

механизмов логического и образного мышления, то есть логического вывода и функций поиска/распознавания.

Список литературы

- Кафтанников и др. 2003 — *Кафтанников, И. Л.* Перспективы использования онтологий в учебном процессе / И. Л. Кафтанников, С. Е. Корovin // *Educational Technology and Society*. — 2003. — № 6 (3).
- Исмаилова и др. 2004 — *Исмаилова, Л. Ю.* В поисках утраченной онтологии. Категорный подход / Л. Ю. Исмаилова, С. В. Косиков // *Технологии информационного общества — Интернет и современное общество* : тр. VII Всерос. объединенной конф. (Санкт-Петербург, 10–12 нояб. 2004 г.) — СПб. : Изд-во Филол. ф-та СПбГУ, 2004. — С. 56–61.
- Ной и др. — *Ной, Н. Ф.* Разработка онтологий 101: Руководство по созданию Вашей первой онтологии / Н. Ф. Ной, Д. Л. МакГиннесс ; Стэнфордский ун-т. — Стэнфорд (Калифорния). — 94305.
- Овдей и др. 2004 — *Овдей, О. М.* Обзор инструментов инженерии онтологий / О. М. Овдей, Г. Ю. Проскудина // *Электронные библиотеки: перспективные методы и технологии, электронные коллекции* : тр. 6-ой Всерос. науч. конф. — RCDL2004. — Пушкино, 2004.
- Кучуганов и др. 2001 — *Кучуганов, В. Н.* Система визуального проектирования баз знаний / В. Н. Кучуганов, И. Н. Габдрахманов // *Информационные технологии в инновационных проектах* : тр. III междунар. науч.-техн. конф. — Ижевск, 2001. — С. 140–143.
- Кучуганов 2002 — *Кучуганов, В. Н.* Семантика графической информации // *Известия ТРТУ* : Тематич. вып. «Интеллектуальные САПР» : материалы междунар. науч.-техн. конф. — Таганрог : Изд-во ТРТУ, 2002. — № 3 (26). — С. 157–166.
- Кучуганов 2005 — *Кучуганов, В. Н.* Визуальное моделирование текстов / В. Н. Кучуганов // «Интеллектуальные системы» (AIS'05) и «Интеллектуальные САПР» (CAD–2005) : тр. междунар. науч.-технич. конф. — М. : ФИЗМАТЛИТ, 2005. — Т. 4. — С. 104–114.
- Непейвода 1989 — *Непейвода, Н. Н.* Некоторые семантические конструкции конструктивных логик схем программ / Н. Н. Непейвода // *Вычислительные системы*. — Вып. 129. — Новосибирск, 1989. — С. 49–66.
- Kuchuganov 1994 — *Kuchuganov, M.* A predicate logic of well-founded actions / M. Kuchuganov // *Lecture Notes in Computer Science*. — Vol. 813. — 1994. — P. 218–226.

Распознавание рукописных текстов

А. В. Кучуганов, Г. В. Лапинская
Ижевский государственный технический университет, Ижевск,
Россия

The problem of text recognition in documents processing and information input into the computer is rather actual. Our work offers an approach to the problem solution of hand-written text recognition, consisting of that the scanned sample of the hand-written text vectorized, initial elements of the received image are represented as the count which elements are compared to the reference samples stored in base.

The described model is incorporated in the working version of program HWTR (hand-written text recognition) for recognition of hand-written text.

При обработке документов и вводе информации в компьютер весьма актуальной является задача распознавания текста. Распознавание печатных текстов компьютером — область, сегодня достаточно хорошо исследованная. Что же касается распознавания рукописного текста, то качественный рывок еще впереди. Хотя уже есть несколько отработанных направлений. Во-первых, это системы распознавания форм, заполненных печатными буквами от руки, которые применяются во многих областях. Во-вторых, это распознавание отдельных рукописных букв, написанных особым пером на специальном экране (touch-screen), которое широко применяется в карманных компьютерах и электронных записных книжках. Эти распознающие системы демонстрируют достаточно высокую точность, приближающуюся к точности клавиатуры.

Чтение компьютером слитных букв, то есть обычного письма, сегодня мало разработано. Но исследовательские проблемы, которые надо решить на этом пути, чрезвычайно интересны.

Типичная задача автоматического распознавания образов формулируется примерно так: все множество подлежащих обработке изображений некоторым способом разбивается на конечное число классов, называемых образами [Бутакон 1987]. Автоматическому устройству, снабженному механизмом восприятия образов, предъ-

является изображением. Устройство должно решить, к какому классу оно относится.

Представлялось вполне естественным, что поскольку необходимо устройство, принимающее решение, то алгоритмы его работы должны опираться на статистическую теорию решения. Объекты каждого класса (образы) характеризовались некоторой совокупностью признаков, каждый из которых мог принимать одно из определенного множества значений. Возникла задача: с помощью семейства поверхностей разбить пространство признаков на непересекающиеся области и сопоставить каждую область одному из образов. Если это сделано, то процесс распознавания представляется достаточно простым: для предъявленного изображения нужно вычислить значения всех признаков, иными словами, определить некоторую точку в пространстве признаков; затем выявить, в какую из выделенных областей попадает точка. Изображение относится к тому образу, которому соответствует найденная область.

