SLAVISTICA VILNENSIS 2018 63. 339–351. ISSN 2351-6895. Online ISSN 2424-6115
DOI: https://doi.org/10.15388/SlavViln.2018.63.11860

Александр Валерьевич Кучуганов, Денис Рашидович Касимов, Валерий Никонорович Кучуганов, Павел Петрович Осколков

Ижевский Государственный Технический Университет им. М.Т. Калашникова (Россия)

Автоматизация распознавания старославянских рукописей с помощью дескрипционных логик

В статье предлагается подход к распознаванию старославянских символов, включающий этапы тернарной сегментации, выделения границ цветовых областей, синтеза двух вариантов скелетона, формирования нечеткого графа, создания логического описания в нечеткой дескрипционной логике и распознавания с помощью подсистемы формальных автоматических рассуждений. Приводятся результаты предварительных экспериментов, которые подтверждают перспективность предложенного подхода[1].

Ключевые слова: распознавание, скелетон, структурные элементы, нечеткий граф, дескрипционная логика, старопечатные тексты, старославянский язык, древнерусский язык.

1. Введение

Автоматизация процесса перевода старославянских текстов в электронную форму, позволяющую применять компьютерные методы обработки текстовой информации, существенно упрощает проведение исторических и лингвистических исследований. Традиционные OCR-системы не справляются с распознаванием древних текстов. Специализированные методы и системы демонстрируют недостаточную надежность распознавания — в среднем не выше 80% на манускриптах неплохого качества.

Авторы работы *Feature Selection for Classification of Old Slavic Letters* [Bande et al. 2014] предлагают два способа распознавания старославянских кириллических символов: на основе “деревьев решений” и путем нечеткой классификации. Оба метода оперируют статистическими и простыми структурными признаками символов (свыше 20 признаков).

Экспериментальная система распознает символы каждым из способов со средней точностью и полнотой 70–80%. Следует, однако, отметить сложность настройки системы простым пользователем (не экспертом). При этом возможность простой и быстрой настройки системы (в частности, под конкретного писца) является весьма эффективным способом повышения качества распознавания.

В другой работе — *Word-Based Adaptive OCR for Historical Books* [Kluzner et al. 2009] — описана методика распознавания целых слов, а не отдельных символов исторических текстов. Сначала производится выделение отдельных слов на изображении, затем выполняется кластеризация полученного множества фрагментов-слов таким образом, чтобы в каждом кластере были представлены экземпляры одного и того же слова. После кластеризации производится распознавание изображений слов с помощью традиционного OCR-средства. К примеру, книгу 18 века, написанную немецким готическим шрифтом, разработанная программная система корректно распознала на 86,6%, что на 4,1% выше результатов традиционной OCR-системы. Однако, подходы пословного распознавания плохо применимы к кириллическим текстам, т.к. слова в них не отделены друг от друга и между буквами имеется относительно большое расстояние. Выделение отдельных слов в таких текстах представляет непростую задачу даже для человека-неспециалиста.

2. Предлагаемый подход

Предлагаемый нами подход к распознаванию старославянских символов схематично представлен на рис. 1.

Сначала производится скелетизация изображений символов, состоящая из следующих этапов: 1) тернарная сегментация изображения; 2) выделение границ черных областей и границ объединения серых и

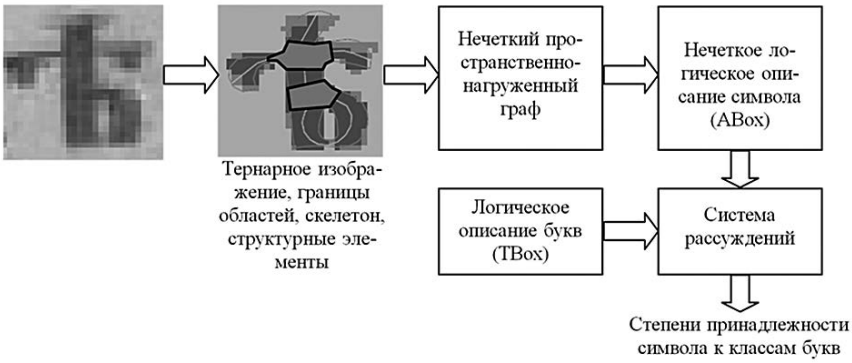


Рис. 1. *Процесс описания, анализа и распознавания символов*

черных областей; 3) синтез двух вариантов скелетона: черного и серо-черного; 4) аппроксимация скелетона; 5) идентификация областей перекрестков [Кучуганов 2012].

