

Исследование изображений печатных знаков разных шрифтов с целью выявления признаков, необходимых и достаточных для распознавания знаков компьютером

Э.М. Пинт, И.Н. Петровнина, И.И. Романенко, К.А. Еличев
Пензенский государственный архитектурно-строительный университет

Аннотация: До настоящего времени задача распознавания печатных знаков разных шрифтов с высокими быстродействием и надежностью не решена. Коллектив авторов провел исследование изображений печатных знаков разных шрифтов на предмет выявления признаков, необходимых и достаточных для распознавания знаков компьютером. Такими признаками оказалась определенная комбинация направлений, образуемая элементами контуров знаков. По разработанной программе для компьютера изображение каждого печатного знака, хранящееся в запоминающем устройстве компьютера, масштабно преобразовывалось, ликвидировались нехарактерные признаки знаков, определенные дефекты и выделялся контур знака. Контур представлялся в виде последовательности определенных направлений, образуемых его элементами. Все печатные знаки русского алфавита разных шрифтов распознавались по определенной методике, несмотря на разные типы печатных шрифтов, на определенные возможные дефекты.

Ключевые слова: печатный знак, направление, стандартный вид, распознавание, контур, шрифт, признак, засечки, читающее устройство, информация.

В настоящее время еще не решена проблема распознавания печатных и дорожных знаков различных шрифтов. Эту функцию должно выполнять читающее устройство, воспринимающее печатный текст с носителя информации и распознающее с использованием компьютера печатные знаки различных шрифтов.

Читающие устройства могут использоваться для автоматической обработки статей, книг, отчетов, экспериментальных данных, для автоматизации процесса поиска информации в фонде литературы, для обработки сводок, таблиц и других сведений, имеющих на промышленных предприятиях, на транспорте, для обработки банковских документов и т. д.

При реализации второй основной функции читающего устройства – распознавание отдельных знаков алфавита встречаются следующие трудности:

- устройство должно различать относительно большое количество знаков алфавита: букв, цифр и пр.;
- устройство должно воспринимать различные шрифты, т. е. уметь читать одну и ту же букву или цифру, но напечатанную с другой высотой, шириной, с другим наклоном, более жирно, более узко и т.д.;
- устройство должно воспринимать буквы или цифр, напечатанные с различного рода пропусками в изображении, с изменениями толщин линий, радиусов закруглений и пр.

Как следует из анализа существующих методов распознавания печатных знаков, некоторые методы для распознавания требуют определенной стандартизации знаков текста. Другие методы позволяют распознавать стилизованные по форме или специальным образом отмеченные знаки, т.е. не могут быть использованы для чтения реального печатного текста, т.к. требуют реформы системы печати.

Лучшие из существующих методов позволяют распознавать печатные знаки при определенных вариациях формы знаков, при определенных изменениях ориентации, при определенных дефектах, но количество распознаваемых знаков оказывается ограниченным и замыкается в основном в пределах одного шрифта [1-5].

В настоящей статье описывается выбор необходимой и достаточной комбинации информационных признаков печатных и дорожных знаков различных шрифтов для распознавания знаков компьютером.

Цифровой или буквенный знак независимо от алфавита и типа шрифта представляет собой рисунок, состоящий из отдельных частей, представляющих собой прямые или кривые линии, соединенные в определенных сочетаниях.

Сочетание отдельных частей знака будем называть контуром знака. Полное описание внешнего вида любого из печатных знаков содержит

большое количество избыточной информации, повторяющейся в разных печатных знаках и ненужной для распознавания.

Чтобы распознавать печатные знаки разных шрифтов определенного алфавита, необходимо выделить существенные признаки знаков. Как нетрудно видеть, контур любого знака можно представить в виде последовательности бесконечно малых, равных по модулю векторов в прямоугольной системе координат (x, y) . Модуль бесконечно малого вектора обозначим:

$$|\overline{\Delta}| = \lim \sqrt{(x_k - x_{k-1})^2 + (y_k - y_{k-1})^2}, \quad (1)$$

при $k \rightarrow n$, где n – номер координат x и y , соответствующий количеству бесконечно малых векторов.