Традиционно существует три метода распознавания (или типа классификаторов): шаблонный, признаковый и структурный [Абраменко 2000]. Шаблонные классификаторы преобразуют исходное изображение символа в набор точек и затем накладывают его на шаблоны, имеющиеся в базе системы. Шаблон, имеющий меньше всего отличий, и будет искомым. У этих систем достаточно высокая точность распознавания дефектных символов (склеенных или разорванных). Недостаток — невозможность распознать шрифт, хоть немного отличающийся от заложенного в систему (размером, наклоном или начертанием). Признаковые классификаторы по каждому символу вычисляют набор чисел (признаков) и сравнивают эти наборы. Но так как набор признаков никогда полностью не соответствует объекту, то заведомо часть информации о символе будет теряться. Структурные классификаторы хранят информацию о топологии символа. Этот способ тоже имеет свои недостатки: как только вы представите «разорванную» из-за дефектов печати букву, она уже не подойдет под свое описание.

В структурно-пятенном эталоне изображение представляется в виде пятен, связанных парными отношениями. Наглядно это можно представить себе в виде теннисных шаров, нанизанных на резиновый жгут. Шары могут сдвигаться относительно друг друга.

Такую связку подвижных шаров можно «натянуть» на различные изображения одного символа, и система становится менее зависимой от шрифтов и дефектов. Технология распознавания с помощью структурно-пятенных эталонов получила название «фонтанное преобразование» (от английского font — шрифт).

Проанализировав возможности компьютерного анализа рукописного текста, мы пришли к выводу, что его перед обработкой необходимо векторизовать. Для этого была создана программа AQRResearcher, которая предназначена для перевода растрового изображения рукописного текста в векторное, при этом информация о геометрических примитивах записывается в файл, который в последствии будет обрабатываться [Кучуганов 2001: 227–230].

После обработки изображения рукописного текста с помощью программы AQRResearcher мы получили представление введенного текста в виде отрезков прямых и дуг. Для анализа этих данных была создана база эталонов, в которой хранится информация о каждой букве в отдельности.

Сущность разработанного нами подхода к распознаванию заключается в следующем: исходные элементы изображения представляются в виде графа [Кучуганов 1985], который описывается следующими параметрами: координаты начала элемента; координаты конца элемента; форма элемента (дуга, прямая); направление прямой или направления векторов от центра к концам дуги, которые определяются в зависимости от наклона почерка и могут принимать восемь значений, представленных на рис. 1.

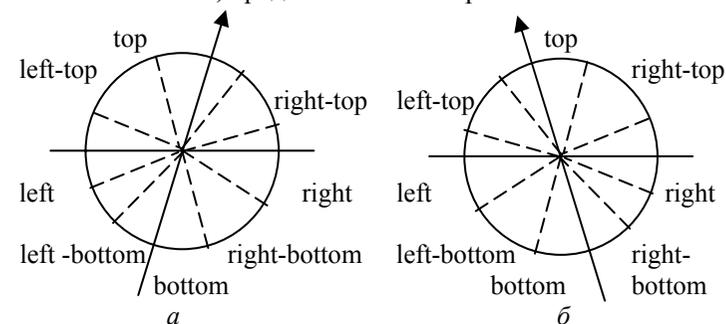


Рис. 1. Значения направления углов векторов от центра к концам дуги: a — для правонаклонного почерка, b — для левонаклонного почерка

Для представления в памяти компьютера этот граф удобно за-
дать двумя матрицами:

$$M_1 = \begin{pmatrix} \varphi_{11}, & \dots, & \varphi_{l1} \\ \varphi_{12}, & \dots, & \varphi_{l2} \\ \dots & \dots & \dots \\ \varphi_{1n}, & \dots, & \varphi_{ln} \end{pmatrix}; M_2 = \begin{pmatrix} AH_{11}, & \dots, & AH_{k1}, & AK_{11}, & \dots, & AK_{m1} \\ AH_{12}, & \dots, & AH_{k2}, & AK_{12}, & \dots, & AK_{m2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ AH_{1n}, & \dots, & AH_{kn}, & AK_{1n}, & \dots, & AK_{mn} \end{pmatrix},$$

где каждой строке матрицы M_1 соответствует строка матрицы M_2 .

Каждая строка матрицы M_1 описывает свойства одной вершины графа исходных элементов текста с помощью некоторых параметров $\varphi_1, \dots, \varphi_l$. В матрице M_2 AH и AK — адресная часть строки. Здесь записываются номера строк матрицы M_1 , в которых находятся вершины, связанные ребром с началом и концом данной вершины соответственно. Можно считать, что в AH записаны адреса ребер, входящих в данную вершину, а в AK — исходящих.

