Электронное моделирование

УДК 004.896+656:25

М. Н. Василенко, д-р техн. наук
Р. А. Ковалев
Кафедра «Автоматика и телемеханика на железных дорогах»,
Петербургский государственный университет путей сообщения
Императора Александра I

МЕТОДЫ ВЫДЕЛЕНИЯ ТЕКСТОВЫХ ВЫРАЖЕНИЙ ПРИНЦИПИАЛЬНЫХ ЭЛЕКТРИЧЕСКИХ СХЕМ ЖЕЛЕЗНОДОРОЖНОЙ АВТОМАТИКИ И ТЕЛЕМЕХАНИКИ

Автоматизированное распознавание принципиальных электрических схем железнодорожной автоматики и телемеханики – актуальная и сложная задача. Алгоритм ее решения разделяется на несколько специализированных алгоритмов. Основными подзадачами можно назвать выделение и распознавание: структуры принципиальных электрических схем, текста, штампа и пр. В данной статье описывается подход к обработке текста на принципиальных электрических схемах железнодорожной автоматики и телемеханики.

Текстовая информация на этих схемах является крайне важной, без ее анализа невозможно произвести полный перевод данных сканированного изображения в электронный вид. Предлагаются алгоритм выделения текстовой информации со схемы с применением алгоритма кластеризации для выделения групп символов и метод составления разделенных уникальных выражений (лексем) в группе.

Описывается анализ полученных вариантов текстовых выражений и способы выбора наиболее корректных вариантов. Приводится общая схема процесса обработки принципиальных электрических схем с учетом методов анализа текстовой информации, предлагаемых в статье.

Описан прототип программы, реализующий разработанные методы, и его использование на тестовой выборке из 300 печатных и затем отсканированных принципиальных электрических схем разного качества.

электронный документооборот; техническая документация; распознавание образов; распознавание текста; принципиальные схемы

Введение

Достаточно большое количество документации, в том числе в ОАО «РЖД», в настоящее время представлено в бумажной форме. Одним из на-

правлений развития ОАО «РЖД» является переход на электронный документооборот технической документации, и прежде всего, перевод принципиальных электрических схем (ПЭС) в электронный вид [1–10].

Текстовую информацию на ПЭС можно разделить на нескольких основных типов:

– собственно информация (атрибуты) элементов ПЭС;

– информация о логических связях документов (межлистовые переходы);

– информация на штампе;

– прочая информация.

В процессе распознавания атрибутная информация крайне важна и если возложить ее обработку на оператора, это сделает общий процесс перевода малоэффективным. Однако выполнить распознавание текста в слабоструктурированном документе с точностью, не требующей последующего вмешательства человека, крайне сложно. Пример такой информации приведен на рис. 1.