Соответственно направление бесконечно малого вектора:

$$\varphi_i = \lim \arccos \frac{x_k - x_{k-1}}{|\overline{\Delta}|}, \quad (2)$$

при $k \rightarrow n$ знак представляется в виде суммы элементарных векторов, полученной по мере обхода контура знака, начиная с вектора $\overline{\Delta}_{\min}$ и кончая вектором $\overline{\Delta}_{\max}$:

$$|\overline{\Delta}|_{\overline{\Delta}=\overline{\Delta}_{\min}}^{\overline{\Delta}=\overline{\Delta}_{\max}} \sum m e^{j\varphi_i}, \quad (3)$$

где: $m = 1, 2, 3, 4, \dots, n$;

$\varphi_i = \varphi_1, \varphi_2, \varphi_3, \dots, \varphi_t$.

Начальный вектор $\overline{\Delta}_{\min}$ мы будем определять концевой точкой знака, определяется как конечный вектор, завершающий обход контура знака. При составлении такой последовательной суммы векторов могут появиться «перекрестия» – точки контура, в которых начинается несколько векторов, ориентированных соответственно по разным направлениям.

В случае возникновения перекрестия, авторы условились обходить контур по направлению с наименьшим номером из номеров: $\varphi_1, \varphi_2, \varphi_3, \dots, \varphi_t$. Оставшиеся не обойденными части контура условились обходить поочередно, начиная с перекрестия с младшим номером.

Для выделения существенных характерных признаков знаков вместо бесконечно большого количества направлений авторы ограничились сравнительно небольшим количеством направлений: от $\varphi_{\text{нач}}$ до $\varphi_{\text{кон}}$.

При этом следовало ограничиться таким количеством направлений φ_g , чтобы знаки, представленные в виде суммы количеств векторов, распознавались друг относительно друга. В таком случае существенные признаки не будут потеряны.

Для выделения существенных характерных признаков, не зависящих от масштаба элементов знака, надо было провести масштабные преобразования, т.е. условились считать количество векторов m по любому из направлений φ_g равным единице ($m = 1$).

Тогда масштабное преобразованный знак представляется в виде суммы:

$$\left| \bar{\Delta} \right|_{\Delta=\Delta_{\min}}^{\Delta=\Delta_{\max}} \sum e^{j\varphi_g}, \quad (4)$$

Если знаки, представленные таким образом, будут распознаваться друг относительно друга, то существенными характерными признаками знака будут направления элементов знака φ_g , составляющие для каждого знака определенную специфическую последовательность в порядке обхода по контуру.

Например, если взять за основу 16 направлений (рис. 1), то направления элементов контура буквы «Л», начиная с левой концевой точки, будут такими, как изображены на рис. 2.

алфавита для определения типа знака [7]. Стандартные виды образовывались после обхода идеально напечатанных контуров букв и цифр русского алфавита. При этом учитывались особенности контуров знаков выбранных для распознавания печатных шрифтов.

Для анализа использовались два компьютерных шрифта с отличающимися по форме знаками: знаки одного шрифта имели дополнительные декоративные украшения – «засечки», знаки другого шрифта их не имели (рис. 3). При этом формы контуров знаков выбранных шрифтов повторяются в основном во всех остальных шрифтах русского алфавита для одинаковых по смыслу знаков.

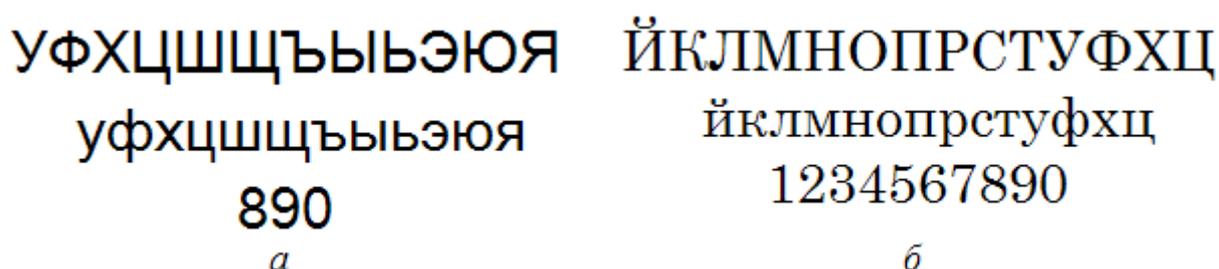


Рис. 3 – Компьютерные шрифты:

а – знаки без засечек; б – знаки с засечками

Вначале выяснялось, какие оптимальные два направления должны быть выбраны из всех возможных комбинаций двух направлений для распознавания наибольшего количества печатных знаков (как нетрудно заметить: по одному какому-либо направлению невозможно распознать какой-либо знак).

Причем, при выборе оптимальных направлений необходимо было учитывать не только максимальное количество определяемых знаков, но и то, какие знаки определяются. Последнее было особенно важно, так как, как известно, буквы имеют разную повторяемость в тексте P_i .

В связи с вышесказанным для выбора K оптимальных направлений ($K = 2$ для двух направлений) из всех возможных комбинаций K направлений

была введена величина ω , которая для оптимальных направлений должна быть максимальной.

$$\omega = b_k \sum_{i=1}^{b_k} P_i, \quad (5)$$

где b_k – количество распознаваемых букв для K направлений.

Таким образом, ω учитывает не только количество распознаваемых букв, но и их вероятность появления в тексте.

Из всех возможных комбинаций двух направлений для анализа были выбраны комбинации, изображенные на рис. 4, так как они отражают направления основных элементов знака: перекладин, стоек, наклонных линий.

Наиболее перспективными с точки зрения распознаваемости оказались последние шесть направлений (рис. 4, 23 – 4, 28).

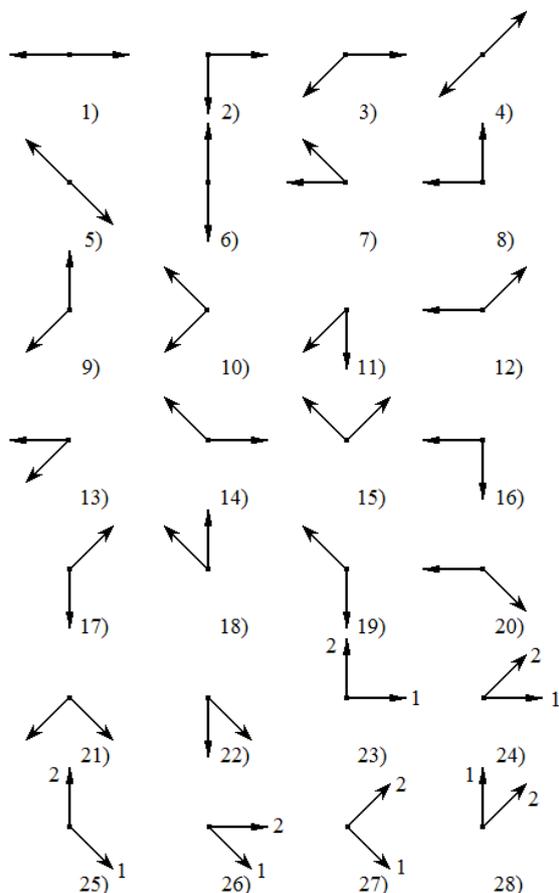


Рис. 4 – Комбинации двух направлений

Для выбранных шрифтов максимальное значение ω ($\omega = 0,0332$) было получено для комбинации, изображенной на рис. 4, 24. При распознавании букв сразу двух шрифтов значения ω определялись по формуле:

$$\omega = b_k \sum_{i=1}^{b_k} \frac{P_i}{2} \quad (6)$$

При выборе оптимальной комбинации трех направлений учитывались выбранная оптимальная комбинация двух направлений и направления основных элементов знака: перекладин, стоек, наклонных линий. В соответствии с этим анализировались комбинации направлений, изображенные на рис. 5.

