Мальсагова Вероника
студентка
Фомичева Татьяна Леонидовна
канд. экон. наук, доцент
ФГОБУ ВО «Финансовый университет
при Правительстве Российской Федерации»
г. Москва

РАСПОЗНАВАНИЕ ТЕКСТА. СИСТЕМЫ МАШИННОГО ЗРЕНИЯ

Аннотация*: в статье представлен обзор систем машинного зрения для оптического распознавания символов (OCR-optical character recognition) в изображении. Методы машинного зрения применяются в различных областях: от медицинской визуализации до дистанционного зондирования, регулирования производственных процессов, обработки документов, нанотехнологий и мультимедийных баз данных. Оптическое распознавание символов является частью решения важнейших прикладных задач (восстановление документов, распознавание номеров автомобилей, публикация текста на веб-странице, оцифровка книг и др.), и является исследуемой проблемой в области машинного и компьютерного зрения) [4].*

Ключевые слова*: машинное зрение, распознавание, Тессеракт, Окропус.*

Abstract*: the article provides an overview of the library machine vision for optical character recognition (OCR) in the image. Machine vision methods are used in various fields: from medical imaging to remote sensing, industrial control, document processing, nanotechnology and multimedia databases. Optical character recognition is part of the solution of the most important applied problems (document recovery, recognition of car numbers, publication of text on a web page, digitization of books, etc.), which is an investigated problem in the fields of machine and computer vision) [4].*

Keywords*: machine vision, recognition, Tesseract, OCRopus.*

Machine vision covers all industrial and non-industrial applications in which a combination of hardware and software provides operational control for devices in performing their functions based on image capture and processing. Industrial vision systems require greater reliability and stability compared to similar educational systems and, as a rule, they are much cheaper than those used for state / military purposes. Therefore, industrial machine vision involves low cost, acceptable accuracy, high reliability, mechanical and thermal stability of the used hardware components [2]. The purpose of a machine vision system is to create a model of the real world from images. Using computer vision for optical text recognition has attracted the attention of researchers in many fields of application and has been used to solve many problems.

The terms «machine vision» and «computer vision» appeared many years ago, when the situation was very different from today. In the early days, computer vision implied the study of «vision» and the possible design of related software, while machine vision meant studying not only the software, but also the hardware environment, as well as the methods for acquiring the images needed for real applications, so this there was a much more engineering-oriented approach. Nowadays, computer technology has advanced so much that a significant proportion of real-time applications can be implemented on simple personal computers. This and many other changes in knowledge in this area have led to a significant convergence between the terms, as a result of which they can be used almost interchangeably [1].

Machine vision systems rely on digital sensors that are protected within the industrial cameras with special optics for imaging so that computer hardware and software can process, analyze and measure various characteristics for decision making. Other components of the machine vision system are lighting, a lens, a device for analysis and processing, and a device for using data.

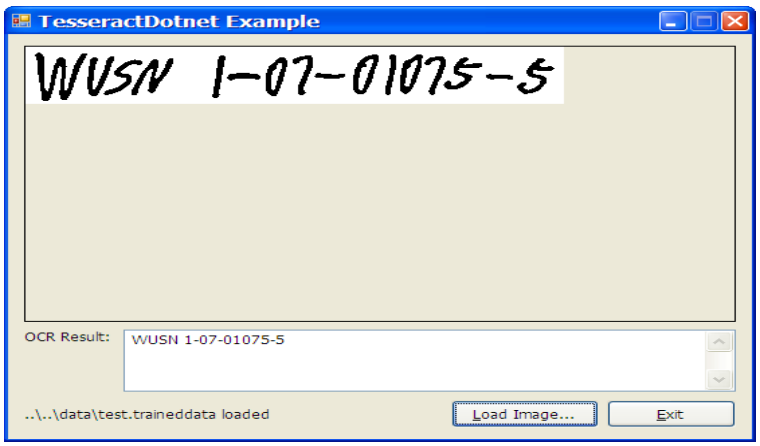
Lighting indicates the item being tested, allowing its features to stand out so that they can be clearly seen with the camera. The lens captures the image and presents it to the sensor in the form of light. The sensor in the machine vision camera converts this light into a digital image, which is then sent to the processor for analysis. Vision

processing consists of algorithms that scan the image and extract the necessary information, conduct the necessary verification and make a decision. Finally, communication is usually carried out using a discrete input-output signal or data sent via a serial connection to a device that registers information or uses it [1].

The most two popular free OCR applications are OCR-opus and Tesseract, which are discussed in more detail.