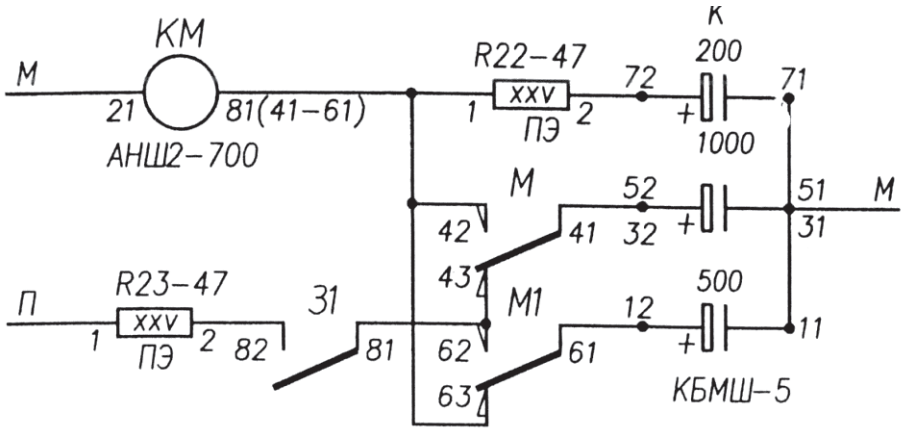

Рис. 1. Текстовая информация элементов ПЭС

В данной работе ставится задача выделения текстовой информации для дальнейшей обработки с помощью известных алгоритмов распознавания текста. Имеющиеся средства распознавания и выделения текстовой информации из документов оперируют в основном документами, выполненными по правилам построения текстовых документов, содержащих абзацы, блоки текста, связные предложения и числовые форматы, а также таблицы. Наиболее известные представители такого программного обеспечения – ABBYY FineReader, Adobe Acrobat, ScanSnap, Tesseract. Однако данные продукты не позволяют в полной мере решить поставленную задачу, так как не учитывают особенности структуры ПЭС, расположения текстовой информации относительно элементов ПЭС и смысловой составляющей текстовых выражений. Авторы статьи ставили цель – разработать простой и эффективный метод выделения текстовой информации ПЭС, основанный на алгоритмах кластеризации по признакам расположения и размера текстовых символов.

1 Анализ и выделение графических объектов принципиальных электрических схем

Для применения системы распознавания в отделах документации необходимо минимизировать трудоемкость внесения и исправления атрибутной информации элементов. В определенных случаях по данной информации можно восстановить недоброкачественно отрисованные элементы, что будет полезно при реализации процедур последующего анализа. Примером такой обработки может быть следующая цепочка рассуждений: распознана информация о нескольких типах реле → произведен анализ текстовой информации → получена текстовая информация, соответствующая определенной марке реле → выбрано реле по соответствию марке. Данный пример является простым случаем латентно-семантического анализа в системах распознавания текста.

Дополнение системы распознавания ПЭС информацией об атрибутах основывается на геометрической области привязки и логическом сопоставлении с определенным шаблоном. Такое представление является интуитивно понятным и может быть использовано в качестве основы для более сложных алгоритмов поиска. Предлагается рассмотреть реле КМ, изображенное на рис. 1. Следующая информация является важной при переводе данного элемента в электронный вид:

– элемент является нейтральным реле постоянного тока;
– реле имеет два вывода, обозначенные номерами 21 и 81;
– указан монтажный адрес;
– указано наименование;
– указана марка.

Система распознавания анализирует графическую составляющую примитивов данного отображения элемента и в качестве результата предоставляет информацию о том, что данный элемент – реле постоянного тока. Это первичная информация анализа, остальные данные вычисляются постфактум. На рис. 2 представлена первичная информация, полученная при распознавании.

Расположение атрибутной информации на рис. 2 намерено опущено.

Анализ тестовых данных начинается с выделения связанных символов – лексем. На рис. 3 представлен пример области, содержащей текстовую информацию некоторых элементов.

Основное отличие текстовых фрагментов ПЭС от блоков текста прочих типов документов заключается в их геометрической привязанности к расположению графических элементов документа, для описания которых они используются. Таким образом, как правило, при поиске и распознавании лексем текстовые фрагменты ПЭС можно рассматривать по отдельным областям, что во многом ускоряет процесс распознавания. Для выделения областей поиска

