Mail.Ru на РОМИП 2008. Алгоритм поиска нечетких дубликатов в коллекции изображений

© Ян Кисель

Mail.Ru
kisel@corp.mail.ru

Аннотация

В статье описывается алгоритм поиска нечетких дубликатов в коллекции изображений, используемый в поисковой системе GoGo компании Mail.Ru.

1. Введение

В этом году мы принимали участие в новой для РОМИП дорожке поиска нечетких дубликатов в коллекции изображений. Наши эксперименты в ней базировались на алгоритмах, используемых для фильтрации нечетких копий в поисках по видео и по изображениям поисковой системы GoGo.

2. Вычисление характеристик

Для поиска одинаковых изображений в наборе сперва соберём данные о каждом изображении, предварительно преобразовав его. А именно — изменим размер изображения до стандартного (использовался размер 96x96 пикселей) отступив определённый бордюр (это избавляет от искажений по краям). Далее произведём сглаживание изображения, применив к каждой точке оператор Гаусса — это стандартное действие для уменьшения помех.

Теперь, имея сглаженное изображения получим сеть из средней яркости отдельных цветовых компонент (RGB), разбив изображение на блоки 6x6 пикселей (т.е. 16x16 блоков при обозначенном размере

изображения): это данные для отдельного метода сравнения (фильтра), назовём его colorgrid.

Далее, получим еще одну цветовую характеристику, более простую для будущего сравнения: гистограмму либо вектор цветовой когеренции (CCV, Color Coherence Vector) — взяв лишь старшие разряды для цветовых каналов (использовалось 2 бита). Применительно к ставнению, гистограмма позволяет отфильтровать изображения на этапе до colorgrid, однако эта дискретная характеристика не должна приниматься как опорная — не смотря на использования 75% процентов яркости канала (2 старших бита), это подразумевает стабилизацию яркости изображения, что не всегда прослеживается в реальных примерах.

Теперь, после расчёта цветовых характеристик, возьмём отпечаток изображения в формате ЧБ (для простоты кодирования [01] и соответственно сравнения). Для выполнения этого шага необходима коррекция цвета изображения, т.к. изъятие пороговых яркостей на исходных цветах изображения не даст его контуров; итак, применяем нормализацию изображения — интенсивность каждого цветового канала распределяем по всему диапазону (от отсутствия до максимума), после чего производим выравнивание гистограммы — статистически более важным её участкам отдаём предпочтение, растягивая для них тоновый диапазон. В результате в контрасте теряют менее важные участки гистограммы. Для получения отпечатка изображения вновь уменьшаем изображение (до размера 32x32), и кодируем битовой последовательностью чёрно-белое изображение.

3. Группировка

На этапе сбора информации об изображении также необходимо вычислить групповой признак, в дальнейшем он будет использоваться как разделитель — сравниваться между собой будут лишь изображения (вернее, характеристики) внутри одной группы. Разбиение необходимо т.к. сложность используемых алгоритмов сравнения зависит от кол-ва изображений более чем линейно, в худшем случае O(n2).

Для группировки был применён простой метод — 64x разрядное целое число, вмещающее данные о цветовом тоне отдельных блоков, а также их среднюю интенсивность по цветовым каналам. Следует отметить, что количество блоков должно быть тем меньше, чем большие изменения (небольшие сдвиги, или локальные

перепады контрастности ввиду сжатия) присутствуют в наборе изображений (определяется визуально).

Гораздо более стабильный к случайным искажениям метод — кодирование перепадов яркости и соотношений яркости отдельных блоков между собой; можно утверждать, что излишнее сжатие не повлияет на эту характеристику (однако, данный метод был использован лишь для тестирования).

4. Поиск дубликатов

К изображениям каждой группы последовательно применяем фильтры. Фильтруя изображения, получаем новый набор и так далее. Само-собой, вначале желательно применить быстрые и резкие (разбивающие на множество подгрупп) фильтры, например сравнивая изображения по отпечатку.

При очень большом размере группы применяется метод кластерного объединения — при нахождении первой пары, производится поиск изображений похожих на первое; второй вариант — нахождение пар похожих картинок, путём сравнения «каждый с каждым». Второй метод несмотря на ресурсоёмкость и относительную сложность обладает одним замечательным свойством — он зачастую стягивает группы похожих изображений в большую группу, отмечая одинаковыми даже те изображения которые при сравнении один к одному (т.е. первым методом) не считались бы таковыми. Эта полезная особенность используется для нахождения дубликатов среди изображений, являющихся кадрами видео-роликов: при большом количестве кадров очень отличающегося качества или кадров на которых последовательно запечетлено движение объекта/камеры этот метод «стянет» их в одну группу.

Суть методов сравнения напрямую исходит от сбора их данных, и рассматривать их нет смысла, лишь перечислим:

- colorgrid, т.е. сравнение по цветовой сетке
- сравнение отпечатков изображений
- сравнение гистрограмм/CCV

Все эти методы обладают возможностью досрочного прекращения сравнения, поэтому сами по себе не несут сложности даже при сравнении нескольких десятков тысяч изображений в одной группе.

Также стоит отметить простое исключение — не сравнивать изображения с существенно отличающимися соотношениями сторон, т.к. реальные преобразования[1] изображений не изменяют его.

5. Заключение

Представленный метод ориентирован на сравнение изображений различающихся размерами и качеством или степенью сжатия. Лишь методика сравнения изображений по гистограммам и CCV не жёстко связана с позицией камеры, тогда как методы *colorgrid* (при большой сетке) и особенно метод отпечатков, очень чувствительны к смещениям.

Данный метод показал свою эффективность и в сравнении изображений видеороликов (превью) - методика выделения картинки для превью схожая пости у всех видеохостингов; в общем случае это выделение первого ключевого кадра не ниже определённой «насыщенности» картинки.

На тестовом наборе метод также показал свою стабильность — среди найденных групп достаточно большая точность, тогда как неучтённые дубликаты (визуально) различаются по критериям не учтённым в представленномом методе (либо присекаемым в нём, таким как большие смещания камеры).

---

[1] под реальными подразумеваются преобразования зафиксированные в поисковой системе GoGo при поиске по картинкам и видео, а это в основном пропорциональное сжатие/растяжение и изменение формата (GIF,JPEG,PNG) и степени сжатия изображения