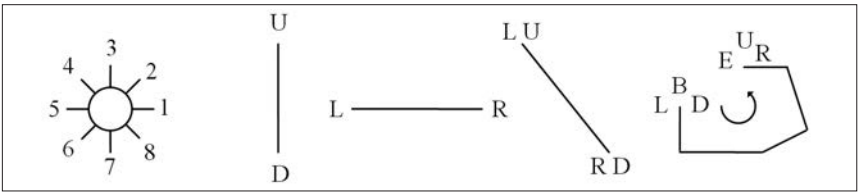
Затем выделяются конструктивные элементы (мачты, перекладины и т.п.) символа путем анализа формы отдельных цепочек элементов скелетона, а также соединительные элементы (области перекрестков).

Формируется нечеткий атрибутивный пространственно нагруженный граф изображения символа, вершины и ребра которого содержат количественные и качественные значения пространственных атрибутов, характеризующих конструктивные элементы символа.

Система имеет базу знаний о символах в форме TBox-описания в нечеткой дескрипционной логике f-SHIN [Straccia 2001]. Используемые элементы:

1. **Атомарные концепты**: КонЭл (конструктивные элементы), СоедЭл (соединительные элементы), ПрямойФормы (объекты с прямой формой), КруглойФормы, ОвальнойФормы, ИзвилистойФормы, ДугообразнойФормы, VобразнойФормы, UобразнойФормы, Замкнутый (замкнутые объекты), ВертОриентации (объекты, имеющие ориентацию “Север” или “Юг”), ГоризОриентации, НаклВлОриентации, НаклВпОриентации, Длинный (объекты большой длины), СреднейДлины, Короткий, Толстый (объекты большой ширины на протяжении всей длины), СреднейТолщины, Тонкий, ТолстыйВСередине (объекты большой ширины в середине) и др.
2. **Роли**: соед1L (соединительный элемент точкой 1 примыкает к точке L конструктивного элемента), соедL1 ≡ соед1L– (роль соед1L является обратной по отношению к роли соедL1), соедDL (конструктивный элемент точкой D примыкает к точке L другого конструктивного элемента), соед7B1E (соединительный элемент точками 7 и 1 примыкает соответственно к точкам B и E конструктивного элемента), соед7B1E ≡ соед1E7B, соедB7E1 ≡ соед7B1E– и т.п.

На рис. 2 представлены коды точек примыкания элементов. Точки примыкания соединительных элементов кодируются значениями их ориентации относительно центроидов элементов. Точки примыкания конструктивных элементов кодируются в зависимости от взаимного расположения по вертикали и горизонтали (верхняя или нижняя, левая или правая), а также по принципу выделения начальной и конечной точек в соответствии с направлением движения по элементу против часовой стрелки.



а) б)

Рис. 2. *Коды точек примыкания: а) у соединительных элементов; б) у конструктивных элементов*

3. **Производные концепты**: Линия ≡ КонЭл ∩ ¬Замкнутый ∩ ПрямойФормы; Мачта ≡ Линия ∩ ВертОриентации; Перекладина ≡ Линия ∩ ГоризОриентации; НаклВлЛиния ≡ Линия ∩ НаклВлОриентации; НаклВпЛиния ≡ Линия ∩ НаклВпОриентации; Петля ≡ КонЭл ∩ ¬Замкнутый ∩ ДугообразнойФормы; Круг ≡ КонЭл ∩ Замкнутый ∩ КруглойФормы; Овал ≡ КонЭл ∩ Замкнутый ∩ ОвальнойФормы и др. В таблице 1 представлен фрагмент описания букв и вспомогательных структур.