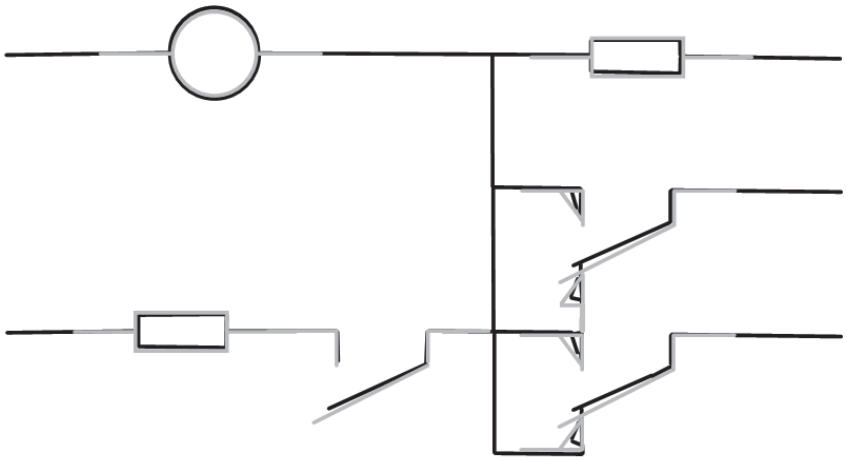

Рис. 2. Первичная информация при распознавании ПЭС

можно воспользоваться различными методами кластеризации, используя геометрические координаты потенциальных символов лексем в качестве параметров. Одним из распространенных и эффективных методов является метод кластеризации DBSCAN (Density-based spatial clustering of applications with noise) [11]. Алгоритм DBSCAN был предложен Мартином Эстером, Гансом-Питером Кригелем и их коллегами в 1996 г. как решение проблемы разбиения данных на кластеры произвольной формы. Идея, положенная в основу алгоритма, заключается в том, что внутри каждого кластера наблюдается типичная плотность точек (объектов), которая заметно выше, чем плотность снаружи кластера, а также плотность в областях с шумом ниже плотности любого из кластеров.

Используя алгоритм DBSCAN при разбиении информации на рис. 3 на области получаем 9 зон (рис. 4). Такое разбиение гарантирует цельность лексем и дальнейшая обработка отдельных кластеров может быть произведена параллельно, чем достигается высокая производительность процесса.

Обычно при построении строк (lines) используют несколько параметров этих строк и один из алгоритмов сравнения. Одним из возможных методов

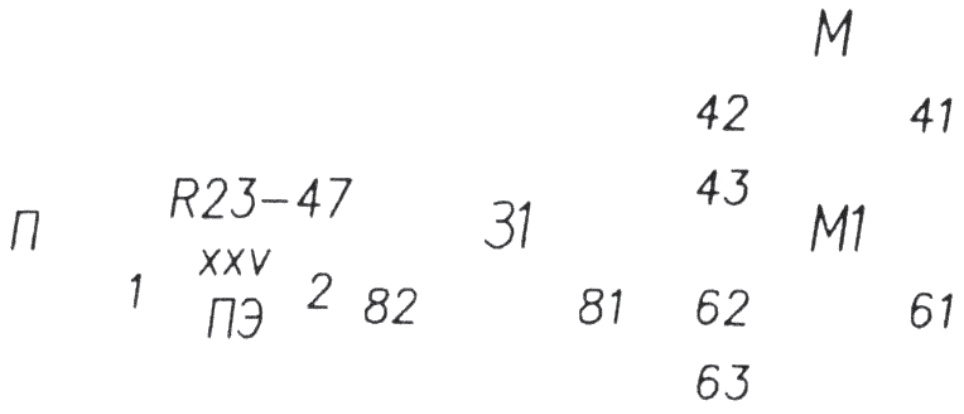

Рис. 3. Тестовая информация схемы ПЭС

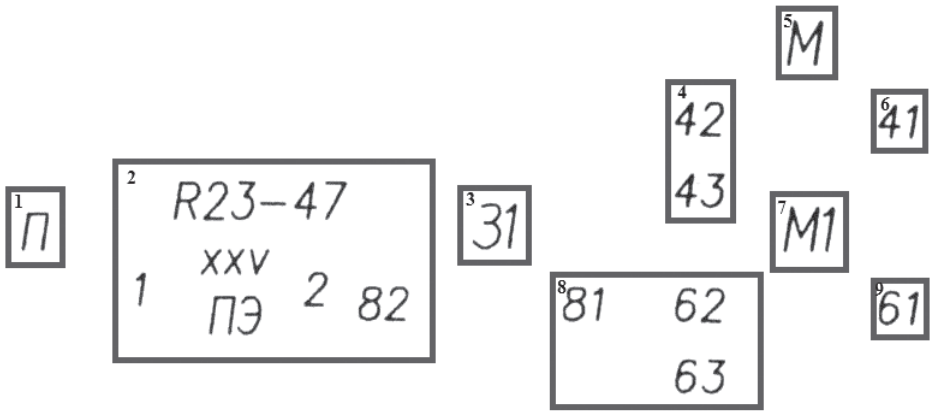

Рис. 4. Результат разбиения текстовых данных

является метод линейной зависимости двух параметров (уравнение прямой на плоскости с угловыми коэффициентами):

$$y = a + bx.$$

Два символа в строке образуют прямую на плоскости. Остальные символы строки проверяются на превышение заданного порога отклонения от данной прямой. Ограничением данного подхода является случай вертикальных строк. Использование общего уравнения прямой ($Ax + By + C = 0$) позволяет обрабатывать также вертикальные строки.

