Поиск изображений в больших БД с использованием коэффициента корреляции цветовых гистограмм

Е.А.Башков, проф., д.т.н.
Н.С.Шозда, ассистент
Кафедра Прикладной Математики и Информатики
Донецкий Национальный Технический Университет

В статье рассматриваются проблемы, связанные с реализацией поиска изображений на основе их содержимого, в частности, по цветовой информации, представленной в изображении. Анализируются существующие проблемы и разработки, предлагается модифицированный алгоритм поиска, позволяющий ускорить процесс поиска изображений по образцу и улучшить качество поиска.

Ключевые слова: *поиск изображений по образцу, цветовая гистограмма, коэффициент корреляции цветовых гистограмм*

1. ПРОБЛЕМА ПОИСКА ИЗОБРАЖЕНИЙ ПО СОДЕРЖИМОМУ

Разработка новых технических средств, позволяющих представлять информацию в виде изображений, привела к тому, что работа с таким видом данных оказалась в центре внимания многих исследователей. С распространением Internet наблюдается тенденция к накоплению информации, представленной в виде изображений, и к созданию баз данных (БД) изображений. Следует отметить, что, в зависимости от области применения, в такие БД включаются как произвольные изображения, так и изображения ограниченного класса, и очень актуальной для таких БД является задача поиска изображений, визуально сходных с заданным. Примером использования БД, содержащих изображения ограниченного класса, могут служить биометрические системы. Однако все алгоритмы, применяемые в таких системах, ориентированы на специфику обрабатываемых изображений, чувствительны к колебаниям в освещении, изменению положения камеры, а также к применению камер различных типов. Что же касается произвольных цветных изображений, то задача их успешного сопоставления до сих пор успешно не решена, хотя имеется насущная потребность в ее решении, поскольку в настоящее время существует значительное количество библиотек оцифрованных изображений, объем которых постоянно увеличивается. Так, база данных виртуальной библиотеки Нового Южного Уэльса [1] в настоящий момент содержит 308 тыс. оцифрованных изображений, и в ближайшее время разработчики планируют дополнить ее еще 1,3 млн. изображений. Цифровая коллекция Эрмитажа [2] включает 4500 изображений, и в будущем планируется поместить в эту коллекцию изображения всех экспонатов музея. Коллекция живописи Национальной галереи искусств США [3] содержит более 100 тыс. изображений, и поиск в коллекции возможен только по названию, теме и автору. Цифровая библиотека VCL Antropomorphic Image Library [4] содержит 120 тысяч уникальных изображений. Цифровая астрономическая библиотека [5] содержит более 100 тыс. изображений. Кроме упомянутых, существуют и другие цифровые библиотеки изображений, однако поиск по содержимому реализован только в цифровой коллекции Эрмитажа.

Весьма активно в последнее время развивается подход, называемый контекстным поиском изображений. В соответствии с этим подходом в БД выполняется поиск изображений, визуально сходных с заданным изображением-образцом. Разработки в этом направлении ведутся в США Колумбийским университетом, фирмами IBM и Virage [6,7,8]. Разработка фирмы IBM используется для поиска изображений в цифровой коллекции Эрмитажа.

2. ФОРМУЛИРОВКА ЗАДАЧИ ПОИСКА ИЗОБРАЖЕНИЙ ПО ИХ СОДЕРЖИМОМУ

В целях формализации решения задачи контекстного поиска изображений авторами сформулирована ее общая постановка.
Пусть имеется БД, содержащая V изображений, V>0. Каждое изображение P_k, k=1, 2,..V, представляет собой матрицу M*N, элементы которой хранят цвета соответствующих пикселей изображения; M и N – ширина и высота изображения соответственно.
Пусть содержимое каждого изображения из БД P_k, k=1, 2,..V, характеризуется скалярной либо векторной величиной F_k, k=1,2,..V. Имеется также изображение – образец поиска, цветовое содержимое которого характеризует аналогичная величина $F_{обр}$. В этом случае задача поиска изображений по образцу представляет собой формирование последовательности Q_k, k=1, 2,..V_1, V_1<=V, изображений из БД, визуально сходных с образцом и расположенных в порядке убывания этого сходства. Место каждого изображения в такой упорядоченной последовательности определяется в результате выполнения алгоритма сортировки, исходными данными для которого служат числовые характеристики степени сходства каждого изображения из БД с образцом поиска (называемыми расстояниями между изображениями):

