EURASIAN JOURNAL OF ACADEMIC RESEARCH
Innovative Academy Research Support Center
UIF = 8.1 | SJIF = 5.685 www.in-academy.uz

PROBLEMS WITH TEXT RECOGNITION IN IMAGES

Egamberdiev Nodir Abdunazarovich
Associate Professor of Tashkent University of Information Technology named after Muhammad al-Khwarizmi
Berdiev Maruf Ramshiddin ogli
Third year student of Tashkent University of Information Technology named after Muhammad al-Khwarizmi
Mahmudov Islomjon Dilmurod ogli
Student of Tashkent University of Information Technology named after Muhammad al-Khwarizmi
https://www.doi.org/10.5281/zenodo.8262911

ARTICLE INFO

Received: 11th August 2023
Accepted: 17th August 2023
Online: 18th August 2023

KEY WORDS

Text Recognition, Artificial Intellige, Classification, OCR.

ABSTRACT

Text recognition in images is a crucial task with widespread applications, from document digitization to augmented reality. However, this article sheds light on the persistent challenges that plague this process. Factors like poor image quality, diverse fonts, complex layouts, and occlusions hinder accurate text extraction. Through a comprehensive analysis of these problems, this article underscores the need for robust solutions, exploring recent advancements in image processing, deep learning, and AI techniques. Addressing these challenges is essential to unlock the full potential of text recognition and enable seamless integration into various technological landscapes.

Image texts are widely used in our daily life as an important form of expression of human language. The problem of extracting text information from visual clues has been around for years attracted wide attention. Great progress has been made in processing characters printed on clean backgrounds, such as scanning document pages. Text pixels from a document scan can be easily separated from the background. Although more and more information is being digitized today, visual texts are included in many forms of digital media such as images and videos. Compared to text in documents, text in mass media contains less but often important information about the content of the mass media. They usually provide important product names, locations, brands, game results, date and time, which is useful information for understanding and indexing these images and videos. However, text in images and videos can be placed on an arbitrary background or placed on the surface of objects in the scene with different font, size, color, alignment, motion, and lighting conditions, making it very difficult to extract the text. The goal of research on text detection and recognition in images and videos is to find the right way to extract different types of text from complex images or videos. This not only expands the application of OCR system to a wider range of multimedia fields, but also helps people to better understand the mechanism of visual text detection and recognition. Effective descriptions are needed to index and retrieve images and

videos. Display content based on low-level features such as color and text for the user to index and retrieve [1-2].

Difficult to enter as keywords, these low-level features are difficult to perform as efficiently as text-based indexing and retrieval methods. Most high-level content images, such as human faces and bodies, physical objects, and movements in images and videos, are not only difficult to automatically capture, but also difficult to describe and match. Image-based indexing and retrieval systems need effective descriptors to enjoy their success. As a form of high-level visual content, text conveys clearer and clearer meaning and easier description than low-level and other high-level visual features.

There are two facts that demonstrate the ease of indexing and retrieval using a text-based search engine and an OCR system. Images and texts are a powerful source of information. Captions in news videos usually provide the name, location, subject, date, and time of the people involved. Titles often also provide an abstract or summary of the program. Provides information such as titles, actors, producers, contributors, etc., which are displayed at the beginning or end of movies. Captions in sports programs often include the names of teams and players. Textual information can also be found on the scene, such as player numbers and names, team names, brand names, locations, and commercials. The maps, figures and tables shown in the videos contain a lot of text about locations, temperatures, certificate elements.

The titles, logos and names of the programs shown in the image are important for the explanation of the program types and names. Even converting paper documents such as vehicle licenses and brand names, CD covers, books and magazines in video streaming or images contain valuable information. More and more websites choose images to enhance the visual impact of their headers. Therefore, these media texts serve as important resources for indexing, annotation, and other content - oriented processing. Text-based search and matching technologies are also well developed. Text-based search engines are well developed to solve the problem of automatic and efficient document retrieval. Document OCR systems provide high recognition rates for machine-printable text on plain paper.