2 Построение возможных вариантов текстовых выражений и выбор наиболее корректных вариантов

Выбор лексем предлагается делать по принципу построения дерева путей от корневого узла-символа во всех направлениях. Такой подход позволяет определить все возможные лексемы по принципу обхода в глубину дерева символов и анализа их расположения относительно позиции в каждой лексеме.

Процедура start-line (root, S, D_{max}) задействует все символы множества S, выбирая те, которые расположены на расстоянии, меньшем D_{max} от символа root. На данном этапе под символом понимается связанная область пикселей изображения (рис. 5).

Параметр D_{max} определяет максимальное расстояние между символами лексемы. Таким образом, дальнейший подбор символов лексемы осуществляется по принципу выбора символов, имеющих минимальное отклонение от образующей прямой и лежащих на расстоянии не более D_{max} от предыдущего символа (рис. 6).

Аргумент w_{max} определяет максимальное отклонение координаты символа лексемы от образующей прямой.

start-line (root, S, $D_{\max}$):
для каждого символа S_i множества S:
1. Если $\sqrt{(x_{root} - x_{S_i})^2 + (y_{root} - y_{S_i})^2} > D_{\max}$, то перейти к п. 4.
2. $y_i = \dfrac{x_{S_i} y_{root} - x_{root} y_{S_i} - x(y_{root} - y_{S_i})}{x_{S_i} - x_{root}}$.
3. Добавить уравнение y_i в список начальных узлов лексем L.
4. Перейти к следующем элементу множества S.

Рис. 5. Псевдокод алгоритма start-line

propagation (y_i (arg), S, $D_{\max}$, $w_{\max}$):
1. Пометить все символы S белым.
2. Если в множестве S нет белых символов, то перейти к п. 9.
3. S_j – любой белый символ в множестве S.
4. Пометить S_j черным.
5. Если $\left| y_i(x_{Sj}) - y_{Sj} \right| > w_{\max}$, то перейти к п. 2.
6. S_{last} – последний добавленный символ i лексемы списка L, если $\sqrt{(x_{S_{last}} - x_{S_j})^2 + (y_{S_{last}} - y_{S_j})^2} > D_{\max}$, то перейти к п. 2.
7. Добавить S_j в i лексему списка L.
8. Перейти к п. 2.
9. Закончить обход.

Рис. 6. Псевдокод алгоритма propagation

Процедура propagation работает по принципу поиска в глубину (Depth-first search, DFS) [12] на некотором множестве объектов V со связями, принадлежащими множеству E. Алгоритм имеет сложность max (|V|,|E|).

В алгоритмах на рис. 5 и рис. 6 используется метод вычисления уравнения прямой по двум точкам, что не позволяет рассматривать вертикальное направление при анализе лексем. Приведенный выше метод замещения уравнения прямой с угловым коэффициентом позволяет решить данную проблему, однако возможно дополнить алгоритм проверкой на параллельность образующей прямой оси OY и выполнить преобразование координат символов.

На рис. 7 рассмотрен пример работы алгоритмов построения лексем в кластере № 2 (рис. 4). В качестве ключевого выбран символ 3. Линиями отмечены образующие лучи и минимальные расстояния символов лексем к образующим лучам. Точками отмечены координаты символов.