$$d_k = f(F_{обр}, F_k), k = 1,2,..V . \quad (1)$$

Таким образом, задача контекстного поиска сводится к вычислению значений d_k, k=1,2,..V и их последующей сортировке.
При поиске изображений по их цветовому содержимому данная общая постановка задачи нуждается в следующих дополнениях.
Пусть имеется базовый набор цветов Colors[1..Cmax], $0 \le Cmax \le 2^R$, R–разрядность цвета пикселя, и каждый пиксель $P_k[i,j]$, k=1, 2,..V, $1 \le i \le M$, $1 \le j \le N$ произвольного изображения может принимать любой из Cmax цветов базового набора. Данное требование является обязательным, при его несоблюдении сравнение изображений невозможно.
В качестве характеристик цветового содержимого изображения используются точечные и гистограммные оценки. К первым относятся средний либо преобладающий цвет точек изображения, ко вторым– цветовые гистограммы и

бинарные цветовые векторы. Цветовая гистограмма изображения представляется вектором H [1.. Cmax], каждый элемент которого вычисляется как:

$$H[i] = \frac{K_i}{M \times N},\ i = 1,2..C\max \quad , \qquad (2)$$

где K_i - количество точек изображения, имеющих цвет Colors[i]. Каждый элемент построенной таким образом ЦГ характеризует вероятность того, что цвет произвольного пикселя изображения совпадает с i–м цветом из базового набора, а сама ЦГ – распределение вероятности значений цвета пикселей. Если количество цветов изображения не совпадает с Cmax, то перед построением цветовой гистограммы дополнительно должна быть решена задача приведения цветовых систем, то есть замены цветов пикселей изображения наиболее близкими цветами из базового набора.

Для решения данной задачи существует ряд методов, однако, как показывает анализ [9], в процессе контекстного поиска применим только один, заключающийся в отбрасывании младших разрядов каждой цветовой составляющей – это единственный метод, позволяющий получить после преобразования один и тот же набор цветов.

Таким образом, при использовании для представления цветового содержимого изображения цветовых гистограмм в качестве основных этапов контекстного поиска изображений можно выделить следующие:

1. Приведение цветов к базовому набору (Q-quantization). Данный этап заключается в замене цветов, представленных в изображении, наиболее близкими цветами из базового набора, по которому строится гистограмма.
2. Построение гистограммы цветов (H-histogram). На этом этапе рассчитываются вероятности появления каждого из базовых цветов в изображении.
3. Сравнение изображений (C-comparison). Данный этап выполняется путем вычисления расстояний между ЦГ изображения–образца и гистограммами всех изображений из БД.
4. Сортировка (S) изображений по возрастанию вычисленных на этапе C значений.

При занесении нового изображения в БД для него последовательно выполняются этапы Q и H. При поиске, когда в качестве образца поиска используется изображение из БД, выполняются этапы C и S. Если же образцом поиска является новое изображение, то выполняются этапы Q, H, C и S.

3. ПОДХОДЫ К КОНТЕКСТНОМУ ПОИСКУ ИЗОБРАЖЕНИЙ

Кроме цветового содержимого, поиск изображений может основываться также на рассмотрении таких параметров, как форма и текстура. При поиске по текстурному содержимому для его представления используются текстурные гистограммы.

Наибольший интерес исследователей вызывает поиск по гистограммным признакам, в частности, по цветовым гистограммам. Исследования в данной области направлены на усовершенствование метрик для сравнения цветовых гистограмм изображений с целью повышения качества поиска. Остановимся более подробно на проблемах, возникающих при решении данной задачи.

