М.Д. Долгушин, А.А. Лесив, И.А. Крупкин, В.А. Шапцев

Тюменский государственный университет, г. Тюмень

УДК 004.896

ИССЛЕДОВАНИЕ АВТОМАТИЗИРОВАННОЙ ВЕРИФИКАЦИИ СХЕМ ТРУБОПРОВОДОВ

Аннотация. На примере одного формата проектной документации по трубопроводам предлагается технология автоматизированной верификации документов, а именно проверки принципиальных схем водоводов и соответствующего текста.

Ключевые слова: цифровая технология, трубопровод, схема, изображение, текст, автоматизированная верификация.

Введение

В связи с увеличением количества цифровой документации в нефтяной отрасли [1] актуальна и возможна более широкая автоматизация ее верификации. Здесь существуют 2 направления: обработка принципиальных схем трубопроводов для сравнения их со стандартами и требованиями заказчика и проверка орфографии и адекватности текстов.

В [2, 3], в частности, рассматривается теория комплексной проверки качества технической документации: обсуждается методология обработки документов, отмечается корреляция ошибок в производстве и в проектировании. Работ по автоматизации верификации проектной документации найдено не было. Автоматизированная оцифровка схем трубопроводов в [4] описана только на уровне консольного интерфейса, и нет проверки орфографии текста.

В то же время имеется множество цифровых технологий (ЦТ), выполняющих верификацию бухгалтерской документации (например, [5]). Соответствующие методы могут быть использованы в верификации проектной документации.

Ниже приводятся результаты исследования созданного программного модуля.

1. Новый модуль поддержки верификации проектной документации

Приведем сведения о исследуемом здесь программном модуле поддержки верификации текста и принципиальных схем трубопроводов. Модуль реализует автоматическую классификацию изображений принципиальных схем методом логистической регрессии [6], автоматическое распознавание текста на них с помощью нейронной сети LSTM [7], встроенной в используемое приложение Tessaract [8], и проверку орфографии основного текста текстовым редактором MySpell встроенным в библиотеку PyEnchant [9].

Логистическая регрессия – статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путем его сравнения с логистической кривой.

Для обучения алгоритма автоматической классификации изображения создан набор данных из документации Роснефти, находящейся в открытом доступе на сайте ТЭК-Торг [10]. Готовый набор таких данных не обнаружен. В классификации используется система векторизации изображений «Гистограмма направленных градиентов» (ГНГ) [11].

LSTM – тип рекуррентной нейронной сети, способный обучаться долгосрочным зависимостям.

Для распознавания текстов на изображении применяется система оптического распознавания символов Tessaract, настроенная на русский язык.

Для поиска орфографических ошибок применена библиотека PyEnchant со словарем на русском языке. В этой библиотеке реализованы методы проверки орфографии и исправления ошибок (предлагаются варианты слов из словаря). Существует возможность дополнять этот словарь.

2. Эксперимент

2.1. Классификация принципиальных схем трубопроводов

Примеры из набора данных для оценки качества классификации приведены на рис. 1 и 2 [12]. Он разбит на обучающую и тренировочную выборки. В них изображения разделены на классы: «корректный» (рис. 1) и «некорректный» (рис. 2). В обучающей выборке – 60 изображений, в тренировочной – 16.

Изображения считываются и сжимаются до размерности 200x200. Затем на каждой схеме выделяются признаки методом ГНГ – HOGDescriptor(), встроенным в библиотеку OpenCV [13]. На полученных векторных представлениях изображения обучается классификатор логистической регрессии, реализованный в библиотеке машинного обучения sklearn [14].

Получена точность классификации 97%. Высокий процент связан с выделяющимися признаками некорректных изображений, такими как: размытие, искривление, гауссовский шум. Из-за малого объема обучающей выборки нет гарантированной устойчивости классификации в случаях меньших различий выборок (образов схем трубопроводов).

