УДК 004.932.75'1

ПОСТРОЕНИЕ УНИВЕРСАЛЬНЫХ КЛАССИФИКАТОРОВ ТЕКСТОВЫХ ОБРАЗОВ РУССКОГО ЯЗЫКА НА БАЗЕ СВЕРТОЧНЫХ НЕЙРОСЕТЕЙ

Н.Н. КУЗЬМИЦКИЙ

Брестский государственный технический университет
ул. Московская, 267, Брест, 224017, Беларусь

Поступила в редакцию 30 марта 2015

Представлена методика построения универсальных классификаторов текстовых образов, основанная на трехуровневом комитете сверточных нейросетей, генерации и учете специфических особенностей начертания образов, зависящих от способов их синтеза. Эффективность методики подтверждена созданием классификаторов символов русского языка, точность которых превышает уровень ведущего коммерческого аналога при распознавании образов представительной базы, созданной в ходе проведения исследования.

Ключевые слова: распознавание образов, сверточная нейронная сеть, обучение, комитет, универсальность, способ синтеза, генерация, база.

Введение

Многие практические системы автоматической обработки текстовых данных сталкиваются с необходимостью распознавания образов, отличающихся как по стилю начертания, так и по способу синтеза: системы обработки банковских документов (договоров, квитанций), почтовой корреспонденции, заявлений, опросных бланков и др. (примеры представлены на рис. 1). Особенности стиля (наклон, размер и др.) могут быть в значительной мере нивелированы удачным выбором признаков, инвариантных к простым пространственным искажениям. Однако адаптивный подбор пороговых коэффициентов при существенном искажении образа может быть серьезно затруднен. Преимуществом классификаторов на основе сверточных нейросетей (СНС) [1], является автоматическая настройка фильтров, извлекающих высокоуровневые признаки входного сигнала путем чередования этапов свертки и подвыборки по аналогии с функционированием зрительной системы млекопитающих. Гораздо более сложной задачей является преодоление «проблемы хрупкости» классификаторов, основанных на методах машинного обучения, заключающейся в падении точности при распознавании образов с отличным от их тренировочных способом синтеза (шрифтовых, рукописных и др.), что в равной мере справедливо и по отношению к СНС [2].

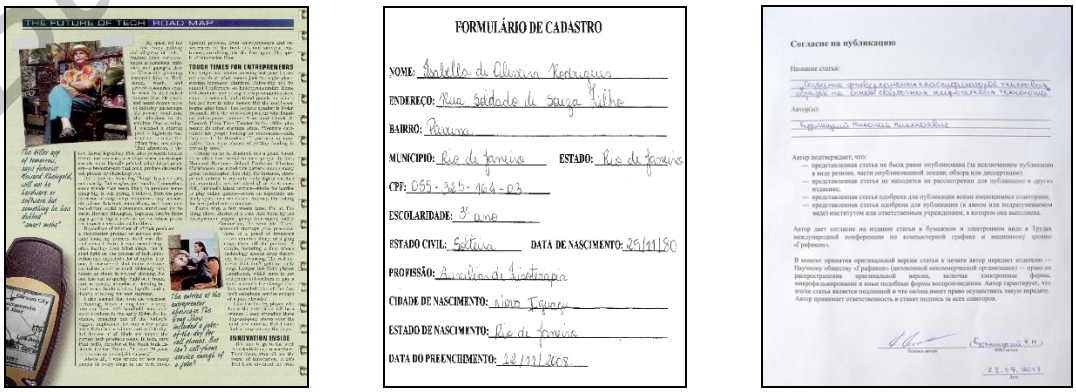

Рис. 1. Примеры изображений, содержащих текстовые образы различного способа синтеза

Обзор литературы по теме исследования

Обзор литературы выявил недостаток комплексных исследований в данной области, связанный с распространенной как в научной, так и коммерческой сферах убежденностью в неактуальности тематики, ввиду ее давнего и успешного разрешения. Вместе с тем немаловажным фактором также является объективная сложность задач обработки применительно к текстовым образам русского алфавита (в частности, исследуемого распознавания). Данный вывод не относится к области OCR (optical character recognition, оптическое распознавание образов), основанной на обработке печатного текста, и представленной успешно зарекомендовавшими себя в реальных приложениях программными продуктами [3]. Но, в перечисленных выше приложениях приходится иметь дело с гораздо менее эталонными и стилистически разобщенными рукопечатными и рукописными образами, распознавание которых традиционными в OCR шаблонными методами неэффективно.

