

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра информационных систем управления

Аннотация к дипломной работе

**«Разработка методов распознавания текста с использованием
N-грамм для коррекции результатов»**

Кусок Екатерина Николаевна

Научный руководитель – кандидат физико-математических наук, доцент
Образцов Владимир Алексеевич

Минск, 2017

Реферат

Дипломная работа, 50 с., 11 рис., 1 табл., 34 источника, 1 приложение.

Ключевые слова: N-ГРАММА, РАСПОЗНАВАНИЕ ТЕКСТА, OCR, OPENCV, TESSERACT, ASP.NET.

Объектом исследования является задача распознавания текста на изображении.

Цель работы – изучение способов распознавания текста на изображении и последующей его коррекции, изучение и анализ существующих подходов к решению сформулированной задачи, изучение методов и алгоритмов работы библиотеки компьютерного зрения и библиотеки оптического распознавания символов, разработка и исследование моделей и алгоритмов оцифровки изображений текста и коррекции полученных результатов. Для решения задачи применяются методы обработки изображений, OCR распознавания, коррекции результатов распознавания с применением символьных N-грамм. Выполнена программная реализация разработанных алгоритмов и проведены компьютерные эксперименты.

Результаты работы – модели, алгоритмы, программный инструментарий и методика его применения для распознавания текста на изображении.

Полученные результаты могут использоваться в пользовательских приложениях для распознавания текста на изображении, а также для оцифровки и индексирования бумажных фондов библиотек или архивов.

Рэферат

Дыпломная работа, 50 с., 11 мал., 1 табл., 34 крыніцы, 1 прыкладанне.

Ключавыя словы: N-грама, распазнаванне тэксту, OCR, OPENCV, TESSERACT, ASP.NET.

Аб'ектам даследавання з'яўляецца задача распазнавання тэксту на малюнку.

Мэта работы - вывучэнне спосабаў распазнавання тэксту на малюнку і наступнай яго карэкцыі, вывучэнне і аналіз існуючых падыходаў да вырашэння сфармуляванай задачы, вывучэнне метадаў і алгарытмаў работы бібліятэкі кампутарнага гледжання і бібліятэкі аптычнага распазнавання сімвалаў, распрацоўка і даследаванне мадэляў і алгарытмаў лічбавання малюнкаў тэксту і карэкцыі атрыманых вынікаў. Для вырашэння задачы прымяняюцца метады апрацоўкі малюнкаў, OCR распазнавання, карэкцыі вынікаў распазнавання з ужываннем знакавых N-грам. Выканана праграмная рэалізацыя распрацаваных алгарытмаў і праведзены кампутарныя эксперыменты.

Вынікі работы – мадэлі, алгарытмы, праграмны інструментар і метадыка яго прымянення для распазнавання тэксту на малюнку.

Атрыманыя вынікі могуць выкарыстоўвацца ў карыстацкіх прыкладаннях для распазнавання тэксту на малюнку, а таксама для аблічбоўкі і індэксавання папярковых фондаў бібліятэк або архіваў.

Abstract

Diploma thesis, 50 p., 11 pic., 1 table, 34 sources, 1 appendix

Keywords: N-GRAMME, RECOGNITION OF TEXT, OCR, OPENCV, TESSERACT, ASP.NET.

The object of the work is the task of recognizing text on an image.

The purpose of the work is to study the ways of text recognition on the image and its subsequent correction, to study and analyze existing approaches to solving the formulated problem, to study methods and algorithms of the computer sight library and the optical character recognition library, to develop and study models and algorithms for digitizing text images and correcting received results. Methods of image processing, OCR recognition, correction of recognition results using symbolic N-grams are applied to solve the problem. The software implementation of the developed algorithms was carried out and computer experiments were done.

The results of the work are models, algorithms, software tools and methods of its application for text recognition on the image.

The obtained results can be used in user applications for recognizing text on an image, as well as for digitizing and indexing paper collections of libraries or archives.