Combining text-based search engine and OCR technologies seems to be a good solution for indexing and retrieving images and videos. However, document OCR systems typically require high resolution and contrast and distribute uniform background text with a normal gray value as input. Current optical character recognition technologies cannot recognize text that is not printed on a clean background. Similarly, document parsing methods that attempt to segment entire pages into separate segments typically require binary input and involve a specific document order. For example, a newspaper or a technical article. In images and videos, the background can be any indoor or outdoor scene, and the text can vary in font, size, color, 3D position and movement, lighting conditions, and shading. Therefore, current OCR cannot be directly applied to recognize text in images or videos. Thus, there is a high demand for systems that can find, extract and recognize unlimited text from any background. The main types of text in images can be roughly divided into scene text and overlay text, depending on the source of the text. Scene text is text that is part of a scene, text that is superimposed on the scene. Caption can come from a recorded scene or photo. Examples of image texts include the name on a T-shirt, commercials in sports stadiums, and road signs. This type of text may

contain important information for indexing or may be random (advertising) and not suitable for indexing. Such a text is more difficult to identify, distinguish and recognize according to the nature of the scene. For example, movement, lighting, affine transformation and cramming of texts. On the other hand, superimposed text usually contains relevant information. Some researchers like to use the term "graphic text" to describe the same concept. In the case of news, for example, it is usually created to provide the viewer with basic information about the current content of the program. Therefore, such text is very important for indexing, as well as known general.

It is also easy to isolate because it appears under constraints. Image text can be characterized by the following properties:

Variation in size, font, color, direction, style, alignment, even within words.

– Part of the text may be closed.

- With complex movement in the image.

– With changing lighting conditions.

– With variability of transformations.

– Deformation, if on a flexible surface

Embedded text usually has the following characteristics:

– Text is always in the foreground and never in the foreground.

– Usually with a stable lighting condition, it is independent of the scene

- Text of pixels values are distributed according to limited rules.

– Size, font, spacing and orientation are constant in the text area.

Text is usually horizontally oriented. The following additional features are typically observed for text added to video images:

is stationary or linear in horizontal or vertical direction

– The background is the same for moving text

– The same text appears in several consecutive frames.

- Low resolution of the text.

Since clean text processing systems have produced very good results, researchers hope to design dirty text processing architectures based on existing technologies. There are different ways of using existing OCR technologies in text recognition. The simplest method is to clean the input images or videos with various segmentation technologies, which can be directly used as the input of the clean OCR system, because it does not need to change anything in the traditional OCR system. Unfortunately, most research shows that it is impossible to segment pixels of text without knowing where and what the characters are. The trade-off is to segment the text - like texture instead of doing actual text segmentation in the first step. And then pure OCR technologies are applied to recognize text from these text - like texture regions or reject them. Some researchers give their own definitions. For example, he defined a similar texture as a horizontal rectangular structure of accumulated sharp edges. Other researchers prefer to define this concept implicitly in their assumptions or algorithms. There are some differences between text segmentation in images and videos. One obvious difference is that videos offer signs of objects moving. We discuss typical architectures for still images and video [5-10].

A key point in designing a text- like text segmentation algorithm is to find a way to measure the difference between text pixels and "background" pixels. Due to the different definitions of text- like text, most of the existing research prefers to design a measure for each specific application rather than to systematically search for a solution. By studying the architecture of their algorithms, we hope to paint a clear picture of the ideas of these studies. One of the first algorithms presented by, it aims to extract and recognize scene texts from images. The algorithm consists of three steps. In the first step, the input images are roughly divided into regions using an adaptive boundary-based image segmentation method. Second, character candidate regions are selected by examining the features under the following assumptions:

– the gray level of the text characters region has a high contrast against the background;
– the segmentation width of the text characters is the same;
– the gray level of segmentation of text characters is the same;
– the spatial frequency of segmentation of text characters is the same.

Applies a recognition process to combine individual parts of a single text character and extract character pattern candidates. The corresponding text characters of the text in this algorithm can be distorted in 3-dimensional space under uncontrolled lighting conditions. In addition, text characters can have different sizes, heights, positions, fonts, formats, and gray levels. From this three-step process algorithm, we can clearly find three types of features that are used to measure similar text in this algorithm. In the final step, the algorithm applies a recognition process to combine individual parts of a single text character and extract character pattern candidates. The corresponding text characters of the text in this algorithm can be distorted in 3-dimensional space under uncontrolled lighting conditions. In addition, text characters can have different sizes, heights, positions, fonts, formats, and gray levels.

