УДК 004.9

РАЗРАБОТКА МЕТОДОВ АВТОМАТИЧЕСКОЙ СТРУКТУРИЗАЦИИ И ДЕФРАГМЕНТАЦИИ ИЗОБРАЖЕНИЙ ТЕКСТОВЫХ ДОКУМЕНТОВ

Горшков Данила Андреевич[1], Ершов Николай Михайлович[2]

[1]Студент;
Государственный университет «Дубна»;
141980, Московская область, г. Дубна, ул. Университетская, 19;
e-mail: gorshkovdanila1999@gmail.com.

[2]Доцент;
Государственный университет «Дубна»;
141980, Московская область, г. Дубна, ул. Университетская, 19;
e-mail: ershovnm@gmail.com.

Работа посвящена исследованиям способов сегментации изображений и методов автоматического распознавания стилей форматирования заданного текстового блока. Целью работы является разработка методов автоматической структуризации и дефрагментации изображений текстовых документов, т.е. нужно произвести сегментацию текстового фрагмента на изображении с дальнейшей программной реализацией автоматической классификации текстовых фрагментов. В работе предложен алгоритм сегментации изображений, основанный на пороговой сегментации. Такой алгоритм позволяет достичь достаточно точной сегментации изображений. Проводится обзор разработанных методов распознавания стилей, описывается актуальность проводимого исследования. Также проводится численное исследование методов. Описывается программная реализация предложенных алгоритмов и методов, с использованием языка программирования Python, демонстрируются примеры работы программы на изображениях с текстовыми блоками. Были проведены тесты разработанных методов на двух выборках текстовых изображений, для этого были сгенерированы тестовые изображений с помощью языка программирования Python, библиотеки ImageMagic и системы компьютерной верстки Latex. Проведенное тестирование показало перспективность предложенных подходов и методов к структуризации и классификации текстовых блоков.

Ключевые слова: сегментация текстовых блоков, распознавание стилей форматирования.

Для цитирования:
Горшков Д. А., Ершов Н. М. Разработка методов автоматической структуризации и дефрагментации изображений текстовых документов // Системный анализ в науке и образовании: сетевое научное издание. 2021. № 2. С. 56–66. URL : http://sanse.ru/download/436.

DEVELOPMENT OF METHODS FOR AUTOMATIC STRUCTURING AND DEFRAGMENTATION OF IMAGES OF TEXT DOCUMENTS

Gorshkov Danila A.[1], Ershov Nikolay M.[2]

[1]Student;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: gorshkovdanila1999@gmail.com.

[2]Assistant professor;
Dubna State University;
19 Universitetskaya Str., Dubna, Moscow region, 141980, Russia;
e-mail: ershovnm@gmail.com.

The work is devoted to the research of image segmentation methods and methods of automatic recognition of formatting styles of a given text block. The aim of the work is to develop methods for automatic structuring and defragmentation of images of text documents, i.e. it is necessary to segment a text fragment on an image with further software implementation of automatic classification of text fragments. The paper proposes an image segmentation algorithm based on threshold segmentation. This algorithm allows you to achieve fairly

accurate image segmentation. The review of the developed methods of style recognition is carried out, the relevance of the research is described. A numerical study of the methods is also carried out. The paper describes the software implementation of the proposed algorithms and methods using the Python programming language, demonstrates examples of the program's operation on images with text blocks. Tests of the developed methods were carried out on two samples of text images, for this purpose test images were generated using the Python programming language, the ImageMagic library and the Latex computer layout system. The conducted testing showed the prospects of the proposed approaches and methods for structuring and classifying text blocks.

Keywords: segmentation of text blocks, recognition of formatting styles.

For citation:
Gorshkov D., Ershov N. Development of methods for automatic structuring and defragmentation of images of text documents. System Analysis in Science and Education, 2021;(2):56–66(In Russ). Available from: http://sanse.ru/download/436.

Введение

С резким развитием компьютерных технологий и коммуникационных технологий в современном обществе большинство документов составляется на компьютерах, в связи с этим люди все меньше используют традиционные системы хранения документов (бумага и т.д.), но задача создания полностью электронного документооборота еще далека до полной реализации.

Задача перевода информации с бумажных на электронные носители актуальна не только в рамках потребностей, возникающих в системах документооборота. Современные информационные технологии позволяют нам существенно упростить доступ к информационным ресурсам, накопленным человечеством, при условии, что они будут переведены в электронный вид. Наиболее простым и быстрым является сканирование документов с помощью сканеров. Результатом работы является цифровое изображение документа – графический файл. Но для сортировки и навигации по изображениям нужно выделять признаки текстовых фрагментов. Таким образом перед нами встает задача сегментации отсканированного документа, т.е. необходимо определить границы текстовых блоков – это позволяет описать макет страницы и использовать это описание при поиске и классификации отсканированных документов. Кроме того, в дальнейшем выделенные блоки можно будет передать для дальнейшей обработки, например для распознавания стилей форматирования.

