О применимости методов распознавания к исследованию статистических свойств множеств символов

М.Б. Гавриков, Н.В. Пестрякова

Аннотация. Описываются эффективные приложения вероятностного высокоточного метода распознавания к задаче анализа статистических свойств множеств символов.

Ключевые слова: классификация, полиномиальная регрессия, печатные и рукопечатные символы.

Введение

Изучение статистических свойств множеств изображений является весьма актуальной задачей [1]. Предметом исследования могут быть, например, обучающие выборки. В таком случае требуется определить, является ли данное множество статистически независимым. Если его объем недостаточен для проведения обучения, возникает необходимость решить проблему искусственного расширения обучающего множества изображений [2]. В то же время избыточность этого объема может привести к неоправданному увеличению времени обучения, а также потере качества из-за накапливания ошибок. Поскольку при тестировании метода необходимо иметь представление о степени близости между обучающими и распознаваемыми выборками, последние также подлежат изучению [6]. Следует учесть, что до сих пор нет общепризнанного способа проведения подобного анализа. В настоящей работе описаны подходы к численному решению указанной задачи, основанные на использовании метода распознавания символов, разработанного авторами [3, 5].

Далеко не каждый метод распознавания пригоден для исследования статистических свойств множеств символов. Это обусловлено тем обстоятельством, что не все способы распознавания имеют вероятностную природу. Описанный в данной работе метод является вероятностным, поскольку в его основе лежит восстановленный с большой степенью достоверности некоторый неизвестный вероятностный закон, в соответствии с которым распределены элементы обучающей последовательности символов, моделирующей датчик случайных векторов. Степень достоверности этого приближения соответствует точности распознавания на обучающем множестве. Ее высокий уровень позволяет использовать данный метод для анализа статистических свойств множеств символов [4].

1. Метод распознавания

Алгоритм позволяет по растру изображения определить, какому символу из некоторого множества с K элементами он соответствует. Нормализованный растр состоит из $N=N1\times N2$ серых пикселов. Все пикселы растра перенуме-

[1]Работа выполнена при финансовой поддержке РФФИ (гранты №13-07-00262 а, № 13-07-12176 офи_м).

рованы в диапазоне $1 \le i \le N$. Яркость i-го пиксела запоминаем в i-ой компоненте вектора $\mathbf{v} \in \mathbf{R}^N$. Для серого растра она лежит на отрезке [0,1].

Отождествим k-й символ с базисным вектором $\mathbf{e}_k=(0\ldots1\ldots0)$ (1 на k-м месте, $1 \le k \le K$) из $\mathbf{R}^K$. Обозначим $Y=\{\mathbf{e}_1,\ldots,\mathbf{e}_K\}$.

Пусть $p_k(\mathbf{v})$ – вероятность того, что растр изображает символ с номером k, $1 \le k \le K$. Распознанным считается символ с порядковым номером k_o, где

$$p_{ko}(\mathbf{v})=\max p_k(\mathbf{v}),\ 1 \le k \le K. \quad (1)$$

Приближенные значения компонент $(p_1(\mathbf{v}),\ldots,p_K(\mathbf{v}))$ представляются в виде многочленов от координат $\mathbf{v}=(v_1,\ldots,v_N)$:

$$p_k(\mathbf{v}) \cong c_0^{(k)} + \sum_{i=1}^{N} c_i^{(k)} v_i + \sum_{i,j=1}^{N} c_{i,j}^{(k)} v_i v_j + \ldots,$$
$$1 \le k \le K. \quad (2)$$

Суммы в правых частях равенств (2) конечные и определяются выбором базисных мономов. А именно, если

$$\mathbf{x}(\mathbf{v})=(1, v_1, \ldots, v_N, \ldots)^{\mathrm{T}}$$

конечный вектор размерности L из выбранных и приведенных в (2) базисных мономов, упорядоченных определенным образом, то в векторном виде соотношения (2) можно записать так:

$$\mathbf{p}(\mathbf{v}) = (p_1(\mathbf{v}),\ldots,p_K(\mathbf{v}))^{\mathrm{T}} \cong A^{\mathrm{T}}\mathbf{x}(\mathbf{v}), \quad (3)$$

