УДК 621.374

ПОВЫШЕНИЕ НАДЕЖНОСТИ СПОСОБА ОПРЕДЕЛЕНИЯ ТИПОВ ПЕЧАТНЫХ СИМВОЛОВ РАЗНЫХ ШРИФТОВ

Пинт Э.М., Петровнина И.Н., Романенко И.И., Еличев К.А.

*ФГБОУ ВО «Пензенский государственный университет архитектуры и строительства», Пенза,
e-mail: office@pguas.ru*

Рассмотрено устройство, обеспечивающее высокую надежность (достоверность) определения видов и типов символов печатных текстов. Литературный обзор способов идентификации видов и типов символов показал, что современные разработки позволяют распознавать ограниченный объем определяемых типов символов и замыкаются в пределах одного шрифта (образа). Цель исследования: создание способа, дающего возможность с достаточной точностью устанавливать вид печатного символа вне зависимости от качества печати текста, дефектов, художественных украшений и размытости образа. Разработанное устройство базируется на персональном компьютере, объединенном с системой ввода информационной базы. Система ввода имеет устройство, преобразующее образную информацию с носителя в виде электрических сигналов-импульсов, поступающих в банк. Банк состоит из кодированных ячеек-звеньев, каждая из которых устанавливает определенную часть образа (изображения). Найдена оптимальная комбинация направлений, образуемая частями изображения символа, отражающая необходимые и достаточные признаки для его распознавания. Начиная с концевой заполненной ячейки изображения символа, находящегося в запоминающем устройстве, осуществляется окантовка изображения с целью выделения направлений, входящих в найденную комбинацию направлений. Образуется код символа, состоящий из номеров направлений, который преобразуется в ячейку, что позволяет определить тип символа. Представленные методы распознавания образов способствуют повышению достоверности определения типов печатных символов. Выбрана комбинация направлений, отражающая признаки печатных символов, которые выявлялись после окантовки формы символа. Определено количество ячеек-звеньев матрицы запоминающего устройства. Установлены правила для стирания содержимого ячеек-звеньев, определяющих лишнюю толщину элементов изображения символа. На примерах изложены методики исправления измененной ориентации частей символа и ликвидации дефектов в изображениях символов. Реализация предложений повышения надежности определения типов печатных символов позволила устанавливать типы символов, несмотря на разные шрифты, на изменения размеров и ориентации частей символа, на наличие декоративных украшений и дефектов.

Ключевые слова: печатный символ, компьютер, программа, направления, комбинация направлений, окантовка, код символа

IMPROVING THE RELIABILITY OF THE METHOD FOR DETERMINING TYPES OF PRINTED SYMBOLS OF DIFFERENT FONTS

Pint E.M., Petrovnina I.N., Romanenko I.I., Elichev K.A.

Penza State University of Architecture and Construction, Penza, e-mail: office@pguas.ru

A device is considered that provides high reliability (reliability) of determining the types and types of characters of printed texts. A literature review of the methods for identifying types and types of characters showed that modern developments allow us to recognize a limited amount of defined types of characters and are closed within a single font (image). The purpose of the study: to create a method that makes it possible to establish the appearance of the printed symbol with sufficient accuracy, regardless of the print quality of the text, defects, art decorations and blurriness of the image. The developed device is based on a personal computer integrated with the information base input system. The input system has a device that converts figurative information from the carrier in the form of electrical signals-pulses entering the bank. The bank consists of coded cell units, each of which establishes a certain part of the image (image). The optimal combination of directions, formed by the parts of the symbol image, reflecting the necessary and sufficient signs for its recognition, is found. Starting from the terminal filled cell of the image of the symbol located in the storage device, the image is edged to select the directions included in the found combination of directions. A symbol code is formed, consisting of direction numbers, which is converted into a cell, which allows you to determine the type of symbol. The presented methods of pattern recognition contribute to increasing the reliability of determining the types of printed characters. A combination of directions is selected that reflects the signs of the printed characters that were detected after the edging of the shape of the character. The number of cells-links of the matrix storage device is determined. Rules are established for erasing the contents of cell units that determine the excess thickness of the image elements of a symbol. The examples describe methods for correcting the changed orientation of parts of a symbol and eliminating defects in symbol images. The implementation of proposals to improve the reliability of determining the types of printed characters allowed establish types of characters, despite different fonts, on changes in the size and orientation of parts of the character, on the presence of decorative ornaments and defects.

