КЛАСТЕРИЗАЦИЯ ИЗОБРАЖЕНИЙ МЕТОДОМ ДЕНДРОГРАММ

Башков Е.А., Вовк О.Л.
Кафедра ПМиИ, ДонНТУ
vovkolga@ukrtop.com

Abstract

Bashkov E.A., Vovk O.L. The Images Clusterization by Dendrogram Method. *This article is devoted to consideration of the new algorithm of image partition into regions by color features. Developed algorithm is used as one of the stages of content-based images retrieval. Comparison with the most used in this area k-means algorithm is realized by processor time feature.*

Введение

Стремительное развитие вычислительных мощностей и постоянное снижение их стоимости сделали возможным хранение больших объемов оцифрованных изображений. В настоящее время электронные коллекции изображений используются все чаще и чаще. Наиболее перспективные области применения баз данных изображений следующие [1]:

- криминалистика (хранение архивов визуальных доказательств);
- медицинская диагностика (определение наиболее точного диагноза путем сравнения с уже установленными диагнозами);
- культурное наследие (тенденция в работе современных музеев – формирование электронных коллекций изображений с целью сохранения культурного наследия и его пропаганды за счёт обеспечения удалённого доступа);
- журналистика и реклама (иллюстрация статей и рекламных проектов);
- интеллектуальная собственность (примером может служить процесс регистрации новых торговых марок предприятий).

В связи с многообразием отраслей использования коллекций оцифрованных изображений очевидна необходимость разработки эффективного механизма доступа к информации таких баз данных. На сегодняшний день наиболее распространенным является поиск изображений по текстовым описаниям [1]. Существенным недостатком этого метода является неоднозначность соответствия между визуальным содержанием и текстовым описанием изображения, которое является субъективным. Поэтому возникает проблема организации средств поиска изображений по визуальному содержимому.

Поиск изображений по визуальному содержимому [2] – набор технологий для извлечения из базы данных изображений, наиболее подобных заданному изображению-образцу по некоторому набору

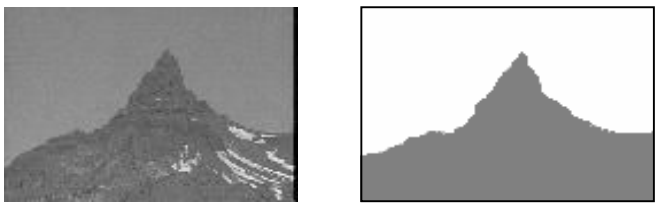
числовых значений характеристик сравниваемых изображений. Характеристики изображений можно анализировать на уровне пикселей, блоков (сегментов), регионов (кластеров) или изображений. Предлагается алгоритм выделения кластеров (объектов изображения) для обеспечения эффективного поиска изображений в базах данных на уровне регионов.

1. Обзор методов сравнения характеристик изображений

Наиболее простым и распространённым, но наименее эффективным является алгоритм поиска изображений по цветовым гистограммам [2, 3], предполагающий сравнение характеристик на уровне всего изображения. Данный метод основан на построении распределения цветов изображения или гистограммы. Основными недостатками данного метода принято считать пренебрежение местонахождением, формой и текстурой того или иного объекта.

Методами более высокого уровня поиска изображений принято считать методы поиска по “цветовой планировке” [2]. Такие методы предполагают сравнение цветовых характеристик изображений на уровне блоков. Т.е. изображение предварительно разбивается на блоки (сегменты) заданного размера и для каждого из сегментов вычисляется среднее значение каждой из цветовых компонент. Анализ осуществляется путём вычисления расстояний между блоками изображений. Однако методы этой группы также не учитывают характеристики формы объектов изображений.

На сегодняшний день наибольший интерес исследователей вызывают методы, основанные на сравнении визуальных примитивов отдельных кластеров (областей, регионов) изображения [4, 5]. Пример выделения отдельных регионов изображения приведен на рис. 1.