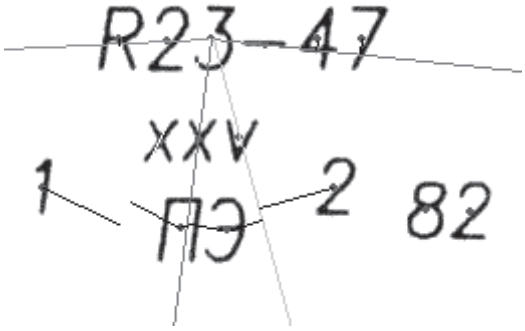

Рис. 7. Работа алгоритмов построения лексем

По окончании процесса поиска всех возможных лексем, когда каждый символ рассматривается в качестве потенциального корневого узла, производится выборка наиболее подходящих наборов лексем. Данный анализ может быть выполнен в несколько этапов с использованием различной информации, экспертных систем и нейронных сетей. Эффективный и простой метод – двухэтапный выбор по критериям относительного расположения и поиска по словарю. Согласно данному методу, последовательно выбираются сонаправленные, с учетом погрешности, лексемы и производится поиск лексем по предметному словарю. Словарь в простейшем случае является списком возможных масок атрибутов элементов ПЭС. Так, к примеру, информацию о резисторе R23-47 дает маска «R.*». Такая простая проверка не является достаточной, но если в словаре найдено сразу несколько соответствий для набора сонаправленных лексем, то вероятность корректности данного набора является удовлетворительной. На рис. 8 показан результат работы описанного метода.

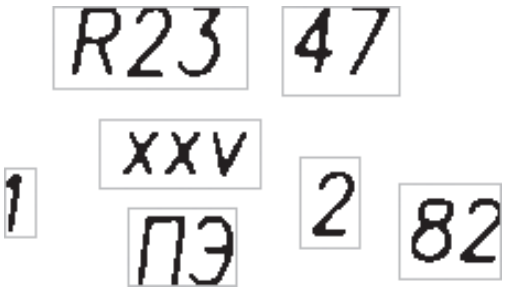

Рис. 8. Результат работы алгоритмов построения и выбора лексем

Для использования словаря необходимо перевести связанные области пикселей в соответствующие им символы языка. Задача распознавания отдельных символов известна и имеет множество решений, которые могут быть использованы при распознавании текстовой информации ПЭС. К примеру, неплохо показал себя метод, предложенный Д. В. Зуевым, он основан на обу-

чении нейронной сети и опробован на документации железнодорожной автоматики [9]. Данный метод поиска отдельных текстовых выражений ПЭС позволяет вычислять угол поворота символов строки относительно остальной структуры ПЭС и использовать эту информацию при распознавании отдельных символов строки.

Улучшить качество предлагаемого метода может использование информации о близлежащих элементах. В самом простом случае данная информация способна ограничить словарь поиска, что позволит сократить количество ошибочно выбираемых записей словаря, относящихся к элементам ПЭС, не являющихся атрибутными носителями распознаваемых выражений, но имеющих похожие маски атрибутов. На практике такие случаи встречаются довольно часто.

Для повышения эффективности метода возможен учет геометрических параметров элементов ПЭС. Текстовая информация располагается на схеме рядом с соответствующим ей элементом и характеризуется углом поворота. Такое дополнение анализа дает еще более точный результат распознавания текстовой информации и дополнительно проверяет элементную базу распознанной ПЭС. На рис. 9 представлен общий подход выделения текстовой информации ПЭС.