Statistical and distributional properties of the pixel values of the character and its local background, such as uniform gray level segmentation of text characters and high contrast against the background. Features of the 2D spatial distribution of characters in a word or sentence, for example, the same spatial frequency of text characters, text alignment. Shape properties of the characters, the same width of the text symbol. In the algorithm, three different features are used one by one to ensure the fast operation of the algorithm. At the same time, some strict limitations are applied in the algorithm, such as gray level uniformity, character segmentation width and spatial frequency, which limit the applicability of the algorithm. Another typical algorithm of text segmentation in images is proposed by four steps. First, a texture segmentation scheme is used to roughly divide the images into text regions and background, assuming that the text-like texture has certain statistical properties along the Gaussian scale. Second, text regions are defined under constraints derived from heuristics assumed for text lines, such as height similarity spacing and alignment (linear alignment with fixed spacing between text lines). Third, the text region candidates are binarized according to the distribution characteristics of the pixel values. Finally, the text string candidates are refined by applying the same type of constraints used in the second step with stricter standards. In contrast to this method, this algorithm focuses on the placement of the text string, but not the characters of the first step, which is often used in other media text OCR systems.

Statistical and distributional properties of the pixel values of the character and its local background, such as uniform gray level segmentation of text characters and high contrast against the background. Features of the 2D spatial distribution of characters in a word or sentence, for example, the same spatial frequency of text characters, text alignment. Shape properties of the characters, the same width of the text symbol. In this algorithm, three different features are used one by one to ensure fast operation of the algorithm. At the same time, some strict limitations are applied in the algorithm. For example, gray level uniformity, character segmentation width, and spatial frequency limit the applicability of this algorithm. Another typical algorithm for segmentation of text in images was proposed by, it consisted of four steps. First, a text segmentation scheme is used to roughly segment images into text regions and backgrounds, assuming that text-like text has certain statistical properties along a Gaussian scale. Second, text regions are defined under constraints derived from heuristics assumed for text lines, such as height similarity (characters are the same size in a word or sentence), spacing, and alignment (by fixing the space between text lines linear smoothing). Third, the text region candidates are binarized according to the distribution characteristics of the pixel values. Finally, the text string candidates are refined by applying the same type of constraints used in the second step with stricter standards. Unlike the method, this algorithm focuses on the localization of the text string (words or sentences), but not the characters of the first step, which is often used in other media text OCR systems.

Both of these algorithms first attempt to approximate text region candidates without considering the shape features of the characters, and then apply tighter constraints to identify and segment these candidates. There are other algorithms for detecting text regions in grayscale and color images using texture analysis and connected component. The first, as we discussed above, focuses on identifying, placing, or segmenting text from images with complex backgrounds. Most of them consider the text to be:

can have different sizes, but must keep the same font size as a word;

should be aligned on a horizontal line. Some algorithms, such as the spatial variance method, can be extended to find text in any linear alignment;

should have a high contrast with the local background in gray value or color value ;

allows you to change the text in a certain way. A built - in OCR system can make such changes. There is no explicit discussion of text sizes in these algorithms, and some algorithms limit text sizes to certain ranges. Most videos are often used to represent moving scenes with a stream of image frames. Some motion information about objects in the scene is hidden between continuous frames. Both everyday experience and research in psychology show that motion can be used to discriminate between different physical objects in the visual system. Therefore, many researchers have tried to use motion data to improve text segmentation in video streams. This work is usually done in two stages. The first step is to calculate the motion in the video. Pixels or blocks of pixels are shown in one frame, then similar pixels and blocks of pixels are found in subsequent frames according to a certain dimension.