Целью работы является разработка методов автоматической структуризации и дефрагментации изображений текстовых документов, т.е. нужно произвести сегментацию текстового фрагмента на изображении с дальнейшей программной реализацией автоматической классификации текстовых фрагментов.

Сегментация изображений

Сегментация текста является важным этапом в анализе изображений документа и не имеет общепринятого решения [1, 2]. Чтобы сегментировать текст из изображения, необходимо определить все возможные области текстовых блоков. Сегментация текста на изображении обычно используется для определения объектов и границ (линий, кривых и т.д.) на изображениях. Это подразумевает процесс маркировки, который присваивает одну и ту же метку для пространственного выравнивания объектов, то есть пикселей, связанных компонентов или характерных точек, так что группа пикселей с аналогичной меткой разделяет определенные визуальные характеристики. Результат сегментации изображения – это набор сегментов, которые вместе покрывают все изображение, или его отдельную часть. Каждый из пикселей в определенной области похож в отношении некоторых характерных свойств, таких как цвет, интенсивность или текстура. Соседние области могут значительно различаться по одним и тем же характеристикам [3].

Перед выполнением сегментации нам необходимо привести изображение в определенный формат, чтобы в дальнейшем произвести над изображение последующую обработку. Предварительная обработка [4, 5, 6] включает оцифровку, удаление шума [7], бинаризацию, нормализацию. Этап предварительной обработки дает нам изображение с достаточно высоким объемом информации. Следующий

этап в процессе анализа документа – сегментация. Сегментация может быть разделена на подкомпоненты [8, 9]. Сегментация является важным этапом, поскольку степень, которой можно достичь при разделении слов, строк или символов, напрямую влияет на скорость распознавания текстовых блоков [10].

Далее, на основе полученных меток, текст делится на различные логические области, каждая из которых представляет заранее определенный набор семантики [8]. Идеальной ситуацией было бы сегментировать область на изображении, которая являлось бы текстовой строкой. Завершение сегментации текстовых строк предоставляет важную информацию для последующих этапов анализа текстовых блоков на изображении, например таких как обнаружение и исправление перекоса, извлечение текстовых функций и распознавания символов. Хоть некоторые методы обнаружения текстовых строк успешно применяются в печатных документах, обработка рукописных документов остается ключевой проблемой при оптическом распознавании символов.

Кроме того, необходимость сегментации превалирует над возможностью уменьшения сложности при реализации эффективной системы распознавания символов. Сегментация имеет применение в различных областях, таких как машинное зрение, обнаружение объектов, сегментация медицинских изображений [13] и задачах распознавания. *Content-based image retrieval* (*CBIR*) – это одно из применений методов компьютерного зрения к проблеме поиска цифровых изображений в больших базах данных на основе синтаксических характеристик изображений (таких как цвет, текстура, форма) [14].

Существуют различные факторы, которые могут препятствовать процессу сегментации текстовых блоков на изображении [1, 12]. Вот некоторые из них:

- Качество изображения: Качество изображения важный фактор для сегментации текстовых блоков. Наличие шума на изображении приводит к снижению точности и эффективности сегментации [16].
- Рукописный или печатный документ: Большинство методов сегментации текстовых строк основаны на предположении, что расстояние между соседними текстовыми строками является точным, а также что текстовые строки являются одинаково прямыми. Однако эти предположения не характерны для рукописных документов. В случае рукописного документа сегментация текстового блока на изображении является основной проблемой. В приоритете находиться печатный текстовый документ. Для такого документа сегментация – простая задача из-за симметричности документа. Определены интервалы между строками, словами и символами, что устраняет проблемы, возникающие при работе с рукописными документами.
- Ориентация текстового блока на изображении: для рукописного текста, если отдельные строки не прямые или если присутствует перекос, общая сложность сегментации текста возрастает [6, 15].
- Текстурированный документ: Наличие текстуры, например изображений, узоров и др. на изображении делает задачу сегментации многогранной.
- Тип текста: Курсивный текст создает дополнительный трудности при сегментации символов из-за наличия слитности.

Сегментация текста на изображении может быть достигнута на трех уровнях [1, 6, 8, 11]. По мере продвижения по разным уровням иерархии сегментации текста мы получаем более мелкие детали. Использование всех трех уровней не обязательно. Сегментация на любом из этих уровней напрямую зависит от характера применения. Чем больше деталей требуется для изображения, тем выше уровень сегментации. Различные уровни иерархии (см. рис. 1).