2.2. Чтение текста с изображения принципиальной схемы

Для чтения текста на изображении используются: модуль pytessaract, распознающий текст в изображении с помощью Tessaract OCR; методы OpenCV считывания и повышения качества изображения символов медианной фильтрацией [15]. Пример верифицируемого текста представлен на рис. 1, где над изображением расположен некорректный заголовок.

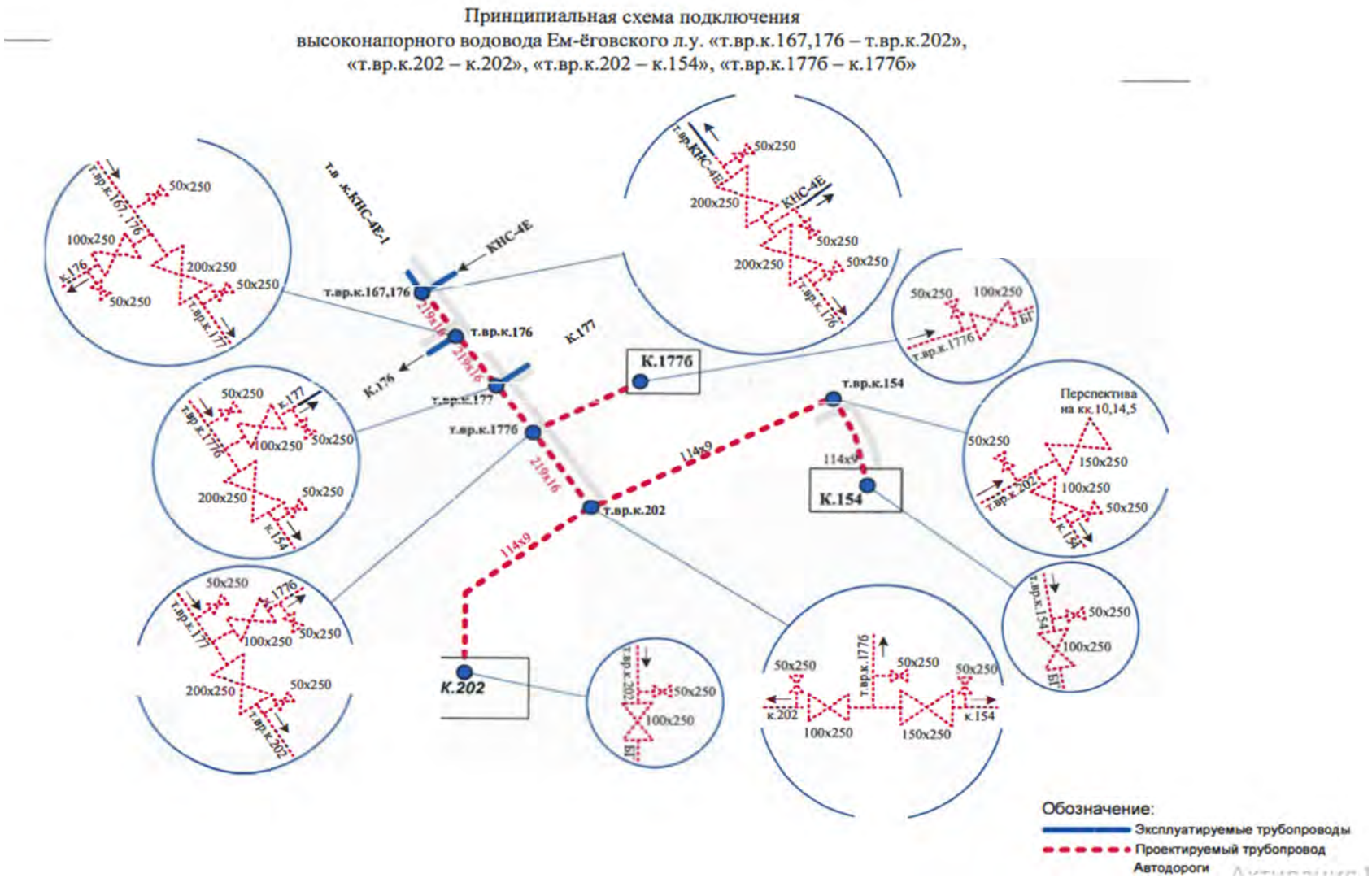

Рис. 1. Пример «корректного» изображения

Рис. 2. Пример «некорректного» изображения

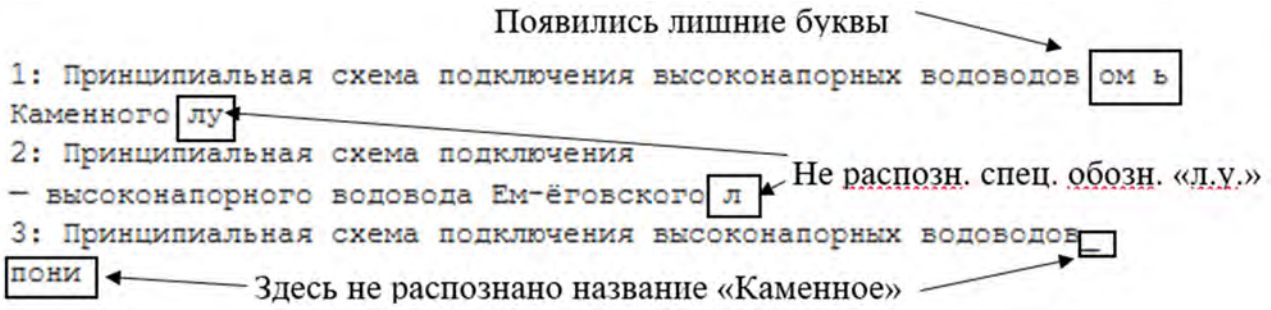

Рис. 3. Результат распознавания текста на изображении

На рис. 3 представлен результат распознавания текста на изображении, где номера 1-3 обозначают исследуемые изображения из выборки. Видны ошибки вследствие низкого разрешения изображений и небольшого размера части текста. Поэтому требуется либо разработка значительно более точного алгоритма распознавания текста, либо повышение качества изображений. Может помочь и увеличение шрифта надписей.

2.3. Проверка орфографии текста

Для проверки орфографии использована библиотека PyEnchant на языке Python. Данная библиотека реализует взаимодействие с различными алгоритмами проверки орфографии. В исследовании используется алгоритм проверки орфографии из