где A – матрица размера $L \times K$, столбцами которой являются векторы $\mathbf{a}^{(1)},\ldots, \mathbf{a}^{(K)}$. Каждый такой вектор составлен из коэффициентов при мономах соответствующей строки (2) (с совпадающим верхним индексом), упорядоченных так же, как в векторе $\mathbf{x}(\mathbf{v})$. Следовательно, приближенный поиск вектора вероятностей $\mathbf{p}(\mathbf{v})$ сводится к нахождению матрицы A.

Значение A вычисляется приближенно в процессе обучения, используя содержащиеся в некоторой базе данных наборы пар векторов $[\mathbf{v}^{(1)},\mathbf{y}^{(1)}], \ldots, [\mathbf{v}^{(J)},\mathbf{y}^{(J)}]$ ($\mathbf{v}^{(j)}$ образ символа с каким-либо номером k ($1 \le k \le K$) и его базисный вектор $\mathbf{y}^{(j)}= (0\ldots1\ldots0)$, где 1 стоит на k-м месте, $1 \le j \le J$):

$$A \cong \left(\frac{1}{J}\sum_{j=1}^{J} \mathbf{x}^{(j)}(\mathbf{x}^{(j)})^{\mathrm{T}}\right)^{-1}\left(\frac{1}{J}\sum_{j=1}^{J} \mathbf{x}^{(j)}(\mathbf{y}^{(j)})^{\mathrm{T}}\right). \quad (4)$$

При получении правой части (4) применяется следующая рекуррентная процедура, где A_0 и G_0 заданы:

$$A_j = A_{j-1} - \alpha_j G_j \mathbf{x}^{(j)}[A_{j-1}^{\mathrm{T}}\mathbf{x}^{(j)} - \mathbf{y}^{(j)}]^{\mathrm{T}},\ \alpha_j = 1/J. \quad (5)$$

$$G_j = \frac{1}{1-\alpha_j}\left[G_{j-1} - \alpha_j \frac{G_{j-1}\mathbf{x}^{(j)}(\mathbf{x}^{(j)})^{\mathrm{T}} G_{j-1}}{1+\alpha_j((\mathbf{x}^{(j)})^{\mathrm{T}} G_{j-1}\mathbf{x}^{(j)} - 1)}\right]$$
$$1 \le j \le J,$$

$$G_j \equiv D^{-1},\ D = \mathrm{diag}\,(\,\mathrm{E}\{x_1^2\}, \mathrm{E}\{x_2^2\},\ldots, \mathrm{E}\{x_L^2\}).$$

Здесь $x_1, x_2,\ldots,x_L$ - компоненты вектора $\mathbf{x}(\mathbf{v})$. Получаемые оценки могут выходить за рамки отрезка [0,1] из-за того, что используемый метод является приближенным. Отрицательные значения искусственно обнулялись, а те, которые были больше 1, делались равными 1.

Обучение и распознавание проводилось на серых растрах, состоящих из $N=16\times16=256$ пикселов, для рукопечатных и печатных символов (цифр, русских и латинских букв). При этом использовались различные модификации вектора $\mathbf{x}(\mathbf{v})$.