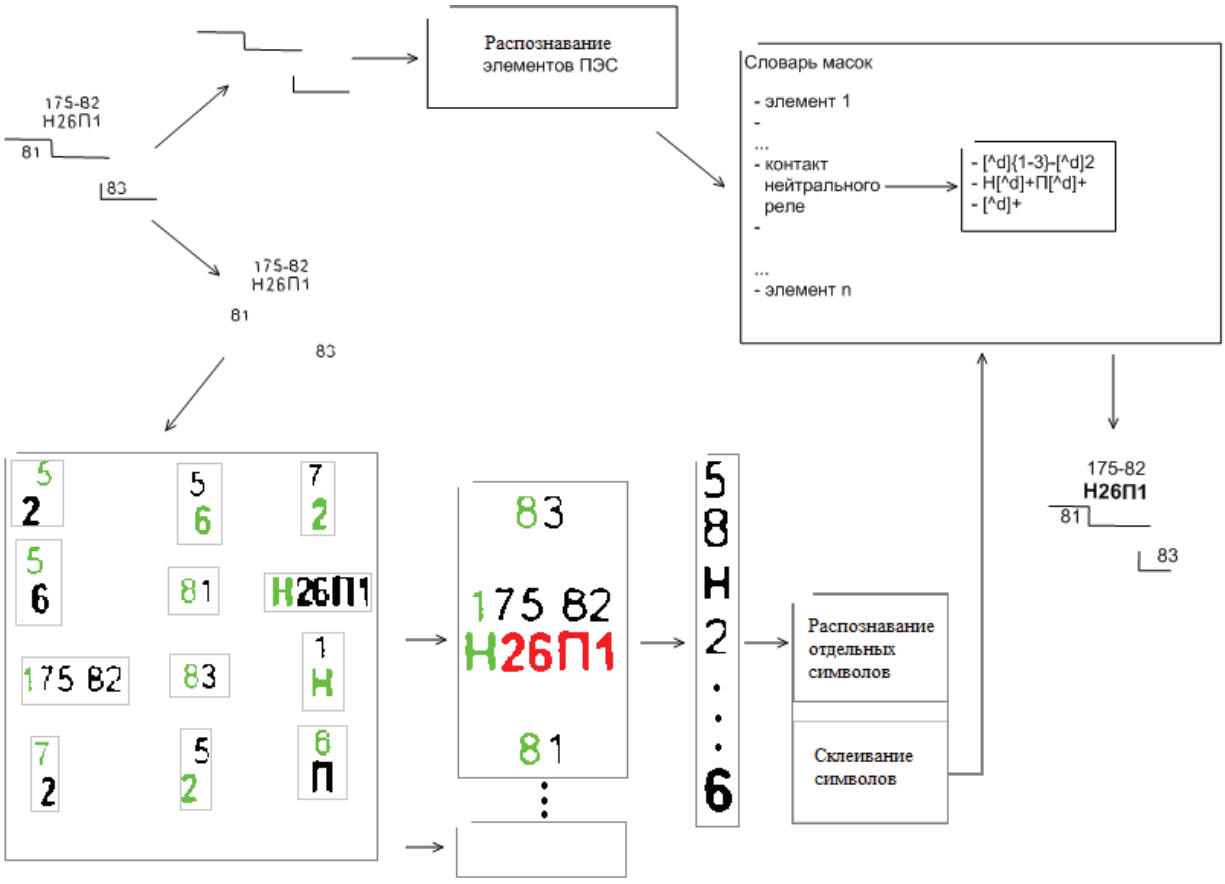

Рис. 9. Общая схема выделения и распознавания текста

Заключение

Описанные в статье методы и алгоритмы могут быть использованы как часть решения общей задачи распознавания ПЭС железнодорожной автоматики и телемеханики. Они не требуют введения дополнительных метрик и нечувствительны к повороту изображения ПЭС.

Возможность практического применения предложенного подхода требует больших вычислительных ресурсов, что для современных средств вычисления не является большой проблемой, однако оптимизация предложенных методов необходима для применения в реальных приложениях.

Экспериментальный анализ работы предложенных алгоритмов поиска и выбора лексем показал себя работоспособным на объеме тестовых документов около 300 отсканированных печатных ПЭС. Доля ошибок составила около 15 %. Большая часть пришлась на случаи наложения символов и больших скоплений связанного шума, возникающих, как правило, из-за дефектов процессов печати и сканирования.

Библиографический список

1. Балуев Н. Н. Проблемы внедрения отраслевого формата / Н. Н. Балуев, М. Н. Василенко, В. Г. Трохов, Д. В. Седых // Автоматика, связь, информатика. – 2010. – № 3. – С. 2.
2. Булавский П. Е. Оценка качества технической документации на системы ЖАТ / П. Е. Булавский // Автоматика, связь, информатика. – 2011. – № 8. – С. 37–39.
3. Булавский П. Е. Электронный документооборот технической документации / П. Е. Булавский, Д. С. Марков // Автоматика, связь, информатика. – 2012. – № 2. – С. 2–5.
4. Булавский П. Е. Синтез формализованной схемы электронного документооборота систем железнодорожной автоматики и телемеханики / П. Е. Булавский, Д. С. Марков // Известия Петербургского университета путей сообщения. – 2013. – № 2. – С. 108–115.
5. Денисов Б. П. Автоматизация проектирования систем железнодорожной автоматики и телемеханики на базе АРМ- ПТД версии 6 / Б. П. Денисов, Н. И. Рубинштейн, С. Н. Растегаев, Н. Ю. Воробей // Актуальные вопросы развития систем железнодорожной автоматики и телемеханики : сб. науч. тр. ; под ред. Вл. В. Сапожникова. – СПб. : Петербургский гос. ун-т путей сообщения, 2013. – С. 66–74.
6. Василенко М. Н. Электронный документооборот в хозяйстве СЦБ / М. Н. Василенко, В. Г. Трохов, Д. В. Зуев // Автоматика связь, информатика. – 2014. – № 8. – С. 2–3.
7. Матушев А. А. Программный комплекс для распознавания монтажной технической документации / А. А. Матушев // Известия Петербургского университета путей сообщения. – 2015. – № 1. – С. 105–109.