A vector pointing from the location of a pixel or block in the first frame to the next frame is used to represent motion data. Second, this motion data can be used to improve our text segmentation based on various assumptions about the text and its background. This type of media reported the OCR system. First, a text segmentation step is used to generate a binary

image that represents the text that appears in the video. Standard OCR software packages are then used to directly recognize the segmented text. The segmentation stage consists of several processing steps. Color images are first processed with a region-based algorithm (split-merge) to divide the entire image into larger homogeneous regions. Then, the region images are binarized with a local color contrast threshold. Some heuristic information about the height and width of the possible text regions is then used to remove unsuitable candidates. In addition, potential words or lines of text are extracted by estimating the writing direction under the assumption that the text is aligned on a horizontal line. In addition, it is assumed that the text appears in the same position or in several consecutive frames depending on the linear movement. Therefore, text candidates that do not appear in the required number of consecutive frames are deleted. Correspondence between multiple frames is calculated by a simple region-matching algorithm and checked on five consecutive frames.

can be a binary image that shows the separated characters in their original place. The advantage of this method is the ease of introducing the existing clean OCR program into the new system. For some applications, e.g. recognition rates for titles and credits range from 41% to 76%, which is lower than other methods discussed in the following subsections. Since few resources are working on this research line, it is still difficult to say whether such an architecture can give a satisfactory result in text recognition. Provides an architecture for using motion data in a different way. The recognition algorithm performs video OCR on captions from news videos that contain low-resolution characters and a complex background with wide variation. In the first step, the algorithm determines approx.

A text region is defined as a horizontal rectangular structure of clustered sharp edges. The second step of the algorithm is aimed at improving image quality. There are two methods used in this step. One method uses sub-pixel interpolation to produce high-resolution images. The second method, called multi-frame integration, uses motion information to enhance the background from the complex background. As an integrated OCR algorithm, characters are extracted by four special line element filters and projection profile analysis. The algorithm stores multiple segmentation results due to the difficulty of character segmentation. In the last step, several segmentation results are used as inputs to the OCR system. Recognition is improved using dictionary - based post-processing.

The algorithm assumes that news headlines are not always high-contrast compared to the moving background. Pixel-based motion analysis is mainly used for contrast enhancement. There are also ways to view video as standalone frames. The example algorithm uses such an architecture. The algorithm uses a neural network to classify each input image frame into a class of text pixels and non-text pixels. After smoothing the classified image frame, text pixels are extracted by applying binarization. When video is processed as independent frames, it can almost be considered still images. The two proposed motion analysis methods are pixel-based or rectangular block -based, while the pixel-based method considers the text pixels as stationary, and the block-based method constrains whether the text movement should be top-down or linear. The OCR system has been researched for many years. Some OCR systems can archive high recognition rates, especially for machine-printed characters on clean backgrounds. Since most of these systems require the input image as a

binary still image or an easily binarized still image, they cannot directly recognize text in an image or video.

Ways to create an image OCR system based on text segmentation technologies presented in the last section. One approach focuses on improving preprocessing in a pure OCR system with precise text detection, positioning, and segmentation technologies to obtain more accurate text segmentation. When the relevant parts of the text texture are fully extracted from the images, they are used as input to the clean OCR system to produce the final text data. These relevant parts of the text texture can be text, lines, words or single character blocks depending on different application systems. Applied architecture is a good example. The input image is first cleaned by four- step processes, then commercialized for recognition OCR software is used. In this way, the new system can use existing OCR technologies. The disadvantage of this method is that recognition is limited by the performance of existing OCR technologies. Although text segmentation processes can clean up images to some extent, the segmented images produced by the process are still unsuitable for traditional OCR. They are usually noisy, have low resolution and character display changes.

The final recognition performance of the system must depend on the noise adaptation of the pure OCR technologies used. Another way to create an OCR system is to develop new character recognition technologies to accommodate the strong noise caused by complex backgrounds. Sato's algorithm is a good example. In the article, a comparison is made between the newly developed recognition algorithm and traditional recognition software. When applied to the news video stream, the recognition rate of the algorithm developed by the authors reaches from 76.2% to 89.8%., which is much higher than traditional OCR's recognition rate of 38.9% to 53.2%. As we discussed in the last section, the goal of text segmentation is to separate the pixels of the image frames into two classes. Ideally, one class consists of pixels of text. Another class contains non - text pixels. A large number of binarization methods have been proposed. They are typically developed for specific applications (eg, mailing address, checks) that allow strong constraints on document structure (eg, location, character size).