Keywords: printed symbol, computer, program, directions, combination of directions, edging, symbol code

Попытки создания устройства, определяющего типы печатных символов, предпринимались в разное время. Распознающее устройство может использоваться для ускоренной обработки печатной информации, для поиска нужной литературы, для анализа печатных текстов и т.д. Устройство

должно выполнять две функции: первая функция – восприятие печатных символов с носителя текстов, вторая – определение типов печатных символов.

Реализация первой функции читающего устройства особых затруднений не вызывает.

При реализации второй функции главной проблемой является то, что устройство должно распознавать типы печатных символов не одного, а различных шрифтов, то есть оно должно определять один и тот же печатный символ (букву или цифру), но напечатанный с разным масштабом (высотой и шириной), с разным наклоном частей символа, с более жирными или более узкими частями символа, с возможными дефектами изображения символа и т.д.

До настоящего времени решение задачи определения типов символов печатных текстов, когда символы различных шрифтов значительно отличаются по своей форме и способу написания, до конца не завершено. Цель проведенных исследований заключается в создании устройства, которое позволило бы с высокой надежностью определять типы печатных букв и цифр с перечисленными выше особенностями формы символов.

В ходе анализа литературных источников установлено существование шести способов определения типов символов печатных текстов. Первый способ основан на определении типов печатных символов с использованием трафаретов. Было предложено устройство [1, 2], в котором изображение печатной цифры проектировалось на диск, содержащий вырезанные трафареты цифр, за которым был расположен фотоэлемент. При полном затемнении фотоэлемента переключалось реле и определялся тип цифры.

При использовании второго способа символы определялись с помощью специальных отметок или стилизованной формы. Типы цифр устанавливались созданной машиной [3–5], при этом цифры изображались в виде вертикальных полосок. Для реализации третьего способа авторы исследований [6–8] выделяли следующие особенности элементов формы символов: количество прямых и кривых частей, расположение этих частей друг относительно друга и т.д. (рис. 1, а). Кроме

того, для определения печатных символов подсчитывалось количество пересечений изображения символа линиями, идущими по горизонтали, расположенными на разной высоте, определялось число концевых точек изображения и т.д.

Четвертый способ определения типов печатных символов основан на использовании наличия или отсутствия черного поля изображения символа в данном фиксированном месте. Созданная машина [4, 9, 10] определяет типы цифр при делении изображения печатной цифры на пять частей и фиксации наличия или отсутствия в частях черного поля.

Для реализации пятого способа автор работы [5, 11, 12] использовал интегральные характеристики формы печатных символов. Так, в разработанном устройстве для определения типов печатных символов использовались проекции площади изображения печатного символа на горизонтальную (рис. 1, б) и вертикальную оси.

Наконец, типы печатных символов определялись с помощью обучающихся электронно-вычислительных машин (ЭВМ) [6]. При этом изображения печатных символов заносятся в запоминающее устройство ЭВМ с помощью рядов фотоэлементов. Алгоритм обучения ЭВМ состоял из следующих частей:

1. Проведение плоскостей, разделяющих один тип символа относительно другого типа

для предъявленных машине символов (рис. 2, а, б, в).

2. Выбрасывание лишних плоскостей (рис. 2, г).

3. Выбрасывание лишних кусков плоскостей (рис. 2, г).

4. Определение типов новых символов.

Все перечисленные выше способы определения типов печатных символов в настоящее время не используются, так как позволяют определять лишь ограниченное число типов символов, в лучшем случае в пределах одного шрифта.

Рис. 1. Вид элементов фигуры, используемых для различения символов по элементам формы:
а – третий способ определения типов символов печатных текстов;
б – пятый способ определения типов символов печатных текстов на основе закона изменения проекции площади фигуры символа при проектировании ее на горизонтальную ось

Рис. 2. Пояснения алгоритма обучения ЭВМ

Рис. 3. Выбор направлений для определения типов печатных символов:
а – найденная схема направлений с числами для перехода на сопредельные ячейки;
б – передвижение по любому из восьми направлений из заполненной ячейки на соседние

Материалы и методы исследования

Разработанное устройство состоит из компьютера, воспринимающей, распознающей и запоминающей частей (ВРЗЧ). Совокупность электрических импульсов, снимающих образ с печатного носителя, поступает в установленном порядке в прямоугольную матрицу считывающего устройства. Матрица разбита на кодированные ячейки и каждому образу соответствует определенная совокупность ячеек в матричной оболочке [13].