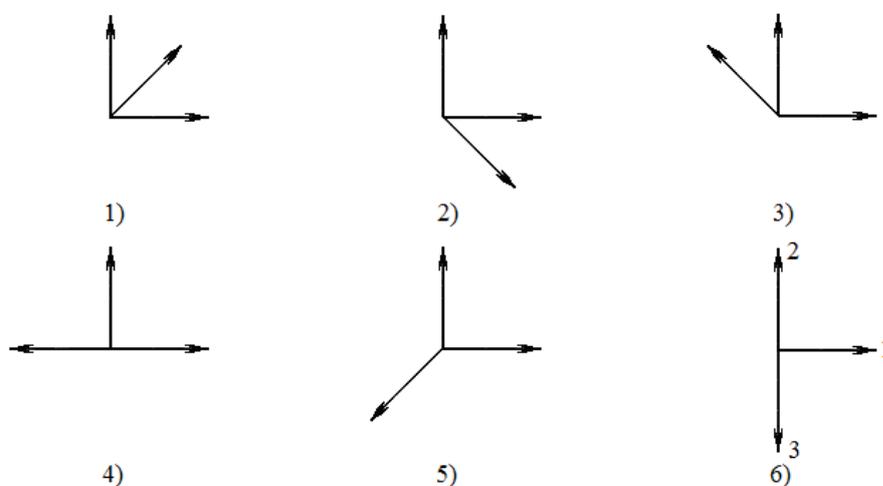


Рис. 5 – Комбинации из трех направлений

Максимальное значение ω ($\omega = 5,7155$) получилось для комбинации, изображенной на рис. 5, 6.

Основываясь на выбранной оптимальной комбинации трех направлений и направлений основных элементов знаков, были рассмотрены комбинации четырех направлений (рис. 6), из которых по наибольшей величине ω была выбрана оптимальная комбинация, изображенная на рис. 6, 5.

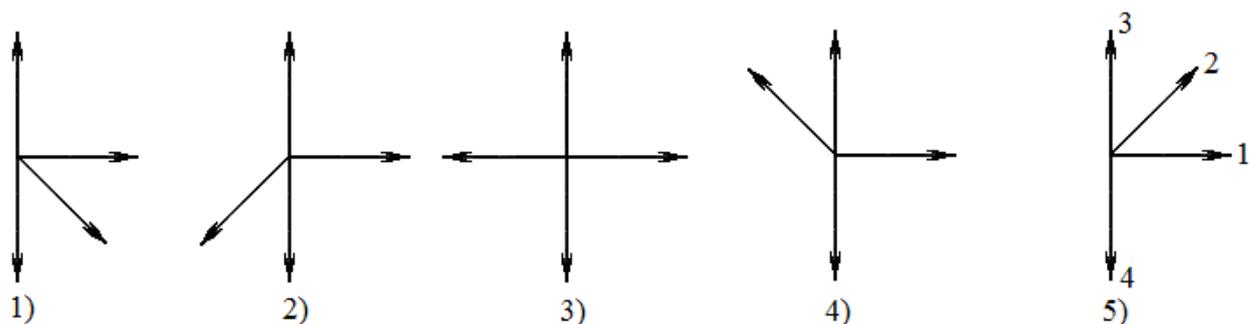


Рис. 6 – Комбинации из четырех направлений

Для этой комбинации направлений распознавалось 79 букв выбранных шрифтов. Аналогично ранее изложенным соображениям анализировались комбинации из пяти направлений (рис. 7), из которых была выбрана оптимальная комбинация, соответствующая рис. 7, 4.

