Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и информационных технологий

# Оптическое распознавание текста с помощью нейронных сетей

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 421 группы
специальности 09.03.01 «Информатика и вычислительная техника»
факультета компьютерных наук и информационных технологий
Вдошкиной Дианы Андреевны

Научный руководитель
профессор, д.ф.-м.н. _____ В.А. Молчанов
дата, подпись

Заведующий кафедрой
доцент, к.ф.-м.н. _____ Л.Б. Тяпаев
дата, подпись

Саратов 2017

# ВВЕДЕНИЕ

Мир стремительно развивается, однако, до сих пор бумага является общепринятым носителем информации. Бумажный документооборот преобладает в большинстве организаций, хотя, он имеет ряд недостатков, такие, как:

- расходы на закупку бумаги - трата денег;
- архивы документов занимают большую площадь;
- поиск в бумажных архивах занимает много времени, а иногда - вообще невозможен;
- бумага имеет свойство желтеть и рассыпаться от времени;

Чтобы избавиться от вышеуказанных проблем организации переходят на электронный документооборот. Наиболее простым и быстрым способом перейти на систему электронного документооборота является сканирование документов с помощью сканеров. Результат работы является цифровое изображение документа – графический файл. К сожалению, такой вид файла не является предпочтительным, поскольку требует больших затрат на хранение и передачу информации. Следовательно, возникает необходимость перевести графический файл в текстовое представление. Это может быть реализовано получением данных посредством оптического распознавания текста.

В настоящее время распознавание текста является всё более обсуждаемой и актуальной темой.

Оптическое распознавание символов (англ. optical character recognition, OCR) — механический или электронный перевод изображений рукописного, машинописного или печатного текста в текстовые данные — последовательность кодов, использующихся для представления символов в компьютере.

Оптическое распознавание текста позволяет редактировать текст, осуществлять поиск слова или фразы, хранить его в более компактной форме, демонстрировать или распечатывать материал, не теряя качества, анализировать информацию, а также применять к тексту электронный перевод, форматирование или преобразование в речь. Оптическое распознавание текста является исследуемой проблемой в областях распознавания образов, искусственного интеллекта и компьютерного зрения.

Целью данной работы является создание программы на языке Java в среде Intellij Idea Community Edition 17.1.2, позволяющей распознавать текст в изображении.

В работе рассмотрены основы оптического распознавания текста и нейронных сетей, этапы алгоритмов распознавания, а также познакомиться с одной из моделей классификатора – классификатор на основе искусственной нейронной сети.

Для достижения поставленной цели решены следующие задачи:

- изучена теория нейронных сетей;
- исследованы способы распознавания образов;
- рассмотрены различные алгоритмы, необходимые для реализации программы.
- реализован алгоритм оптического распознавания текста на языке программирования Java в среде Intellij Idea Community Edition 17.1.2;
- реализован графический интерфейс пользователя.

Бакалаврская работа состоит из введения, 5 разделов, заключения, списка использованных источников и приложения. Общий объём работы – 61 страницы, из них 45 страниц – основное содержание, включая 24 рисунка, список использованных источников из 22 наименований.

## ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении ставятся цели и задачи бакалаврской работы.

В первом разделе работы «Архитектура системы распознавания» рассматривается система распознавания символов, а также даётся описание различных подходов: шаблонного, структурного и контекстного распознавания.

Результаты, полученные О.А. Славиным и Ю.В. Титовым в работе «Динамическое построение функций сравнения с идеальным образом в задаче адаптивного распознавания текстовых символов» изложены во втором разделе «Алгоритм построения функции сравнения». В данном разделе предложен алгоритм построения функции сравнения распознаваемого символа с идеальными образами $\overline{S_i}$ символов $S_i$ из коллекции альтернатив $X = \{(S_1, P_1), \dots, (S_n, P_n$. Идеальными образами являются объекты, обладающими свойством наименьшей средней близости по отношению ко всем элементам рассматриваемого множества образов в смысле некоторой определенной функции близости.

Общая схема алгоритма построения функции сравнения:

1) Идеальные образы $\overline{S_i}$ накладываются друг на друга, и ищется наилучшее их взаимное расположение. Для двух растров наилучшим называется такое положение, при котором расстояние между ними по интегральной метрике минимально. Вначале на первый растр накладывается второй, затем на первый растр накладывается третий, четвертый и т.д.
2) В найденном наилучшем положении, если это необходимо, растры $\overline{S_i}$ приводятся к одинаковому (расширенному) размеру.
3) Строится матрица $M' = ||m'_{jk}||$, каждый элемент которой равен максимальной разности соответствующих элементов растров $\overline{S_i}$:

$m'_{jk} = max_i\left(\overline{S_i}(j,k)\right) - min_i\left(\overline{S_i}(j,k)\right)$, где $\overline{S_i}(j,k)$- значение растра $\overline{S_i}$ в точке $(j,k)$.