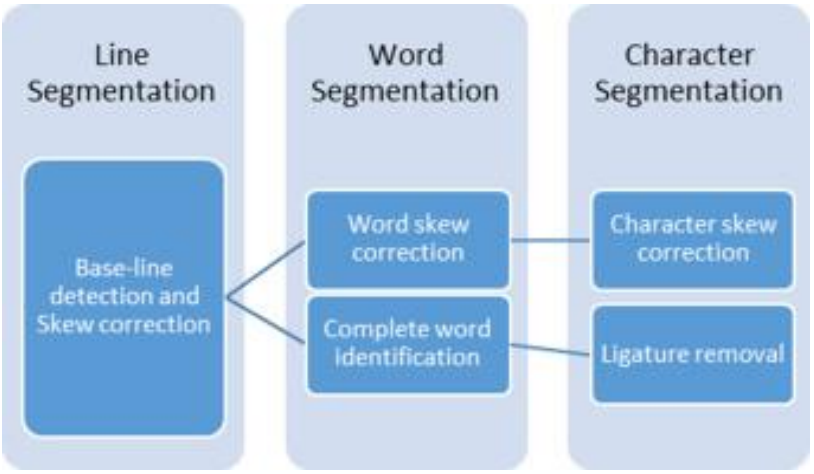

Рис. 1. Уровни сегментации

Сегментация с использованием порогового значения

Основной принцип работы пороговой сегментации состоит в том, чтобы разделить пиксели изображения на несколько категорий путем установки различных характеристических пороговых значений. Общие особенности, включая: оттенки серого или цветовые характеристики исходного изображения или характеристики преобразования. Предполагая, что исходное изображение f (x, y), изображение делиться на две части по определенным критериям [17]. И изображение после сегментации отображается как:

$$g(x,y) = \begin{cases} b_1 & f(x,y) < t \\ b_2 & f(x,y) > t \end{cases}. \quad (1)$$

В формуле $f(x, y)$ исходное изображение, $g(x, y)$ изображение сегментации и t это порог. Метод сегментации пороговых значений с двумя значениями часто используется для сокращения объема данных и упрощения обработки из-за требований к скорости.

Метод глобального порога заключается в вычислении оптимального порога сегментации, который может быть единственным пороговым или много-пороговым значением. Метод включает два типа: точечный и региональный.

Для изображений с очевидными различиями в уровне серого используется бимодальный метод. В этом методе обычно для определения порога используется гистограмма. Изображения, сравниваемые с фоном, обычно имеют бимодальную гистограмму. Объект создает один из пиков, а фон создает другой пик на гистограмме. И граница между объектом и фоном образует впадину между двумя пиками [18].

Когда порог сегмента находиться на дне впадины, сегментация изображения может достичь наилучших результатов. Метод прост и удобен в использовании, но его нельзя использовать для изображения с пиком плоской волны. Более того, когда на изображение влияет шум, нижняя часть гистограммы заполняется или целевой объект находиться близко к пику фона, поэтому нижнюю часть сложно обнаружить.

Но в 1979 году японским ученым был предложен метод наиболее взаимо-вариантной вариации, который представляет собой метод определения адаптивного порога, также называемый *OTSU* [19]. Это характеристика изображения в градациях серого, разделенная на фон и две целевые части. Чем больше разница между фоном и целью, тем больше разница между двумя частями изображения [20]. Метод основан на принципе наименьших квадратов, а его алгоритм прост, что является удобным и выполнимым методом выбора порога. Алгоритм работает следующим образом:

Сначала находятся самые высокие оттенки серого в изображении (L - 1);

Затем каждый уровень серого от 0 до L - 1 принимается в качестве порогового значения K;

После вычисляется вероятность порога, и вероятность суммируется по весу.

И вычисляется среднее значение μ_i, градаций серого пикселя. Показать это можно как:

$$\omega_1 = \sum_{i=0}^{k} p_i . \quad (2)$$

$$\omega_2 = \sum_{i=k+1}^{L-1} p_i . \quad (3)$$

$$\mu_1 = \frac{1}{\omega_1} \sum_{i=0}^{k} i p_i . \quad (4)$$

$$\mu_2 = \frac{1}{\omega_2} \sum_{i=k}^{L-1} i p_i . \quad (5)$$

В (4) рассчитано общее значение серого пикселя на изображении μ, можно показать это как:

$$\mu = \sum_{j=0}^{i-1} i p_i . \quad (6)$$

Также в (4) вычисляется дисперсия σ_B между двумя классами, и максимальный порог T дисперсии находиться по следующей формуле:

$$\sigma^2 = \omega_1(\mu_1 - \mu)^2 + \omega_2(\mu_2 - \mu)^2. \quad (7)$$

Выделение строк в заданном текстовом блоке

Принцип работы данного метода заключается в том, что анализируются взаимные расположения ограничительных рамок каждого контура всех букв текстового фрагмента и анализируя Y координаты определяем лежат ли эти ограничительные рамки в одной строке или не лежат (см. рис. 2, 3). Все прямоугольники, которые лежат на одной строке, мы объединяем в один прямоугольники получаем разбиение текста на строки (см. рис. 4).