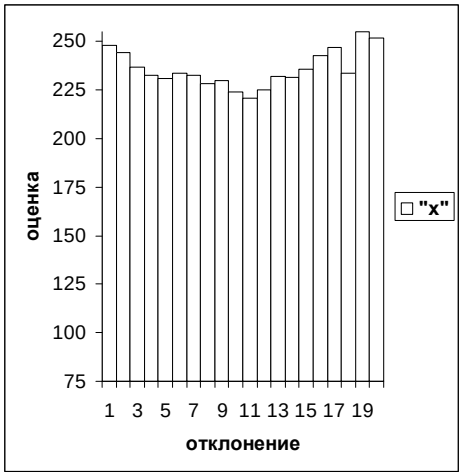
2. Статистический анализ базы рукопечатных цифр

Как обучение, так и распознавание проводились на базе рукопечатных цифр из 174778 элементов. Далее берутся целочисленные оценки 1, 2, …, 255. После умножения оценки на 255 старый диапазон оценок [0,1] переходит в новый [0,255]. Затем проводится дискретизация: [0,1]→1, (1,2]→2, …, (254,255]→255.

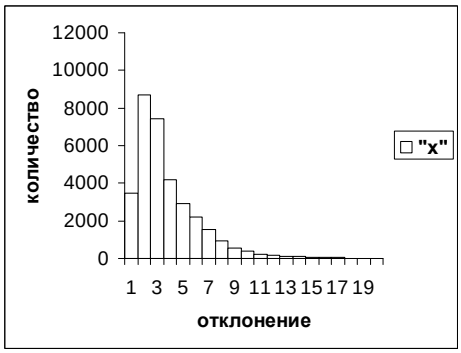
Зависимость средней оценки от отклонения между растрами изображений символа и его «среднестатистическим» растром качественно соответствует изображенной на Рис.1а для «1» и Рис.2а для остальных цифр, но уровень шумов существенно выше.

Для *среднестатистического растра* k-го символа яркость в любом пикселе с номером i равна среднему арифметическому значений яркости i-х пикселов по всем J_k имеющимся в базе растрам изображений символа:

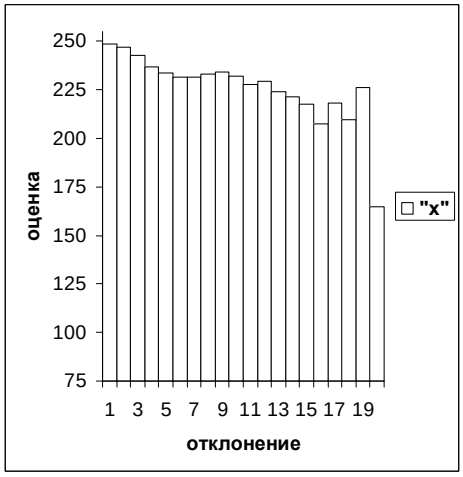
$$v_i^{k,cp} = \left(\sum_{j=1}^{J_k} v_i^{k,j}\right) / J_k. \quad (14)$$



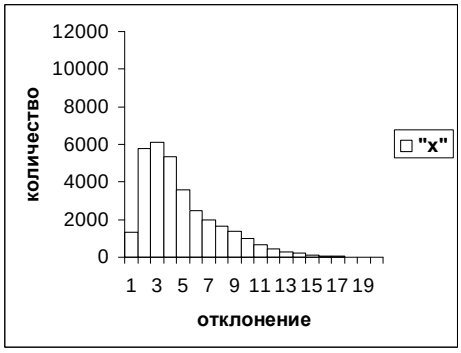
а)



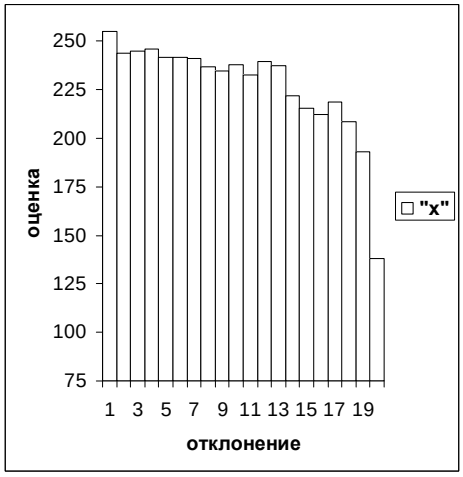
б)