Tesseract

Tesseract [3] was developed as private software in the laboratories of Hewlett Packard from 1985 to 1994. Over the next decade, Tesseract was not in demand, and only in 2005, the source code of the application was opened by Hewlett Packard together with the University of Nevada in Las Vegas. In 2006, the project was acquired by Google. Tesseract is able to recognize almost all UTF-8 encoded characters, including Cyrillic. However, when the text is recognized, dictionaries are used, so the language of the recognized text should be unified. Tesseract is compatible with three operating systems: Windows, Linux, and Mac OS. The result of the program using Tesseract will be of very low quality if the input images are not processed in accordance with the requirements: at least 20x20 pixels in size, any rotations or distortions of the characters must be adjusted, otherwise the text will not be recognized, low-frequency changes in brightness should be filtered without loss of quality, otherwise the binarization phase of the image will destroy most of the processed text, and dark borders must be removed manually, as they may be incorrect o interpreted as symbols. The highest recognition results are achieved with a clear separation of the text from the background. It is usually extremely difficult to guarantee this in practice, therefore, to solve specific problems, other methods are used in which various classifiers and text detectors are trained. However, in some tasks, using OCR Tesseract may be useful. For example, scans of documents or books often do not contain a lot of noise. Picture 1 shows an example of text recognition in an image using the Tesseract library.



Pic. 1. The result of using Tesseract on the.NET platform *OCRopus*

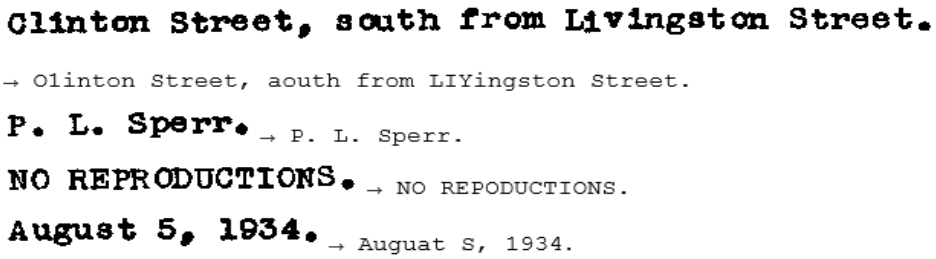
OCRopus [6] has been specifically designed for use in large-scale digitization projects of books such as GoogleBooks and InternetArchive, and supports a large number of languages and fonts. The first version appeared in 2007. OCRopus provides greater than Tesseract accuracy in recognizing complex areas of images containing text by linking them. Dictionaries can be used to highlight components. For example, in the word «oldest» the characters «d» and «e» are associated, and in the word «register» – «g» and «i». The process of converting an image into text is divided into three stages: binarization, segmentation and recognition [5].

At the binarization stage, the image is converted to monochrome. A variety of adaptive threshold processing is applied here, and perhaps an offset estimate is made to rotate some distorted characters.

Segmentation divides the image into several components. The letters of each component present in the binary image are scaled to calculate their sizes. Too large or too small components found in the image are deleted because they are not considered letters. Then, the derivative of the Gaussian kernel is used to determine the upper and lower boundaries of the remaining components. Horizontal blur allows you to align the upper borders of the letters.

Recognition occurs using LSTM-networks (long short-term memory). The input signals of the network are columns of pixels, which for each image are fed into the network one at a time, from left to right. The output signals contain points for each possible letter. The result of text recognition in the image using the OCRopus library

is shown in picture 2. As the results show, some characters are not recognized correctly.



Pic. 2. The result of using OCRopus

OCR libraries based on Tesseract and OCRopus, as a rule, are cross-platform, actively supported and freely distributed, that is, they do not require licensing for use in their own projects. A significant drawback is the inability to correctly recognize all the characters of the image in the presence of noise. In addition, processed images should have the highest possible resolution. Although OCR is not a new technology, it can still be used to solve some application problems. More advanced systems, the models of which are based on convolutional and recurrent neural networks, still do not allow obtaining ideal recognition accuracy.

References

1. Davies E.R. Computer and Machine Vision: Theory, Algorithms, Practicalities (4th ed.) / Academic Press. – P. 410–411. ISBN 9780123869081.

2. Jain R. Machine Vision / R. Jain, R. Kasturi, B.G. Schunck. – McGraw-Hill, Inc., 1995. ISBN 0-07-032018-7.

3. Kay A. Tesseract: Open-Source Optical Character Recognition Engine [Электронный ресурс]. – Режим доступа: http://www.linuxjournal.com/article/9676

4. Schantz H.F. The history of OCR, optical character recognition / Manchester Center, Vt.: Recognition Technologies Users Association. ISBN 9780943072012.

5. How to Digitize Texts with Open-Source Command-Line Optical Character Recognition (OCR) Software [Электронный ресурс]. – Режим доступа: https://hdw.artsci.wustl.edu/articles/154

6. OCRopy: Python-based tools for document analysis and OCR [Электронный ресурс]. – Режим доступа: https://github.com/tmbdev/ocropy