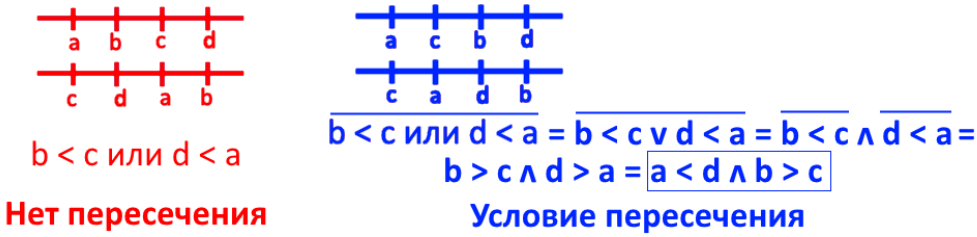

Рис. 2. Условие пересечения ограничительных рамок

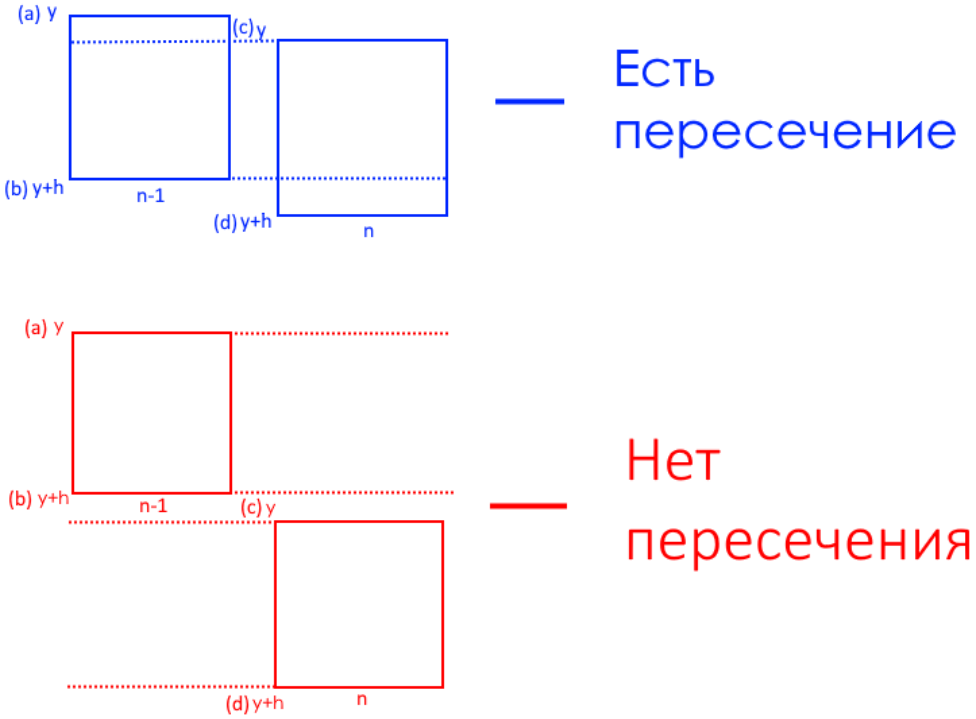

Рис. 3. Графическое представление условия пересечения ограничительных рамок

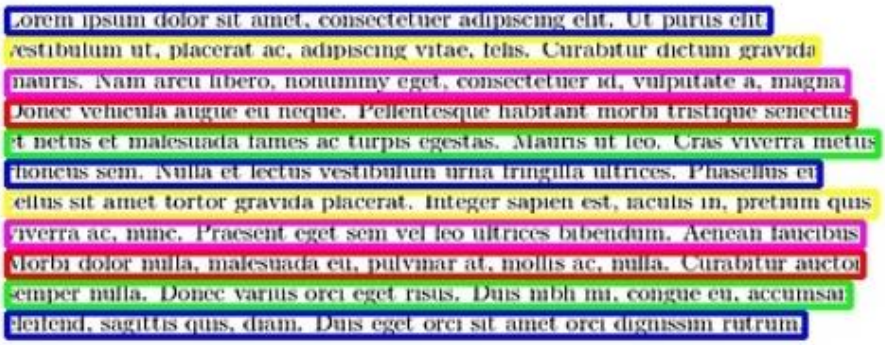

Рис. 4. Пример отрисовки ограничительных рамок для каждой строки

Метод распознавания абзацного отступа

Был разработан метод определения отступа текстовых фрагментов. Проанализировав *X* координаты каждого ограничительного прямоугольника строки и сравнив координаты первой строки со второй определяем отступ текстового фрагмента.