4) Вторая матрица $M = ||m_{jk}||$ заполняется штрафными коэффициентами. В каждом пикселе штраф вычисляется как неубывающая функция $f(x)$ от значения соответствующего элемента $M'$, т.е. $m_{jk} = f(m'_{jk})$.
5) Ищется наилучшее положение растра $R$ на каждом $\overline{S_i}$, при этом используется интегральная метрика.
6) В найденном в пункте 5) положении для каждого $\overline{S_i}$ вычисляется штраф за несовпадение с растром $R$. В пикселе с координатами $(j,k)$ штраф определяется как произведение $m_{jk} \times \left|\overline{S_i}(j,k) - r_{jk}\right|$ соответствующего элемента штрафной матрицы $M$ и модуля разности значений растра $R$ и идеального образа $\overline{S_i}$. Считается сумма штрафов по всем пикселям растра $R(m,n)$:
   $\sum_{j=1}^{m}\sum_{k=1}^{n} m_{jk} \times \left|\overline{S_i}(j,k) - r_{jk}\right|$
7) Из множества альтернатив $S_i$ результатом распознавания считается тот символ, для идеального образа $\overline{S_i}$ которого в пункте 5) получен минимальный штраф.

В третьем разделе «Искусственные нейронные сети» излагаются общие понятия теории нейронных сетей. Искусственная нейронная сеть (ИНС) – математическая модель, построенная по принципу организации и функционирования биологических нейронных сетей – сетей нервных клеток живого организма. Биологический нейрон — сложная система, математическая модель которого до сих пор полностью не построена. Введено множество моделей, различающихся вычислительной сложностью и сходством с реальным нейроном. Одна из важнейших — формальный нейрон. Так же были рассмотрены

различные функции активации: жесткая ступенька, логическая функция (сигмоида), SOFTMAX-функция.

Выбор функции активации определяется:

1. Спецификой задачи.
2. Удобством реализации на ЭВМ, в виде электрической схемы или другим способом.
3. Алгоритмом обучения (некоторые алгоритмы накладывают ограничения на вид функции активации, которые необходимо учитывать).

Формальные нейроны могут объединяться в сети различным образом. Один из самых распространённых видов сети – многослойный персептрон.

Работа многослойного персептрона описывается формулами:

$$WS_{il} = \sum_{i} w_{ijl} x_{ijl}\,; \tag{1}$$

$$OUT_{jl} = F\left(WS_{jl} - B_{jl}\right); \tag{2}$$

$$x_{ij(l+1)} = OUT_{il}, \tag{3}$$

где индексом $i$ обозначается номер входа, $j$ – номер нейрона в слое, а $l$ – номер слоя,

$x_{ijl}$ – $i$-ый входной сигнал $j$-го нейрона в слое $l$;

$w_{ijl}$ – весовой коэффициент $i$-го входа $j$-го нейрона в слое $l$;

$WS_{jl}$ – взвешенная сумма $j$-го нейрона в слое $l$;

$OUT_{jl}$– выходной сигнал $j$-го нейрона в слое $l$;

$B_{jl}$ – пороговый уровень $j$-го нейрона в слое $l$.

В четвертом разделе «Методы обучения» описываются алгоритмы обучения персептрона.

Общая схема обучения персептрона:

1. Инициализировать веса и параметры функции активации в малые ненулевые значения;

2. Подать на вход один образ и рассчитать выход;

3. Посчитать ошибку $E^s$ , сравнив $d^s$ и $y^s$.

4. Изменить веса и параметры функции активации так, чтобы ошибка $E^s$ уменьшилась.

5. Повторять шаги 2-4 до тех пор, пока ошибка не перестанет убывать или не станет достаточно малой.

В четвертом разделе рассматривается обучение сети обратного распространения, которое предполагает выполнение следующих операций:

1) Выбрать очередную обучающую пару из обучающего множества; подать входной вектор на вход сети.
2) Вычислить выход сети.
3) Вычислить разность между выходом сети и требуемым выходом (целевым вектором обучающей пары).
4) Подкорректировать веса сети так, чтобы минимизировать ошибку.
5) Повторять шаги с 1) по 4) для каждого вектора обучающего множества до тех пор, пока ошибка на всем множестве не достигнет приемлемого уровня.

