УДК 004.896+656:25

Седых Д. В.,
Матушев А. А.
Кафедра «Автоматика и телемеханика на железных дорогах»,
Петербургский государственный университет путей сообщения
Императора Александра I

МЕТОДЫ РАСПОЗНАВАНИЯ СТРУКТУРЫ МОНТАЖНЫХ СХЕМ ЖЕЛЕЗНОДОРОЖНОЙ АВТОМАТИКИ И ТЕЛЕМЕХАНИКИ

В данной статье говорится о модуле распознавания структуры монтажной технической документации железнодорожной автоматики и телемеханики. Данный модуль предлагается в качестве дополнения и улучшения специализированного программного комплекса распознавания монтажной технической документации. Работа модуля условно делится на пять этапов распознавания. Первый – определение типа распознаваемого документа. Второй – распознавание базовых элементов документа, которые используются в работе последующих алгоритмов. В зависимости от типа распознаваемого документа на последующих этапах используются разные алгоритмы. Третий этап – сортировка полученных элементов и нахождение закономерностей в документе – это необходимо для выявления не распознанных на предыдущем этапе элементов схемы. Далее полученные данные сравниваются с известными шаблонами монтажной документации. В случае, если документ не относится к известным типам монтажной документации, используются найденные закономерности. На последнем этапе идет поиск пропущенных элементов монтажной схемы, если таковые имеются. Подробно рассмотрены алгоритмы распознавания табличных структур документов, в частности алгоритм поиска базовых элементов и алгоритм нахождения пропущенных элементов. Приведено сравнение эффективности рассмотренного в статье модуля с существующим наиболее технологичным аналогом. Показаны результаты тестирования модуля на различных видах монтажных документов.

электронный документооборот; программный комплекс распознавания монтажной технической документации; распознавание структуры монтажных схем

Введение

В современном мире все больше организаций переходят на электронный документооборот. ОАО «РЖД» – не исключение. Электронный документооборот технической документации в дистанциях сигнализации, централизации и блокировки (СЦБ) реализован с помощью программного комплекса ведения технической документации (АРМ-ВТД), разработанного ООО

«ИМСАТ» [1–7]. Согласно распоряжению ОАО «РЖД» № 1299р вся новая документация в отделы технической документации дистанций СЦБ должна поступать в электронном виде и редактируемом формате. Однако в дистанциях по-прежнему хранится большой объем старой технической документации в бумажном виде. В настоящее время перевод бумажной документации в электронный вид осуществляется вручную, а это большие затраты времени и высокая вероятность ошибок. Для устранения этих недостатков предназначен специализированный программный комплекс распознавания бумажной документации, созданный на основе искусственных нейронных сетей. Данный комплекс был предложен Зуевым Д. В. в [8] и показал высокое качество распознавания документов. Однако используемые в программном комплексе алгоритмы распознавания структуры данных дают неточные результаты при распознавании монтажной технической документации. С целью повышения эффективности распознавания необходимо разработать модуль распознавания структуры технического документа.

1 Описание модуля распознавания структуры технического документа

Предлагаемый модуль для распознавания структуры документа решает несколько важных задач. В первую очередь необходимо автоматически определить тип поступающих в модуль документов. Эта операция необходима для отсеивания документов, не относящихся к монтажной документации. Основной монтажной документацией железнодорожной автоматики являются монтажные схемы, и данный модуль предназначен для их распознавания. Поскольку монтажные схемы весьма разнообразны, целесообразно разбить их на несколько различных типов. Второй этап работы модуля – распознавание базовых элементов, которые для разных типов документов будут различаться. Базовые элементы позволяют в дальнейшем восстановить общую структуру монтажной схемы. Затем идет сортировка полученных данных. Сортировка необходима для определения пропущенных элементов, а также позволяет определить закономерности расположения элементов.

В монтажных схемах существует несколько шаблонов структур. Например, часто встречающаяся структура – монтажная полка статива. В большинстве случаев она представляет собой таблицу на восемь подмест. Сравнение отсортированных элементов с известными шаблонами позволяет более точно определить пропущенные элементы схемы. Если определить принадлежность схемы к одному из шаблонов не удалось, для заполнения пробелов в схеме используются найденные в процессе сортировки закономерности. Таким образом, работу модуля можно разделить на несколько этапов (рис. 1). Результатом работы модуля является распознанная структура

Рис. 1. Основные этапы работы модуля распознавания структуры документа

документа в специальном формате. Данная структура передается в программный комплекс для непосредственного распознавания данных в монтажных документах.