Метод распознавания типа выравнивания текста

Был разработан метод определения типа выравнивания текстовых документов. Все ограничительные прямоугольники были разбиты на 3 части (см. рис. 5). После чего было подсчитано количество элементов в каждой части. На основе этих данных была построена гистограмма, по которой можно определить тип выравнивания текстовых фрагментов (см. рис. 6).

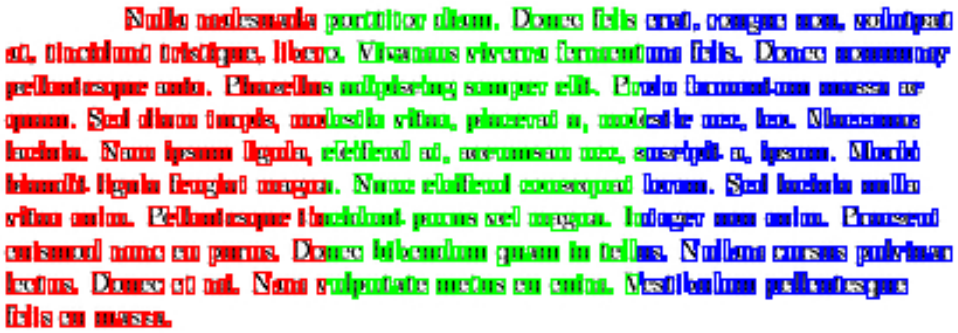

Рис. 5. Разбиение текстового блока на 3 части

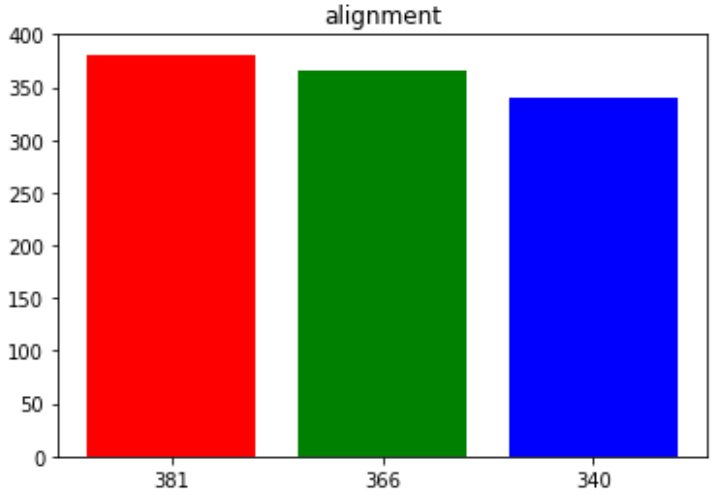

Рис. 6. Гистограмма выравнивания текстовых фрагментов

Определение размера шрифта

Был разработан метод определения размера шрифта текстовых фрагментов. С помощью координат по оси Y ограничительных прямоугольников каждого символа были найдены высоты каждого прямоугольника (см. рис. 7). Затем была взята медиана от всех высот и тем самым был найден размер шрифта текстовых фрагментов.

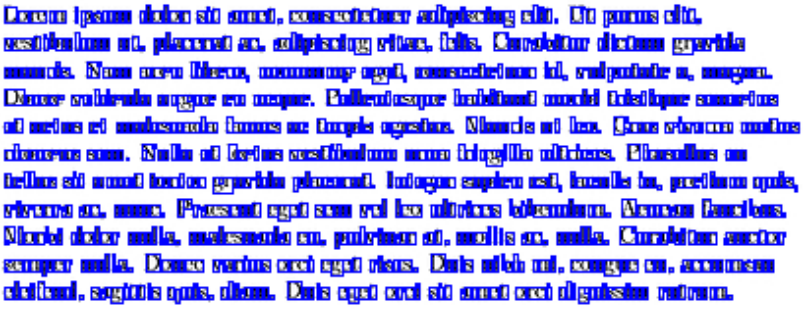

Рис. 7. Текстовый блок с ограничительными прямоугольниками каждого символа

Определение межстрочного интервала

Был разработан метод определения межстрочного интервала текстовых фрагментов. Были взяты координаты по оси Y каждого ограничительного прямоугольника строки, тем самым были найдены высоты каждой строки (см. рис. 8). Затем взяв медиану по высотам всех строк был определен межстрочный интервал текстовых фрагментов.