В пятом разделе «Практическая часть» описывается созданная в ходе проделанной работы программа, которая конвертирует отсканированные документы в текстовый формат. Программа была реализована на языке Java в среде Intellij Idea Community Edition 17.1.2. Код программы приведен в приложении А.

Опишем алгоритм оптического распознавания текста, реализованный в программе:

На вход алгоритма подаётся JPEG изображение. На этапе предварительной обработки изображение переводится в двухцветный черно-белый формат.

Сегментация происходит по следующему алгоритму:

1. Ищется горизонтальная линия U, такая, что она расположена как можно выше на изображении и при этом содержит хотя бы один черный пиксель.
2. Ищется горизонтальная линия D, такая, что она расположена как можно выше на изображении, но ниже U, и при этом не содержит черных пикселей.
3. Полоса изображения между U и D считается строкой, в которой ищутся отдельные символы.
4. Символом считается максимальный по ширине сегмент строки (т.е. прямоугольник), такой, что на каждой его вертикальной линии содержится хотя бы один чёрный пиксель.
5. Если найденный символ заметно шире среднестатистического, то в нём ищется линия разреза (вертикальная линия в центральной части символа, содержащая минимальное количество черных пикселей), и символ делится по ней на два.
6. После этого алгоритм повторяется, но уже только для той части изображения, что расположена ниже линии D.

Полученные прямоугольники масштабируются (приводятся к размеру 40x40 пикселей) и подаются на вход классификатора. Сегментатор также отвечает за расстановку пробелов и переводов строки.

Последовательность символов, распознанных классификатором, пробелов и переводов строки и является результирующим текстом. Словарная или контекстная обработка не проводится.

В пятом разделе также рассмотрен принцип работы классификатора, основанном на искусственной нейронной сети – однослойном персептроне.

Опишем процедуру обучения ИНС. Для её реализации необходимо иметь большое количество образцов всех символов алфавита. Перед началом обучения в качестве весов и пороговых значений устанавливаются небольшие случайные значения. Обучение состоит в многократном обучении сети отдельными образцами.

Пусть выбран $i$-ый символ алфавита. Тогда на вход нейронной сети подаётся случайный образец $i$-го символа. Если выполняются неравенства (4) и (5):

$$OUT_j < UL \tag{4}$$

$$OUT_i > RL \; \forall j \neq i \tag{5}$$

где $UL$ и $RL$ – константы unrecognitionLimit и recognitionLimit соответственно, то считается, что нейронная сеть уже знает этот символ. Иначе веса и пороговые уровни корректируются по следующим формулам:

$$w_{jk} = w_{jk} - 2x_k\left(EOUT_j - OUT_j\right); \tag{6}$$

$$t_j = t_j - 2\left(EOUT_j - OUT_j\right), \tag{7}$$

где $w_{jk}$ – вес $k$-го входа $j$-го нейрона;

$x_k$ – $k$-ый входной сигнал;

$t_j$ – пороговый уровень $j$-го нейрона;

$OUT_j$ – выходной сигнал $j$-го нейрона;

$EOUT_j$ – ожидаемое значение выходного сигнала $j$-го нейрона.

$$EOUT_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \quad (8)$$

Коррекция проводится многократно до тех пор, пока указанные выше неравенства (4) и (5) не начнут выполняться.

# ЗАКЛЮЧЕНИЕ

Программы распознавания текстовых документов являются сложными программными средствами, реализующими большое число наукоемких алгоритмов. Современные возможности сканирования документов и реализации алгоритмов распознавания делают возможным автоматизировать ввод документов в компьютер.

Приведем результаты исследования методов построения программ оптического распознавания текста. Во-первых, для реализации программ OCR легче и полезнее шаблонное описание. Однако, этот вид описания не допускает описывать объекты, которые имеют большую степень непостоянства, то есть изменчивости. Во-вторых, структурный метод хоть и сложнее для исполнения, но может применяться даже для распознавания рукописных текстов, в то время как шаблонное только для печатных.

В настоящее время, все известные российский программы распознавания, такие как: CuneiForm, Autor и FineReader сочетают в себе два метода – структурный и шаблонный, что необходимо для обеспечения надёжности распознавания.

Для полноты работы системы оптического распознавания текста необходимо, чтобы объект был представлен и обработан целиком. После чего, все фразы должны быть проверены в словарях.