В частности, в [4] авторы, используя линейный дискриминантный анализ, метод главных компонент и признаки в виде дескрипторов функций длины хорды, получили среднюю ошибку распознавания рукописных и печатных символов русского языка на уровне 15 %. При этом в работе не был указан объем выборок и используемое число классов. Авторы [5] предложили структурно-признаковый метод, основанный на скелетизации символа и его разбиение на блоки. Точность распознавания 1600 примеров (50 образов для каждого класса) составила 96 %, однако исследовались только рукопечатные заглавные символы, объединенные в 30 классов. В [6] был представлен подход на основе нейросетей типа неокогнитрон. Тестовая выборка также была составлена из набора рукопечатных заглавных символов русского алфавита (исключая символы 'Й', 'Ё', 'Щ', 'Ъ') и арабских цифр. Использовалось по 100 примеров каждого класса, общий объем базы – 3900 примеров, точность распознавания достигла уровня 91,5 %.

Анализ существующей литературы позволил определить следующие проблемы:

1) исследования проводились либо на не декларируемом множестве образов, либо на выборках, объем которых недостаточен для объективной характеристики методов;

2) в рассмотренных подходах анализу подвергались только выборки заглавных символов русского языка, число классов в которых отличалось от исходных 33-х, рукописные и прописные символы в неустановленном количестве применялись лишь в [4];

3) эффективность представленных в работах классификаторов не достаточна для их применения в распознавании текстовых образов, произвольного способа синтеза.

Таким образом, задача создания универсальных классификаторов текстовых образов русского языка является весьма актуальной. Сверточные нейронные сети, успешно зарекомендовавшие себя в классификации цифр и заглавных символов английского языка [2], представляются перспективной альтернативой походам с эвристическим подбором признаков и шаблонным сравнением с эталонами. Однако для исследования их эффективности в рассматриваемой задаче необходима представительная база образов различного типа.

Формирование базы образов символов русского языка

В связи с отсутствием общедоступных баз русскоязычных образов возникла необходимость создания подобной базы. При этом для исследования универсальности представители классов должны быть однозначно классифицируемы и являться различными по способу синтеза. Так в качестве источника печатных образов выступили шрифты операционных систем Windows, а также нескольких графических редакторов. Всего было собрано около 3000 файлов шрифтов ttf-формата, из которых удалены схожие по начертанию, а также чрезмерно художественные, не применяемые в реальных приложениях, итоговое число использованных шрифтов составило 205.

Для сбора рукопечатных и рукописных образов была разработана модель формы, представленная на рис. 2, предназначенная для индивидуального заполнения респондентами. Форма включает блоки для ввода регистрационных данных, обособленных образов цифр, заглавных и прописных символов русского языка, а также их последовательностей (слов). Использование последних обусловлено стремлением к получению максимально разнообразной выборки, учитывая, что обособленные и слитные символы могут значительно отличаться.