Исходное изображение

W_{H2Q2} 3-я

Рис. 1 - Пример кластеризации изображения

Методы, основанные на кластеризации, позволяют сравнивать изображения на уровне характеристик отдельных объектов. К визуальным характеристикам объектов принято относить [4, 6]: характеристики цвета, характеристики текстуры и характеристики формы. Учёт всей

совокупности этих характеристик делает поиск изображений устойчивым к масштабированию, повороту и перемещению объектов изображения. Однако чтобы процесс поиска был наиболее результативным необходимо использовать наиболее эффективный метод кластеризации.

2. Анализ алгоритмов кластеризации изображений

Все методы кластеризации изображений предлагается условно можно разделить на две группы: статистические методы кластеризации и методы кластеризации, основанные на выделении перепадов яркости.

При решении проблемы кластеризации каким-либо методом из первой группы, подразумевается предварительное представление изображения в виде статистической выборки. После чего над пикселями изображения производятся операции, подобные операциям над элементами выборки.

Наиболее распространенным статистическим методом принято считать метод k-means (k-средних) кластеризации [2, 7, 8]. Процесс кластеризации согласно этому методу состоит из следующих этапов:

1) Сегментация изображения на квадратные блоки заданного размера и вычисление для каждого блока средней цветовой характеристики.
2) Произвольное распределение блоков изображения на k кластеров (начальное значение количества кластеров k=2).
3) Расчёт центра для каждой из характеристик каждого кластера в соответствии с формулой (1):

$$\hat{x}_j = \frac{\sum x_i}{R_j}, \qquad (1)$$

где R_j – количество точек j-ого региона, x_i – значения характеристик блоков, входящих в регион.

4) Перегруппировка блоков внутри кластеров – вычисление для каждого блока расстояния до центра каждого из кластеров (см. формула (2))

$$D(k) = \sum_i \min_{1 \le j \le k} (x_i - \hat{x}_j)^2 \qquad (2)$$

и отнесение блока к тому из кластеров, до которого расстояние по всем характеристикам блока (индекс i) минимальное.

5) После включения каждого нового блока в кластер необходимо произвести пересчет центров кластеров.
6) Далее происходит наращивание k и повторяются этапы 2) – 5).

Критерии прекращения увеличения количества кластеров:

- разброс значений *D(k)* внутри каждого кластера становиться меньшим некоторого параметра θ (подобранного экспериментально);
- результат группировки при увеличении количества кластеров значительно не изменяется, т.е. $D(k) - D(k-1) < \theta$;
- k выходит за пределы (экспериментально в [2, 7] установлено, что не эффективно бить изображение на число кластеров большее шестнадцати), т.е. $k>16$.

Отметим основные недостатки этого алгоритма. Во-первых, отсутствует правило задания параметра θ для критериев окончания. В результате практических испытаний этого алгоритма установлено, что для разных изображений параметр θ существенно отличается и от её подбора в каждом конкретном случае зависит качество и скорость кластеризации. Во-вторых, так как начальная стадия – сегментирование изображения на квадратные области заданного размера, то края объектов получаются “ступенчатыми”.

Первый из отмеченных недостатков связан с отсутствием эффективного критерия окончания процесса кластеризации, определенным шагом в этом направлении можно считать работу [9]. В ней в качестве критерия окончания процедуры разбиения на регионы предлагается использовать статистическую значимость нулевой гипотезы.

Среди модификаций k-means алгоритма особое место занимают алгоритмы fuzzy-кластеризации [10, 11]. Авторы этих методов предполагают представление объектов изображения в виде нечетких множеств и вводят функции принадлежности этим множествам, на основании которых определяют блоки изображения в тот или иной кластер.

Также среди статистических методов кластеризации необходимо отметить метод кластеризации по максимальной дисперсии [12], авторы которого предлагают технологию выбора количества кластеров до начала процесса кластеризации.

При кластеризации путем выделения перепадов яркости проблема выделения границ объектов решается путем локализации на изображении резких перепадов яркости цвета [13]. С этой целью вычисляется градиент функции интенсивности в каждой точке изображения, после чего подавляются значения меньше установленного порога. За основу принято брать метод Собеля [14].

