Алгоритм поиска нечетких дубликатов на основе простых признаков.

© Добров Г. Б., Пятков Е. А.
ВМК, МГУ им. Ломоносова
wslc@rambler.ru, gpyatkov@gmail.com

Аннотация

В статье предложен метод поиска изображений по визуальному подобию. Для решения задачи использовались сравнительно простые и быстро вычислимые признаки, основанные на анализе цветовых гистограмм и яркостных переходов на изображении.
Также описан алгоритм кластеризации, представляющий собой модификацию алгоритма FOREL.

1. Введение

Целью работы было создание программы, умеющей довольно быстро находить нечеткие дубликаты в большой коллекции. При этом понятие нечеткого дубликата понималось в довольно широком смысле: два изображения считались дубликатами, если у них совпадал центральный, действующий объект и место действия.

Задача поиска дубликатов в коллекции является, по сути, задачей кластеризации, где кластер – группа похожих картинок. Для большинства алгоритмов кластеризации основной операцией является подсчет расстояния между двумя объектами. В задачах анализа изображений существуют алгоритмы, способные качественно считать меру сходства между объектами. К таковым, в первую очередь, относятся алгоритмы выделения и сопоставления ключевых точек, такие как SIFT [4] и SURF [2]. Основным недостатком этих алгоритмов является их высокая ресурсоемкость, что особенно заметно в большой коллекции изображений.

Именно поэтому основные усилия при создании системы были направлены на исследование возможности использовать простые, быстро извлекаемые из изображений признаки для кластеризации.

2.Выделенные признаки

Выделенные признаки можно разбить на несколько групп.

Первой группой признаков были гистограммы. Для их построения использовались две разные палитры: HSV и RGB. Гистограммы для RGB-палитры строились следующим образом: цветовое пространство разбивалось на 32 равные части и считалось, сколько пикселей изображения попадает в каждую из частей. Для HSV-палитры гистограммы строились отдельно по каждой из компонент. Несмотря на то что признаки, полученные из разных гистограмм, основаны на цветовых характеристиках отдельных пикселей, они неплохо дополняли друг друга, так как HSV и RGB палитры связаны нелинейным преобразованием и отвечают за разные свойства объектов.

Ко второй группе можно отнести признаки, основанные на подсчете граничных пикселей. Для вычисления этих признаков изображения переводились в черно-белый формат, а затем находились резкие яркостные переходы. После чего из изображения вверху и внизу выделялись полоски высотой по 20% от общей, и для этих полос считалось количество «граничных» пикселей. Было замечено, что в одном кластере боковые края изображения могут весьма сильно различаться (например, небольшой поворот камеры), но нижний и верхний край обычно остаются без изменений.

В работе также использовать ключевые точки, выделенные алгоритмом SURF. Ключевые точки выделялись для весьма ограниченного количества картинок и использовались как вспомогательный признак для алгоритма кластеризации.

Так же была идея строить кластерные структуры на основе вышеописанных граничных или ключевых точек, а затем центры кластеров или относительные расстояния между ними использовать как признаки для алгоритма кластеризации. Однако идея не оправдала себя, так как не удалось подобрать достаточно устойчивый алгоритм кластеризации, чтобы центры не сильно менялись при изменении количества или положения точек.

3. Алгоритм кластеризации

В качестве базового алгоритма кластеризации был выбран алгоритм FOREL [5]. Данный алгоритм выбирает случайный объект и предполагает, что это центр кластера, затем находит все объекты, что находятся не далее, чем на заданное расстояние от центра кластера, и по формуле вычисления центра масс пересчитывает центр. Процесс повторяется итерационно, пока не стабилизируется, после чего объекты, попавшие в кластер, изымаются из множества, и поиск кластеров повторяется на остатке объектов.

Как видно, данный алгоритм не предполагает попадание одного объекта в разные кластеры, однако нам не кажется, что в данной задаче это существенно повлияло на результат из-за большого размера кластеров.

