Информационная система для создания и управления электронными коллекциями графических документов ♣

А.А. Рогов, К.А. Рогова, П.В. Кириков, М.Ю. Быстров

Петрозаводский государственный университет
rogov@psu.karelia.ru, ksushar@mail.ru, lispad@gmail.com, maksimkab@yandex.ru

Аннотация

Разрабатываемая информационная система позволяет создавать коллекции графических документов, хранить различные наборы графических (основанных на цветовосприятии и текстурных характеристиках) и текстовых параметров для каждого графического документа, выполнять классификацию и поиск по различным комбинациям параметров, а так же на основе сходства изображений. Система разрабатывается под Интернет.

1 Введение

В настоящее время все большую популярность получают электронный коллекции графических документов. Обычные пользователи создают свои фотоальбомы (не только на дома, на локальном компьютере), но и в сети Интернет; научные сообщества используют большие объемы графической информации в своих исследованиях. К сожалению, сейчас нет единой системы, которая позволила бы не только хранить изображения в определенном порядке, но и классифицировать графическую информацию по выделенным параметрам и осуществлять поиск.

Целью данной работы является создание информационной системы, позволяющей создавать, управлять и анализировать коллекции графических документов. Особенностью создаваемой информационной системы является возможность хранения и использования иерархии документов и значений наборов признаков, зависящих от типа документа. Создаваемое в коллекции дерево для хранения иерархии графических объектов может иметь неограниченное число уровней и неограниченное число узлов на каждом уровне. Не смотря на то что, создаваемая система предназначена в первую очередь для научных исследований, подобный способ описания графических объектов позволяет применять систему для создания каталога запчастей/товаров с иллюстрациями, произведений художников и т.д.

Труды 11[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск, Россия, 2009.

2 Общие характеристики разрабатываемой системы

Информационная система предназначена для размещения на WEB-сервере и обеспечивает удобный доступ через сеть Интернет с любого подключенного к сети устройства. Она пригодна для решения большого ряда задач, среди которых можно выделить:

- создание и редактирование иерархической структуры коллекций графических документов;
- хранение различных наборов параметров для каждого графического документа;
- классификация и поиск графических документов по различным комбинациям параметров, а так же на основе сходства текстур, цветового восприятия и т.д.;
- описательная статистика коллекции;
- разделение доступа к системе.

Система логически разделена на две части – пользовательскую и административную и предоставляет простой и удобный интерфейс. Для ввода информации будут пригодны графические документы в различных форматах, автоматически осуществляться их приведение к единому стандарту, и в автоматизированном режиме предоставлять возможность выделять объекты на них.

Главное отличие предлагаемой системы от существующих систем создания электронных фотоальбомов состоит в возможности приписывать графическому документу набор индивидуальных признаков и осуществлять поиск по выделенной комбинации признаков. На основе признаков автоматически производится классификация объектов коллекции с целью поиска наиболее близких между собой. Кроме того, пользователю предлагается статистическая информация о наличии в коллекции объектов с выделенным набором признаков и анализ выделенных признаков статистическими методами.

Информационная система реализуется на php с использованием web-сервера apache и сервера баз данных mysql. В данный момент прототип информационной системы апробируется на основе коллекции графических документов петроглифов Северной Фенноскандии.

3 Административная и пользовательская составляющие системы

Для зарегистрированных пользователей системы, в административном разделе заложены дополнительные функции работы с графической информацией: пополнение информации, редактирование и удаление; создание набора признаков; присвоение признаков изображениям, создание иерархии рисунков и т.д.

Рассмотрим работу этого модуля на примере создаваемой коллекции петроглифов Северной Фенноскандии. В Северной Фенноскандии выделены несколько крупных местонахождения наскальных рисунков, среди которых мы рассматриваем Норвегию, Мурманскую область и Карелию. В каждом месте выделяют более мелкие группы, потом еще более мелкие и т.д., пока не доходят до сюжетных схем и отдельных петроглифов. Наглядным представлением такого расположения являются иерархические наборы карт, схем и рисунков. Для этого необходимо загрузить карты в систему и выделить соответствующие точки местонахождения. Точно так же можно работать со схемой петроглифов: квадратной областью выделяется петроглифов и в базу данных заносится вся необходимая информация о нем. Таким образом, с этой часть системы может работать специалист в любой области, без необходимости изучения языков работы с базами данных. Основными пользователями этого раздела будут создатели коллекции. Для больших научных коллективов (сетевого научного сообщества) возможна распределенная работа научного коллектива без принудительной синхронизации получающейся базы данных.