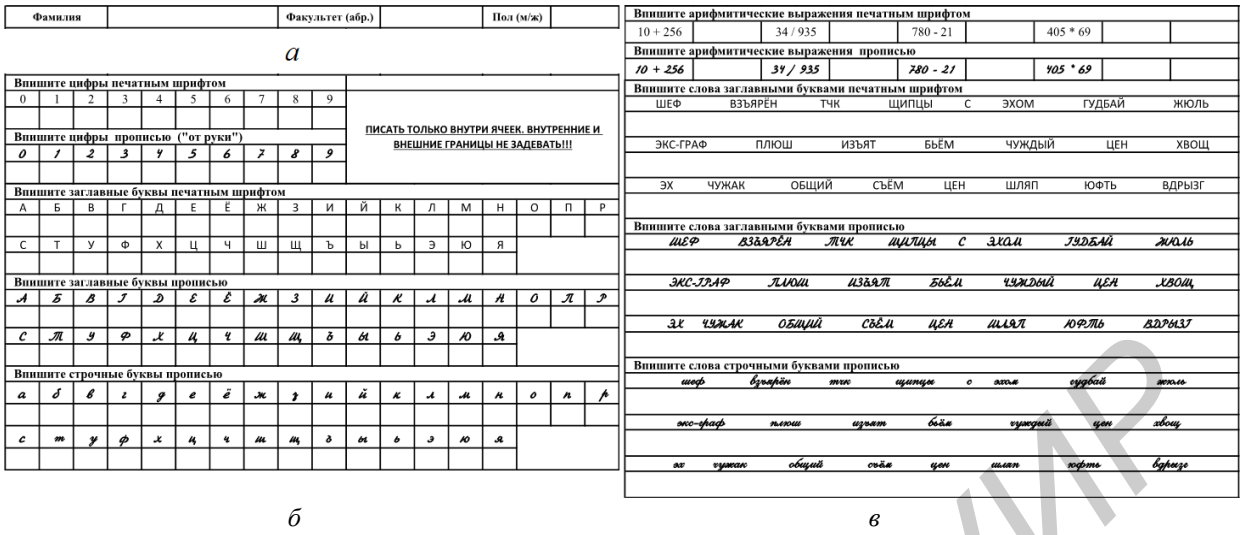

Рис. 2. Фрагменты изображения формы, используемой для создания базы: блоки для ввода регистрационных данных (*а*), обособленных текстовых образов (*б*) и их последовательностей (*в*)

Процесс обработки одной формы включал следующие этапы:

1) синтез изображения формы, например, с помощью сканера;

2) выделение блоков с веденными респондентами данными;

3) сегментация слов на отдельные образы, с «ручной» проверкой качества.

Для реализации второго этапа был использован следующий алгоритм:

1) форма приводится к строго вертикальной ориентации: изображение формы бинаризуется (например, с помощью метода Отсу); выделяется самый крупный связный сегмент (каркас); выполняется его поворот на оптимальный угол;

2) фоновые регионы, ограниченные каркасом формы, группируются в строки, для этого: регионы сортируются в порядке сверху вниз; рассчитывается величина пересечения каждого с текущими строками; регион добавляется к той строке, с которой имеется максимальное пересечение по вертикали (минимальное – 50 %) либо инициирует новую строку;

3) в цикле выполняется анализ каждой строки: если число ее регионов (ячеек) больше, чем предусмотрено моделью, происходит объединение наиболее подходящих (размер которых минимален), при этом удаляются внутренние «отростки» границы ячеек (фрагменты символов, восьмисвязные с нею) с целью приведения их формы близкой к выпуклому четырехугольнику.

В рамках описываемого исследования было собрано 250 образцов форм, заполненных студентами Брестского государственного технического университета (БрГТУ). Номинальное число образов полученной базы составило: цифры – по 750 для каждого класса (всего 7500), заглавных букв – по 2000 (всего 66000), прописных – по 1000 (всего 33000), итого 106500. Отметим, что реальное число образов по классам несколько отличалось от номинального ввиду ошибок, допущенных респондентами: пропуск символов, нарушение регистра, вписывание вместо требуемого иного символа, невозможность однозначного распознавания образа.