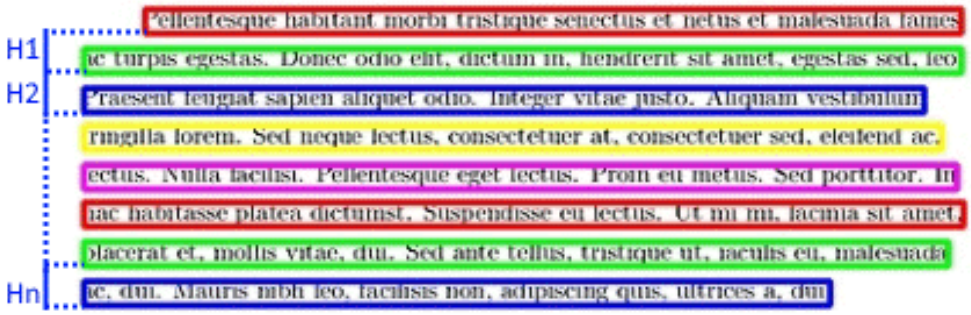

Рис. 8. Пример нахождения высоты каждой строки

Распознавание типа шрифта

Был разработан метод определения типа шрифта текстовых фрагментов (см. рис. 9). Взяв серединную координату по оси *X* ограничительного прямоугольника каждого символа, была построена гистограмма, показывающая количество элементов по оси *Y* на каждой серединной координате *X* (см. рис. 10, 11). И если на гистограмме преобладает значение равное нулю, то шрифт является моноширинным, а если преобладают значения большие нуля, то шрифт является пропорциональным.

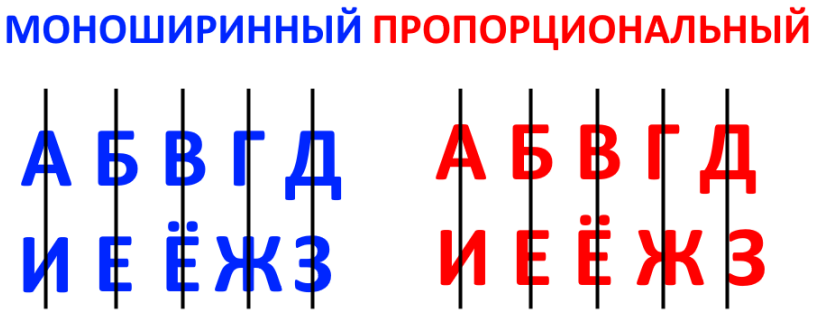

Рис. 9. Пример моноширинного и пропорционального шрифта

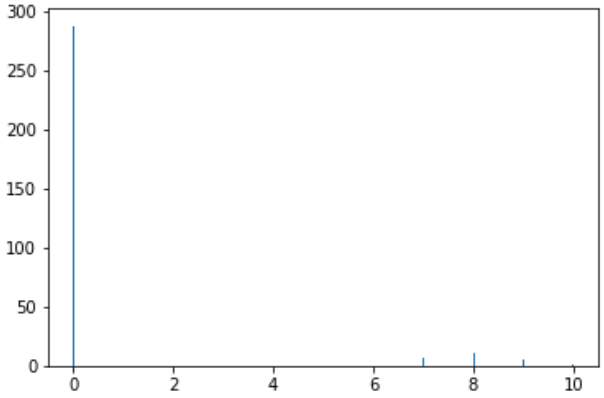

Рис. 10. Гистограмма массива моноширинного шрифта

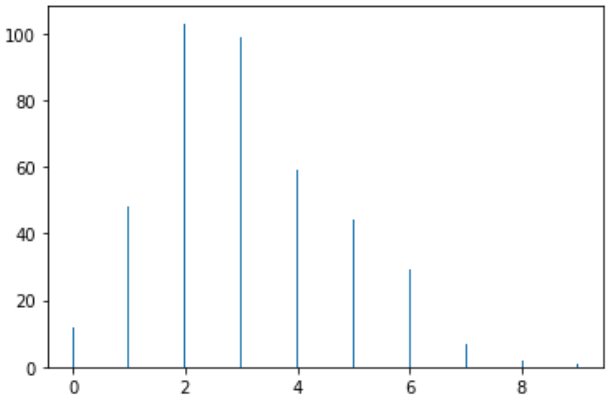

Рис. 11. Гистограмма массива пропорционального шрифта

Программная реализация

Была выполнена реализация методов сегментирования текстовых фрагментов, а также методы автоматической классификации текстовых фрагментов. В качестве основного языка программирования, для реализации методов, был выбран язык *Pyhon*. Для данного языка были использовали такие библиотеки как: *OpenCV*, *Numpy* и *Matplotlib*. Также были использованы такие модули языка *Python* как: *Glob*, *Random*, *Copy*, *Sys* и *Math*.