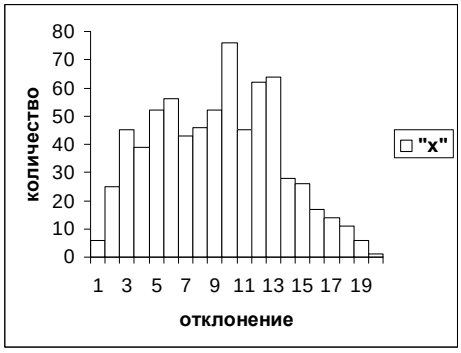
в)



г)



д)



е)

Рис. 1. Поведение средней оценки (а, в, д) и количества образов (б, г, е) при отклонении от среднестатистического вектора для «1»

Расстояние между растрами $\mathbf{v}=(v_1,\ldots,v_N)$ и $\mathbf{u}=(u_1,\ldots,u_N)$ определяем суммированием модулей разности значений яркости в i-х пикселах:

$$\|\mathbf{v}\text{-}\mathbf{u}\| = \sum_{i=1}^{N} |v_i - \mathrm{u}_i| . \qquad (15)$$

Отклонения между растрами распознанных верно изображений символа и его среднестатистическим растром находятся на отрезке [v_true_min, v_true_max] ($35{,}41 \le v\_\text{true\_min} \le 50{,}76$ и $101{,}70 \le v\_\text{true\_max} \le 173{,}80$ для всех цифр).

Делим [v_true_min, v_true_max] (аналог оси абсцисс Рис. 1а, 2а) на 20 равных частей. На каждом участке вычисляем среднюю оценку (ось ординат Рис. 1а, 2а). Для «1» оценка монотонно убывает, затем увеличивается (велика зашумленность) до значения 255 на наибольшем удалении. Для остальных цифр есть тенденция к убыванию, но значителен шум.

Диаграмма числа правильно распознанных изображений из каждой части отрезка [v_true_min, v_true_max] для всех символов аналогична изображенной на Рис.1б, 2б.

Для неправильно распознанных образов отклонения между растрами изображений символа и его среднестатистическим растром лежат на отрезке [v_false_min, v_false_max] ($52{,}22 \le v\_\text{false\_min} \le 61{,}62$ и $93{,}89 \le v\_\text{false\_max} \le 131{,}38$). Неправильные оценки вдвое меньше правильных. Для каждого символа v_true_min < v_false_min, но диапазон [v_true_min, v_true_max] мало отличается от [v_false_min, v_false_max]. Поскольку доля ошибок низка, распределение количества изображений, распознанных как верно, так и неверно, схоже с результатами, полученными для правильного распознавания.

Зависимость средней оценки от отклонения между полиномиальными векторами **x,** построенными по растрам изображений символа, и его среднестатистическим вектором представлена на Рис.1а для «1», а на Рис.2а для «3» и аналогично для остальных цифр.

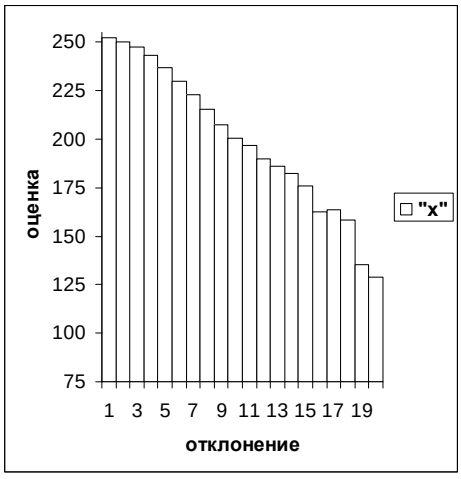
Для *среднестатистического полиномиального вектора* k-го символа значение в i-й компоненте равно среднему арифметическому i-х компонент векторов по всем J_k имеющимся в базе изображениям символа:

$$x_i^{k,cp} = \Big(\sum_{j=1}^{J_k} x_i^{k,j}\Big) / J_k . \qquad (16)$$

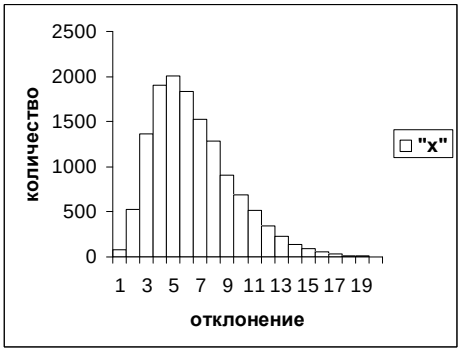
Расстояние между векторами $\mathbf{v}=(v_1,\ldots,v_L)$ и $\mathbf{u}=(u_1,\ldots,u_L)$ определяем как сумму по L компонентам модуля разности значений в i-х компонентах:

$$\|\mathbf{v}\text{-}\mathbf{u}\| = \sum_{i=1}^{L} |v_i - \mathrm{u}_i| . \qquad (17)$$

Отклонения между полиномиальными векторами распознанных верно изображений символа и его среднестатистическим вектором ле-



а)



б)

Рис. 2. Поведение средней оценки (а) и количества образов (б) при отклонении от среднестатистического вектора для «3»

жит на отрезке [x_true_min, x_true_max] ($2004 \le x$_true_min ≤ 2798 и $4954 \le x$_true_max ≤ 7917 для всех цифр).

Делим отрезок [x_true_min, x_true_max] (оси абсцисс на Рис.1а, 2а) на 20 равных частей. Для изображений с полиномиальными векторами, попавшими в каждый такой участок, вычисляем среднюю оценку распознавания (оси ординат на Рис. 1а, 2а). Для «1» она сначала убывает, а затем увеличивается до 255 на предпоследнем участке, а для остальных цифр убывает монотонно. Уровень шумов существенно ниже, чем для растров.

На Рис.1б, 2б по оси ординат отложено число правильно распознанных изображений из каждой части отрезка [x_true_min, x_true_max] для символов «1» и «3» (по остальным цифрам – аналогично).

Отклонения между полиномиальными векторами неправильно распознанных изображений символа и среднестатистическим вектором этого символа находятся на отрезке [x_false_min, x_false_max] ($2913 \le x$_false_min ≤ 3491 и $4909 \le x$_false_max ≤ 6437). Для каждого символа x_true_min < x_false_min, но диапазон [x_true_min, x_true_max] отличается от [x_false_min, x_false_max] не очень существенно. Следовательно, поскольку доля ошибок мала, распределение числа образов, распознанных как верно, так и неверно, каждой из цифр схоже с Рис.1б, 2б.

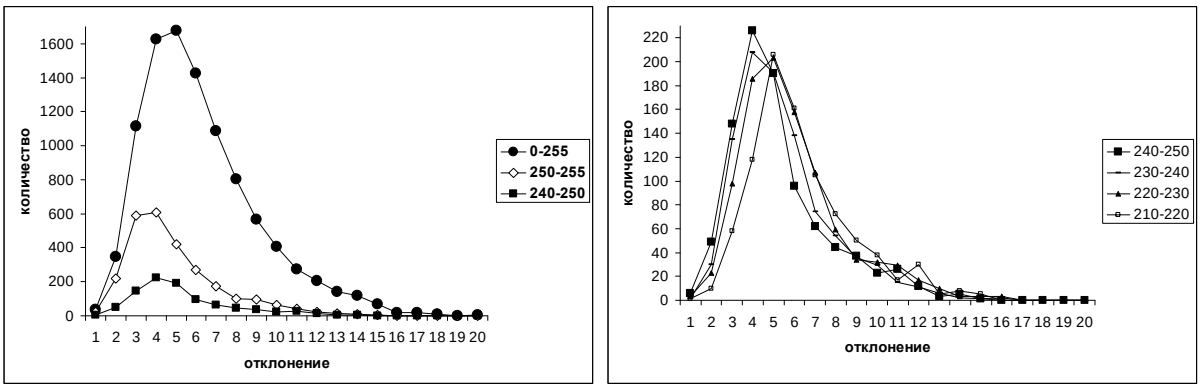
Чтобы сравнить поведение оценки в терминах растров и векторов, совместим отрезки [v_true_min, v_true_max] и [x_true_min, x_true_max]. Точке v_true соответствует x_true = x_true_min + (v_true - v_true_min)·(x_true_max - x_true_min) / (v_true_max - v_true_min). Для символов, отличных от «1», до 1/2 или 1/3 величины максимального отклонения от 0 средняя оценка по векторам выше, чем по растрам. На отдаленных участках ситуация противоположная.