Методика построения универсальных классификаторов текстовых образов

В исследовании рассматривалась задача распознавания символов произвольного типа (печатные, рукопечатные и рукописные) полного алфавита (заглавные и прописные) русского языка, для которого более весомым фактором, по сравнению, например, с английским, является значительное отличие образов в рамках одного класса (рис. 3, *а*). Это привело к необходимости расширения количества классов до числа их принципиально разных представлений. Существует также и противоположная проблема – малое отличие образов некоторых классов (рис. 3, *б*), которая решалась путем их объединения в одном. Проведенная перегруппировка позволила сформировать 32 класса для заглавных букв и 28 – прописных. Отметим ее важность для сверточной нейросети, используемой в качестве основной модели классификаторов. Похожие образы разных классов, как и чрезмерно отличные одного, могут негативно повлиять на обучение: увеличить его продолжительность, потребовать расширения архитектуры сети.

а

б

в

Рис. 3. Примеры образов русского языка с разным начертанием в пределах класса (*а*), низким уровнем межклассового отличия (*б*), отличным начертанием в пределах класса (*в*)

Для проведения экспериментальной работы в базе необходимо выделить обучающее и контрольное множество, сохранив уникальность образов в каждой. Очевидным являлось ее разделение на две половины по 125 форм каждая, в результате размер обучающего для заглавных букв составил 33000, прописных – 16500 (для контрольного аналогично). Однако полученного объема множеств недостаточно для создания СНС-архитектуры LeNet-5 [1]. Вариант их расширения за счет контрольных не рассматривался, т.к. последние нужны для объективного анализа свойств классификаторов. При этом во множествах необходимо обеспечить одинаковое представительство классов, включая их отличные представления (подклассов) (рис. 3, *в*), т.к. нарушение баланса представительства может снизить обобщающие способности сети (например, объединенный класс символов 'Е'/'Ё' имеет четыре варианта начертания). Для достижения баланса была применена генерация, позволяющая на основе волновых искажений и поворотов в диапазоне ±10° из примеров обучающей части базы и печатных шрифтов синтезировать новые прообразы. Каждый подкласс в равной мере участвовал в синтезе прообразов, центрированных в изображении размером 32×32 пикселя, а минимальный уровень их попарного различия составил 25 %. Число генерируемых прообразов определялось с учетом возможности обучения нейросетей на стандартном оборудовании, в результате их число для каждого класса заглавных букв составило 2520, прописных – 2880.

Расширенное с помощью генерации обучающее множество необходимо разделить на тренировочную и тестовую выборки: образы сортировались в порядке уменьшения их среднего отличия от других в классе, каждый шестой попадал в тестовую. В результате для заглавных и прописных размер тренировочной части составил 67200, тестовой – 13440. Создание классификаторов на их основе проводилось по схеме «регулярное масштабирование + селекция», описанной в [2]. В результате для заглавных был построен комитет с усредняющим голосованием, СНС в котором обучались на образах размеров: (h = 20, 24, w / h), (h = 20, w = 10), (w = 20, w / h) (для прописных аналогично). Для каждой сети выполнялась перегенерация тренировочной и тестовой выборки с профильным для нее размером образов. Средняя точность СНС на их тренировочных частях была на уровне 97,91 %, тестовых – 97,59 %.

В связи с тем, что в созданных сетях использовались классы, объединяющие несколько исходных (например, 'З'/'Э'), комитеты необходимо было дополнить соответствующими «составными» СНС. Создание выборок для их обучения выполнялось аналогично многоклассовым: для одного класса генерировалось 12000 образов с минимальным отличием в 20 %, которые также разделялись на тренировочную и тестовую части в соотношении 1:6. В результате были построены 11 СНС для следующих пар символов: 'З'/'Э', 'Г'/'Т', 'Е'/'Ё', 'И'/'Й', 'Ш'/'Щ', 'Ь'/'Ъ', 'Л'/'П', 'б/д', 'е'/'ё', 'г'/'ч', 'и'/'й'. Их средняя точность на тренировочных множествах была на уровне 99,25 %, тестовых – 99,11 %. Полученные классификаторы являлись комитетами из многоклассовых СНС, объединенных в подкомитете, и составных, применяемых при выборе в качестве ответа метки объединенного класса (схема использования приведена на рис. 4).