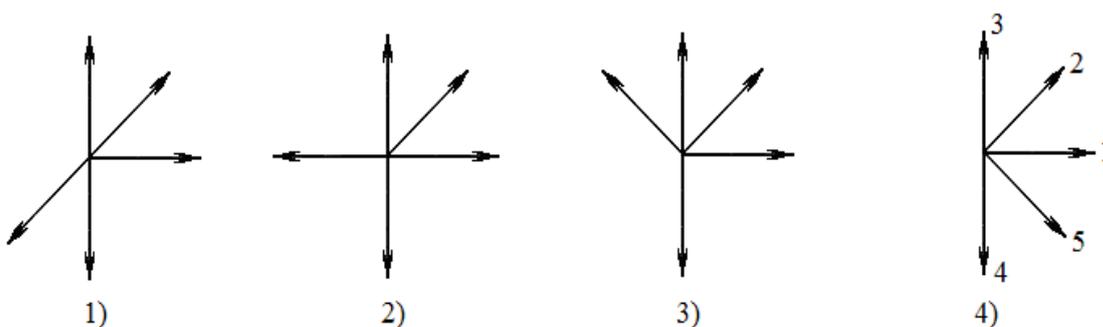


Рис. 7 – Комбинации из пяти направлений

Аналогично из комбинаций шести направлений (рис. 8) была выбрана оптимальная комбинация, изображенная на рис. 8, 1.

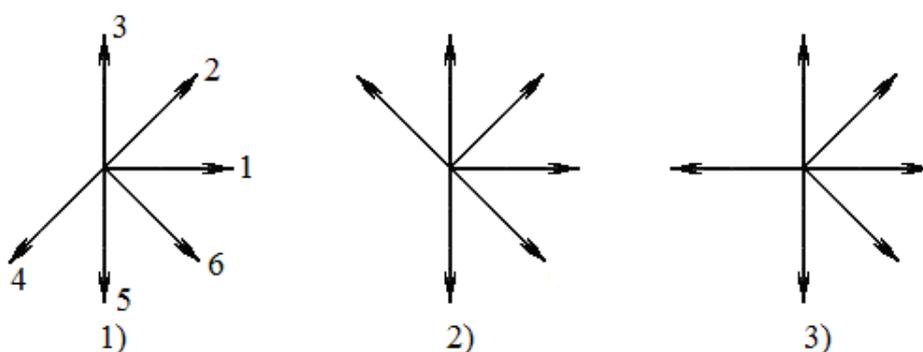


Рис. 8 – Комбинации из шести направлений

Из предпочтительных комбинаций семи направлений (рис. 9) по большей величине ω была выбрана комбинация, изображенная на рис. 9, 2.

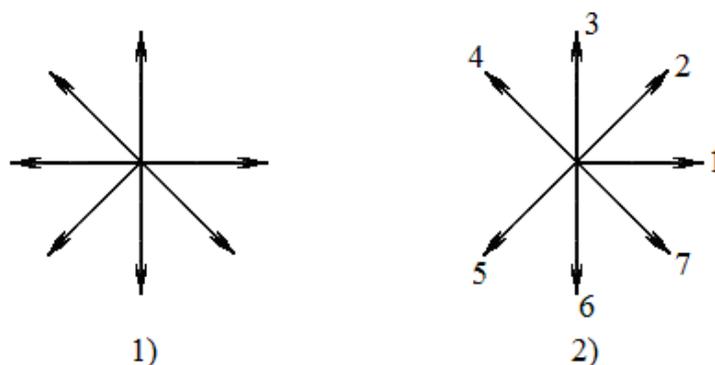


Рис. 9. Комбинации из семи направлений

Наконец, с введением последнего направления (рис. 10) все знаки выбранных печатных шрифтов распознавались.

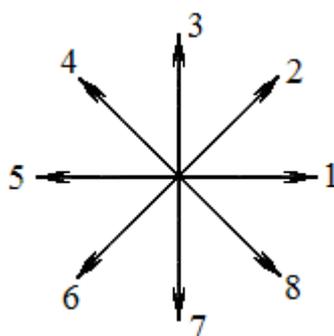


Рис. 10 – Выбранная комбинация направлений

Итак, анализ контуров знаков привел к тому, что для выбранной комбинации направлений (рис. 10) все знаки рассматриваемых печатных шрифтов, представленные в виде последовательности направлений, получаемой в результате обхода по контуру начиная с концевой точки, распознаются друг относительно друга независимо от масштаба элементов знака, от разных шрифтов, то есть выбранные направления действительно являются оптимальными, поскольку несут необходимые и достаточные признаки для распознавания всех знаков рассмотренных печатных шрифтов.