Для обычных пользователей система является открытой и доступна через Интернет любому интересующемуся. Доступ к функциям системы в таком случае ограничен лишь «чтением» и «анализом». То есть, пользователь может работать с системой как с обычным сайтом, классифицировать и искать изображения по выделенным признакам, но изменение и дополнение информации для него невозможно.

4 Типы признаков изображений

Все признаки можно разделить на 2 части. Первая - значения признаков вводит администратор системы. При этом, для вычисления некоторых признаков возможна частичная автоматизация. Во второй группе признаки получаются в автоматическом или авторизованном виде и касаются параметров цветовосприятия и текстур изображения.

В первой группе каждое изображение может быть описано по параметрам следующего вида:

- количественные характеристики изображений;
- номинальные переменные с отношением порядка и категоризированные переменные, кроме того, отдельные признаки могут иметь более сложную фасетную структуру, которую можно описать с помощью графа;
- текстовые описательные признаки (они не используются при статистическом анализе);
- фрагменты изображений: взаимосвязь и повтор объектов, описываются с помощью ориентированных и неориентированных графов.

Приведем пример фасетной структуры на основе признаков петроглифов Северной Феноскандии. Примерами признаков для лосей/оленей, птиц и лодок являются следующие:

<u>Для лосей и оленей:</u>

1. голова
   - o удлиненная
   - o укороченная
   - o нормальная
2. уши
   - o наличие
   - o отсутствие
3. холка
   - o наличие
   - o отсутствие
4. рога
   - o отсутствие
   - o лося
   - o оленя
   - o ни те, ни другие
5. шея
   - o короткая
   - o длинная
   - o нормальная
6. шея
   - o утолщенная
   - o узкая
   - o нормальная
7. серьга на шее
   - o наличие
   - o отсутствие
8. корпус
   - o массивный
   - o грузный
   - o линейный
9. изгиб спины
   - o внутрь
   - o вне
   - o отсутствует

10. изгиб живота
    - o внутрь
    - o вне
    - o отсутствует
11. изгиб передней пары ног
    - o внутрь
    - o вне
    - o отсутствует
12. изгиб задней пары ног
    - o внутрь
    - o вне
    - o отсутствует
13. хвост
    - o отсутствует
    - o короткий
    - o удлиненный
14. животное обращено
    - o вправо
    - o влево

Для птиц:

1. Фигура
    - o реалистичная
    - o схематизированная
2. степень выбивки
    - o контурное с заполнением
    - o контурное без заполнения
    - o силуэтная
3. ориентация фигуры
    - o правая
    - o левая
4. лапы
    - o одна лапа
    - o две лапы
    - o нет лап
5. хвост
    - o наличие
    - o отсутствие
6. шея
    - o длинная
    - o короткая
7. шея
    - o прямая
    - o изогнутая
8. шея (угол между головой и нижней линией туловища)
    - o прямой
    - o тупой
    - o острый
9. клюв
    - o не выделен
    - o длинный
    - o короткий

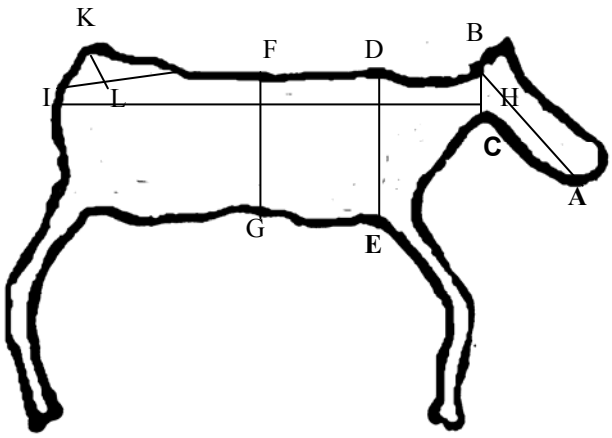
Для лодок:

1. Фигура
    - o силуэтные
    - o контурные
2. длина
    - o длинная
    - o короткая
3. борта
    - o высокие
    - o низкие
4. пассажиры
    - o наличие
    - o отсутствие
5. пассажиры
    - o в виде столбиков
    - o реалистично
6. Носовое украшение
    - o отсутствие
    - o в виде головы лесного животного
    - o в виде головы птицы
7. киль
    - o отсутствует
    - o один
    - o два
8. угол наклона форштвеня к корпусу лодки
    - o прямой
    - o тупой
    - o острый
9. корпус
    - o прямой
    - o изогнутый
    - o расширяющийся
10. ориентация фигуры
    - o правая
    - o левая

Возможно вычисление некоторых параметров изображения в полуавтоматическом режиме с использованием встроенного функционально программируемого калькулятора. Проиллюстрируем его работу на примере изображений лосей и оленей.

Определение некоторых признаков, таких как толщина шеи, тип головы, тип корпуса и т.д. визуально по рисунку не всегда является точным и часто зависит от эксперта. Для того, чтобы избежать двусмысленности восприятия предлагается использовать формулы отношения между длинами соответствующих отрезков.

На изображение петроглифа лося/оленя выделяют следующие отрезки, приведенные на рисунке: BA - длина головы, HI - длина корпуса, KL - длина хвоста, BC - толщина шеи, DE, FG - толщина передней и средней частей корпуса. Для



того, чтобы отметить эти отрезки на рисунке в программе, необходимо мышкой отметить точки начала и конца отрезков. После этого автоматически высчитываются длины этих отрезков. Чтобы определить значение признака, необходимо сравнить длины соответствующих отрезков. В этом случае формулы задаются пользователем. Имеется возможность использовать все арифметические операции, а так же операции сравнения. При выполнении всех действий учитывается вычислительная погрешность.

Используя подобный калькулятор происходит формализация восприятия изображения и упрощается ввод характеристик и они становятся более точными.

Выделение фрагментов изображения и их взаимосвязь происходит с помощью специального графического интерфейса при помощи мыши прямо в окне браузера. Данный признак описываются с помощью ориентированных и неориентированных графов.

Во второй группе признаков выделим:

- характеристики текстуры;
- характеристики цветовосприятия.

Одним из стандартных способов представления цветовой характеристики изображения является цветовые гистограммы. Для ее построения пространство всех цветов разбивается на подмножества так, чтобы схожие цвета попали в один интервал. Для каждого интервала подсчитывается количество пикселей, чей цвет принадлежит данной области. Для анализа гистограмм используются различные метрики, например, сумма модулей разностей значений элементов гистограмм для каждой цветовой области [3]. Вместо гистограммы можно брать вектор цветовой когеренции [2]. Другим вариантом представления является статистическая модель: рассматривается статистическое распределение различных цветовых каналов. Сравнение распределений является оценкой схожести [4]. Кроме того, можно рассматривать не только одномерные распределения, но и трехмерные, учитывая все взаимосвязи между каналами (ковариации). Рассматриваются интервалы наиболее часто встречающихся цветов, размеры одноцветных цветовых фрагментов, перевод цветных изображений в бинарное и их анализ.

Для анализа текстур, одним из применяемых методов является анализ независимых компонент. С его помощью выделяют фильтры, которые признаны отражать основные направления текстур для той базы изображений, на основе которой они строятся [3]. Кроме этого используется спектр фрактальной размерности Реньи [3, 4].

Возможно вычисление некоторых параметров изображения в полуавтоматическом режиме с использованием встроенного функционально программируемого калькулятора. Проиллюстрируем его работу на примере изображений лосей и оленей.

Определение некоторых признаков, таких как толщина шеи, тип головы, тип корпуса и т.д. визуально по рисунку не всегда является точным и часто зависит от эксперта. Для того, чтобы избежать двусмысленности восприятия предлагается использовать формулы отношения между длинами соответствующих отрезков.

