

**RESEARCH OF POSSIBILITIES OF APPLICATION INFYREADER  
OCR-SYSTEMS IN PATTERN RECOGNITION PROBLEMS OF  
GRAPHICAL OBJECTS OF SCIENTIFIC AND TECHNICAL  
DOCUMENTATION**

*Annatation. We investigate a method of rapid and qualitative recognition of scientific and technical documentation containing the object information in the form of formulas, digital images, various types of graphs. The method is based on the use of software InftyReader OCR-system developed by laboratory staff Masakazu Suzuki at the University mathematics graduate of Kyushu and the research group on mathematical processing InftyProject data [1-6]. This method makes it possible to recognize monochrome images 400-600 dpi, and quickly and accurately convert various types of graphical information in the text required a custom format.*

*Keywords: OCR, image recognition containing mathematical formulas, TeX, transformation of paper materials in digital format.*

**Introduction.** There are many programs for OCR (Optical Character Recognition), which recognizes optical objects (images of various image formats) and convert them into text files. At the same time the majority of OCR-systems will not recognize the files containing mathematical formulas and other different kinds of graphic information. However, scientific and technical documentation, mathematical texts, most of them contain complex multi-index formula multilevel matrix notation, various images: charts, graphs and the like are used as Cyrillic or Greek, Latin alphabets, are difficult to digitize.

InftyReader is a software for recognizing print media, including complex documents with mathematical formulas and various kinds of graphical objects, which has user capabilities that most OCR systems do not correspond to [2,3].

InftyReader reached more than 200 000 pages of digitizing specialized monographs in mathematics and showed a high probability of detection of complex technical expressions.

InftyReader uses text recognition engines from two leading manufacturers, three different mechanisms for foreign languages in order to maintain high recognition quality with multilingual documents. Therefore, this math software is also useful for documents. Recognition results can be easily output to various file formats. Using this software, you can easily process mathematical expressions in printed materials on your PC, save them in the form of journals, print on paper and in textbooks, create web pages and much more.

### Investigation of OCR-systems InftyReader opportunities

InftyReader recognizes the scanned data from printed documents, converts them into text data. The result can be displayed as a Microsoft Office file Word 2007, LaTeX file, XHTML file with MathML, Human Readable Tex file (Tex without numerical formula) and IML file for InftyEditor, which is convenient to fix and edit recognition results. PDF files created from text programs such as LaTeX, WORD, easier to recognize. The recognition process consists of four phases: page layout analysis, pattern recognition formulas and text structure analysis of mathematical expressions and manual override. Layout analysis is to transfer the image to the internal digital format, clearing it from the "digital garbage" and to identify the constituent elements of the page - tables, pictures, text blocks. At the stage of recognition of text is separated by mathematical formulas. Structural phase reduces to the analysis of mathematical formulas and presenting them in some internal format suitable for display and export to an external file. According developers data, the average error rate (in text and formulas) does not exceed 1-2%, and in modern computers image processing speed of a few ppm. Simultaneous use OCR-mechanisms ExpressReaderPro (Toshiba Corporation) and WinReader (MediaDrive Corporation) to improve character recognition results in text fields. In InftyReader have the opportunity to recognize the table, including the mathematical expression in the cells. The ability to convert PDF-files into LaTeX source and XHTML (MathML), including mathematical expressions.

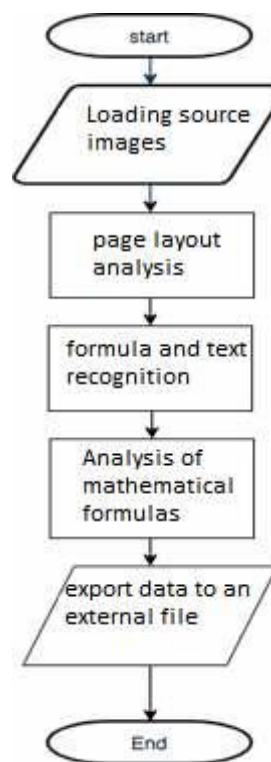


Figure 1

## Application examples

As an example of InftyReader consider the identification and recovery in those two pages (398 and 399) of a scientific article shown in Fig. 2 and 4 graphically tiff format.