Aspell. Этот алгоритм проверяет правописание слов в тексте посредством сравнения со словарем, добавленным в модуль. Для проверки русского языка был добавлен русский словарь из LibreOffice. Отсутствующие слова можно добавить. С опытом работы модуля увеличится точность проверки орфографии за счет расширения словаря.

```
['[!] ERROR: визде']
['[.] ']
['[.] Это правильный, но не визде тект']
['[.] ']
['[.] HOW ABOUT:', '0: визе | 1: виде | 2: виз де | 3: девиз']
>> help
['[.] 0..N:\treplace with the numbered suggestion']
['[.] R0..RN:\talways replace with the numbered suggestion']
['[.] i:\t\tignore this word']
['[.] I:\t\talways ignore this word']
['[.] a:\t\tadd word to personal dictionary']
['[.] e:\t\tedit the word']
['[.] q:\t\tquit checking']
['[.] h:\t\tprint this help message']
['[.] ------------------------------------------------']
['[.] HOW ABOUT:', '0: визе | 1: виде | 2: виз де | 3: девиз']
>> e
[.] New Word: везде
['[!] ERROR: тектс']
['[.] ']
['[.] Это правильный, но не везде тект']
['[.] ']
['[.] HOW ABOUT:', '0: текст']
>> 0
["[+] Replacing 'тектс' with 'текст'"]
```

Рис. 4. Пример проверки орфографии в выражении:
«Это правильный, но не визде тект»

На рис. 4 представлен пример проверки орфографии сообщения через консольный интерфейс.