Рассмотрим каждый из этапов на приведенной схеме.

2 Определение общей структуры документа

Условно монтажные схемы можно разделить на два базовых типа – табличный и схематический. В документе табличного типа данные представлены в виде таблиц, в документах схематического типа данные находятся на схематическом изображении (например, верхние клеммные панели). На данном этапе модуль определяет тип документа. Чтобы определить тип документа, необходимо найти на схеме связные области. Это можно сделать с помощью кластерного анализа [9, 10]. Из найденных областей выбирается наиболее крупная область, которая проверяется на наличие в ней таблицы. Отличительная особенность таблиц – множество горизонтальных и вертикальных линий, образующих ячейки. Если таблицу обнаружить не удалось, проводится проверка на наличие в схеме отличительных особенностей известных вариантов схематической документации. Тип документа определяет критерии базовых элементов и методы их сортировки. На текущий момент в модуле реализована возможность распознавания только табличных типов документов, распознавание схематических типов находится в разработке.

3 Распознавание базовых элементов

На следующем этапе работает алгоритм поиска по заданным критериям базового элемента. Например, для таблицы под базовым элементом понимается ячейка, которая с точки зрения элементов симметрии – прямоугольник, для его определения необходимо знать высоту (h) и длину (l), а также координаты верхнего левого угла (x, y).

В [11] был рассмотрен алгоритм поиска ячейки на основе нахождении ее углов (рис. 2). Алгоритм, записанный с помощью языка логических схем алгоритмов (ЛСА) [12], представлен ниже:

$$\downarrow^{6} Ap_1 \uparrow^{1} p_2 \uparrow^{2} p_3 \uparrow^{3} p_4 \uparrow^{4} p_5 \uparrow^{5} B \downarrow^{5} C \downarrow^{1}\downarrow^{2}\downarrow^{3}\downarrow^{4} p_6 \uparrow^{6} O. \qquad (1)$$

Действие алгоритма основано на поиске черных пикселей (оператор A). От найденного черного пикселя (условие p_1) ведется поиск верхнего левого угла (условие p_2) (см. рис. 2 вверху слева), найденный угол позволяет определить координаты (x, y). Далее с помощью датчиков черных пикселей ищутся правый верхний угол и нижний левый угол (условия p_3 и p_4) (см. рис. 2 вверху справа, внизу слева). Нахождение этих двух углов позволит определить длину (l) и ширину (h) прямоугольника соответственно. Параметры полученной ячейки сравниваются с массивами однотипных ячеек (условие p_5). Если параметры совпадают, то она записывается в соответствующий массив (оператор C). Если же размеры не совпадают, создается новый тип ячеек (оператор B). После нахождения всех черных пикселей (условие p_6) алгоритм заканчивает работу и выдает результат (оператор O). В результате работы данного алгоритма мы получаем около 80 % распознанных элементов (рис. 3).