398

MASAKAZU SUZUKI

*Proof.* — We have  $q_1\delta_1 = p\delta_0$  by the corollary to Proposition 5. Therefore, it is sufficient to prove (2) for  $k \geq 2$ . Set  $\sigma = i_k$ , and let us consider the surface  $M_\sigma$  obtained by the  $(\sigma - 1)$ -th blowing up in the process to get  $M$  from  $M_1$ . We may say that  $M_\sigma$  is the surface obtained by the blowing down of  $L_{h+1}, L_h, \dots, L_{k+1}$  successively from  $M$ . Let  $\pi_\sigma : M \rightarrow M_\sigma$  be the contraction mapping. As in the previous sections, let us denote the proper images of  $\bar{C}, \bar{C}_k, E_i$  in  $M_\sigma$  by  $\bar{C}^{(\sigma)}, \bar{C}_k^{(\sigma)}, E_i^{(\sigma)}$  respectively. By Theorem 3,  $\bar{C}_{k+1}^{(\sigma)}$  intersects transversely  $E_\sigma^{(\sigma)}$  at the same point  $Q = \pi_\sigma(L_{k+1} \cup \dots \cup L_{h+1})$  as  $\bar{C}^{(\sigma)}$ . Hence, the functions  $f$  and  $g_{k+1}$  on  $M_\sigma$  have the same indetermination point  $Q \in E_\sigma^{(\sigma)}$ . Let

$$P_f^{(\sigma)} = \sum_{i=0}^{\sigma} \nu_i E_i^{(\sigma)}, \quad P_{g_{k+1}}^{(\sigma)} = \sum_{i=0}^{\sigma} \bar{\nu}_i E_i^{(\sigma)}$$

be the pole divisor of  $f$  and  $g_{k+1}$  on  $M_\sigma$  respectively. Let  $\bar{\delta}_0, \bar{\delta}_1, \dots, \bar{\delta}_k$  be the order of the pole of  $g_{k+1}$  on  $E_{j_0} (= E_0), E_{j_1} (= E_1), \dots, E_{j_k}$ . We have  $\bar{\delta}_0 = \bar{\nu}_{j_0}, \bar{\delta}_1 = \bar{\nu}_{j_1}, \dots, \bar{\delta}_k = \bar{\nu}_{j_k}$ . The coefficients  $\nu_i, \bar{\nu}_i$  ( $i = 0, 1, \dots, \sigma$ ) are the solutions of the following equations:

$$\sum_{j=0}^{\sigma} (E_i^{(\sigma)} \cdot E_j^{(\sigma)}) \nu_j = \begin{cases} 0 & (i \neq \sigma) \\ d_{k+1} & (i = \sigma), \end{cases}$$

$$\sum_{j=0}^{\sigma} (E_i^{(\sigma)} \cdot E_j^{(\sigma)}) \bar{\nu}_j = \begin{cases} 0 & (i \neq \sigma) \\ 1 & (i = \sigma). \end{cases}$$

Hence, by Lemma 4, we have  $\nu_i = d_{k+1} \bar{\nu}_i$  for all  $i = 0, 1, \dots, \sigma$ . In particular,

$$\delta_i = \bar{\delta}_i \cdot d_{k+1}, \quad (i = 0, 1, \dots, k).$$

Therefore, in order to prove (2), it is sufficient to prove

$$(3) \quad q_k \bar{\delta}_k \in \mathbb{N} \bar{\delta}_0 + \mathbb{N} \bar{\delta}_1 + \dots + \mathbb{N} \bar{\delta}_{k-1}.$$

By Theorem 3,  $\bar{C}_k^{(\sigma)}$  intersects  $E_{j_k}^{(\sigma)}$  transversely and does not intersect other components  $E_i^{(\sigma)}$  ( $i \neq j_k$ ). We have

$$\begin{aligned} \bar{\delta}_k &= (P_{g_{k+1}}^{(\sigma)} \cdot \bar{C}_k^{(\sigma)}) \\ &= (\bar{C}_{k+1}^{(\sigma)} \cdot \bar{C}_k^{(\sigma)}) \\ &= (\bar{C}_{k+1}^{(\sigma)} \cdot P_{g_k}^{(\sigma)}). \end{aligned}$$