Особенное поведение оценки для «1» породило гипотезу, что база единиц составлена из двух подбаз. Чтобы выделить их, нашли изображения, чьи векторы удалены от x_true_min более чем на 2/3 величины x_true_max-x_true_min. По ним построили среднестатистическй вектор $\mathbf{x}_1$. Для этих 714 образов при отклонении от $\mathbf{x}_1$ оценка монотонно падает (Рис.1д). По оставшимся 32388 изображениям построили среднестатистический вектор $\mathbf{x}_2$, при удалении от которого оценка также падает (Рис.1в). Для этих подбаз распределения числа изображений (Рис.1е, г) оказались схожими с распределениями для полных баз символов (Рис.1б, 2б).

Для символа «8» показано, как монотонное убывание средней оценки распознавания при отклонении от среднестатистического вектора соотносится с распределениями числа верно распознанных изображений для различных оценок. Рассмотрены следующие диапазоны оценок: [255, 250), [250, 240), [240, 230), [230, 220), [220, 210), [210, 200), [200, 190), [190, 180), [180, 170), [170, 160), [160, 150), [150, 140), [140, 130), [130, 120). Изображений с более низкими оценками мало. На Рис. 3а, б приведены распределения числа образов с оценками внутри диапазонов с более высокими оценками (для низких аналогично). На Рис.3а также показано распределение для всего спектра оценок [0, 255].

На каждой из частей отрезка [x_true_min, x_true_max] средняя оценка получается суммированием оценок 1, 2, 3, …, 254, 255 с весами, определяемыми средней (по этой части) вероятности оценки. Монотонное убывание средней оценки соответствует наличию организационной структуры.

Среднестатистические растры и векторы распознаются правильно для всех символов. Любое изображение распознается как перечень из десяти альтернатив для каждого из символов с соответствующей оценкой. Альтернативы нумеруются по мере убывания оценок. Для правильно распознанного образа оценка $0^{ой}$ альтернативы есть оценка распознавания. Соотношение между оценками $0^{ой}$ и $1^{ой}$ альтернативы говорит о «контрастности» распознавания (она тем больше, чем больше различаются оценки $0^{ой}$ и $1^{ой}$ альтернативы). Для каждой цифры оценка распознавания среднестатистического растра ниже, чем вектора. Оценка $1^{ой}$ альтернативы для любого среднестатистического растра выше, чем оценка $1^{ой}$ альтернативы среднестатистического вектора как того же символа, так и другого. Значит, среднестати-



а) диапазоны [0, 255], [255, 250), [250, 240) б) диапазоны [250, 240), [240, 230), [230, 220), [220, 210)

Рис. 3. Число верно распознанных образов для различных оценок символа «8».

стический растр любого символа имеет меньшую контрастность, чем среднестатистический вектор. Разброс оценок по всем символам при распознавании среднестатистических растров равен 229-105=124. Он намного выше, чем у векторов (240-219=21).

Растры верно распознанных изображений любого символа могут находиться дальше от его среднестатистического растра, чем растры неверно распознанных образов (аналогично для полиномиальных векторов).

Среди верно распознанных изображений 87,50% растров наименее удалены от среднестатистического растра «своего» символа (для разных символов их доля варьируется в диапазоне 0,729 – 0,991). Аналогично, полиномиальные векторы 88,40% изображений наименее удалены от среднестатистического вектора «своего» символа (их доля 0,710 – 0,977).