8. Василенко М. Н. Развитие электронного документооборота в хозяйстве АТ / М. Н. Василенко, В. Г. Трохов, Д. В. Зуев, Д. В. Седых // Автоматика связь, информатика. – 2015. – № 1. – С. 14–16.
9. Зуев Д. В. Синтез объектной нейросетевой модели распознавания образов и ее применение в задачах железнодорожной автоматики : дис. … канд. техн. наук : 05.13.18 / Зуев Денис Владимирович. – СПб., 2013. – 122 с.
10. Матушев А. А. Распознавание структуры монтажных схем ЖАТ / А. А. Матушев, Д. В. Седых // Автоматика, связь, информатика. – 2015. – № 10. – С. 4–7.
11. Ester M. A density-based algorithm for discovering clusters in large spatial databases with noise / M. Ester, H.-P. Kriegel, J. Sander, X. Xu, E. Simoudis, J. Han, U. Fayyad // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). – AAAI Press. – Pp. 226–231.
12. Левитин А. В. Алгоритмы. Введение в разработку и анализ / А. В. Левитин. – М. : Вильямс, 2006. – 576 с.

Mikhail N. Vasilenko,
Roman A. Koval`ov
«Automation and remote control on railways» department,
Emperor Alexander I St. Petersburg state transport university

Methods of textual expression selection of elementary electric diagrams of railway automation and remote control

Computerized recognition of fundamental electrical circuits (FES) railway automation and remote control (RARC) is an urgent and difficult task. The decision can be reasonably divided into the decision of the individual sub-tasks. Thus, the general recognition algorithm is divided into several specialized algorithms and the decision becomes more simple and straightforward. The main sub-tasks are selection and recognition of the FES structure, of the text, of the stamp and of other information. The article describes an approach for text processing on the FES RARC.

Text information of the FES is extremely important and without its analysis it is impossible to provide a complete transmission of the data of scanned FES RARC image into electronic form. The article proposes an algorithm for selection of text information of FES, using the clustering algorithm to select groups of symbols, as well as method of preparation of unique separated expressions (lexical units) within the group.

The article also describes the problem of analysis of obtained versions of text expressions and methods of selecting the most correct options. At the end of the article there is a general flow chart of the FES processing method, taking into account the methods of analysis of textual information proposed in the article.

Description of the study ends with the development of a software prototype, that implements the methods, mentioned in the article, and its using on a test

sample of 300 printed and then scanned FES of varied quality. The paper also provides the conclusions.

technical documentation; image recognition; text recognition; elementary electric diagrams