Figure 2 - The original page image 398 in tiff format

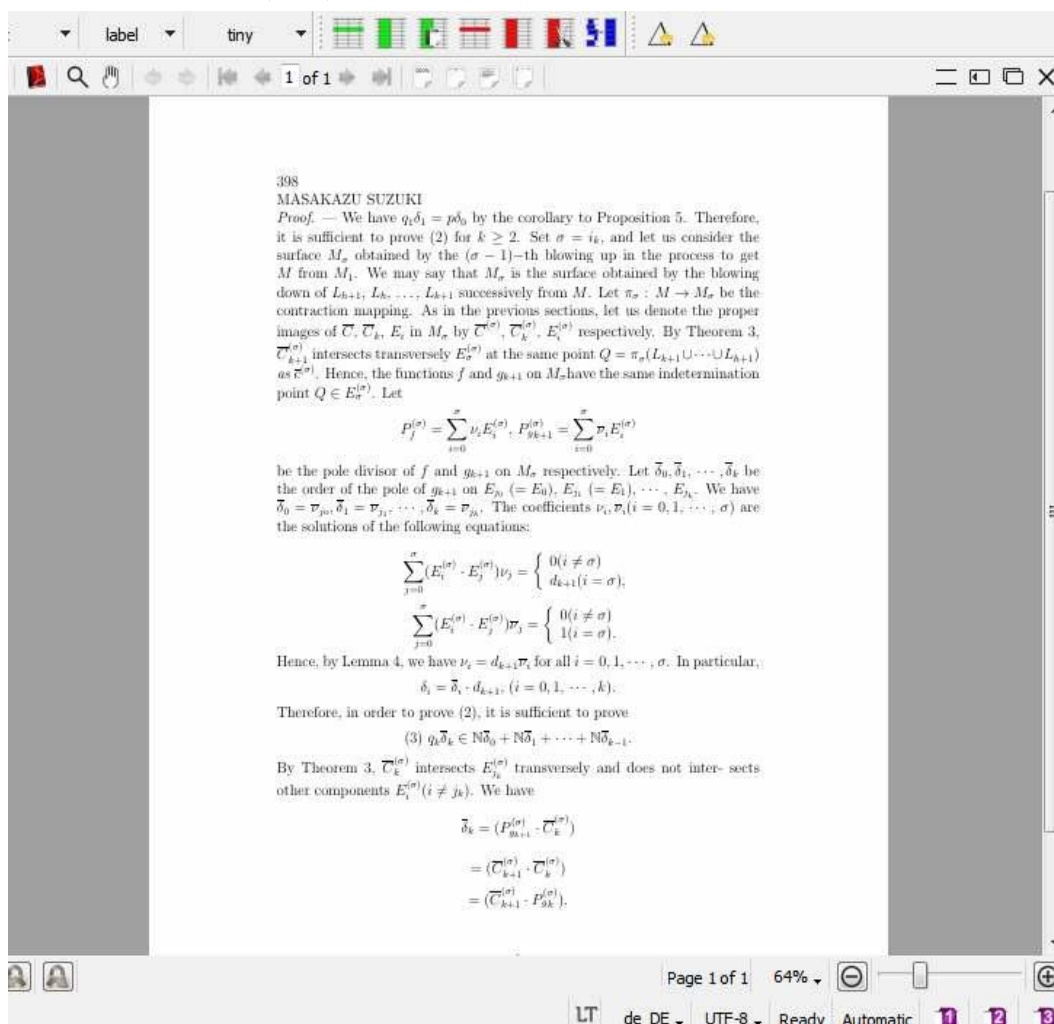


Figure 3 - The compiled file in TeX TeXStudio, recognized and restored in InftyReader

This implies that  $g_k$  has the pole of order  $\bar{\delta}_k$  on  $E_\sigma^{(\sigma)}$ . On the other hand, by Lemma 1,  $g_{k+1}$  has the pole of order  $q_k \bar{\delta}_k$  on  $E_\sigma^{(\sigma)}$ . Hence,  $E_\sigma^{(\sigma)}$  is neither the zero nor the pole of  $\Phi = \frac{g_{k+1}}{g_k}$ . Further,  $\Phi$  is holomorphic in a neighborhood of  $Q$  and  $\Phi(Q) = 0$ . Therefore,  $\Phi$  is not constant on  $E_\sigma^{(\sigma)}$ .

Now, set  $\psi = g_{k+1} - g_k^{q_k}$ . Then,

$$\frac{\psi}{g_k^{q_k}} = \Phi - 1$$

is also a non-constant function on  $E_\sigma^{(\sigma)}$ . Therefore,  $\psi$  has also the pole of order  $q_k \bar{\delta}_k$  on  $E_\sigma^{(\sigma)}$ . On the other hand, since

$$\deg_y(\psi) < n_{k+1} = n_k q_k, \quad n_k = \deg_y(g_k),$$

by the division of  $\psi$  by  $g_k^{q_k-1}$ , we get

$$\psi = c_1 g_k^{q_k-1} + \psi_1$$

with  $\deg_y(c_1) < n_k$ ,  $\deg_y(\psi_1) < n_k(q_k - 1)$ . Dividing  $\psi_{i-1}$  by  $g_k^{q_k-i}$  successively for  $i = 2, \dots, q_k - 1$ , we get

$$\psi_{i-1} = c_i g_k^{q_k-i} + \psi_i,$$

where  $\deg_y(c_i) < n_k$ ,  $\deg_y(\psi_i) < n_k(q_k - i)$ . Thus, setting  $c_{q_k} = \psi_{q_k-1}$ , we get