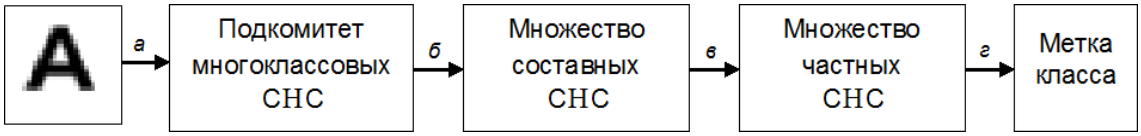

Рис. 4. Схема применения классификатора текстовых образов: предобработка (*а*), расчет начальной метки (*б*), ее корректировка для объединенного класса (*в*), выбор итоговой из меток похожих классов (*г*)

В табл. 1 представлена точность созданных классификаторов (COM1_RUS_big, COM1_RUS_lit) и ведущего коммерческого аналога [7], полученная при оптимальных значениях его параметров. В экспериментах использовались множества, содержащие в равной мере образы 33 символов алфавита (за исключением Rus_font): Rus_gen – сгенерировано из обучающей части базы, содержит 82500 примеров; Rus_cell/Rus_word – подмножества образов, обособленных ячейками/сегментированных из слов форм контрольной части, содержащие по 11000/22000 заглавных (5500/11000 прописных).

Таблица 1. **Точность (в %) классификаторов заглавных (big) и прописных (lit) образов русского языка**

Классификатор	**Rus_big_font**	**Rus_big_cell**	**Rus_big_gen**	**Rus_big_word**
COM1_RUS_big	99,07	97,56	96,51	95,85
COM2_RUS_big	99,26	98,11	97,64	96,37
ABBYY FlexiCapture 10	94,81	89,08	88,75	75,21
	Rus_lit_font	**Rus_lit_cell**	**Rus_lit_gen**	**Rus_lit_word**
COM1_RUS_lit	99,10	92,29	93,72	91,03
COM2_RUS_lit	99,30	95,25	94,95	92,75
ABBYY FlexiCapture 10	87,62	82,05	79,38	77,34

Данные таблицы позволили сделать следующие выводы:

1) имеется зависимость увеличения сложности распознавания от типа образов в порядке: машинописные, рукопечатные, синтезированные, рукописные;

2) прописные символы обладают большей вариативностью начертания, в результате качество их распознавания было ниже чем заглавных в среднем на 4,1 %;

3) средняя точность созданных классификаторов была выше уровня коммерческого на 10,30 % и 12,44 %, что продемонстрировало их большую универсальность.

Детальное изучение ошибок на множествах Rus_big_gen, Rus_lit_gen показало, что их значительная часть была допущена при перекрестном распознавании пар классов, похожих по начертанию: 'А'/'Д', 'У'/'Ч', 'И'/'Н', 'п'/'р', 'в'/'ь', 'т'/'м'. Для таких пар целесообразным являлось создание "частных" СНС, обучение которых проводилось по аналогии с составными. В результате были сформированы классификаторы COM2_RUS_big, COM2_RUS_lit, состоящие из многоклассового подкомитета, составных и частных сетей, применяемых по схеме, приведенной на рис. 4. Их тестирование (табл. 1) показало увеличение точности распознавания, что подтверждает эффективность частных сетей, как экспертов на узком подмножестве классов. При этом предложенная структура классификаторов является весьма гибкой и может быть в случае необходимости дополнена членами на любом из трех уровней.

В последней части исследования рассматривалась задача создания классификатора полного алфавита русского языка. Отметим, что подобные классификаторы в ряде приложений являются более актуальными, чем зависящие от регистра. В частности, основная масса печатных шрифтов содержит как прописные, так и заглавные образы в их традиционном представлении. Кроме того, респонденты при заполнении форм, подобных используемой в исследовании для создании базы, даже при наличии требования к соблюдению регистра поля зачастую ненамеренно нарушают его из-за естественных особенностей почерка.