5. Анализ признаков

С помощью специального модуля корреляционного анализа введенные признаки можно проверить на статистическую независимость с помощью критерия χ^2 Пирсона. Для этого выделяют признаки, группы объектов и задают уровень значимости. Кроме того, для анализа признаков используются методы описательной статистики.

6. Алгоритмы классификации и кластеризации

Для различных задач и типов признаков используются различные методы классификации и кластеризации. Для этого создаются модули статистического анализа документов. Рассмотрим первую группу признаков. Признак можно описать как $f : X \to D_f$, где D_f - множество допустимых значений признака, тогда, если заданы признаки $f_1,\ldots, f_n$, то вектор $x = (f_1(x),\ldots, f_n(x))$ называется признаковым описанием графического документа. Ввиду различия множеств допустимых значений для различных признаков для корректной работы алгоритма выполняется нормировка значений признаков: $\tilde{f}_j = \frac{f_j(x) - \min(f_j)}{\max(f_j) - \min(f_j)}$. В этом случае значение каждого признака будет лежать в пределах $[0,1]$. В качестве меры расстояния между документом и эталоном берётся эвклидово расстояние $(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$. Для классификации объектов коллекции на основе признаков применяются методы дискриминантного и кластерного анализов. Например, методом иерархического кластерного анализа (метод ближайшего соседа).

Классификация по цветовому и текстурному анализу может быть осуществлена несколькими методами. Примерами являются поиск по цветовым моментам и цветовым гистограмм [1]. Кроме того, существуют методы поиска нечетких дубликатов 12]. Поиск нечетких дубликатов позволяет предположить, являются ли два объекта частично

одинаковыми или нет. Частично одинаковые изображения могут образовывать один кластер. Кроме того, схожесть изображений по степени цветового восприятия может быть осуществлена на: сравнительной площадь белого, при наличии большого фрагмента определенного цвета и т.д.

7. Алгоритмы поиска изображений

Управление типами документов, а также наборами признаков, общих для всех графических документов коллекции и уникальных для каждого типа, позволяет хранить разнообразные данные о каждом графическом документе, организовать поиск документов по типу, признаку или набору признаков, задав точное значение или границы варьирования значения каждого признака и точность поиска (количество совпадений признаков для номинальных и категоризированных и интервалы изменения для количественных переменных).

Поиск по изображениям предназначен для поиска изображений, похожих на данное или на его фрагмент. На вход подается исследуемое изображение, а на выходе должны появиться изображения из базы данных, наиболее похожие на исходное. Для поиска похожих графических объектов на основе текстур изображений – методы нейронных сетей и геометрического программирования.

Рассмотрим его на примере наскальных изображений северной Фенноскандии. На сегодняшний день все новейшие материалы по петроглифам представляют собой набор цветных фотографий. Определенную сложность поиска создает фактическое отсутствие некоторых частей изображения. Поиск также осложняется тем, что часто невозможно определить, где верх, а где низ изображения. При этом, требование, что при поиске необходимо только совпадение контура изображения, позволяет упростить поиск, а значит, изображение петроглифа можно рассматривать, как бинарное (скале соответствует белый цвет, а петроглифу - черный). В зависимости от выбранных параметров поиска (точность поиска, процент совпадений элементов изображений) будет найдено одно или несколько изображений. Для поиска используются сеть адаптивного резонанса и структурный метод поиска.

В результате поиска, пользователю предоставляется доступ к информации о кодовом номере, месторасположении, характерных признаках найденного петроглифа и петроглифах, близких к нему по ранее описанным признакам.

8 Технические особенности реализации

Для создания и функционирования информационной системы использовалось свободно распространяемое программное обеспечение, и использование системы не нарушает лицензионных соглашений третьих сторон. В качестве WEB-сервера был использован сервер Apache/1.3.23, интерпретатор PHP 4.4.9, сервер баз данных MySQL 3.23.49. Одним из возникших вопросов был вопрос о средстве хранения графических документов. Было рассмотрено два варианта: хранение изображений в файлах и в базе данных в биде BLOB. Преимущество первого способа - более высокая скорость работы, второго - более простой контроль над целостностью данных и контролем доступа к ним. Были произведены вычислительные эксперименты, которые показали что, скорость чтения файлов сравнима со скоростью извлечения информации из БД (порядок 10^{-2} сек). При этом время обработки (масштабирования) изображения СБ2-модулем интерпретатора PHP имеет порядок 10^{-1} сек. Таким образом, ввиду незначительности потерь в скорости доступа при большем числе положительных сторон был выбран метод хранения в БД. В данной реализации используется тип MEDIUMBLOB позволяющий хранить до 16 мегабайт данных.