В работе рассмотрены основы оптического распознавания текста и нейронных сетей, этапы алгоритмов распознавания, а также одна из моделей классификатора – классификатор на основе искусственной нейронной сети.

В ходе работы написана программа на языке Java в среде Intellij Idea Community Edition 17.1.2, демонстрирующая один из способов распознавания текста в изображениях.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Арлазаров В.Л., Куратов П.А., Славин О.А. Распознавание строк печатных текстов. - М.: Эдиториал УРСС, 2000.

2 Технологии оптического распознавания текстов // WolfPromotion - волчий сайт рекламиста URL: http://travin.msk.ru/arc/OCR.html (дата обращения: 22.02.2017).

3 Уоссермен Ф. Нейрокомпьютерная техника : Теория и практика. - М.: Мир, 1992.

4 Заенцев И.В. Нейронные сети. Основные модели. - Воронеж: Изд-во Воронежского гос. ун-та, 1999.

5 Широбоков В.А. Системы распознавания текстов. Автореферат работы магистра // Портал магистров Донецкого национального технического университета URL: http://masters.donntu.edu.ua/2005/fvti/shirobokov/diss.htm (дата обращения: 23.01.2017).

6 Славин О. А., Титов Ю. В. Динамическое построение функций сравнения с идеальным образом в задаче адаптивного распознавания текстовых символов // Информационные технологии и вычислительные системы. 2007. № 1. С. 3–12.

7 Арлазаров В. Л., Котович Н. В., Славин О. А. Адаптивное распознавание // Информационные технологии и вычислительные системы. 2002. № 4. С. 11–22.

8 Lebourgeois F., Henry J. L. An Evolutive OCR System Based on Continuous Learning // Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96), December 1996. P. 272-277

9 Славин О.А., Корольков Г.В., Болотин П.В. Методы распознавания грубых объектов. В сб. "Развитие безбумажных технологий в организациях", 1999, с. 331-355

10 Бойцов Л. М. Классификация и экспериментальное исследование современных алгоритмов нечеткого словарного поиска [Электронный ресурс] / Л. М. Бойцов // Труды 6–ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL2004, Пущино, Россия, 2004. – Режим доступа : http://www.rcdl.ru/papers/2004/paper27.pdf.

11 Wang J., Jean J. Segmentation of merged characters by neural networks shortest path. Pattern Recognition 27,Vol. 5 (1994), pp. 649-658

12 Minsky M. L, Papert S. 1969. Perseptrons. Cambridge, MA: MIT Press. (Русский перевод: Минский М. Л., Пейперт С. Персептроны. – М: Мир. – 1971.)

13 Нейронные сети и распознавание символов// Geektimes URL: https://geektimes.ru/post/113245/ (дата обращения: 23.01.2017).

14 Багрова И. А., Грицай А. А., Сорокин С. В., Пономарев С. А., Сытник Д. А. Выбор признаков для распознавания печатных кириллических символов // Вестник Тверского Государственного Университета 2010 г., 28, стр. 59-73

15 Melin P., Urias J., Solano D., Soto M., Lopez M., Castillo O., Voice Recognition with Neural Networks, Type-2 Fuzzy Logic and Genetic Algorithms. Engineering Letters, 13:2, 2006.

16 Травин А. Технологии оптического распознавания текстов // Электронный офис. 1996. Ноябрь.

17 Хайкин С. Нейронные сети: полный курс, 2е издание. : Пер. с англ. М. Издательский дом «Вильямс», 2006.

18 Терехов С. А. Лекции по теории и приложениям искусственных нейронных сетей. Лабораттория Искусственных Нейронных Сетей НТО-2. Снежинск. ВНИИТФ.

19 Дробков А.В., Семёнов А.Б. Исследование одного метода распознавания рукопечатных символов // Вестник Тверского государственного университета, серия «Прикладная математика». Тверь, 2009. №15. С. 15-26.

20 Rosenblatt F. 1962. Principles of neurodynamics. New York: Spartan Books. (Русский перевод: Розенблатт Ф. Принципы нейродинамики. – М.: Мир., 1965.)

21 Almeida L. B. 1987. Neural computaters. Proceedings of NATO ARW on Neural Computers, Dusseldorf. Heidelberg: Springer-Verlag.

22 Burr D. J. 1987. Experiments with a connecnionlist text reader. In Proceedings of the IEEE First International Conferense on Neural Networks, eds. M. Caudill and C.Butler, vol. 4, pp. 717-24. San Diego, CA: SOS Printing.