База опытно-экспериментальной работы

Петрозаводский государственный университет

Основные этапы исследования

- 1. Сбор данных
- 2. Анализ данных
- 2. Разработка алгоритма
- 3. Программная реализация алгоритма

Практическая значимость

Описанный в работе алгоритм позволяет частично автоматизировать задачу структурирования информации получаемой из заданного документа или веб-страницы.

Описание алгоритма

Алгоритм состоит из следующих этапов:

- 1. Предварительная обработка изображения
- 2. Формирование вектора частот пикселей фона по строкам матрицы изображения.
- 3. Выявление участков с пониженной дисперсией построчных частот и объединение последовательно расположенных элементов
- 4. Фильтрация полученных участков на основе порогового значения дисперсии для суммы "полезных" пикселей по столбцам

Описание алгоритма (Предобработка)

- 1. Преобразование изображения к оттенкам серого (grayscale)
- 2. Определение насыщенности фона изображения по пороговому значению: объявляется переменная равная О, выполняется цикл по каждому пикселю, если его значение больше заданного порога (по умолчанию 150) то к значению переменной добавляется 1, иначе вычитается.
- 3. Бинаризация изображения (инвертированная если значение переменной из шага 2 больше нуля)

Описание алгоритма (Предобработка)

5. Применение операции свёртки со следующими масками:

$$\begin{pmatrix} -1 & 2 & -1 \ -1 & 2 & -1 \ -1 & 2 & -1 \end{pmatrix}$$

$$egin{pmatrix} -1 & -1 & -1 \ 2 & 2 & 2 \ -1 & -1 & -1 \end{pmatrix}$$

Рис. 2а: Маска для удаления горизонтальных прямых

Рис. 26: Маска для удаления вертикальных прямых

Описание алгоритма (Предобработка)

	CNN	ЕМ для 3-х компонент	BIC
Итерация 1	49.88	91.55	169.85
Итерация 2	49.49	86.3	161.37
Итерация 3	47.06	85.15	181.11
Итерация 4	47.52	86.77	198.37
Итерация 5	48.49	86.51	194.03
Среднее время	48.49	87.26	180.95

Как видно из таблицы V разработанная модель машинного обучения позволила ускорить процесс определения вида распределения цен на инстанс.

Рис. За: Исходное изображение

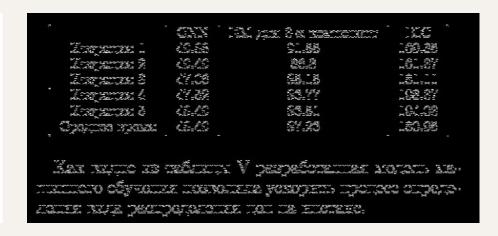


Рис. 36: Предобработанное изображение

Описание алгоритма

Для каждой строки матрицы изображения выполняется расчёт частоты пикселей с 0-м значением, следующим образом:

$$k_i = 1 - rac{(\sum x_{i,j})}{255 \cdot W}$$
 , где $x_{i,j}$ - значение i -го пикселя строки j , W - ширина изображения

$$K = \left(egin{array}{c} \kappa_1 \ dots \ \end{array}
ight)$$
 - вектор построчных частот, где h высота изображения

Описание алгоритма

Далее определяется пороговое значение для определения зон с пониженной дисперсией построчных частот, как несмещённая оценка дисперсии по всем элементам вектора K:

$$t_1=S^2(K)$$

Затем по элементам вектора K осуществляется проход окном, фиксирующим несмещённую оценку дисперсии v_l для фиксированного количества элементов: p, с шагом равным p. Если $0 \leqslant v_l \leqslant t_1$, то глобальный индекс начальной строки в окне записывается в вектор V:

$$V=egin{pmatrix} d_1\ dots\ d_n \end{pmatrix}$$
, где n - количество окон удовлетворяющих условию, d_i - номер начальной строки окна

Описание алгоритма (Фильтрация)

Если индекс конца одного окна отличается от индекса начала следующего на 1, то такие окна объединяются

Для каждого полученного окна определяется следующая матрица:

$$K_{horizontal} = egin{pmatrix} s_{d_1,1} & \cdots & s_{d_1,w} \ dots & \ldots & dots \ s_{d_n,1} & \cdots & s_{d_n,w} \end{pmatrix}$$
, где $s_{d_l,i} = rac{\sum\limits_{j=d_l}^{d_l+p} x_{i,j}}{255}$

По каждой строке определяется несмещённая оценка дисперсии, если полученная оценка больше порогового значения t_2 (по умолчанию $t_2 = 10$), то соответсвующее данной строке окно классифицируется как таблица.