После анализа были определены необходимые и достаточные признаки для определения типов печатных символов [14]. Ими оказались направления (1–8), входящие в выбранную комбинацию направлений (рис. 3, а). Направления комбинации образуются элементами контуров печатных символов. По разработанной для компьютера программе изображение символа, находящееся в матрице компьютера, начиная с концевой точки, обходится по контуру символа в соответствии с выбранной комбинацией направлений. С целью определения основного направления соседние заполненные ячейки матрицы опрашиваются по выбранным направлениям комбинации, начиная с заполненной концевой ячейки

изображения символа. Главное направление устанавливалось числом заполненных ячеек-звеньев (то есть вес), превышающим число ячеек, определяющих наибольшую толщину отдельных частей напечатанного знака. Согласно заданному алгоритму программы, происходит переход с помощью чисел (рис. 3, а) на соседнюю заполненную ячейку в найденном основном направлении, определяя параметр (номер) основного направления и т.д. В дальнейшем процесс повторялся.

В итоге образуется код печатного символа, состоящий главным образом из последовательности номеров основных направлений. Таким образом, код символа создавался при обходе по главным направлениям фигуры символа. Переход от одной заполненной ячейки на другую соседнюю осуществлялся следующим образом. Пусть заполненная концевая ячейка изображения символа (или другая заполненная ячейка) имеет адрес 001001 (рис. 2, б).

Переход от данной ячейки на соседнюю, например, в направлении 1 совершается за счет добавления к адресу ячейки 001001 числа 001000 (рис. 2, б). Образуется число 010001, то есть определяется адрес соседней заполненной ячейки матрицы в направлении 1, на которую и происходит

переход. Таким образом, код символа формируется из номеров основных направлений и определенных неосновных.

В дальнейшем производилось сравнение сокращенного (упрощенного) кода символа с эталонными кодами. Эталонный код, то есть ряд направлений, создаваемый путем обхода по направлениям элементов образцового очертания знака, был создан для каждого типа печатного символа. В процессе сравнения определялось количество неидентичных номеров для каждого эталонного кода. Тип печатного символа определялся по минимальному количеству несопадений номеров [13–15].

Для обеспечения высокой надежности распознавания (определения) типов печатных знаков (символов) были приняты следующие меры.

1. Для составления кода печатного знака использовались необходимые и достаточные признаки (определенные направления, образуемые частями изображения печатного символа), позволяющие с высокой надежностью устанавливать типы и виды печатных символов разных шрифтов.

2. В процессе определения типа печатного символа по разработанному алгоритму производилось так называемое стирание по толщине, то есть приведение толщины отдельных частей знака к наименьшей в одну элементарную ячейку матрицы, и таким образом символы с разной толщиной частей (в определенных пределах) определялись в равной степени.

3. Образы печатных символов с помощью воспринимающей системы разработанного устройства заносились в определенную матрицу – звено ВРЗЧ устройства персонального компьютера, включающую 32×32 ячеек и обеспечивающую распознавание всех типов печатных символов разных шрифтов, кроме букв «ш» и «щ».

4. Полученный после окантовки изображения символа его код упрощался, то есть происходила ликвидация неосновных направлений, возникающих из-за изменения ориентации символа в целом или его отдельных частей из-за возможных дефектов, из-за наличия декоративных украшений в символах, из-за определенных шрифтов и т.д.

5. Эталонные коды составлялись с учетом характерных признаков печатных символов разных шрифтов. В эти признаки входили главным образом основные направления, а также определенные суммы, возникающие при переходе во время окантовки изображения символа от его одной части к другой.

6. Определение типов печатных знаков происходило путем сравнения номеров на-

правлений упрощенного кода символа с номерами направлений всех эталонных кодов.