Соответственно, среди неверно распознанных изображений всего лишь 53,35% растров наименее удалены от среднестатистического растра символа, получившего наивысшую оценку (для полиномиальных векторов 51,53%). Причем не для каждого из символов их доля равна или превышает величину 0,5). Однако для векторов таких символов больше (9 против 7).

3. Сравнительный статистический анализ баз рукопечатных и печатных цифр

Как обучение, так и распознавание проводились на одной и той же базе: для рукопечатных цифр – из 174778 элементов, а для печатных цифр – из 5496 элементов. Использовались оценки 1, 2, …, 255.

Для печатных и рукопечатных символов построены диаграммы средней оценки распознавания в терминах растров (векторов) при делении отрезка [v_true_min, v_true_max] ([x_true_min, x_true_max]) на 5 равных частей с учетом малого объема базы печатных цифр по сравнению с рукопечатными.

Для печатных цифр диаграммы имеют общую тенденцию к убыванию (схожи с Рис.2а), причем для растров (в отличие от векторов) высок уровень шумов. Для рукопечатных цифр уменьшение числа отрезков деления с 20 до 5 приводит к понижению уровня шумов для растров и векторов.

Распределения числа распознанных верно изображений в терминах растров (векторов) для печатных и рукопечатных цифр аналогичны.

Для обоих типов написания оценки неправильного распознавания значительно меньше, чем оценки правильного распознавания.

Для любого рукопечатного символа v_true_min (x_true_min) меньше, чем v_false_min (x_false_min) любого другого или того же символа, а для печатных символов значительно меньше. При этом следует учесть, что для печатных символов распознано неверно лишь по два изображения для «1» и «3»

Для произвольного символа G диапазон отклонений между растрами (векторами) изображений символов, отличных от G, и среднестатистическим растром (вектором) G по

рассматриваемой базе находится от минимального ┐*v*_min до максимального ┐*v*_max (соответственно от ┐*x*_min до ┐*x*_max).

Для печатных цифр *v*_true_min (*x*_true_min) произвольного символа меньше, чем ┐*v*_min (┐*x*_min) того же или какого-либо другого символа. Для рукопечатных цифр выполняется закономерность, являющаяся частным случаем приведенной: *v*_true_min (*x*_true_min) некоторого символа меньше, чем ┐*v*_min (┐*x*_min) того же символа.

При рукопечатном написании для *v*_false_max (*x*_false_max) некоторого символа и ┐*v*_max (┐*x*_max) произвольного символа имеем: *v*_false_max < ┐*v*_max (*x*_false_max < ┐*x*_max). Для печатных это также выполняется, но неправильно распознались лишь по два образа цифр «1» и «3».

Для любой рукопечатной цифры (кроме «1») при сравнении *v*_true_max (*x*_true_max) этого символа и ┐*v*_max (┐*x*_max) того же самого или любого другого символа (включая «1») выполняется: *v*_true_max <┐*v*_max (*x*_true_max < ┐*x*_max). Это выполняется и для каждого печатного символа.

Кроме того, для рукопечатных цифр каждая из трех «минимальных» величин *v*_true_min, *v*_false_min, ┐*v*_min (*x*_true_min, *x*_false_min, ┐*x*_min) некоторого символа меньше любой «максимальной» величины *v*_true_max, *v*_false_max, ┐*v*_max (*x*_true_max, *x*_false_max, ┐*x*_max) того же или какого-либо другого символа. Для печатных цифр выполняется закономерность, являющаяся частным случаем приведенной: каждая из трех «минимальных» величин *v*_true_min, *v*_false_min, ┐*v*_min (*x*_true_min, *x*_false_min, ┐*x*_min) некоторого символа меньше, чем любая «максимальная» величина *v*_true_max, *v*_false_max, ┐*v*_max (*x*_true_max, *x*_false_max, ┐*x*_max) того же символа.