The binary method uses a single threshold value applied to the entire image, which is also called the global thresholding method. They can be divided into non - parametric, parametric and other methods. Nonparametric methods are described in Nonparametric Methods. This method calculates the between- class and intra-class covariance ratio for each potential threshold, where the two classes represent the foreground and background pixels. The goal is to find the limit that maximizes this ratio, which is then chosen as the limit. It is described in a similar way to maximize the entropy of two classes. Another method is based on conservation of momentum. It finds the threshold that best preserves the moment statistics in the captured binary images compared to the original grayscale. Initial moments are calculated from the intensity histogram.

Parametric Methods A parametric threshold method based on a minimum error bound is reported. It is based on a pattern recognition technique where the foreground and background intensity distributions are modeled as normal probability density functions. The threshold is chosen so that the classification error between two classes is minimal. It was found that the error rate with this method is lower than the non - parametric methods.

Adaptive Delimitation One of the possibilities for delimitation of non - homogeneous regions is the use of adaptive delimitation methods. These methods typically analyze local windows across the image where adaptive thresholding is performed. The main problem of these methods is the choice of window size. The window should be large enough to contain the number of pixels in the window, but not be too large to moderate the non - uniform background intensity. Various methods were evaluated for documents, such as background images, shadows, light, smears and spot checks. This shows that the performance of the methods depends on the application.

Delineation methods generally perform well on images with homogeneous backgrounds. However, when applied to a non-homogeneous background, as is often the case for text in images and videos, the results are often unsatisfactory, as noted by several researchers Scale-space method: 36 image pixels presented a scale-space algorithm for clustering into regions. Three second-order Gaussian derivatives on three different scales are used as a filter for calculation. The size of the set of feature vectors for each pixel in the image. Text string regions are defined by clustering 9- dimensional feature vectors into a class of interest. In other words, one of the nine classes is the text string region class.

scaling approach used in the algorithm is also suitable for dealing with changes in text size. proposed a method for placing text regions based on edge detection, which is called the spatial dispersion method. It calculates the local spatial variance along each horizontal line over the entire input grayscale image. (this method assumes that the text is horizontally aligned.) An edge detector is then used to find important horizontal edges. Finally, edges in opposite directions are paired with the bottom and top borders of the text line. The main drawback for the phase dispersion method is that it sometimes halves extended characters such as σ, π. He also proposed a method for embedding text in color images, called the linked component method. This method reduces the RGB color values from 24 bits to 6 bits by first reserving the upper 2 bits for the R, G and B value. Color-reduced images are separated into mono-color components. The color corresponding to the highest number of pixels in the image is taken as the background. The color components are then connected to the text string using heuristics. In the next step, the algorithm uses the OCR system to recognize the text in the connected component. The detected connected component is expanded to find missing characters in the neighborhood. This algorithm can use some form of character information, such as a block adjacency graph, to perform connected component analysis. It works for both grayscale images and color images, but requires the characters to have a specific color from the background. A major drawback of the linked components algorithm is that it cannot recognize linked symbols.

The complexities surrounding text recognition in images are undeniable, presenting formidable barriers despite its significance in modern applications. This article has elucidated the multifaceted challenges posed by image quality, font diversity, layout intricacies, and occlusions. However, advancements in image processing and deep learning offer promising avenues for overcoming these obstacles. As technology evolves, collaborative efforts between academia and industry hold the key to innovative solutions that can empower accurate and efficient text recognition. Addressing these challenges is not only vital for improved user

experiences but also for unlocking the full potential of text-based information in our increasingly visual digital landscape.

References:

[1] P. Havali, J. Banu. Deep Convolutional Neural Network for Image Classification on CUDA Platform, ScienceDirect, 2019, Pages 99-122.

[2] Dilnoz Muhamediyeva, Nadir Egamberdiyev. An application of Gauss neutrosophic numbers in medical diagnosis // International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities http://www.icisct2021.org/ ICISCT 2021, November 3-5, 2021.

[3] Kamilov M.M., Khujaev OK, Egamberdiev NA The method of applying the algorithm of calculating grades for finding similar diagnostics in medical information systems, International Journal of Innovative Technology and Exploring Engineering, 8-6S, pp. 722-724.