References

1. Baluev N. N., Vasilenko M. N., Trokhov V. G., Sedykh D. V. (2010). Problems of implementation the industry framework [Problemy vnedreniya otraslevogo formata], Automation, communication, information science (Avtomatika, svyaz', informatika), issue 3, p. 2.
2. Bulavsky P. E. (2011). Quality assessment of technical documentation for RARC systems [Otsenka kachestva tekhnicheskoy dokumentatsii na sistemy ZhAT], Automation, communication, information science (Avtomatika, svyaz', informatika), issue 8, pp. 37–39.
3. Bulavsky P. E., Markov D. S. (2012). Electronic document management of technical documentation [Elektronnyy dokumentooborot tekhnicheskoy dokumentatsii], Automation, communication, information science (Avtomatika, svyaz', informatika), issue 2, pp. 2–5.
4. Bulavsky P. E., Markov D. S. (2013). Synthesis of formalized diagram of electronic document management of railway automation and remote control systems [Sintez formalizovannoy skhemy elektronnogo dokumentooborota sistem zheleznodorozhnoy avtomatiki i telemekhaniki], Proceedings of Petersburg transport university (Izvestiya Peterburgskogo universiteta putej soobshcheniya), issue 2, pp. 108–115.
5. Denisov B. P., Rubinstein N. I., Rastegaev S. N., Vorobey N. Yu. (2013). Design automation of railway automation and remote control systems on the basis of ARM-PTD, v. 6 [Avtomatizatsiya proyektirovaniya sistem zheleznodorozhnoy avtomatiki i telemekhaniki na baze ARM-PTD versii 6], Topical issues of development of railway automation and remote control systems: collection of scientific papers (Aktual'nyye voprosy razvitiya sistem zheleznodorozhnoy avtomatiki i telemekhaniki: sbornik nauchnykh trudom), under the editorship of Vl. V. Sapozhnikov. St. Petersburg, Peterburg state transport university (Peterburgskiy gosudarstvennyy universitet putey soobshcheniya), pp. 66–74.
6. Vasilenko M. N., Trokhov V. G., Zuev D. V. (2014). Electronic document management at StsB facilities [Elektronnyy dokumentooborot v khozyaystve STsB], Automation, communication, information science (Avtomatika, svyaz', informatika), issue 8, pp. 2–3.
7. Matushev A. A. (2015). Software complex for recognition of assembly technical documentation [Programmnyy kompleks dlya raspoznavaniya montazhnoy tekhnicheskoy dokumentatsii], Proceedings of Petersburg transport university (Izvestiya Peterburgskogo universiteta putej soobshcheniya), issue 1, pp. 105–109.
8. Vasilenko M. N., Trokhov V. G., Zuev D. V., Sedykh D. V. (2015). Electronic document management development within AT facilities [Razvitiye elektronnogo dokumentooborota v khozyaystve AT], Automation, communication, information science (Avtomatika, svyaz', informatika), issue 1, pp. 14–16.

9. Zuev D. V. (2013). Synthesis of object-based neural network of image recognition and its application for railway automation tasks [Sintez ob'yektnoy neyrosetevoy modeli raspoznavaniya obrazov i yeyo primeneniye v zadachakh zheleznodorozhnoy avtomatiki]: Candidate thesis in Engineering Science (Dissertatsiya na soiskaniye uchenoy stepeni kandidata tekhnicheskikh nauk): 05.13.18 / Zuev Denis Vladimirovich [Place: Peterburg state transport university (Peterburgskiy gosudarstvennyy universitet putey soobshcheniya)]. St. Petersburg, 122 p.
10. Matushev A. A., Sedykh D. V. (2015). Recognition of ZhAT assembly diagram structure [Raspoznavaniye struktury montazhnykh skhem ZhAT], Automation, communication, information science (Avtomatika, svyaz', informatika), issue 10, pp. 4–7.
11. Ester Martin, Kriegel Hans-Peter, Sander Jörg, Xu Xiaowei, Simoudis Evangelos, Han Jiawei, Fayyad Usama M. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, pp. 226–231.
12. Levitin A. V. (2006). Algorithms. An introduction into design and analysis [Vvedeniye v razrabotku i analiz]. Moscow, Williams (Vil'yams), 576 p.

Статья представлена к публикации членом редколлегии В. А. Ходаковским
Поступила в редакцию 29.04.2016, принята к публикации 25.05.2016

ВАСИЛЕНКО Михаил Николаевич – доктор технических наук, профессор кафедры «Автоматика и телемеханика на железных дорогах» Петербургского государственного университета путей сообщения Императора Александра I.
e-mail: vasilenko. m.n@gmail.com

КОВАЛЕВ Роман Александрович – аспирант кафедры «Автоматика и телемеханика на железных дорогах» Петербургского государственного университета путей сообщения Императора Александра I.
e-mail: romanlisper@gmail.com