Имея классификаторы заглавных и прописных символов, очевидной являлась попытка создания классификатора полного алфавита путем интеграции их знаний. В частности, был построен комитет COM1_RUS_all, объединяющий COM2_RUS_big и COM2_RUS_lit, ввиду отличия используемых ими классов образов, по схеме MAX (итоговым является решение наиболее уверенного). Результаты тестирования полученного комитета, представленные в табл. 2, продемонстрировали снижение точности по сравнению с классификаторами, зависящими от регистра. Детальный анализ ошибок позволил выявить следующую проблему: комитет заглавных зачастую обладает большей уверенностью при распознавании прописных образов, не используемых в ходе его обучения (и наоборот). Причиной является как близость начертания некоторых классов, например, '*в*\'В', так и стремление нейросети отнести любой входной образ к одному из известных ей.

Возможными вариантами решения проблемы являлись:

1) создание для сетей дополнительного класса, в который относились все отрицательные примеры (например, для заглавной класс с прописными);

2) разделение классов на группы по какому-либо признаку образов, например, ширине, наличию углов, «дыр» и др., исключающему перекрестные ошибки;

3) формирование общего алфавита из заглавных и прописных классов.

Таблица 2. **Точность (в %) классификаторов полного алфавита русского языка**

Классификатор	Rus_big_font	Rus_big_cell	Rus_big_gen	Rus_big_word
COM1_RUS_all	98,67	97,74	97,03	95,01
COM2_RUS_all	99,04	97,77	97,29	95,17
ABBYY FlexiCapture 10	94,032	88,36	88,48	84,19
	Rus_lit_font	**Rus_lit_cell**	**Rus_lit_gen**	**Rus_lit_word**
COM1_RUS_all	95,90	93,54	92,36	90,87
COM2_RUS_all	98,84	94,62	94,22	92,10
ABBYY FlexiCapture 10	86,05	81,65	76,69	74,98

Первый подход не принес ожидаемого результата, т.к. при обучении сеть не могла настроить признаки, позволяющие относить к одному классу образы с существенно отличным начертанием. Второй столкнулся с большим числом перекрестных ошибок групповых сетей, объединяемых по схеме MAX, из-за несовершенства ручного подбора разделяющих признаков. В рамках третьего была исследована возможность создания алфавита с укрупненным до 24, 30 классами, например, 'УЧуЦ', однако это также не принесло положительного эффекта.

В результате были собраны вместе все 43 заглавных и прописных класса, использованных ранее (рис. 5). При этом были объединены классы с одинаковым начертанием образов в обоих регистрах (например, 'ж'/'Ж'), а также сформирован новый из похожих символов 'П'/'п'/'Л'. Создание итогового классификатора также проводилось по схеме «регулярное масштабирование + селекция», для обучения сетей генерировались тренировочные и тестовые выборки объемом 75250 и 15050 примеров соответственно. Сформированный классификатор (COM2_RUS_all) имел структуру, аналогичную предыдущим, включающую: многоклассовый подкомитет, обученный на образах размеров: (h = 18, 20, 22, w / h), (h = 20, w = 14), (w = 18, w / h), а также составные и частные СНС, созданные ранее. Результаты его тестирования, отраженные в табл. 2, позволили сделать следующие выводы:

1) точность распознавания COM2_RUS_all превышала COM1_RUS_all в среднем на 1 %, что доказало преимущество применения общего алфавита образов;

2) небольшое снижение качества классификатора, по сравнению с COM2_RUS_big / COM2_RUS_lit, на 0,52 % для заглавных и 0,62 % – для прописных, продемонстрировало актуальность подробного анализа различных графических представлений образов;

3) средняя точность COM2_RUS_all для заглавных и прописных выборок (97,31 % и 94,95 %) превышала уровень коммерческого на 8,55 % и 15,10 %, что подтвердило большую универсальность его характеристик.