[4] Muhamediyeva DT, Jurayev Z.Sh., Egamberdiyev NA, Qualitative analysis of mathematical models based on Z-number // Proceedings of the Joint International Conference STEMM: Science – Technology – Education – Mathematics – Medicine. May 16-17, 2019, Tashkent, pp. 42-43.

[5] Egamberdiyev NA, FUZZY REGRESSION ALGORITHM FOR CLASSIFICATION OF WEAKLY FORMED PROCESSES // SCIENCE AND PRACTICE: IMPLEMENTATION TO MODERN SOCIETY MANCHESTER, GREAT BRITAIN, 26-28.12.2020.

[6] D.Mukhamediyeva, N.Egamberdiev, ALGORITHM OF CLASSIFICATION OF MEDICAL OBJECTS ON THE BASIS OF NEUTROSOPHIC NUMBERS, Proceedings of the 4th International Scientific and Practical Conference SCIENCE, EDUCATION, INNOVATION: TOPICAL ISSUES AND MODERN ASPECTS TALLINN, ESTONIA, 4- 5.10.2021, pp. 374-380.

[7] Mukhamedieva DT, Egamberdiev NA, Zokirov J.Sh., Mathematical support for solving the classification problem using neural network algorithms // Turkish Journal of Computer and Mathematics Education. Vol.12 No.10 (2021).

[8] DTMukhamedieva and NAEgamberdiev, APPROACHES TO SOLVING OPTIMIZATION TASKS BASED ON NATURAL CALCULATION ALGORITHMS, Scientific-technical journal, 3(2) 2020, pp. 58-67.

[9] NAEgamberdiev, OTXolmuminov, KhROchilov, ANALYSIS OF CLASSICAL MODELS OF CLASSIFICATION OF SLOWLY FORMED PROCESSES, International Scientific-Online Conference: SOLUTION OF SOCIAL PROBLEMS IN MANAGEMENT AND ECONOMY", Spain, October 7, 2022, pp. 12-16.

[10] NAEgamberdiev, OTXolmuminov, Khrochilov, CHOOSING AN EFFICIENT ALGORITHM FOR SOLVING THE CLASSIFICATION PROBLEM, International Scientific Online Conference: THEORETICAL ASPECTS IN THE FORMATION OF PEDAGOGICAL SCIENCES, October 10, 2022, pp.154-158.

[11] D. Muhamedieva, N. Egamberdiev, O. Kholmuminov, APPLICATION OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES FOR CREDIT RISK ASSESSMENT, "Science and innovation" international scientific journal. 2022, No. 6. Pages 388-395.

[12] D. Muhamediyeva, N. Egamberdiyev, An application of Gauss neutrosophic numbers in medical diagnosis, International Conference on Information Science and Communications Technologies ICISCT 2021, Tashkent, Uzbekistan, 2021, pp. 1-4.
[13] D. Muhamediyeva, N. Egamberdiyev, A. Bozorov, FORECASTING RISK OF NON-REDUCTION OF HARVEST, Proceedings of the 2nd International Scientific and Practical Conference, SCIENTIFIC COMMUNITY: INTERDISCIPLINARY RESEARCH, Hamburg, Germany, 26-28.01.2021, pp. 694-698.
[14] F. Nuraliev, O. Narzulloev, N. Egamberdiev, S. Tastanova, V Mejdunarodnaya nauchno-prakticheskaya konferencija, RECENT SCIENTIFIC INVESTIGATION, Oslo, Norway, April 26-28, 2022, c. 447-451.
[15] Mukhamedieva D.T., Egamberdiev N.A., Podkhody k resheniyu zadach optimizatsii na osnove algoritmov prirodnyx vychisleniy, Scientific and technical journal of Fergana Polytechnic Institute, 2020, Volume 24, No. 2, c. 75-84.
[16] NAEgamberdiyev, MMKamilov, A.Sh. Hamroyev, Development of an algorithm for determining the system of dimensional fixed basis sets for educational selections, Muhammad al-Khorazmi Avlodali, 1(7) 2019, pp. 45-48.
[17] Khojayev OQ, NAEgamberdiyev, Sh.N. Saidrasulov, Algorithm for choosing an effective method for solving the problem of classification, Information Communications: Networks, Technologies, Solutions. 1(49) 2019. Quarterly Scientific and Technical Journal, pp. 39-43.