Таблица 1. *Описание конфигураций (букв и вспомогательных структур)*

Производный концепт	***Примеры экземпляров***	***Определение***
ВерхняяМачта-БуквыЬ		≡ Мачта ∩ СреднейДлины ∩ Толстый ∩ ∃соедD3.(СоедЭл ∩ ∃соед7U.(Мачта ∩ (СреднейДлины ∪ Короткий) ∩ Толстый ∩ ∃соедDL.(Перекладина ∩ СреднейДлины ∩ ТонкийВСередине)) ∩ ∃соед8E.(Петля ∩ СреднейДлины ∩ (Толстый ∪ СреднейТолщины)))
ВерхняяПравая-Перекладина		≡ (Перекладина ∪ НаклВпЛиния) ∩ (СреднейДлины ∪ Короткий) ∩ ТолстыйСлева ∩ ТонкийВСередине ∩ ТолстыйСправа
ВерхняяВерт-МачтаБуквыБ	 	≡ ВерхняяМачтаБуквыЬ ∩ ∃соедUL.ВерхняяПравая-Перекладина
ВерхняяМачта-БуквыЪ	 	≡ ВерхняяМачтаБуквыЬ ∩ ∃соедUR.ВерхняяЛеваяПерекладина

Для распознаваемого символа строится нечеткое ABox-описание. В примере на рис. 3 сказано, что v1 является экземпляром концепта КонЭл с полной уверенностью, экземпляром концепта ВертОриентации – со степенью принадлежности 0.3; между v1 и v2 существует отношение соедRU с полной уверенностью, между v2 и v3 – отношение соедD3 со степенью уверенности 0.8 и т.д.

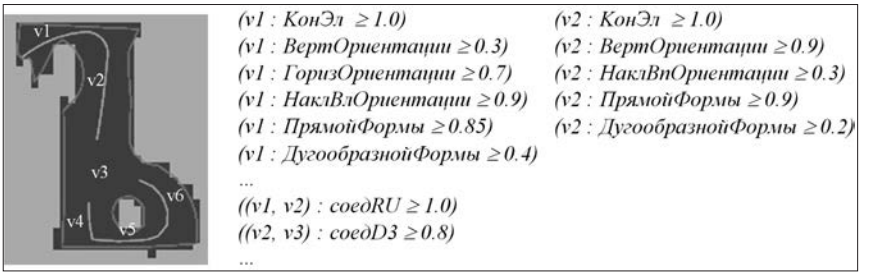


Рис. 3. *Пример распознаваемого символа и его нечеткого логического описания*

Распознавание символа мы сводим к выполнению в подсистеме автоматических рассуждений (reasoner) запросов следующего содержания: “определить элемент символа, который с высокой степенью принадлежности является экземпляром концепта ВерхняяМачтаБуквыЪ”. В примере, представленном на рис. 3, подсистема рассуждений определит, что элемент v2 является экземпляром концепта ВерхняяМачтаБуквыЪ со степенью принадлежности, равной 0,93.