Заключение

Проведенные эксперименты показали работоспособность предложенных принципов. Требуется повышение точности и устойчивость классификации принципиальных схем путем расширения обучающего набора данных и добавления классов, соответствующих проблемам на принципиальных схемах.

В распознавании текстов документации, проверке их структуры, пунктуации, орфографии требуются искусственные нейросети. Это повысит семантическую адекватность обработки текстов.

Исследованный программный модуль станет частью разрабатываемой Web-системы поддержки верификации схем трубопроводов.

СПИСОК ЛИТЕРАТУРЫ

1. Документооборот на нефтегазовых предприятиях. 3 направления для оптимизации. URL: https://neftegaz.ru/analisis/companies/329343-dokumentooborot-na-neftegazovykh-predpriyatiyakh-3-napravleniya-dlya-optimizatsii/ (Дата обращения: 04.05.2021).

2. Антонова А.Ю., Клышинский Э.С. Метод автоматической проверки качества технической документации, МГИЭиМ. URL: https://cyberleninka.ru/article/n/metod-avtomaticheskoy-proverki-kachestva-tehnicheskoy-dokumentatsii/viewer (Дата обращения: 04.05.2021).

3. Tuhacek M., Svoboda P., Quality of Project Documentation, 2019, *IOP Conf. Ser.: Mater. Sci. Eng.* URL: https://iopscience.iop.org/article/10.1088/1757-899X/471/5/052012 (Дата обращения: 04.05.2021).

4. Долгушин М.Д., Цыганова М.С. Разработка и реализация алгоритма цифровизации схемы трубопровода // Математическое и информационное моделирование/ Материалы Всероссийской конференции молодых ученых, г. Тюмень, 1-8 июня 2020 г. – Тюмень: Изд-во ТюмГУ, 2020. – Вып. 18. – С. 56-69. URL: https://library.utmn.ru/dl/PPS/Bidulja_2020_18.pdf/info (Дата обращения: 04.05.2021).

5. Приложение для верификации бухгалтерских проектов. URL: www.jumio.com (Дата обращения: 04.05.2021).

6. Yorko, Открытый курс машинного обучения. Тема 4. Линейные модели классификации и регрессии, URL: https://habr.com/ru/company/ods/ blog/323890/ (Дата обращения: 04.05.2021).

7. Olah C. Understanding LSTM Networks // URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/ (Дата обращения: 29.05.2021).

8. Tesseract User Manual. URL: https://tesseract-ocr.github.io/tessdoc/ (Дата обращения: 04.05.2021).

9. Документация PyEnchant. URL: https://pyenchant.github.io/pyenchant/ (Дата обращения: 07.05.2021).

10. Закупочные процедуры, ТЭК-Торг. URL: https://www.tektorg.ru/rosneft/procedures (Дата обращения: 04.05.2021).

11. Кулинкович В.А. Применение методики гистограмм направленных градиентов для классификации дактилоскопических изображений // Журнал Белорус. гос. ун-та. Математика. Информатика. 2017. № 1. С. 53–60. URL: https://elib.bsu.by/bitstream/123456789/179301/1/53-60.pdf (Дата обращения: 04.05.2021).

12. Долгушин М.Д. Набор данных по принципиальным схемам. URL: https://drive.google.com/drive/folders/1aKZuX971DIcU3dd4Xj-bNO0Vjt432Upl (Дата обращения: 04.05.2021).

13. Документация OpenCV. URL: https://opencv.org (Дата обращения: 04.05.2021).

14. scikit-learn: machine learning in Python. URL: https://www.sklearn.org (Дата обращения: 04.05.2021).

15. Медианная фильтрация. URL: https://ru.bmstu.wiki/Медианная_фильтрация (Дата обращения: 08.05.2021).