Традиционно для сравнения гистограмм используются конъюнкция гистограмм, евклидово, косинусное и квадратичное расстояния, вычисляемые, соответственно, по формулам (3), (4), (5) и (6). Анализируя эти формулы, можно определить область значений для конъюнкции гистограмм, евклидова, косинусного и квадратичного расстояний. Эти сведения приведены в таблице 1.

$$D1 = \sum_{i=1}^{C\max} \left| H_1[c_i] - H_2[c_i] \right| \qquad (3)$$

$$D2 = \sum_{i=1}^{C\max} (H_1[c_i] - H_2[c_i])^2 \qquad (4)$$

$$D3 = 1 - \cos\theta = \frac{D2}{2} \qquad (5)$$

$$D4 = (H_1 - H_2)^T \cdot A \cdot (H_1 - H_2) \qquad (6)$$

Таблица 1

Область значений различных метрик для сравнения цветовых гистограмм

Метрика	Область значений при использовании нормализованных гистограмм
Конъюнкция гистограмм	$0 \le D1 \le C_{\max}$
Евклидово расстояние	$0 \le D2 \le C_{\max}$
Косинусное расстояние	$0 \le D3 \le C_{\max}$
Квадратичное расстояние	$0 \le D4 \le C_{\max}^2$

Легко увидеть, что максимальное значение расстояния между гистограммами определяется числом элементов гистограммы. При использовании же ненормализованных гистограмм значения этих величин сверху практически не ограничены. По этой причине расстояния сами по себе не отражают степень сходства изображений, для которых вычислены; они имеют значение только при сравнении с другими аналогичными величинами, что является существенным недостатком данных метрик. По этой причине поиск с использованием таких метрик по сути заменяется сортировкой изображений в порядке возрастания вычисленного расстояния до изображения-образца.

4. МОДИФИЦИРОВАННЫЙ АЛГОРИТМ ПОИСКА ИЗОБРАЖЕНИЙ В БД

Авторами предлагается другой подход к реализации поиска изображений с использованием цветовых гистограмм, который основан на вычислении коэффициента корреляции цветовых гистограмм изображений:

$$\rho = \frac{\mathrm{cov}(H_1, H_2)}{\sigma(H_1) \cdot \sigma(H_2)},$$

где

$\mathrm{cov}(H_1, H_2) = \frac{1}{C_{\max}} \sum_{i=1}^{C\max} (H_1[i] - \mu(H_1)) \cdot (H_2[i] - \mu(H_2))$ - ковариация цветовых гистограмм H_1 и H_2,

$$\sigma(H) = \sqrt{D(H)} = \sqrt{\frac{1}{C_{\max}} \sum_{i=1}^{C\max} (H[i] - \mu(H))^2}$$ -

среднеквадратичное отклонение элементов цветовой гистограммы,

$$\mu(H) = \frac{1}{C_{\max}} \sum_{i} H[i]$$ - математическое ожидание элементов цветовой гистограммы.

Достоинством данного подхода является то, что коэффициент корреляции является естественной характеристикой наличия линейной корреляции между двумя случайными величинами, и его величина может быть оценена сама по себе, и, таким образом, можно ограничить набор изображений, являющихся результатами поиска, и, таким образом, избежать сортировки всей БД. Так, из результирующего набора изображений должны быть исключены те, для которых вычисленный коэффициент корреляции меньше либо равен нулю. Коэффициент корреляции не удовлетворяет свойствам идентичности, неотрицательности и неравенства треугольника, традиционно предъявляемым к метрикам[7], однако его использование позволяет ограничить число изображений, выводимых в качестве результатов поиска, что является весомым аргументом в пользу данного метода. Более того, данный подход позволяет пользователю задавать минимально допустимую степень сходства образца и результатов поиска.