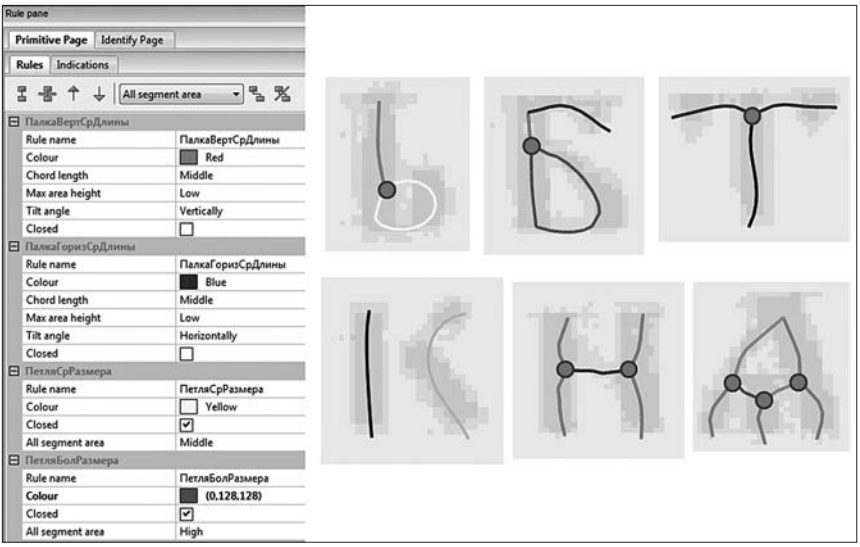
3. Описание реализации

В базе знаний системы имеются логические определения (DL-описания) конструктивных элементов, из которых состоят старославянские буквы. Интерфейс ввода определений конструктивных элементов представлен на рис. 4(а). Определением элемента является набор отличительных признаков и их возможных значений. Значения задаются как в четкой, так и нечеткой форме.

Конструктивные элементы описываются следующими атрибутами: длина хорды, количество сегментов слева и справа от хорды, площадь каждого сегмента, площадь левых/правых сегментов, площадь всех сегментов, максимальная высота сегмента, угол наклона элемента, замкнутость.

Один и тот же элемент может подходить под несколько определений с той или иной степенью уверенности (принадлежности).

Примеры конкретных элементов, соответствующих введенным определениям, показаны на рис. 4(б).



а) б)

Рис. 4. *База знаний (TBox) о конструктивных элементах символов: а) пример DL-описаний конструктивных элементов; б) пример распознавания конструктивных элементов по логическим описаниям*

База знаний системы содержит также логические определения букв. На рис. 5(а) приведен интерфейс ввода определений букв. Буква рассматривается как конструкция из элементов, определенных ранее. В определении буквы указываются входящие в ее состав элементы и способы

а) б)

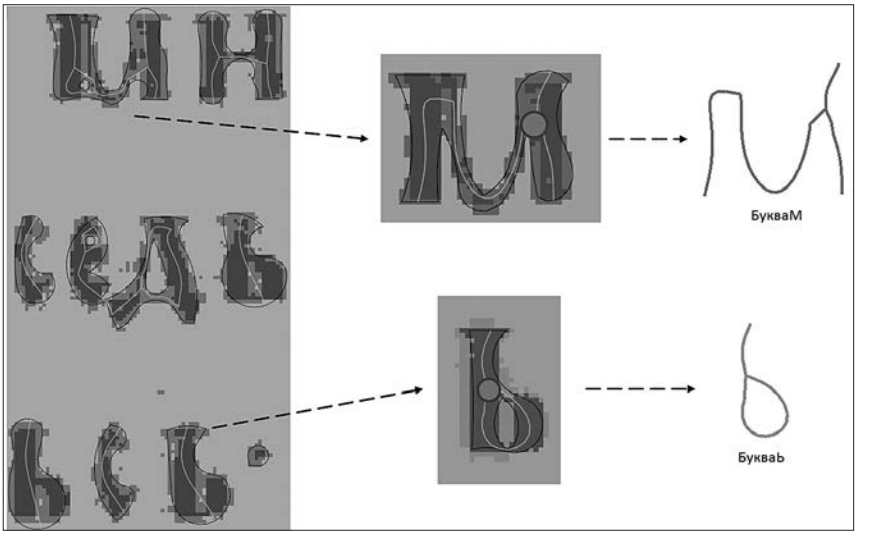
Рис. 5. *База знаний (TBox) о символах: а) пример DL-описаний символов; б) пример распознавания символов по логическим описаниям*

примыкания их друг к другу, т.е. какими точками они друг с другом соединяются.

На рис. 5 (б) приведены примеры конкретных букв, соответствующих введенным определениям. Распознавание осуществляет подсистема логических рассуждений.

Один и тот же символ подсистема рассуждений может отнести сразу к нескольким определениям букв. Каждый вариант распознавания обладает определенной степенью уверенности (принадлежности).

Для нераспознанных символов осуществляется попытка построения альтернативных версий скелетона. Для этого производится адаптивная цветовая сегментация изображения символа, которая позволяет получить более качественный скелетон и успешно распознать символ (рис. 6).