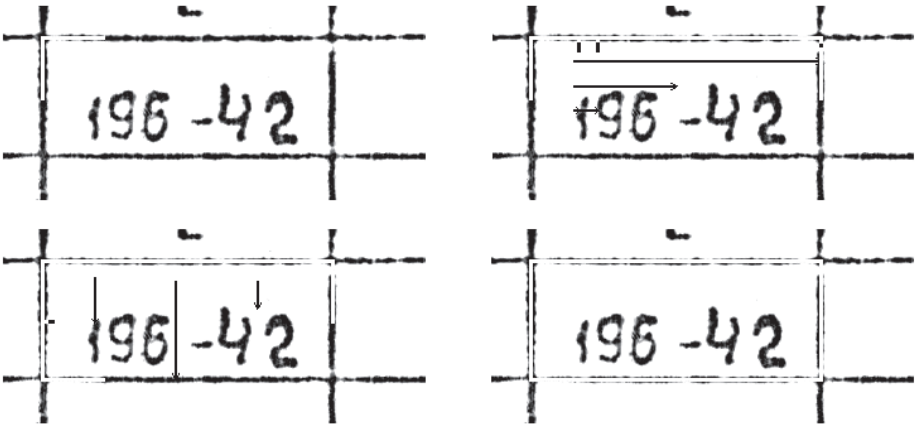

Рис. 2. Этапы поиска ячейки

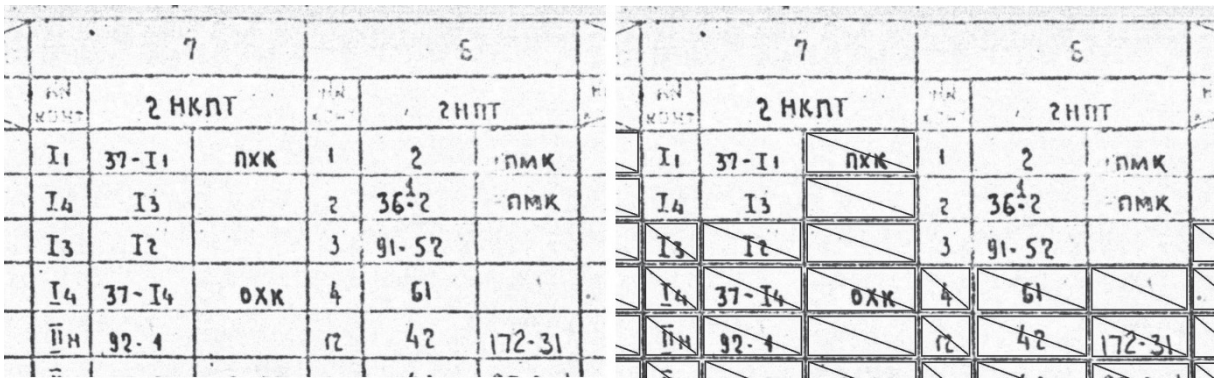

Рис. 3. Фрагмент монтажной схемы (*слева*); распознанные ячейки (*справа*)

4 Сортировка полученных элементов

Для заполнения пропусков, оставшихся после работы алгоритма поиска (1), общий массив найденных элементов необходимо отсортировать и выявить закономерности в размерах ячеек. Перед сортировкой от каждой ячейки идет проверка наличия пробелов сверху и снизу от нее. Если пробелы имеются, то они устраняются с помощью алгоритма:

$$\begin{gathered}\downarrow^{3} Aq_1 \uparrow^{1} q_2 \uparrow^{2} \downarrow^{4} \downarrow^{5} \downarrow^{6} \downarrow^{7} \downarrow^{8} \downarrow^{9} q_3 \uparrow^{3} O \\ \downarrow^{1} q_4 \uparrow^{4} q_5 \uparrow^{5} B_1\omega \uparrow^{6} \downarrow^{2} q_6 \uparrow^{7} q_7 \uparrow^{8} B_2\omega \uparrow^{9},\end{gathered} \tag{2}$$

где A – оператор, выбирающий ячейку из массива; B_1, B_2 – операторы, создающие ячейку на месте пропуска; O – оператор завершения работы алгоритма; q_1, q_2 – условия проверки наличия пробела выше (q_1) или ниже (q_2) текущей ячейки; q_3 – условие проверки того, является ли текущая ячейка последней в массиве; q_4, q_5 – условия проверки наличия ячеек слева и справа над текущей; q_6, q_7 – условия проверки наличия ячеек слева и справа под текущей.

Если все условия выполнены, то на месте пробела записывается ячейка с длиной текущей ячейки и высотой вышестоящих (или нижестоящих) ячеек (пример показан на рис. 4).

