**Лабораторная работа №1**.

**Распознавание символов с помощью функций расстояния.**

Системы оптического распознавания символов (OCR – Optical Character Recognition) предназначены для перевода растрового изображения в текст. Для получения растрового изображения страницы текста обычно используется монохромный сканер с разрешением от 300 dpi, который позволяет получить вполне удовлетворительное с точки зрения распознавания растровое изображение. Программа распознавания должна преобразовать это изображение в текстовый формат, т. е. присвоить каждому символу некоторый код.

Затем производится последовательное распознавание отдельных символов путем сравнения их с имеющимися эталонными образами, полученными ранее в процессе обучения системы. В результате распознаваемому символу присваивается код того эталона, с которым он совпадает наилучшим образом. Все алгоритмы распознавания печатных символов можно разделить на два класса: шрифтовые и безшрифтовые. Шрифтовые алгоритмы используют априорную информацию о шрифте (тип, размер), которым напечатаны буквы, т. е. программе OCR на этапе обучения необходимо предъявить полноценную выборку текста, напечатанного данным шрифтом.

Программа измеряет и анализирует различные признаки каждого символа шрифта и заносит их в свою базу эталонных образов. По окончании этого процесса программа готова к распознаванию символов данного конкретного шрифта. Бесшрифтовые алгоритмы не имеют априорных знаний о шрифте символов, поступающих к ним на вход, т. е. измеряют и анализируют различные признаки, присущие печатным символам как таковым независимо от типа и размера шрифта. В предельном случае для бесшрифтового алгоритма процесс обучения может отсутствовать и характеристики символов задает человек. Более часто база эталонных образов создается в процессе обучения программы на выборке реальных символов. Результаты экспериментов показывают, что разные изображения одного и того же печатного символа практически всегда отличаются друг от друга, причем степень отличия зависит как от качества печатной продукции, так и от технических параметров сканера (число градаций яркости, величина разрешения). В лабораторной работе эти отличия имитируются равномерным зашумлением, уровень которого можно регулировать. Вероятность правильного распознавания символа зависит как от выбранного метода распознавания, так и от степени различия эталонных символов. В лабораторной работе можно попиксельно формировать растровые изображения эталонных символов с разной степенью отличия.

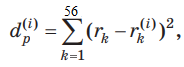
В лабораторной работе не рассматривается первый этап распознавания, т. е. предполагается что бинаризация и сегментация уже проведены и положение фрагментов изображения текста, в которых находятся распознаваемые символы, известно. Таким образом, остается лишь задача последовательного распознавания изображений отдельных символов, которая решается в работе тремя методами (по площади, по вектору признаков и методом маски) путем последовательного сравнения признаков текущего символа с признаками эталонных символов.

В методе распознавания по площади используется единственный признак – площадь символа, т. е. число пикселей в прямоугольном фрагменте известного размера (в работе – квадрат со стороной 28 пикселей), априори относящихся к символу (в работе – светлые пиксели). В результате зашумления распознаваемого символа значение его признака (площади) увеличивается на число ярких пикселей в области фона и уменьшается на число темных пикселей в области эталонного изображения данного символа, что приводит к значительной флуктуации значения признака на разных изображениях одного и того же символа. Расстояние между распознаваемым символом и текущим эталоном определяется



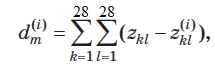
как где ds(i) – расстояние между распознаваемым символом и i-м эталоном по методу площади; S, Si – площадь распознаваемого символа и i-го эталона соответственно.

При распознавании по вектору признаков в качестве отдельного признака используется число светлых пикселей в строке и столбце изображения. Следовательно, число элементов в векторе признаков в лабораторной работе равно 56 (28 строк и 28 столбцов). Расстояние между распознаваемым символом и текущим эталоном определяется как



где dp(i) – расстояние между распознаваемым символом и i-м эталоном по вектору признаков; rk, rk(i) – значение k-го признака распознаваемого символа и i-го эталона соответственно.

При распознавании методом маски последовательно производится сравнение изображений распознаваемого символа и каждого эталона по всем пикселям. В данном случае каждый пиксель можно считать признаком, имеющим значение 0 или 1. Понятно, что число признаков составляет 28\*28=784, а расстояние между распознаваемым символом и текущим эталоном определяется как



где dm(i) – расстояние между распознаваемым символом и i-м эталоном по методу маски; xkl, xkl(i) – значение яркости пикселя с координатами (k, l) распознаваемого символа и i-го эталона соответственно.

При определении расстояния по приведенным выше формулам в лабораторной работе не устанавливается порог достоверности, т. е. распознаваемый символ будет отнесен к ближайшему эталону даже при очень значительном расстоянии до него. На практике экспериментально подбирается значение максимально допустимого расстояния, при превышении которого вычисление расстояния до текущего эталона прекращается, а расстояние между этим эталоном и распознаваемым символом считается неопределенным. Если расстояния до всех эталонов неопределенные, то символ относится к классу нераспознанных символов.

**Порядок выполнения работы**

**Задание 1**

1. Считать данные из файла train.csv\*, содержащий образы цифр (первое значение сама цифра, последующие – значение пикселей образа, 0-белый фон, 1… 256 – элемент цифры).
2. Сформировать эталонные образы цифр заданных в файле по площади.
3. Считать данные из файла test.csv\*, и с помощью ***метода распознавания по площади*** распознать символы, представленные в этом файле.

**Задание 2**

1. Считать данные из файла train.csv\*, содержащий образы цифр (первое значение сама цифра, последующие – значение пикселей образа, 0-белый фон, 1… 256 – элемент цифры).
2. Сформировать эталонные образы цифр заданных в файле по площади.
3. Считать данные из файла test.csv\*, и с помощью ***метода распознавания по признакам*** распознать символы, представленные в этом файле.

**Задание 3**

1. Считать данные из файла train.csv\*, содержащий образы цифр (первое значение сама цифра, последующие – значение пикселей образа, 0-белый фон, 1… 256 – элемент цифры).
2. Сформировать эталонные образы цифр заданных в файле по площади.
3. Считать данные из файла test.csv\*, и с помощью ***метода распознавания по маске*** распознать символы, представленные в этом файле.