Тот факт, что цветовая гистограмма, вычисленная по (2), является нормализованной, и сумма ее элементов равна 1, позволяет в значительной степени сократить временные затраты на вычисление коэффициента корреляции. Кроме того, вычисление среднеквадратичного отклонения (далее – этап D) может быть выполнено ранее, непосредственно после построения цветовой гистограммы, что также способствует ускорению вычислений. На рисунке 1 выполнено сравнение временных характеристик при использовании предлагаемого подхода (коэффициент корреляции ЦГ с предварительным сохранением среднеквадратичного отклонения) и традиционной метрики, обладающей наилучшими временными характеристиками (евклидово расстояние). Результаты данного сравнения свидетельствуют о возможности сокращения времени вычислений за счет использования предлагаемого модифицированного алгоритма. Выполнена экспериментальная проверка эффективности данного алгоритма, которая показала, что результаты поиска получаются такими же, как и при использовании традиционных метрик, однако изображения, не сходные с искомым образцом, в результирующий набор не включаются, то есть метод позволяет получить вполне приемлемые результаты. Тестирование проводилось на коллекции различных картинок, содержащих изображения цветов, животных, пейзажей, абстрактные картины и др. В результате поиска определяются схожие по цветовому содержимому изображения. Кроме того, экспериментально получено пороговое значение коэффициента корреляции, позволяющее отбросить изображения, мало сходные с образцом, однако теоретическое обоснование данной величины пока не получено.

В качестве возможных применений поиска изображений по образцу можно назвать дизайнерские системы принятия решений (например, подбор обоев или тканей в заданной цветовой гамме) либо программы, позволяющие систематизировать большие наборы изображений (определить совпадающие изображения среди всех, имеющихся на диске).

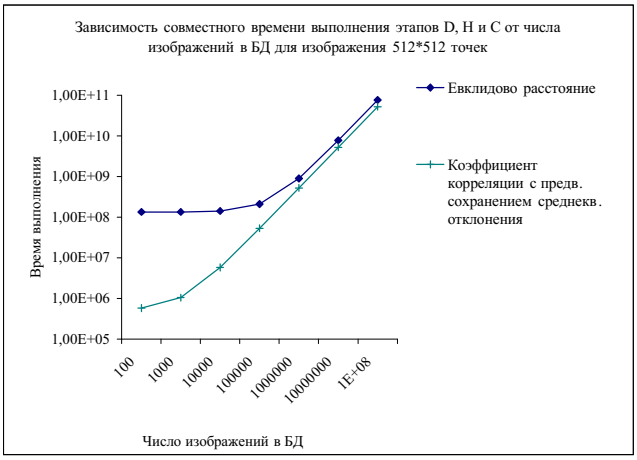

Рисунок 1

5. ЗАКЛЮЧЕНИЕ

Занимаясь разработками в данном направлении, авторы видят цель исследований в создании эффективных методов поиска изображений, позволяющих получать результаты, визуально сходные с образцом, за приемлемое время. В настоящее время ведутся исследования, посвященные поиску изображений по текстурным признакам также с использованием коэффициента корреляции, а также поиску с учетом пространственной информации об объектах, содержащихся внутри изображения.

6. ЛИТЕРАТУРА

[1] http://cdromweb.snsw.gov.au/picman/ welcome.html

[2] http://www.hermitagemuseum.org

[3] http://www/nga/gov

[4] http://us.vclart.net/vcl

[5] www.astronomy.ca/images/

[6] *IBM Almaden Research Center. Query by Image and Video Content: the QBIC System. Computer, September 1995, 23-31.*

[7] J. R. Smith and S.-F. Chang, *"VisualSEEk: A Fully Automated ContentBased Image Query System", ACM Multimedia Conference, Boston, MA Nov. 1996. Demo: http://www.ctr.columbia.edu/VisualSEEk*

[8] Amarnath Gupta. *Visual Information Retrieval Technology- A Virage Perspective*

[9] Башков Е.А., Шозда Н.С. *Алгоритмы дискретизации цветового пространства и их использование в контекстном поиске изображений// Наукові праці Донецького державного технічного університету. Серія "Інформатика, кібернетика та обчислювальна техніка.–Випуск 15.–Донецьк, 2000.– С.192–196*

Сведения об авторах

Евгений Александрович Башков-профессор, доктор технических наук, заведующий кафедрой прикладной

математики и информатики Донецкого национального технического университета.
E-mail: bashkov@r5.dgtu.donetsk.ua
Наталья Стефановна Шозда- ассистент кафедры прикладной математики и информатики Донецкого национального технического университета.
E-mail:shozda@r5.dgtu.donetsk.ua