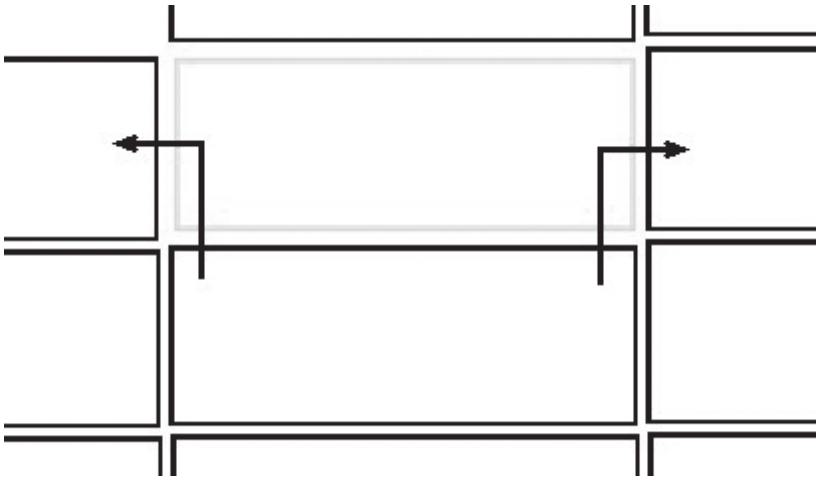

Рис. 4. Нахождение соседней ячейки

Сортировка заключается в разбиении общего массива ячеек на массивы строк – горизонтально сгруппированных ячеек. Места в строках можно назвать столбцами. В процессе сортировки сравниваются координаты текущей ячейки в строке с координатами предыдущей. Если между ними есть пробел, проверяется, не совпадает ли длина пробела с длиной других ячеек строки: если совпадает, то на месте пробела пропущена ячейка (рис. 5). Процесс можно описать алгоритмом:

$$\downarrow^4 Ar_1 \uparrow^1 r_2 \uparrow^2 B\omega \uparrow^3 \downarrow^2 r_3 \uparrow^4 \downarrow^1 \downarrow^3 O, \quad (3)$$

где A – оператор, выбирающий ячейку из строки; B – оператор, создающий ячейку на месте пропуска; O – оператор завершения работы алгоритма; r_1 – условие проверки наличия пробела; r_2 – условие проверки, равна ли длина пробела длине текущей ячейки; r_3 – условие проверки того, является ли текущая ячейка последней в строке.

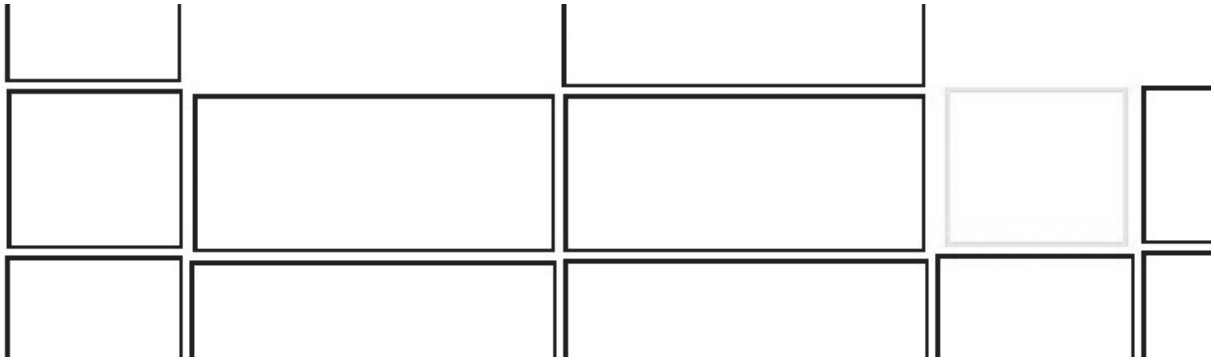

Рис. 5. Ячейка на месте пробела

Для описания закономерностей в таблице используется информация о количестве столбцов в ней. Закономерности выражаются в количестве столбцов, входящих в ячейку. Алгоритм проверяет ячейки строки на повторения и выбирает наиболее часто повторяющуюся комбинацию. Например, если в строке одна ячейка имеет размер в два столбца, а остальные ячейки имеют размер в один столбец, то закономерность данной строки «1». Если же строка будет иметь пробелы и разный размер ячеек, например «1, 2, пробел, 1, пробел, 1, 2, 1, 2», то наиболее часто повторяющаяся комбинация «1, 2».

5 Сравнение полученных данных с шаблонами и поиск недостающих элементов

Следующий шаг – сравнение полученных ячеек с шаблонами известных типов технической документации. Если документ соответствует одному из шаблонов, то найденные на предыдущем этапе закономерности размеров заменяются на закономерности, взятые из соответствующего шаблона.