К недостаткам этой группы методов можно отнести: зависимость результатов кластеризации от освещенности объектов внутри изображения и как следствие – чувствительность кластеризации к поворотам и переносам объектов.

3. Алгоритм кластеризации методом дендрограмм

В основу предлагаемого метода кластеризации положен статистический метод распознавания образов – метод дендрограмм [15]. Однако так как данный метод ранее не использовался для кластеризации изображений, была произведена его существенная модификация для решения проблемы выделения различных объектов изображения.

На подготовительном этапе кластеризации осуществляется переход из стандартного пространства цветов RGB в цветовое пространство, которое обладает свойствами однородности, полноты и компактности [16]. В соответствии с требованием однородности вычисленное подобие цветов должно соответствовать их визуальному подобию. Пространство, обладающее свойством полноты, включает все различные воспринимаемые цвета. Компактность цветового пространства означает, что любой цвет отличается от остальных.

Перечисленным требованиям удовлетворяет пространство HSL [16], в котором H (hue) – оттенок (аналог длины световой волны), S (saturation) – насыщенность и L (luminance) - яркость [17]. Подробные описания и формулы для перехода в это пространство цветов приведены в [18, 19]. Данное пространство является более информативным по сравнению с RGB и потому переход в него ускоряет процесс сходимости алгоритма k-means кластеризации, рассмотренного выше. На рис. 2 представлены результаты работы данного алгоритма на наборе картинок различного качества и размера. В качестве критерия тестирования было выбрано кумулятивное время процессора, затраченное на выделение отдельных регионов изображений.

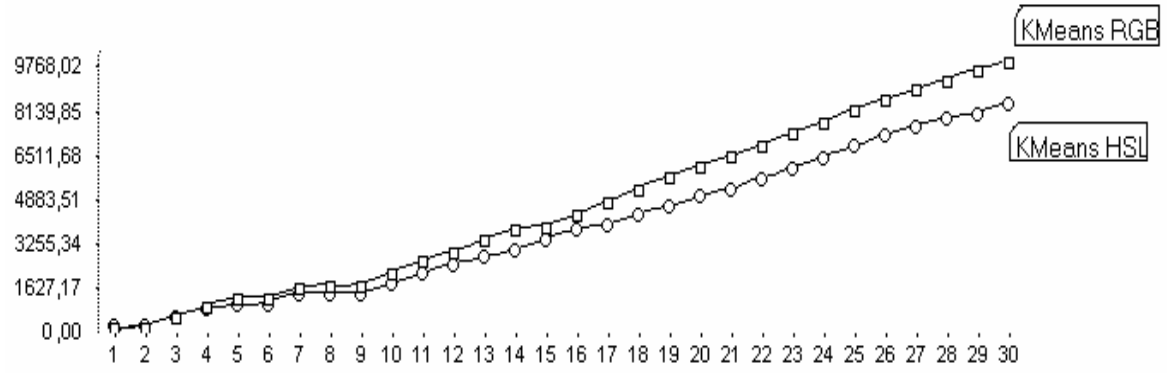


Рис.2 - Графики зависимости процессорного времени работы алгоритма k-means кластеризации от выбора цветового пространства

Причем, следует отметить, что качество кластеризации практически не изменилось (рис. 3) при переходе из одного цветового пространства в другое.

Исходное изображение

Кластеризация в RGB
Время процессора 96 с
Число кластеров 4

Кластеризация в HSL
Время процессора 75 с
Число кластеров 4

Рис.3 - Пример k-means кластеризации для пространств цветов RGB и HSL

После перехода в пространство HSL необходимо осуществить предварительную сегментацию исходного изображения на блоки для увеличения скорости обработки изображения. В алгоритмах, описанных выше, этот этап эквивалентен разбиению изображения на квадратные блоки заданного размера. Произведем сегментацию изображения по цветовому подобию. В качестве критерия объединения точек в блоки выступает компонента H, т.к. она является аналогом световой волны и выражает в пространстве цветов HSL самую визуально информативную характеристику – оттенок чистого цвета [17]. Однако для большей точности сегментирования в блоки объединяются только те точки с одинаковым значением оттенка, у которых квадрат разницы по компонентам насыщенности и яркости не превышает некоторого заданного ε (в ходе ряда экспериментов установлено, что для любых изображений этот параметр принимает значения в диапазоне $0.1 \div 0.3$).