Совпавшие номера ликвидировались, а несовпавшие номера в эталонном коде и в упрощенном коде символа суммировались для каждого эталонного кода, и по наименьшему количеству несовпавших номеров направлений определялся тип и вид печатного образа. Для оценки достоверности установления печатных букв рекомендуется зависимость, учитывающая возможность появления в печатных текстах букв P_i :

$$P = \frac{m}{nS} \sum_{i=1}^s \frac{r}{n} P_i,$$

где m – число правильно прочтенных букв; S – число предназначенных для опознавания букв алфавита; n – количество проверок каждой буквы; r – число безошибочно опознаваемых букв для одного смыслового символа (то есть для одного типа буквы).

Все печатные буквы алфавита точно распознавались разработанным устройством друг относительно друга, за исключением букв «ш» и «щ», максимальная надежность идентификации которых $P_{\max}$ составила 0,98.

Результаты исследования и их обсуждение

В процессе экспериментальных исследований анализировались два компьютерных шрифта. У одного из них печатные символы имели дополнительные украшения формы, а символы другого шрифта их не имели. Была выбрана комбинация направлений (рис. 3, а), отражающая существенные и необходимые признаки печатных символов, которые выявлялись после окантовки формы символа и по которым определялись типы символов.

По результатам исследований определено количество ячеек-звеньев прямоугольной матрицы ВРЗЧ устройства компьютера с учетом размеров заглавных символов и возможных смещений по вертикали и горизонтали изображений символов (32×32 ячеек-звеньев).

Разработаны следующие правила для стирания содержимого ячеек, формирующих лишнюю толщину элементов изображения символа:

– при обходе по направлениям «2», «4», «6», «8» содержимое заполненных ячеек по направлениям «1» и «5» ликвидируется, что устраняет лишнюю толщину наклонных линий символа;

– при окантовке по направлениям «3» и «7» происходит исчезновение содержимого ячеек по направлениям «1» и «5», что

удаляет лишнюю толщину вертикальных линий символа;

– при окантовке по направлениям «1» и «5» содержимое ячеек по направлениям «3» и «7» исчезает, что устраняет лишнюю толщину горизонтальных линий символа.

На рис. 4 представлена окантовка по направлениям букв «Д» различных шрифтов, включая перекрестие; изображены ячейки матрицы, составляющие лишнюю толщину частей знака, содержимое которых ликвидируется при обходе знака по контуру.

Было отмечено, что при обходе контура знака в качестве главного направления избиралось то направление, которое имело количество ячеек-звеньев большее или равное «весу». Вес (количество ячеек) хотя бы на одну ячейку должен быть больше количества ячеек, образующих толщину частей символа. После завершения окантовки контура символа кодовая запись знака, представляющая собой определенную комбинацию номеров направлений, преобразуется для того, чтобы исключить присутствие таких дефектов, как каверны в изображении символа, размыты, несвойственные отклонения вертикальных, горизонтальных, наклонных линий знака.

С целью устранения вышеперечисленных дефектов разработан следующий алгоритм оптимизации. Из кода символа исключаются номера неосновного направления в том случае, если при окантовке до и после

них располагаются номера определенно-го основного направления или определенной суммы.

На рис. 5, I, представлена окантовка по направлениям отклоненных линий знака: горизонтальной (а), наклонной (б), вертикальной (в). В результате окантовки отклоненной горизонтальной линии (рис. 5, I, а) шифр линии представляет собой комбинацию номеров 118181118111 (вес равен 4, что соответствует трем номерам основного направления). После выполнения упрощения код линии выглядит как 111111111, что соответствует идеальной горизонтальной линии.

После окантовки отклоненной наклонной линии (рис. 5, I, б) код линии представляет собой комбинацию номеров 221222121222, а после операции упрощения – 222222222, что равнозначно идеальной наклонной линии. И в результате окантовки отклоненной вертикальной линии (рис. 5, I, в) код линии представляет собой ряд номеров 332332333233, а после операции упрощения – 333333333, что соответствует оптимальной вертикальной линии.

Из-за низкоккачественной печати либо непропечатки различных частей печатного символа, могут оказаться размытыми изображения или появиться такие дефекты, как каверны. Следовательно, необходимо рассмотреть возможность устранения таких дефектов. На рис. 5, II, представлено изображение буквы «Г» с размытиями.