Представление знака в виде векторной суммы, образуемой элементами контура, позволило выявить характерные признаки знака – минимально возможное количество направлений (рис. 10).

Для выбранной оптимальной комбинации направлений авторами были составлены стандартные виды печатных знаков русского алфавита, полученные путем обхода идеально напечатанных контуров знаков анализируемых шрифтов, то есть в эти виды вошли также направления элементов контуров, специфичные для различных анализируемых шрифтов.

Печатные знаки разных шрифтов, но одного смыслового символа отличаются друг от друга толщиной линий, наклоном по отношению к вертикали знака в целом или его отдельных частей, наличием или отсутствием дефектов печати, масштабом элементов знака, наличием или отсутствием декоративных украшений – засечек, но одинаковым, общим для знаков одного смыслового символа является определенный стандартный вид.

Так как для распознавания знака компьютер по разработанной авторами методике проводит сравнение полученной после обхода знака по контуру последовательности направлений со стандартными видами знаков [7, 8], то последовательность для однозначного определения типа знака должна была быть как можно ближе приведенной к стандартному виду этого знака.

Поэтому по разработанной авторами программе компьютер для распознавания знака выделяет его контур, ликвидируя при этом следующее: лишнюю толщину элементов знака, нехарактерные отклонения вертикальных, горизонтальных и наклонных линий знака, определенные возможные дефекты из-за некачественной печати, нехарактерные засечки в форме знаков [9-13].

Таким образом, в результате проведенных авторами исследований были выбраны необходимые и достаточные информативные признаки знаков



различных печатных шрифтов, позволившие компьютеру по разработанной программе с высокой степенью надежности [6] распознавать печатные (дорожные) знаки различных шрифтов.

Литература

1. Lear I., «A machine that reads written, words». The New Scientist, 1959, v. 6-№ 154, pp.34-36.
2. «Direct Reading for Data Processing». Process Control and Automation, 1960, v. 7, № 3, pp. 15-18
3. Переверзев-Орлов В.С., Поляков В.Г. «Универсальный автомат для чтения печатного текста» Доклад на КОИМПАЧТ, 1961, г. Москва, с.12-16.
4. Letter segmenting apparatus for OCR comprising multi-level segmentor operable when binary segmenting fails by Toshio Miyazaki et al, NEC, published June 3, 1980. A detailed description of how an OCR system can identify individual letters in a written sample.
5. Chris Woodford. Last updated: January 2, 2018. URL: explainthatstuff.com/how-ocr-works.html.
6. Пинт Э.М., Яшин А.В., Еличев К.А. Vyhodnocení spolehlivosti tištěného rozpoznávání znaků počítače. // Materiay V medzynaodowey naukowj koferenci «Aktalne problém nowioczesnych nauk-2009. - Prezemys: Nauka I studia, 2009. – a. 103.
7. Галушка В.В., Фатхи В.А. Формирование обучающей выборки при использовании искусственных нейронных сетей в задачах поиска ошибок баз данных // Инженерный вестник Дона, 2013, №2. – URL: ivdon.ru/magazine/archive/n2y2013/1597.
8. Пинт Э.М., Петровнина И.Н., Романенко И.И., Еличев К.А. Заключительный алгоритм рационального метода распознавания

компьютером печатных знаков разных шрифтов и распространение метода на образы, связанные с автоматизацией работы дорожных машин и автомобилей. // Материалы IV международной научно-практической конференции «Перспективные направления развития автотранспортного комплекса», Пенза: Изд. ПГУАС, 2011. – с.165.

9. Пинт Э.М., Петровнина И.Н., Романенко И.И., Еличев К.А. Управление рабочими органами дорожных машин за счет устройства восприятия и распознавания печатных символов и знаков. // Материалы международной научно-практической конференции «Новые дороги России». – Пенза: Изд. ПГУАС, 2011. – с. 232.