а) б) в)

Рис. 6. *Адаптивная цветовая сегментация нераспознанных символов: а) исходный скелетон; б) результат адаптивной сегментации; в) результат распознавания*

Результатом работы модуля графического распознавания являются варианты значений каждого символа (рис. 7).

Эти данные далее передаются в модуль уточнения по грамматическому словарю древнерусского языка. Словарь создан в рамках проекта “Манускрипт” (manuscripts.ru) под руководством В.А. Баранова.

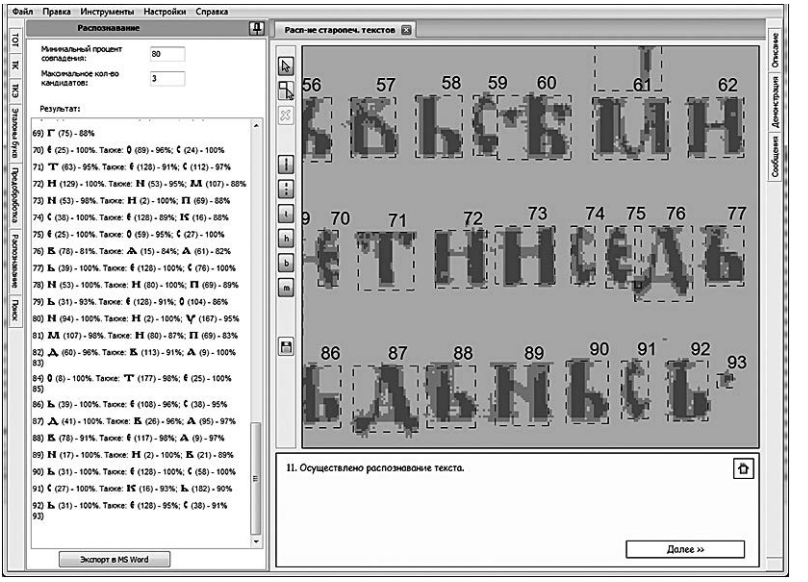


Рис. 7. *Пример результатов графического распознавания символов*

4. Эксперименты

В экспериментальном исследовании использовался текст факсимильного издания Остромирова Евангелия 1056–1057 гг. Тестовый набор содержал 10 страниц данного текста.

В таблице 2 представлены значения показателей полноты и точности распознавания по отдельным буквам, полученные в нашей системе, а также в системе исследователей из Македонии [Bande et al. 2014] (по данным авторов).

Показатели вычислялись следующим образом: пусть S — множество изображений отдельных символов на старопечатных текстах из тестового набора; b — некоторая буква алфавита; $Sb \subset S$ — подмножество символов, являющихся буквой b; $Rb \subset S$ — подмножество символов, которые система отнесла к букве b. Тогда:

$Полнота_b = |Rb \cap Sb| / |Sb|$,

$Точность_b = |Rb \cap Sb| / |Rb|$.

Для нашей системы в таблице 2 дополнительно приведены значения полноты с учетом кандидатов на результат распознавания, которые дают оценку того, какое качество распознавания можно достичь за счет последующего уточнения с помощью словаря древнерусского языка.

Полученные результаты показывают, что предложенный подход распознает символы со средней полнотой 86%, точностью 85%, что не-

Таблица 2. *Сравнительные оценки качества распознавания символов*

Буква	Наш экспериментальный модуль распознавания			Классификатор на основе дерева решений (Македония)	
	Полнота	Точность	Полнота с учетом потенциальных (для последующего уточнения с помощью словаря)	Полнота	Точность
а	0.75	0.87	1	0.5	0.63
б	0.89	0.89	1	0.67	1
в	0.88	0.60	1	0.88	0.64
є	0.81	0.59	1	0.56	1
л	0.85	0.95	1	0.89	1
м	0.80	0.92	0.90	0.8	1
н	0.83	0.89	1	1	0.82
н	0.86	0.74	1	1	0.82
о	0.92	0.90	1	0.86	0.86
с	0.91	0.87	1	1	0.82
т	0.82	0.95	1	0.86	0.58
у	0.79	0.70	0.85	0.44	0.67
ь	0.96	0.92	1	0.71	0.71
ъ	1	1	1	0.79	0.65
ѣ	0.86	0.87	1	нет данных	
ѧ	0.83	1	1	0.4	1
Ср.:	0.86	0.85	0.98	0.76	0.81

сколько выше, чем у известных аналогов. Здесь следует учитывать, что эксперименты в нашей системе и в аналогичных системах проводились на разных тестовых наборах изображений.