9 Преимущества разрабатываемой системы

Первой попыткой создания информационной системы была локальная версия [3]. С ее помощью была создана электронная коллекция петроглифов Карелии. В созданной ранее системе для навигации по каталогу групп петроглифов использовала дерево, жестко прописанное в программном коде системы. Для добавления новой группы необходимо было перекомпилировать продукт, что вызывало определенные трудности. Интерфейс программы содержал графические схемы групп, позволяя пользователю кликом мышки выбирать подгруппу или петроглиф и получать о них информацию. В системе использовались изображения схем в BMP формате, вручную подготовленные в графическом редакторе: необходимо было раскрасить области перехода на следующий уровень в оттенки красного цвета, так чтобы код цвета соответствовал коду подгруппы. Число возможных узлов на уровне было ограничено 255 возможными оттенками красного цвета, при изменении только составляющей красного цвета в палитре RGB.

Использование цветового кодирования представлялось невозможным при создании WEB-реализации информационной системы: размер BMP изображений схем достигал десятков мегабайт, что неприемлемо для загрузки с сервера даже при быстром Интернет-соединении. При конвертации изображения в другие форматы (JPEG/PNG) для сокращения размера происходит частичная потеря цветовой информации, и полученную схему невозможно использовать для навигации теми же средствами, что в локальной системе. Этот метод привносил трудности для добавления новых

разделов в дерево графических документов - возникала необходимость ручной раскраски схем. Решением данного вопроса стал отказ от цветового кодирования. Для навигации используются отмеченные прямоугольные области на изображении, а в административной части администратору системы предлагается графический интерфейс для разметки изображения с помощью мыши прямо в окне браузера. В базе данных сохраняются относительные координаты начала и конца области, что позволяет использовать навигацию при отображении схем в различных масштабах.

Заключение

В настоящее время, кроме создаваемой с помощью системы коллекции петроглифов Северной Фенноскандии, система апробируется на материалах Карельского государственного краеведческого музея при создании коллекции открыток и коллекции фотографий со строительства Беломорско-Балтийского канала.

Литература

[1] Васильева Н., Марков И. Синтез цветовых и текстурных признаков при поиске изображений по содержанию. // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. – Санкт-Петербург: НУ ЦСИ, 2008, С. 135-144.

[2] Кисель Я. Алгоритм поиска нечетких дубликатов в коллекции изображений. // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008. – Санкт-Петербург: НУ ЦСИ, 2008, С. 170-173.

[3] Рогов А.А., Рогова К.А., Спиридонов К.Н., Быстров М.Ю. Система поиска в электронной коллекции изображений петроглифов Карелии. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды 10 Всероссийской научной конференции "RCDL-2008"(Дубна, Россия 7-11 октября 2008г.). - Дубна: ОИЯИ, 2008. С. 246-251.

[4] Рогов А.А., Спиридонов К.Н. Применение спектра фрактальных размерностей Реньи как инварианта графического изображения. // Вестник Санкт-Петербургского университета. Сер. 10. 2008. Вып. 2. С. 30-43.

[5] Sticker M., Dimai A. Color Indexing with Weak Spatial Constraints. In Proceeding of the SPIE Conference, 1996.

The Information System for Graphic Documents Electronic Collections Creating and Administration

A.A. Rogov, K.A. Rogova, P.V. Kirikov, M.Yu. Bystrov

Creating information system allows to develop collections of graphic documents, to storage different sets of grahic (based on color perception and texture characteristics) and text parameters for each document, to carry out classification and search by different parameters combinations and by depictions similarity. The system is designed for Internet.

♣ Исследования поддержаны грантом РГНФ № 08-01-12116в (руководитель Н.В. Лобанова).