На начальной стадии алгоритма дендрограмм изображение задаётся набором блоков – кластеров, для каждого из которых вычисляется среднее значение цветовых компонент. Дальнейшие преобразования кластеров осуществляются путем повторения следующих этапов:

1) По набору кластеров осуществляется построение матрицы расстояний D между ними.
2) Используя матрицу D, определяются кластеры, расстояние между которыми является минимальным. Наиболее “близкие” по расстоянию кластеры объединяются в один кластер.

3) Для нового кластера вычисляются средние значения цветовых характеристик.
4) Пересчитываются расстояния от всех кластеров изображения до полученного кластера.
5) Шаги 2) – 4) повторяются до тех пор, пока не будет достигнуто заданное пользователем количество кластеров, либо не будет удовлетворен критерий окончания кластеризации.

Особенностью метода дендрограмм для кластеризации является формирование матрицы расстояний *D*. При вычислении элементов матрицы традиционное евклидово расстояние заменяется коэффициентами композиционного различия [20]:

$$d_{ij} = q_{ij} \Big/ (m + p_{ij}), \qquad (3)$$

где q_{ij} и p_{ij} – количества цветовых компонент, имеющих соответственно различающиеся и одинаковые значения для *i*-ого и *j*-ого кластеров; *m* – общее число признаков (компонент).

Так как при решении задачи кластеризации трудно определить понятия “одинаковые” и “различающиеся” значения цветовых компонент, вводится новое понятие – степень различия цветовых компонент. Для определения степени различия цветовых компонент двух кластеров, до начала процесса кластеризации определяется максимальное различие для каждой из цветовых составляющих всего изображения:

$$r_k = (\max x_k - \min x_k)^2, \qquad (4)$$

где x_k – *k*-ая цветовая характеристика, $k = 1 \div m$ (в случае рассматриваемой задачи *m*=3).

Тогда степень различия цветовых компонент двух кластеров можно представить в виде:

$$q_{ij} = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \Big/ r_k . \qquad (5)$$

Так как хранение матрицы расстояний (коэффициентов композиционного сходства) требует дополнительных затрат оперативной памяти, требуется произвести минимизацию её объема. Для этого в силу симметричности матрицы необходимо осуществлять хранение не всей матрицы целиком, а её нижней (верхней) треугольной части. Но так как современные средства разработки программного обеспечения не предоставляют возможностей обработки треугольных матриц, то наиболее оптимальным способом представления таких матриц можно считать их представление в виде списка. Однако кроме самих значений в списке необходимо хранить индексы, соответствующие элементам исходной матрицы, а это опять же увеличивает объем используемой оперативной

памяти. Предлагается хранить только один из индексов – номер строки, а номер столбца рассчитывать по следующей формуле:

$$j + \frac{(i-1)i}{2} = num, \qquad (6)$$

в которой: *i,j* – номер строки и столбца исходной матрицы; *num* – номер элемента в списке.

В качестве критерия окончания кластеризации в данной работе предлагается использовать описанные выше коэффициенты композиционного различия [19]. Т.е. процесс кластеризации необходимо останавливать, когда будут объединяться два кластера, у которых коэффициент композиционного различия больше некоторого параметра δ (авторы данной работы предлагают выбирать этот параметр в зависимости от критерия точности кластеризации в диапазоне $0.05 \div 0.3$).

Сравнительный анализ разработанного алгоритма с алгоритмом k-means кластеризации по критерию быстродействия (кумулятивному процессорному времени) приведен на рис. 4, по критерию качества – на рис. 5.