Также был произведен эксперимент по распознаванию текста Постной Триоди (РНБ, Погод. 41) (рис. 8).

База знаний была предварительно пополнена дополнительными вариантами описаний букв, поскольку в Постной Триоди (РНБ, Погод. 41)

Рис. 8. *Фрагмент текста Постной Триоди (РНБ, Погод. 41)*

написание некоторых букв (таких как “веди”, “земля”, “иже”) немного отличается от написаний в Остромировом Евангелии.

Анализировались две версии скелетона: 1) скелетон черного и 2) скелетон серо-черного. Обе версии скелетона представлены на рис. 9. Черным цветом отображены правильно распознанные символы; белым цветом показаны нераспознанные фрагменты скелетонов.

а) б)

Рис. 9. *Результаты распознавания символов: а) символы, правильно распознанные по скелетону черного; б) символы, правильно распознанные по скелетону серо-черного*

В общей сложности на данном изображении программа правильно распознала 287 из 354 символов (81%). Распознавание надстрочных знаков в данной версии системы не реализовано.

Заключение

Таким образом, применение дескрипционной логики позволяет непротиворечиво и структурировано представить синтаксис изображения старославянского символа в удобочитаемом формате (как для человека, так и для компьютера) и производить анализ средствами формальных логических рассуждений.

Новизна метода распознавания состоит в адаптивной тернарной сегментации изображения и синтезе описания на основе аппарата дескрипционной логики.

Дальнейшее совершенствование подхода предполагает: усиление алгоритма адаптивной цветовой сегментации, усиление выразительности дескрипционного описания букв, автоматический синтез эталонных DL-описаний новых вариантов букв, разработку модуля уточнения результатов распознавания на основе словарей древнерусского языка.

Примечание

[1] Работа выполнена при частичной поддержке РФФИ (№ 15-07-08077 А) и Госзадания МОиН РФ (Проект № 625).

Литература

Кучуганов А. В., 2012.: Когнитивный алгоритм построения геометрического остова невыпуклых фигур, *Приволжский научный журнал*, № 3 (23). Н. Новгород, 84–89.

Bande C.M., Klekovska M., Nedelkovski I. et al., 2014: Feature Selection for Classification of Old Slavic Letters, *Control Engineering and Applied Informatics*, 16.4. 81–90.

Kluzner, V., Tzadok A., Shimony Y. et al., 2009: Word-Based Adaptive OCR for Historical Books, *Proceedings of the 10th International Conference on Document Analysis and Recognition* (ICDAR2009). 501–505.

Straccia U., 2001: Reasoning within fuzzy description logics, *Journal of Artificial Intelligence Research*, *14*. 137–166.

Bibliography (Transliteration)

Bande C.M., Klekovska M., Nedelkovski I. et al., 2014: Feature Selection for Classification of Old Slavic Letters, *Control Engineering and Applied Informatics*, 16.4. 81–90.

Kluzner, V., Tzadok A., Shimony Y. et al., 2009: Word-Based Adaptive OCR for Historical Books, *Proceedings of the 10th International Conference on Document Analysis and Recognition* (ICDAR2009). 501–505.

Kuchuganov A. V., 2012.: Kognitivnyy algoritm postroeniya geometricheskogo ostova nevypuklykh figur, *Privolzhskiy nauchnyy zhurnal*, № 3 (23). N. Novgorod, 84–89.

Straccia U., 2001: Reasoning within fuzzy description logics, *Journal of Artificial Intelligence Research 14*. 137–166.