XX – ячейки толщины, содержимое которых уничтожается при окантовке очертания

Рис. 4. Окантовка по направлениям букв «Д» различных шрифтов

Рис. 5. Виды окантовки символов: I – окантовка по направлениям отклоненных линий знака; II – окантовка по направлениям буквы «Г» с размывами линий

В результате окантовки буквы «Г» формируется следующий код буквы: 33242343323334311218121117. После проведения упрощения код буквы представляет собой комбинацию номеров: 333333333311111117. Преобразованный код после сравнения со своим стандартным видом «31» не дает расхождений. (При вычислении числа несовпадений, направление 7 в расчет не принимается, так как количество его номеров меньше веса.) Поэтому рассматриваемые размывы линий знака не влияют на его опознавание.

Заключение

Для надежного определения разработанным устройством типов печатных символов выбраны существенные и достаточные признаки, определено оптимальное количество ячеек-звеньев матрицы ВРЗЧ устройства компьютера.

Для дальнейшего повышения надежности предусмотрены и реализованы меры, позволяющие распознавать типы печатных знаков, вне зависимости от вида шрифта, изменения в определенных пределах размеров символов и ориентации символа в целом или его отдельных частей, от наличия декорирования и определенных дефектов.

Таким образом, разработано новое устройство, которое с высокой надежностью и большим быстродействием решает проблему функции распознавания типов печатных символов. Разработанное устройство позволяет автоматизировать процесс обработки печатной информации.

Список литературы

1. Сочивка В.П. Электронные опознающие устройства. М.: Энергия, 1964. 56 с.

2. David H. Direct Reading for Data Processing. Process Control and Automation. 1960. Vol. 7. No. 3. P. 3–30.
3. Seema A., Rajeshwar D. Pattern Recognition in an Electronic Reader. International Journal of Computer Science and Telecommunications. Vol. 3. Issue 8. August 2012. P. 25–29.
4. Nielsen W.F. System Analysis Using Digital Computers. Electronic industries. 1960. Vol. 19. No. 8. August. P. 212–214.
5. Романов В.П. Интегральные методы опознавания. Конференция по обработке информации, машинному переводу и автоматическому чтению текста / Акад. наук СССР. Ин-т науч. информации. М., 1961. Т. 60. 23 с.
6. Ковалевский В.А. Корреляционный метод распознавания изображений // Журнал вычислительной математики и математической физики. 1962. № 4 (2). С. 684–694.
7. Pradeep J., Srinivasan E., Himavathi S. Diagonal Feature Extraction Based Handwritten Character. International Journal of Computer Applications (0975 – 8887). 2010. Vol. 8. No. 9. P. 17–21.
8. Pradeep J., Srinivasan E., Himavathi S. IJCSIT, Diagonal based feature Extraction for handwritten alphabets recognition system using neural network. International Journal of Computer Science & Information Technology (IJCSIT). 2011. Vol. 3. No. 1. P. 27–37.
9. Sharma N., Patnaik T., Kumar B. Recognition for Handwritten English Letters. International Journal of Engineering and Innovative Technology: A Review. 2013. Vol. 2. Iss. 7. P. 318–321.
10. Singh V., Kumar B., Patnaik T. Feature Extraction Technique for handwritten Text in Various Script. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231–2307. March 2013. Vol. 3. P. 238–241.
11. Woodford C. Optical character recognition (OCR). [Electronic resource]. URL: <http://www.explainthatstuff.com/how-ocr-works.html> (date of access: 23.04.2020).
12. Schroff F., Kalenichenko D., and Philbin J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. P. 815–823.
13. Пинт Э.М., Петровнина И.Н., Романенко И.И., Еличев К.А. Система восприятия и передачи изображений печатных символов с оригинала в матрицу запоминающего устройства компьютера, определяющего типы символов // Инженерный вестник Дона. 2019. № 1. [Электронный ресурс]. URL: <http://www.ivdon.ru/magazine/archive/n1y2019/5591> (дата обращения: 22.04.2020).
14. Пинт Э.М., Петровнина И.Н., Еличев К.А. Анализ формы печатных знаков для выделения существенных признаков и определения знаков компьютером // Нива Поволжья. Пенза: Изд. ПГАУ, 2018. № 3. С. 112–119.
15. Пинт Э.М., Петровнина И.Н., Еличев К.А. Математическая обработка результатов исследования читающего устройства // Наукоедение. 2015. Т. 7. № 1. [Электронный ресурс]. URL: <http://naukovedenie.ru/index.php?p=vol7-1> (дата обращения: 22.04.2020).