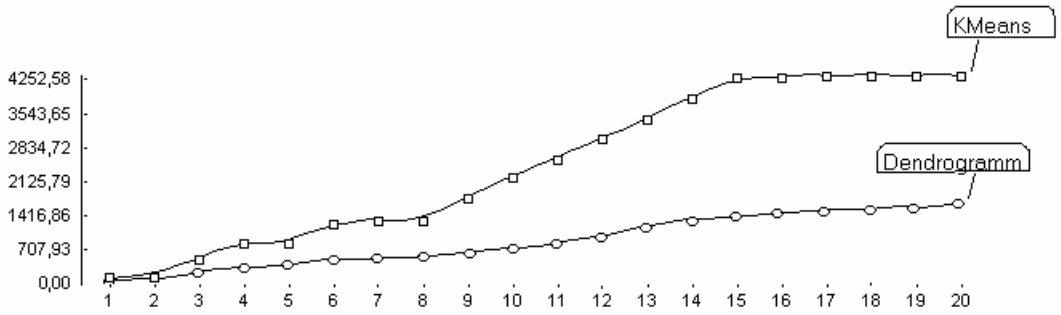
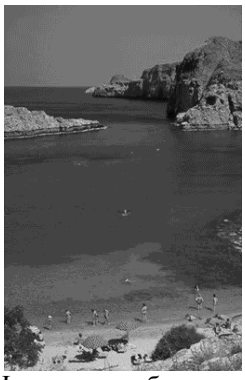
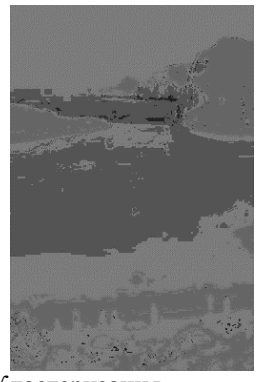


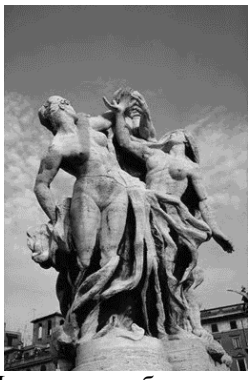
Рис.4 - Графики зависимости процессорного времени работы алгоритма k-means кластеризации и алгоритма кластеризации методом дендрограмм



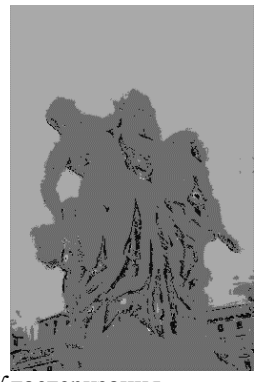
Исходное изображение



Кластеризация
Время процессора 13 с
Число кластеров 4



Исходное изображение



Кластеризация
Время процессора 18 с
Число кластеров 3

Рис. 5 - Примеры работы алгоритма кластеризации по методу дендрограмм

Заключение

В данной работе предложен новый алгоритм кластеризации изображений – алгоритм кластеризации методом дендрограмм. Для анализа эффективности разработанного алгоритма авторы статьи проводят сравнение предлагаемого алгоритма с наиболее распространенным алгоритмом – алгоритмом k-means кластеризации, используемым в большинстве современных систем поиска изображений в базах данных по визуальному содержимому. Однако на ряду с достоинствами метода дендрограмм (меньшие затраты процессорного времени, более высокое качество кластеризации) в рассматриваемой работе отмечается и очевидный недостаток этого метода – необходимость использования оперативной памяти для хранения основной структуры алгоритма (матрицы коэффициентов композиционного сходства). Для преодоления этого ограничения алгоритма производится предварительная сегментация изображения по длине световой волны пикселей (что приводит к уменьшению начального количества кластеров изображения) и предлагается хранение матрицы в виде списка с вычисляемым индексом. Для дальнейшего усовершенствования разработанного алгоритма можно провести анализ эффективности существующих критериев окончания процесса кластеризации и разработать более совершенный критерий, повышающий точность разбиения изображений на регионы.