В качестве среды разработки был выбран *Google Colab* – бесплатный облачный сервис на основе *Jupyter Notebook*.

Исследования проводились на 2 выборках текстовых изображений. Первая выборка состоит из 117 текстовых изображений (см. рис. 12), сгенерированных с помощью языка программирования *Python*, библиотеки *ImageMagic* языка программирования *Python*, *PDF* и системы компьютерной верстки *Latex*.

Вторая выборка состоит из 4 изображений (см. рис. 13), сгенерированных с помощью текстового редактора *Word*, *PDF* и библиотеки *ImageMagic* языка программирования *Python*.

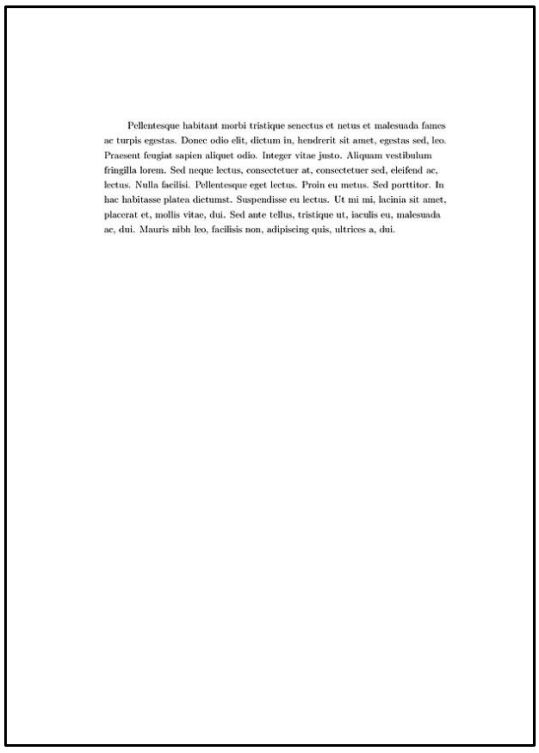
Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Рис. 12. Пример сгенерированного изображения из первой выборки

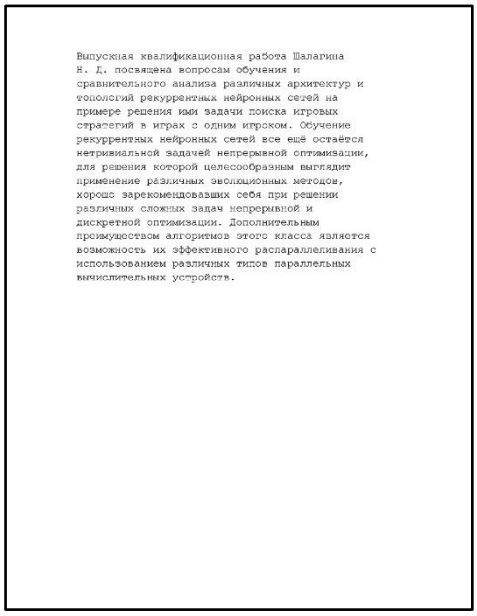

Рис. 13. Пример сгенерированного изображения из второй выборки

Загрузив 117 изображений из первой выборки в программу, и сравнив полученные значения с исходными получаем следующие результаты (см. рис. 14).

```
Процент распознования абзацного отступа: 82.90598290598291
Процент распознования типа шрифта: 100.0
Процент распознования типа выравнивания: 68.37606837606837
```

Рис. 14. Результаты тестирования на первой выборке

Загрузив 4 изображений из второй выборки в программу, и сравнив полученные значения с исходными получаем следующие результаты (см. рис. 15).

```
Процент распознования абзацного отступа: 100.0
Процент распознования типа шрифта: 100.0
Процент распознования типа выравнивания: 100.0
```

Рис. 15. Результаты тестирования на второй выборке

Заключение

В ходе выполнения настоящей работы были изучены методы сегментации текстовых блоков на изображении, а также разработана программная реализация загрузки исходного изображения с текстовым фрагментом с последующей сегментацией букв и строк. Так же была разработана программная реализация методов автоматической классификации текстовых фрагментов. Были проведены тесты разработанных методов на двух выборках текстовых изображений, для этого были сгенерированы тестовые изображений с помощью языка программирования *Python*, библиотеки *ImageMagic* и системы компьютерной верстки *Latex*. Проведенное тестирование показало перспективность предложенных подходов и методов к структуризации и классификации текстовых блоков.