Aleksander Kuchuganov, Denis Kasimov,
Valeriy Kuchuganov, Pavel Oskolkov

Automation of Recognizing Old Slavonic Manuscripts with help of Description Logics

In the paper an approach to recognition of old Slavonic symbols is proposed, which includes the stages of ternary segmentation, detecting edges of areas, synthesis of two skeleton variants, forming a fuzzy graph, creating a logical description in a fuzzy description logic, and recognition with help of a subsystem of formal automatic reasoning. Results of preliminary experiments are provided, which confirm the perspective of the proposed approach.

Keywords: recognition, skeleton, structural elements, fuzzy graph, description logic, old printed texts, Old Church Slavonic language, Old Russian language.

Aleksandr Kučuganov, Denis Kasimov,
Valerij Kučuganov, Pavel Oskolkov

Senųjų slavų rankraščių tekstų atpažinimo automatizavimas deskripcinės logikos pagalba

Straipsnio autoriai siūlo naują senosios slavų kalbos simbolių atpažinimo metodiką, kurios etapai: ternarinė segmentacija, spalvų sričių ribų išskyrimas, dviejų skeletono variantų sintezė, neryškaus grafo formavimas, loginio aprašo kūrimas netikslioje diskripcinėje logikoje ir senųjų simbolių atpažinimas automatinės sistemos pagalba. Straipsnyje pateikiami preliminarių bandymų rezultatai, patvirtinantys siūlomo atpažinimo metodo perspektyvumą.

Reikšminiai žodžiai: atpažinimas, struktūriniai elementai, neryškus grafas, deskripcinė logika, senųjų spausdinių tekstai, senoji slavų kalba, senoji rusų kalba.

Поступило в редакцию: 16 марта 2018 г.
Принято к печати: 15 мая 2018 г.

Александр Валерьевич Кучуганов, к.т.н., доцент кафедры Автоматизированные системы обработки информации и управления Ижевского государственного технического университета им. М.Т. Калашникова.

Aleksander Kuchuganov, Ph.D., Assoc. prof., Department of Automated Data Processing and Control Systems, Kalashnikov Izhevsk State Technical University.

Aleksandr Kučiuganov, technikos mokslų daktaras, Iževsko valstybinio M.T. Kalašnikovo technikos universiteto Automatizuotų informacijos ir valdymo apdorojimo sistemų katedros docentas.

E-mail: aleks_kav@udm.ru.

Денис Рашидович Касимов, к.т.н., доцент кафедры Автоматизированные системы обработки информации и управления Ижевского государственного технического университета им. М. Т. Калашникова.

Denis Kasimov, Ph.D., Assoc. prof., Department of Automated Data Processing and Control Systems, Kalashnikov Izhevsk State Technical University.

Denis Kasimov, technikos mokslų daktaras, Iževsko valstybinio M. T. Kalašnikovo technikos universiteto Automatizuotų informacijos ir valdymo apdorojimo sistemų katedros docentas.

E-mail: kasden@mail.ru

Валерий Никонорович Кучуганов, д.т.н., профессор, заведующий кафедрой Автоматизированные системы обработки информации и управления Ижевского государственного технического университета им. М. Т. Калашникова.

Valeriy Kuchuganov, Habil. prof., Head, Department of Automated Data Processing and Control Systems, Kalashnikov Izhevsk State Technical University.

Valerij Kučiuganov, habilituotas technikos mokslų daktaras, profesorius; Iževsko valstybinio M.T. Kalašnikovo technikos universiteto Automatizuotų informacijos ir valdymo apdorojimo sistemų katedros vedėjas.

E-mail: kuchuganov@istu.ru;

Павел Петрович Осколков, инженер кафедры Автоматизированные системы обработки информации и управления Ижевского государственного технического университета им. М. Т. Калашникова.

Pavel Oskolkov, engineer, Department of Automated Data Processing and Control Systems, Kalashnikov Izhevsk State Technical University.

Pavel Oskolkov, inžinierius, Iževsko valstybinio M.T. Kalašnikovo technikos universiteto Automatizuotų informacijos ir valdymo apdorojimo sistemų katedra.

E-mail: aleks_kav@udm.ru

XIX вѣкъ

1.

Мѣмайнѣ·а·кастѣ

[illegible]

Тибиш жажи
Софийска

Мерларов
сотр

1855 1990