Недостающие элементы находятся с помощью найденных или полученных из шаблонов закономерностей размеров ячеек. Данный процесс описан ниже:

$$\downarrow^4 B \downarrow^2 Cp_1 \uparrow^1 \downarrow^3 D \downarrow^1 p_2 \uparrow^2 p_3 \uparrow^3 p_4 \uparrow^4 O. \quad (4)$$

В каждой строке (оператор *B*) от каждого столбца (оператор *C*) проверяются пробелы между ячейками (условие p_1). Если пробел имеется, то определяется его размер, по которому можно определить число пропущенных элементов и записать на месте пропуска ячейки с параметрами, указанными в последовательностях (оператор *D*). У последнего элемента строки (условие p_2), пробел проверяется от него до края таблицы, определенного с помощью последнего столбца таблицы (условие p_3). Пробел устраняется методом, описанным в операторе *D*. После прохождения всех строк (условие p_4) алгоритм завершает работу (оператор *O*).

На рис. 6 показан пример заполнения пробелов. В девятой строке имеется пробел между ячейками с номерами 15 и 18. Параметры пропуска – два пропущенных столбца. Последовательность данной строки – «1», следовательно, пропущены две ячейки размерностью один столбец. Для определения значения длины ячеек используется длина столбца.

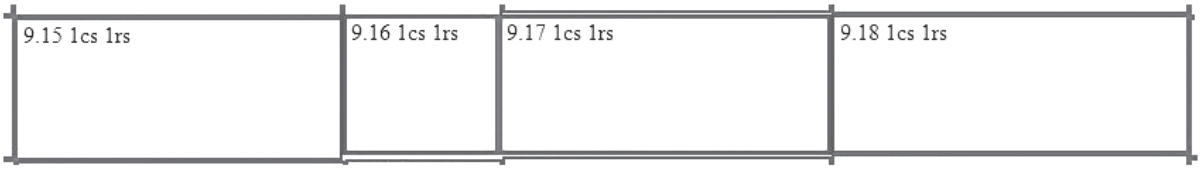

Рис. 6. Нахождение оставшихся ячеек

Закономерности помогают также разбивать или объединять ошибочные ячейки.

Если размер ячейки по закономерности больше одного столбца, то значение длины определяется в зависимости от суммы длин нескольких столбцов (рис. 7). Если же размер ячейки должен быть меньше, то она разбивается.