Рис. 5. Множество классов, используемых классификатором полного алфавита русского языка

Таким образом, трехуровневая архитектура классификаторов доказала свою работоспособность в решении задач распознавания. При этом эффективность применения созданных на ее основе классификаторов может быть увеличена за счет использования контекстной информации (интеллектуальных словарей). Обобщив описанные исследования можно сформулировать методику построения универсальных классификаторов текстовых образов, включающую следующие этапы.

1. Систематизация множества образов:

– классификация образов по способам синтеза существенных для текущей задачи;

– расширение количества классов до числа принципиально различных представлений;

– объединение классов с низким уровнем межгруппового отличия.

2. Формирование тренировочных и тестирующих выборок:

– обеспечение равного представительства вариантов начертания и способов синтеза;

– контроль минимального уровня отличия образов в классах;

– расширение объемов выборок до требуемых с помощью процедуры генерации.

3. Создание классификатора в виде комитета сверточных нейросетей:

– формирование по схеме «регулярное масштабирование + селекция» подкомитета многоклассовых СНС, применяемого для расчета начальной метки класса входного образа;

– построение «составных» СНС, используемых при выборе подкомитетом метки класса, объединяющего несколько исходных текущего алфавита;

– определение на валидационной выборке классов с наибольшим числом перекрестных ошибок и создание «частных» СНС, применяемых в качестве корректоров итоговых решений.

Заключение

Исследована проблема универсальности распознавания на примере русскоязычных текстовых образов, для которых выполнен анализ множества графических представлений, выполнена перегруппировка классов. Создана модель формы и алгоритм ее автоматической обработки, с помощью которых собрана представительная база рукопечетных и рукописных графических образов русского языка, в объеме более 100000 примеров. Разработана методика построения универсальных классификаторов текстовых образов, эффективность которой подтверждена формированием классификаторов полного алфавита и зависимых от регистра, средняя точность которых превышает уровень коммерческих OCR/ICR систем на примерах базы. Созданные с помощью методики классификаторы могут быть использованы в интеллектуальных прикладных системах оцифровки документов и потокового ввода данных.

CONSTRUCTION OF UNIVERSAL RUSSIAN CHARACTERS CLASSIFIERS BASED ON CONVOLUTIONAL NEURAL NETWORKS

N.N. KUZMITSKY

Abstract

Presents the method of constructing universal classifiers of text images based on three-level committee of convolutional neural networks, generation and accounting of specific text patterns features, depending on methods of their synthesis. Efficiency of this method was confirmed by creation classifiers of Russian language characters, accuracy of which exceeds level of leading commercial counterpart in recognition of representative database patterns created in course of research.

Список литературы

1. *LeCun Y., Bottou L.* // Proceedings of the IEEE. 1998. Vol. 86 (11). P. 2278–2324.
2. *Кузьмицкий Н.Н.* // Матер. Междунар. конф. по компьютерной графике и зрению «ГрафиКон'2013». Владивосток, 16–20 сентября 2013 г. С. 234–237.
3. ABBYY: FineReader. [Электронный ресурс]. – Режим доступа: http://www.abbyy.ru/finereader/. – Дата доступа: 01.02.2015.
4. *Запрягаев С.А., Сорокин А.И.* // Вест. Воронежского государственного университета. Сер. Системный анализ и информационные технологии. 2009. № 2. С. 49–58.
5. *Вихров А.Г., Богуш Р.П., Глухов Д.О.* // Вест. Полоцкого государственного университета. Сер. С, Фундаментальные науки. 2010. № 9. С. 35–43.
6. NeuroFace: Садыхов Р.Х., Ваткин М.Е. Алгоритм обучения нейронной сети неокогнитрон для распознавания рукописных символов распознавания рукописных символов [Электронный ресурс]. – Режим доступа: http://neuroface.narod.ru/files/neocog_hand_writ.pdf. – Дата доступа: 25.01.2015.
7. ABBYY: FlexiCapture [Электронный ресурс]. – Режим доступа: http://www.abbyy.ru/flexicapture/. – Дата доступа: 03.02.2015.