Эталонные образцы хранятся в базе тоже в виде графов. Для сравнения известного графа с анализируемым графом строки рукописного текста мы строим параллельно анализируемому графу точно такой же граф, но дополненный двумя параметрами-метками. Первый параметр будет использоваться для обозначения совпавшей вершины из эталонного графа с вершиной из анализируемого графа, а второй — вершины, которые соответствуют соседним вершинам по отношению к совпавшей, чтобы потом сравнивать только с соседями совпавших вершин.

Таким образом, просматривая эталонные образцы и граф анализируемого текста и делая необходимые пометки при совпадении, мы осуществляем распознавание изображения рукописного текста.

Подводя итог проделанной работе, следует отметить, что в разрабатываемой нами программе задается доля вероятности совпадения признаков анализируемого и исходного графов. Это позволяет корректировать процесс распознавания с учетом влияния внешних факторов, таких, как размер почерка, качества бумаги и т. д. Конечно, несмотря на это, существует вероятность того, что мы не сможем распознать какие-то символы рукописного текста.

Чтобы продемонстрировать механизм работы программы, мы удалили из базы эталонов все образцы исходных букв, кроме буквы «и». Эта буква содержит несколько разветвлений и представля-

ет собой достаточно сложный образец для распознавания. После обработки данной строки в программе AQRsearcher данные были сохранены в файл arroxim0.tte. После анализа данных этого файла в нашей программе было обнаружено две буквы «и», остальной текст был заменен знаками «_», так как не был распознан. Для обнаружения ошибок или неточностей работы программы присутствует механизм визуальной отладки, то есть все, что распознается, выделяется на исходном рисунке. Результаты работы программы представлены на рис. 2.



Рис. 2. Пример работы программы распознавания рукописного текста

Ни для кого не секрет, что в настоящее время идет бурное развитие в сфере распознавания рукописных текстов. Такие всемирно известные компании, как Nintendo, Zi Corporation, Paragon Software, Hewlett-Packard, ABBYY, Parascript, LLC и другие, борются за право лидерства в этой области. Разрабатывая данную программу, мы поставили перед собой цель исследовать алгоритмы, приближающиеся по своим возможностям в плане гибкости и параллельной обработки к механизмам зрения биологических систем, в частности, путем разделения и оперативного взаимодействия функций зрительного анализатора и логического (мыслительного) уровня.

Планируется разработать систему адаптации программы к почерку конкретного человека. Это позволит создавать уменьшенные базы эталонов, что, в свою очередь, ускорит процесс распознавания.

Список литературы

- Абраменко 2000 — Абраменко, А. Компьютер читает [Электронный ресурс]. — Электрон. текстовые дан. — 2000. — Режим доступа: <http://www.osgai.narod.ru/>, свободный. — Загл. с экрана.
- Бутаков и др. 2000 — Бутаков, Б. А. Обработка изображений на ЭВМ / Б. А. Бутаков, В. Н. Островский, И. Л. Фадеев — М.: Радио и связь, 1987. — 240 с.
- Кучуганов 2001 — Кучуганов, А. В. Метод адаптивной векторизации фотоизображений / А. В. Кучуганов // 11-й междунар. конф. по

компьютерной графике и машинному зрению Графикон : сб. тр. (Н. Новгород, 10–15 сент. 2001 г.). — Н. Новгород : ННГАСУ, 2001. — С. 227–230.

Кучуганов 1985 — Кучуганов, В. Н. Автоматический анализ машиностроительных чертежей / В. Н. Кучуганов. — Иркутск, 1985.

Базы данных и корпуса текстов средневекового французского языка: подходы, проекты, технологии

А. М. Лаврентьев

Институт филологии Сибирского Отделения РАН, Новосибирск,
Россия
Ecole Normale Supérieure Lettres et Sciences humaines, Лион,
Франция

In this paper we will give a review of the current projects on the Medieval (9th — 15th centuries) French text databases, corpora and electronic editions. We will present: different choices of composition and sources for such databases (full text vs. sample corpora; using critical editions vs. original manuscripts); the principles of markup and the technologies used. Two projects will be considered in more detail: the Base de Français Médiéval (<http://bfm.ens-lsh.fr>) and the Charrette project (<http://lancelot.baylor.edu>)

История текстовых баз данных на французском языке восходит к концу 1950-х гг., когда была начата работа над подготовкой словаря «Тезаурус французского языка» (Trésor de la langue française), охватывающего период с XVI по XX век. Текстовая база, получившая название Frantext, активно развивалась в 60-е — 80-е гг., с начала 90-х гг. появилась возможность работы с ней с использованием CD-ROM, а затем и через Интернет. База Frantext приобрела широкую известность и стала активно использоваться лингвистами, литературоведами и историками. В то же время медиевисты оставались лишенными подобного исследовательского инструмента. Между тем тексты являются практически единственным источником данных о языке и литературе средневековья, и использование информационных технологий при их изучении позволило бы получить принципиально новые результаты.

В конце 70-х и 80-х годах возникают первые проекты компьютерного изучения средневековых французских текстов, однако полученные данные долгое время остаются доступными лишь их