Согласно полученным результатам, для данного метода распознавания различаются мелко-, средне- и крупномасштабные явления.

К мелкомасштабным отнесены те, для которых не используется механизм осреднения (распознавание и выставление оценок образам символов).

При описании среднемасштабных используются среднестатистические растры и векторы, но не рассматривается механизм осреднения оценок (или он является несущественным). Сюда относится получение среднестатистического растра и вектора для каждого символа, относительно которых находят распределения числа образов того же символа для разных диапазонов оценок, а также местонахождение правильно, неправильно распознанных его изображений и «чужих» образов. На среднемасштабном уровне над «хаосом» мелкомасштабных явлений выявлена организационная структура – в расположении правильно, неправильно распознанных, а также «чужих» символов относительно среднестатистического вектора (растра).

Для крупномасштабных ключевыми являются ориентация на среднестатистический растр или вектор определенного символа и использование механизма осреднения оценок. Над «хаосом» мелкомасштабных и среднемасштабных явлений обнаруживается «порядок». Несмотря на то, что на всем диапазоне отклонений от среднестатистического вектора имеются изображения данного символа с различными оценками, количество которых определяется полученным распределением, результат их «коллективного» действия – организационная структура в виде монотонного уменьшения средней оценки при удалении от среднестатистического вектора (растра).

Заключение

Итак, показано, что разработанный авторами вероятностный высокоточный метод распознавания может успешно использоваться для анализа статистических свойств множеств символов. При этом на основе разработанных методик изучена структура базы обучения. Найдены закономерности в поведении оценок распознавания. Исследованы особенности взаимного расположения правильно, неправильно распознанных изображений символа, а также образов «чужих» символов (отличных от данного). Проанализирован механизм формирования средней оценки из оценок отдельных образов. Для рукопечатных и печатных цифр проведен сравнительный анализ полученных результатов.

Литература

1. Миркес Е. М., Нейрокомпьютер. Проект стандарта/ под ред В. Л. Дунина-Барковского. — Новосибирск: Наука, 1999.
2. Никодимов Д.Ю., Старовойтов В.В. Расширение обучающего множества для настройки биометрических систем распознавания. Труды 4 международной конференции «Обработка информации и управление в чрезвычайных и экстремальных ситуациях», 29 ноября - 1 декабря 2004, Минск, Беларусь, с.204 - 209.
3. Гавриков М.Б., Пестрякова Н. В. Метод полиномиальной регрессии в задачах распознавания печатных и рукопечатных символов. //Препринт ИПМ РАН, М., 2004, №22, 12 стр.
4. Гавриков М.Б., Мисюрев А.В., Пестрякова Н.В., Славин О.А. Об одном методе распознавания символов, основанном на полиномиальной регрессии. // Автоматика и Телемеханика. 2006, №2, с. 119-134.
5. Пестрякова Н.В. Структуры в распознавании. Информационные технологии и вычислительные системы. 2009, №1, С. 58-71.
6. Пестрякова Н.В. Динамика качества распознавания при нарастании степени различия баз обучения и распознавания. Информационные технологии и вычислительные системы. 2010, №2, С. 75-82.

Гавриков Михаил Борисович. Старший научный сотрудник ИПМ им. М.В. Келдыша РАН. Окончил МГУ в 1975 году. Кандидат физико-математических наук. Автор более 100 печатных работ и одной монографии. Область научных интересов: математическое моделирование, численные методы, вычислительная плазмодинамика, распознавание образов. E-mail: nadya_p@cs.isa.ru

Пестрякова Надежда Владимировна. Ведущий научный сотрудник ИСА РАН. Окончила МФТИ в 1983 году. Доктор технических наук. Автор более 50 печатных работ и одной монографии. Область научных интересов: математическое моделирование, вычислительная гидродинамика, распознавание образов. E-mail: nadya_p@cs.isa.ru