10. Артемьев И.С., Лебедев А.И., Долгий А.И., Хатламаджиян А.Е., Меерович В.Д. Метод блочного оптического распознавания инвентарных номеров железнодорожных подвижных единиц на основе комитетной нейроиммунной модели классификации // Инженерный вестник Дона, 2014, №1. URL: ivdon.ru/ru/magazine/archive/n1y2014/2259.

11. Пинт Э.М., Романенко И.И., Еличев К.А. Основните елементи на системата за четене. // материал за 8-международна научна практическа конференция Naunovite научни постижения, 2012 том 32, София "Бял ГРАД-БТ", 2012 г. - с. 88.

12. Пинт Э.М., Петровнина И.Н., Романенко И.И., Еличев К.А. Полный алгоритм рационального метода распознавания компьютером печатных знаков разных шрифтов и других символов. // Научно-теоретический журнал Вестник. БГТУ им. В.Г. Шухова, № 1, 2013. – с. 210.

13. Пинт Э.М., Романенко И.И., Еличев К.А. Результаты исследования читающего устройства. // Научно-теоретический журнал «Вестник» № 1. – Белгород: Изд. БГТУ им. В.Г. Шухова, 2014. – с. 182.

References

1. Lear I., «A machine that reads written, words». The New Scientist, 1959, v. 6, № 154, pp.34-36.
2. «Direct Reading for Data Processing». Process Control and Automation, 1960, v. 7, № 3, pp. 15-18.
3. Pereverzev-Orlov V.S., Polyakov V.G. Universalnyj avtomat dlya chteniya pechatnogo teksta [Universal automatic machine for reading printed text], 1961, pp. 12-16.
4. Letter segmenting apparatus for OCR comprising multi-level segmentor operable when binary segmenting fails by Toshio Miyazaki et al, NEC, published June 3, 1980. A detailed description of how an OCR system can identify individual letters in a written sample.
5. Chris Woodford. Last updated: January 2, 2018. URL: explainthatstuff.com/how-ocr-works.html
6. Pint E.M., Yashin A.V., Elichev K.A. (Materiay V medzynarodowey naukowi konferenci «Aktalne problem nowioczesnych nauk-2009». Prezemys Nauka i studia), 2009. p.103
7. Galushka V.V., Fathi V.A. Inzhenernyj vestnik Dona (Rus), 2013, №2. URL: ivdon.ru/magazine/archive/n2y2013/1597.
8. Pint E.M. Petrovnina I.N., Romanenko I.I., Elichev K.A. (Materialy IV mezhdunarodnoj nauchno-prakticheskoy konferencii «Perspektivnye napravleniya razvitiya avtotransportnogo kompleksa»), Penza, PGUAS, 2011. p.165.
9. Pint E.M., Petrovnina I.N., Romanenko I.I., Elichev K.A. (Materialy mezhdunarodnoj nauchno-prakticheskoy konferencii «Novye dorogi rossii»), Penza, PGUAS, 2011. p.232



10. Artem'ev I.S., Lebedev A.I., Dolgij A.I., Hatlamadzhijan A.E., Meerovich V.D. Inzhenernyj vestnik Dona (Rus), 2014, №1. URL: ivdon.ru/ru/magazine/archive/n1y2014/2259.

11. Pint E.M., Romanenko I.I., Elichev K.A. (Materiali za 8-a mezhdunarodna nauchna praktichna konferenciya najnovite nauchni dostizheniya, 2012). V 32, Sofiya «Byal grad-bt», 2012. p.88

12 Pint E.M., Petrovnina I.N., Romanenko I.I., Elichev K.A. (Nauchno-teoreticheskij zhurnal Vestnik. BGTU im. V.G. Shuxova), V 1, 2013. 210 p.

13. Pint E.M., Romanenko I.I., Elichev K.A. (Nauchno-teoreticheskij zhurnal Vestnik. BGTU im. V.G. Shuxova), № 1, 2014. 182 p.