Список литературы

1. Santos R., Clemente G. S., Ing Ren T., Calvalcanti G. Text Line Segmentation Based on Morphology and Histogram Projection // Document Analysis and Recognition, 10th International Conference on, 2009. Pp. 651-655. DOI: 10.1109/ICDAR.2009.183.

2. Likforman-Sulem L., Zahour A., Taconet B. Text line segmentation of historical documents: a survey // International Journal on Document Analysis and Recognition. 2007. № 9. Pp. 123-138. DOI:10.1007/s10032-006-0023-z.

3. Shapiro L. G., Stockman G. C. Computer Vision. New Jersey: Prentice-Hal, 2001. 580 pp. ISBN 0-13-030796-3.

4. Sagar B. M., Shobha G. , Ramakhanth Kumar P. Converting printed Kannada text image file to machine editable format using Database // International Journal of Computers. 2008. Vol. 2. Pp. 173–175.

5. Srihari S. N., Govindaraju V., Shekhawat A. Interpretation of Handwritten Addresses in US Mailstream // *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93)*. 1993. Pp. 291–294. DOI: 10.1109/ICDAR.1993.395729.

6. Maloo M., Kale K. V. Gujarati Script Recognition: A Review // International Journal of Computer Science Issues.2011. Vol. 8, Issue 4, № 1. Pp. 480–489.

7. Patil S. B. Neural Network based bilingual OCR system: experiment with English and Kannada bilingual document // International Journal of Computer Applications. 2011. Vol. 13, № 8. Pp. 6–14.

8. Thungamani M. and Ramakhanth Kumar P. A Survey of Methods and Strategies in Handwritten Kannada Character Segmentation // International Journal of Science Research. 2012. Vol 01, Issue 01. Pp. 18–23.

9. Das M. S., Reddy C. R. K., Govardhan A., Saikrishna G. Segmentation of Overlapping Text lines, Characters in printed Telugu text document images // International Journal of Engineering Science and Technology. 2010. Vol. 2(11). Pp. 6606–6610.

10. Arica N.,Yarman-Vural Fatos T. An Overview of Character Recognition Focused on Off-Line Handwriting // IEEE Transactions. Systems, Man, and Cybernetics. Part C: Applications and Reviews. 2001. Vol. 31. Pp. 216–233. DOI: 10.1109/5326.941845.

11. Marinai S., Nesi P. Projection Based Segmentation of Musical Sheets //Document Analysis and Recognition. 1999. Pp. 515–518. DOI: 10.1109/ICDAR.1999.791838.

12. Junga K., Kimb K. I., Jain A. K. Text information extraction in images and video: A survey// Pattern Recognition. 2004. Vol. 37. Pp. 977–997. DOI: 10.1016/j.patcog.2003.10.012.

13. Pham D. L.; Xu Ch. Prince J. Current Methods in Medical Image Segmentation // Annual Review of Biomedical Engineering. 2000. Vol. 2. Pp. 315–337. DOI: 10.1146/annurev.bioeng.2.1.315.

14. Lew M. S., Sebe N., Djeraba C., Jain R. Content-based Multimedia Information Retrieval: State of the Art and Challenges // ACM Transactions on Multimedia Computing, Communications, and Applications. 2006. Vol. 2, № 1. Pp. 1–19. DOI:10.1145/1126004.1126005.

15. Naramala V. R., Srikrishna A., Bhogapathi R., Babu G. An efficient feature extraction and classification of handwritten digits using neural networks // International Journal of Computer Science, Engineering and Applications. 2011. Vol.1, № .5. Pp. 47–56. DOI: 10.5121/ijcsea.2011.1505.

16. Xiaoyan Z., Yifan S. New Algorithm for Handwritten Character Recognition. Beijing, China.

17. Aboeleneen Z. A, Abdel Azim G. An Improved Image Segmentation Algorithm Based on MET Method[J] // International Journal of Computer Science Issues. 2012. Vol. 9, Issue 5, № 3. Pp. 346–351.

18. Xu C., Huang F., Mao Z. An improved two-dimensional Otsu thresholding segmentation method[J] // Application of Electronic Technique, 2016.

19. Zhao H, Wang Q, Wu W, Yuan N. An Improved Method Research of SAR Images Thresholding Segmentation // Lecture Notes in Electrical Engineering, 2014. Vol. 238. Pp. 1151–1157. DOI: 10.1007/978-1-4614-4981-2-125.

20. Wang Y. Q., Zhuang L. L., Shi C. X. Construction Research on Multi-Threshold Segmentation Based on Improved Otsu Threshold Method // Advanced Materials Research. 2014. Vol.1046. Pp. 425–428. DOI: 10.4028/www.scientific.net/AMR.1046.425.