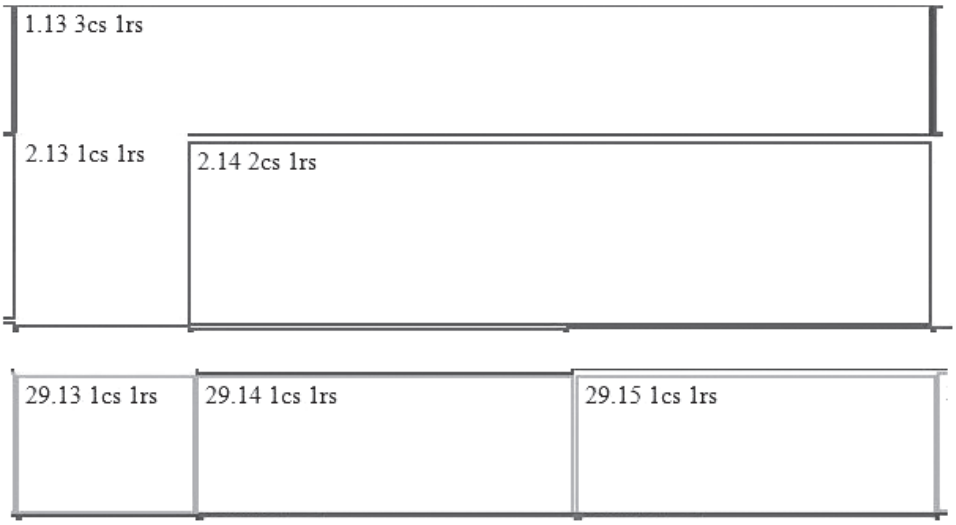

Рис. 7. Длинные ячейки

Заключение

Разработанный авторами модуль показал отличные результаты распознавания табличной структуры монтажных документов. В случае, когда тип документа совпадает с одним из заложенных в модуль шаблонов, работа программы показывает самый высокий процент распознавания. Если таблица документа подходит под стандартный шаблон, удается найти 95–100 % ячеек таблицы в зависимости от качества исходного изображения (пример распознанной схемы приведен на рис. 8). Однако, если таблица не подходит под описание известных шаблонов, поиск недостающих элементов будет проводиться с помощью найденных в процессе сортировки закономерностей. В таком случае возможно небольшое снижение процента распознанных элементов – в среднем при таком режиме работы программы удается найти 80–95 % ячеек. Программа позволяет также изначально указать шаблон для распознаваемого документа, что позволяет более точно распознавать документы низкого качества. Несмотря на снижение качества распознавания неизвестных таблиц, эти результаты выше, чем у сторонних средств распознавания таблиц. Например, программа Fine Reader фирмы ABBYY распознает таблицы целиком; если не находится несколько элементов, вся таблица остается нераспознанной, в среднем данная программа распознает всего лишь 40 % таблиц. Данный подход неприменим для решения задачи распознавания старых документов, на которых возможны различные виды повреждений, нарушающих целостность таблицы. Найденные с помощью предлагаемого модуля базовые элементы имеют нумерацию и координаты, несмотря на возможное наличие пробелов. Это позволяет восстановить поврежденные участки схемы. Данные хранятся в специальном внутреннем формате, из которого впоследствии основной программный комплекс может восстановить монтажную схему в отраслевом формате. По завершении разработки алгоритмов распознавания схематических монтажных схем модуль

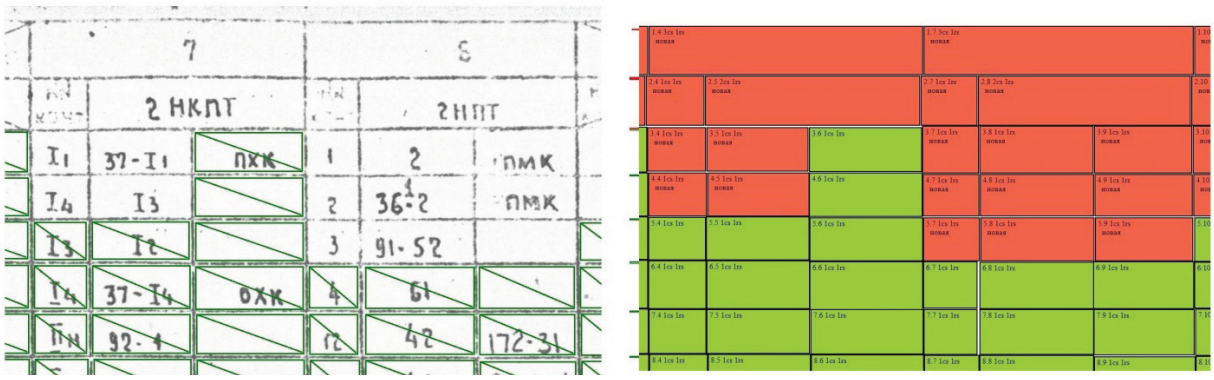

Рис. 8. Ячейки, распознанные с помощью алгоритма поиска (светло-серый); ячейки, распознанные на последующих этапах (темно-серый)

распознавания структуры сможет распознавать большинство известных видов монтажной технической документации. Таким образом, данный модуль существенно повышает скорость и качество работы програмного комплекса распознавания технической документации.

Библиографический список

1. Булавский П. Е. Электронный документооборот технической документации / П. Е. Булавский, Д. С. Марков // Автоматика, связь, информатика. – 2012. – № 2. – С. 2–5.
2. 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.
3. 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.
4. Матушев А. А. Распознавание структуры монтажных схем ЖАТ / А. А. Матушев, Д. В. Седых // Автоматика, связь, информатика. – 2015. – № 10. – С. 4–7.
5. Булавский П. Е. Формализация алгоритмического описания систем обеспечения жизненного цикла железнодорожной автоматики и телемеханики / П. Е. Булавский, Д. С. Марков, В. Б. Соколов, Т. Ю. Константинова // Автоматика на транспорте. – 2015. – Т. 1. – № 4. – С. 418–432.
6. Булавский П. Е. Методика оценки временных характеристик процессов электронного документооборота технической документации / П. Е. Булавский, Д. С. Марков // Автоматика на транспорте. – 2016. – Т. 2. – № 1. – С. 81–94.
7. Василенко М. Н. Методы выделения текстовых выражений принципиальных электрических схем железнодорожной автоматики и телемеханики / М. Н. Василенко, Р. А. Ковалев // Автоматика на транспорте. – 2016. – Т. 2. – № 4. – С. 541–551.
8. 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.
9. Айвазян С. А. Прикладная статистика: Классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. – М. : Финансы и статистика, 1989. – 450 с.
10. Мандель И. Д. Кластерный анализ / И. Д. Мандель. – М. : Финансы и статистика, 1988. – 176 с. : ил.
11. Матушев А. А. Программный комплекс для распознавания монтажной технической документации / А. А. Матушев // Известия Петербургского университета путей сообщения. – 2015. – № 1. – С. 105–109.