Недостатки алгоритма

- На изображении должны находится текстовые или иные элементы не являющиеся таблицами
- Алгоритм не фиксирует ширину таблицы

Примеры работы

	CNN	ЕМ для 3-х компонент	BIC
Итерация 1	49.88	91.55	169.85
Итерация 2	49.49	86.3	161.37
Итерация 3	47.06	85.15	181.11
Итерация 4	47.52	86.77	198.37
Итерация 5	48.49	86.51	194.03
Среднее время	48.49	87.26	180.95

Как видно из таблицы V разработанная модель машинного обучения позволила ускорить процесс определения вида распределения цен на инстанс.

Рис. 4а: Исходное изображение

	CNN	ЕМ для 3-х компонент	BIC
Итерация 1	49.88	91.55	169.85
Итерация 2	49.49	86.3	161.37
Итерация 3	47.06	85.15	181.11
Итерация 4	47.52	86.77	198.37
Итерация 5	48.49	86.51	194.03
Среднее время	48.49	87.26	180.95

Как видно из таблицы V разработанная модель машинного обучения позволила ускорить процесс определения вида распределения цен на инстанс.

Рис. 46: Изображение с найденной таблицей

Примеры работы

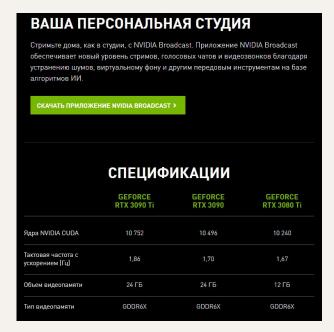


Рис. 5а: Исходное изображение

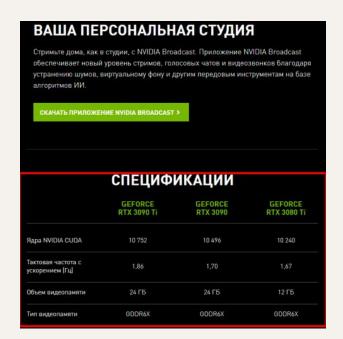


Рис. 56: Изображение с найденной таблицей

Примеры работы

Описание Мужские кроссовки из натуральной замши - это совершенно другие ощущения и абсолютно другой уровень комфорта. Такая обувь максимально удобно садится на ногу и при этом сочетает в себе еще и стиль! Модель представлена в трех цветах: выбирайте хаки, коричневые или синие кожаные кроссовки. Удобная подошва, крепкая шнуровка, современный дизайн со вставками из текстиля. PATROL - это прекрасная синергия всех параметров в одной модели. Для спорта, прогулок, походов хоть на работу, хоть на неформальную вечеринку. Универсальная обувь для любых ситуаций. Характеристики **Е.** Добавить к сравнению Текстиль, Натуральная замша Российский размер (обуви) Материал Материал стельки Текстиль Коллекция Весна-лето 2022 Материал подошвь Филон Материал верха Замша На любой сезон Внутренний материал Мужской Вид застёжки Шнурки Китай Целевая аудитория Взрослая Страна-изготовитель Бренд в одежде и обуви Patrol Коричневый

ногу и при этом сочетает в себе еще и стиль! Модель представлена в трех цветах: выбирайте хаки, коричневые или синие кожаные кроссовки. Удобная подошва. крепкая шнуровка, современный дизайн со вставками из текстиля. PATROL - это прекрасная синергия всех параметров в одной модели. Для спорта, прогулок, походов хоть на работу, хоть на неформальную вечеринку. Универсальная обувь для любых ситуаций. Характеристики 📑 Добавить к сравненик Материал Текстиль, Натуральная замша Российский размер (обуви) Материал стельки Коллекция Весна-лето 2022 Материал подошвы Материал верха Замша Сезон На любой сезон Внутренний материал Текстиль Мужской Вид застёжки Шнурки Китай Целевая аудитория Взрослая Страна-изготовитель Patrol Бренд в одежде и обуви Коричневый

Мужские кроссовки из натуральной замши - это совершенно другие ощущения и абсолютно другой уровень комфорта. Такая обувь максимально удобно садится на

Описание

Рис. 6а: Исходное изображение

Рис. 66: Изображение с найденной таблицей

Список литературы

- 1. Tran, D.N., Tran, T.A., Oh, A., Kim, S.H., Na, I.S.: Table detection from document image using vertical arrangement of text blocks. Int. J. Contents 11(4), 77–85 (2015)
- 2. Nguyen, DD. TableSegNet: a fully convolutional network for table detection and segmentation in document images. IJDAR 25, 1–14 (2022). https://doi.org/10.1007/s10032-021-00390-4

Спасибо за внимание