Мы внесли в алгоритм следующие модификации:

- Существовали как основные признаки, так и дополнительные. По основным признакам строилась сама кластерная структура, а по дополнительным осуществлялась фильтрация объектов. Основными были выбраны 32 признака из RGB-палитры, а признаки, полученные из HSV-палитры и на основе яркостных изменений, были дополнительными.
- В коллекции существовали довольно большие, но изолированные кластеры. Поэтому была добавлена возможность включения приграничных объектов, при условии что на определенном расстоянии вне кластера другие объекты отсутствуют.
- После построения всех кластеров запускалась процедура сращивания. Для этого выбирались два довольно близких кластера, и вычислялась середина между их центрами. Если значительное количество объектов из обоих кластеров было близко к середине, то кластеры объединялись. Такая процедура хорошо проявила себя на кластерах, где основной объект двигался или поворачивалась камера.
- Также существовала процедура сращивания по ключевым точкам. Ключевые точки вычислялись только для центров тех кластеров, у которых были близки гистограммные признаки, но различались яркостные. Кластеры объединялись, если количество соотнесенных точек превосходило определенный порог.

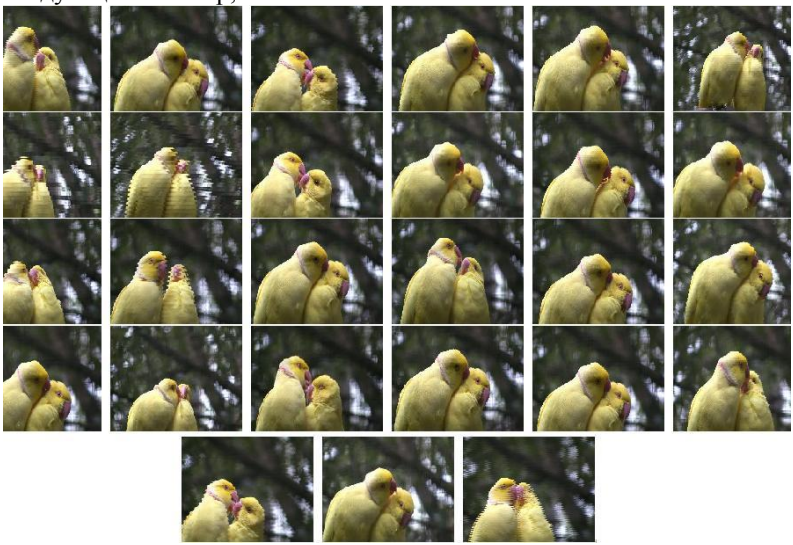
Данные изменения позволили достаточно быстро проводить саму кластеризацию, так как все признаки можно было вычислить на предварительном этапе. Самые трудоемкие вычисления, связанные с соотнесением ключевых точек на разных изображениях, были сведены к минимуму.

4. Результаты и выводы

По проведенным оценкам наша программа поиска дубликатов изображений получила следующие характеристики:

- Полнота – 74%
- Точность – 43%

По полноте программа опережает следующую систему на 15%, однако из всех систем обладает минимальной точностью (минимальный результат у остальных участников 56%). Это связано в первую очередь с тем, что само понятие нечеткого дубликата точно не определено, и, судя по результатам, мы воспринимали его несколько шире, чем асессоры. Мы настраивали параметры следующим образом: было выбрано несколько объемных кластеров, и мы старались, чтобы эти большие кластеры не разбивались на более маленькие. Однако асессоры посчитали, что эти кластеры, а, следовательно, и аналогичные им, на самом деле состоят из нескольких разных сюжетов. Как пример можно привести следующий кластер, составленный системой:



Асессоры посчитали, что точность, показанная нашим алгоритмом на приведенном кластере, равна 66%.

Тем не менее, невозможно отрицать, что использованные признаки не обеспечивают достаточной точности распознавания дубликатов. В дальнейшем мы планируем добавить признаки, опирающиеся на некоторые точечные особенности изображения. Предполагается, что при этом потери по полноте не должны быть существенными, однако точность значительно возрастет.

Литература

[1] R. Brunelli , O. Mich Histograms Analysis for Image Retrieval, 2001

[2] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346--359, 2008

[3] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. In *International Journal of Computer Vision*, volume 60, pages 91-110, 2004.

[4] W.-L. Zhao, C.-W. Ngo, H.-K. Tan and X. Wu. Near-Duplicate Keypoint Identification with Interest Point Matching and Pattern Learning. In *IEEE Transactions on Multimedia*, volume 9(5), pages 1037-1048, 2007.

[5] Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.

Near duplicates detection algorithm based on simple features.

© Dobrov Grygory, Pyatkov Evgeny
Moscow State University
wslc@rambler.ru, gpyatkov@gmail.com

In the article we describe a method for image retrieval by visual similarity. We used relatively simple and fast computable features, based on analysis of color histograms and brightness transitions in the image.

We use the clustering algorithm, which is a modification of the algorithm FOREL.