Литература

1. Eakins J. P., Graham M. E., “A report to the JISC Technology Applications Programme”, Institute for Image Data Research, University of Nothumbria at Newcastle, Jan. 1999, 54 p.
2. Wang J. Z., Li J., Wiederhold G., “SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 9, Sept. 2001, pp. 947-963.
3. Башков Е. А., Шозда Н. С. Поиск изображений в больших БД с использованием коэффициента корреляции цветовых гистограмм. – GraphiCon’2002. – Нижний Новгород, 2002. – с. 458-460.
4. Smeulders A., Worring M., Santini S., Gupta A., Jain R., “Content-Based Image Retrieval at the End of the Early Years”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, Dec. 2000, pp. 1349-1380.
5. Chen C., Wang J. Z., “Large-scale Emperor digital library and semantics-sensitive region-based retrieval”, Proc. International Conference on Digital Library - IT Opportunities and Challenges in the New Millennium, Beijing, July 9-11, 2002, pp. 454-462.

6. Gupta A., Jain R., “Visual Information Retrieval”, Communications of the ACM, vol. 40, no. 5, May 1997, pp. 70-79.
7. Wang J. Z., Du Y., “Scalable Integrated Region-based Image Retrieval using IRM and Statistical Clustering”, Proc. ACM and IEEE Joint Conference on Digital Libraries, Roanoke, VA, ACM, June 2001, pp. 268-277.
8. Hartigan J. A., Wong M. A., “Algorithm AS136: a k-means clustering algorithm”, Applied Statistics, 1979, vol. 28, pp. 100-108.
9. Жданов А. С., Костин В. С., Значимость и устойчивость автоматической классификации в задаче поиска оптимального разбиения. – Институт Экономики и Организации Промышленного Производства СО РАН, http://ie.ie.nsc.ru/~rokos/znach.doc
10. Baraldi A., Blonda P., “A Survey of Fuzzy Clustering Algorithms for Pattern Recognition – Part I”, IEEE Transactions on systems, man, and cybernetics – Part B: Cybernetics, vol. 29, no. 6, Dec. 1999, pp. 778-785.
11. Deer P., “Change Detection using Fuzzy Post Classification Comparison”, PhD thesis, Department of Computer Science, The University of Adelaide, 1998.
12. Veenman C. J., Reinders M. J. T., Backer E., “A Maximum Variance Cluster Algorithm”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 9, Sept. 2002, pp. 1273-1280.
13. Байгарова Н. С., Бухштаб Ю. А., Евтеева Н. Н. Современная технология содержательного поиска в электронных коллекциях изображений. – Институт прикладной математики им. М. В. Келдыша РАН, http://artinfo.ru/eva/eva2000M/eva-papers/200008/Baigarova-R.htm.
14. Sobel I., “An isotropic image gradient operator”, Machine Vision for Three-Dimensional Scenes, Academic Press, 1990, pp. 376-379.
15. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. Учебник для вузов. – М.: ЮНИТИ, 1998. – 1022 с.
16. Башков Е. А., Шозда Н. С. Алгоритмы дискретизации цветового пространства и их использование в контекстном поиске изображений. – Научные труды ДонГТУ. Серия: Проблемы моделирования и автоматизации проектирования, выпуск 15: – Донецк, ДонГТУ, 2000 . – с. 192-196.
17. Руководство по работе с цветом компании X-Rite, http://www.Realcolor.ru.
18. Tkalcic M., Tasic J., “Colour spaces: Project for the Digital signal processing course”, http://ldos.fe.uni-lj.si/docs/documents/20021113135017_markot.pdf
19. Color space FAQ, http://www.neuro.sfc.keio.ac.jp/~aly/polygon/info/color-space-faq.html.
20. Автоматическая классификация и распознавание образов, http://redyar.samara.ru/stat/wavto.html.

Дата надходження до редколегії: 6.10.2003 р.