12. Лазарев В. Г. Синтез управляющих автоматов / В. Г. Лазарев, Е. И. Пийль. – М. : Энергоатомиздат, 1989. – 328 с.

Dmitry V. Sedykh,
Andrey A. Matushev
«Automation and remote control on railways» department,
Emperor Alexander I St. Petersburg state transport university

Methods of assembly diagrams structure recognition of railway automation and remote control

This article describes the structure detection module of recognition of the structure of assembly technical documentation of railway automation and remote control devices. This module is offered as an add-on and an improvement of the specialized software complex for technical documentation recognition. The module operation can be roughly divided into five stages of recognition. The first stage is the determination of the type of the document to recognize. The second step is the recognition of the basic elements of the document, which are used in the following algorithms. Depending on the type of the document to recognize different algorithms are used in subsequent steps. The next step is sorting of obtained components, and also finding patterns in the document, that is necessary for detection of structure elements, non-recognized at the previous stage. Next, obtained data are compared with known templates of assembly documentation. If the document does not correspond to a certain type of assembly documentation, the found patterns are used. At the last stage there is search for missing components of the assembly diagram, if any. The article describes in detail algorithms of recognition of tabular structures of documents, in particular the algorithm of basic elements search and the algorithm of missing elements search. It also gives a comparison of the efficiency of described module with the existing most technological counterpart. The article provides test results for the module for different types of assembly documents.

electronic document management; ARM–VTD; software complex for recognition of assembly technical documentation; assembly diagrams structure recognition

References

1. 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.

2. 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.
3. 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.
4. 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.
5. Bulavsky P. E., Markov D. S., Sokolov V. B., Konstantinova T. Yu. (2015). Formalization of algorithmic description of life-cycle supporting systems of railway automation and remote control [Formalizatsiya algoritmicheskogo opisaniya sistem obespecheniya zhiznennogo tsikla zheleznodorozhnoy avtomatiki i telemekhaniki], Automation on transport (Avtomatika na transporte), vol. 1, issue 4, pp. 418–432.
6. Bulavsky P. E., Markov D. S. (2016). Method of assessment of time parameters of electronic document management processes for technical documentation [Metodika otsenki vremennykh kharakteristik protsessov elektronnogo dokumentooborota tekhnicheskoy dokumentatsii], Automation on transport (Avtomatika na transporte), vol. 2, issue 1, pp. 81–94.
7. Vasilenko M. N., Kovalev R. A. (2016). Methods of textual expression selection of elementary electric diagrams of railway automation and remote control [Metody vydeleniya tekstovykh vyrazheniy printsipial'nykh elektricheskikh skhem zheleznodorozhnoy avtomatiki i telemekhaniki], Automation on transport (Avtomatika na transporte), vol. 2, issue 4, pp. 540–541.
8. 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.
9. Ayvazyan S. A., Bukhshtaber V. M., Enyukov I. S., Meshalkin L. D. (1989). Applied statistics: Classification and dimension reduction [Prikladnaya statistika: Klassifikatsiya i snizheniye razmernosti]. Moscow, Finances and Statistics (Finansy i statistika), 450 p.
10. Mandel' I. D. (1988). Clustering analysis [Klasternyy Analiz]. Moscow, Finances and Statistics (Finansy i statistika), 176 p.

11. 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.
12. Lazarev V. G., Piyl' E. I. (1989). Synthesis of control automata [Sintez upravlyayushchikh avtomatov]. Moscow, Energoatomizdat, 328 p.

Статья представлена к публикации членом редколлегии М. Н. Василенко
Поступила в редакцию 05.05.2016, принята к публикации 08.07.2016

СЕДЫХ Дмитрий Владимирович – инженер кафедры «Автоматика и телемеханика на железных дорогах» Петербургского государственного университета путей сообщения Императора Александра I.
e-mail: sedyhdmitriy@gmail.com

МАТУШЕВ Андрей Александрович – аспирант кафедры «Автоматика и телемеханика на железных дорогах» Петербургского государственного университета путей сообщения Императора Александра I.
e-mail: Dron_90@bk.ru