$$\psi = \sum_{i=1}^{q_k} c_i g_k^{q_k-i}.$$

Here, we have

$$\deg_y(c_i) < n_k = n_{k-1} q_{k-1}, \quad n_{k-1} = \deg_y(g_{k-1}).$$

In the same way, dividing  $c_i$  and its rests by  $g_{k-1}^{q_{k-1}-1}$ ,  $g_{k-1}^{q_{k-1}-2}$ ,  $\dots$ ,  $g_{k-1}$  successively, we get

$$c_i = \sum_{j=1}^{q_{k-1}} c_{ij} g_{k-1}^{q_{k-1}-j}$$

with  $\deg_y(c_{ij}) < n_{k-1}$ . Thus, we have

$$\psi = \sum_{i=1}^{q_k} \sum_{j=1}^{q_{k-1}} c_{ij} g_k^{q_k-i} g_{k-1}^{q_{k-1}-j}.$$

Figure 4 - The original page image 399 in tiff format

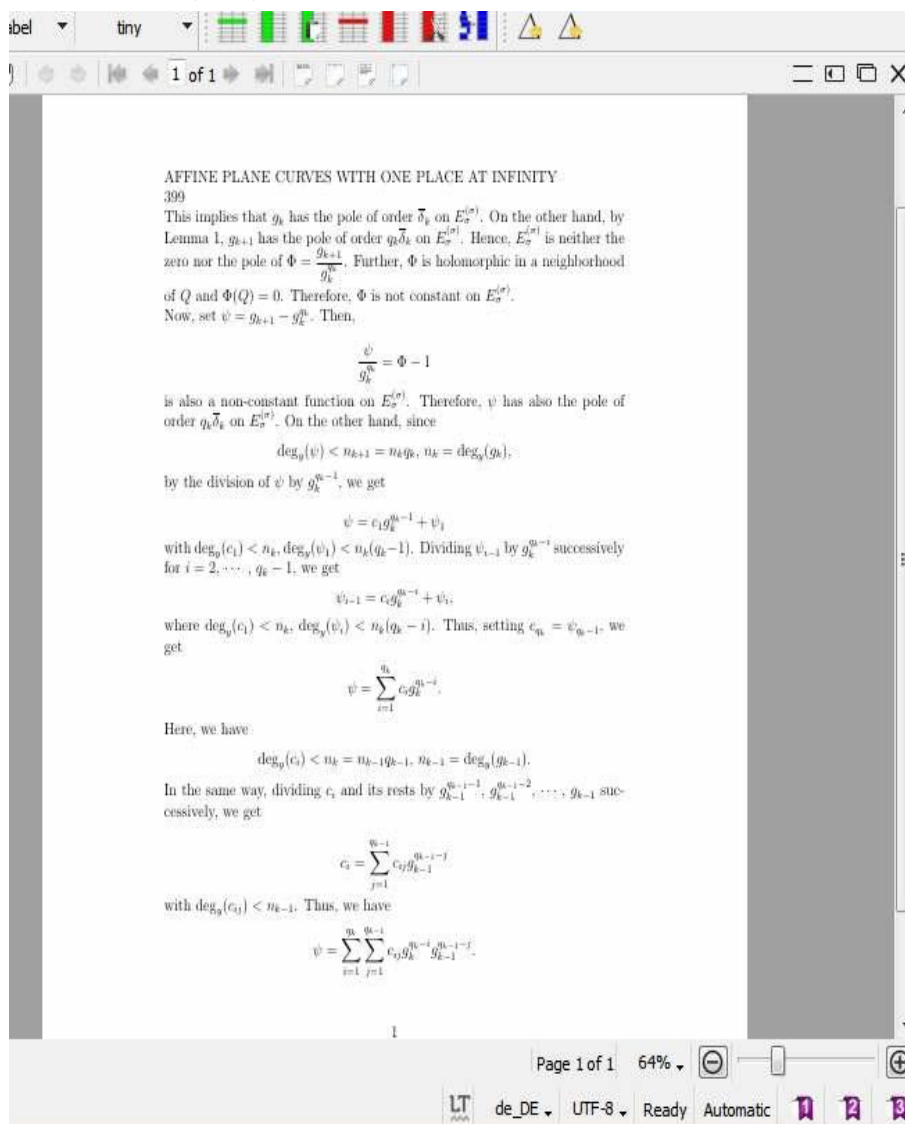


Figure 5 - The compiled file in TeX TeXStudio, recognized in InfyReader

### Image Requirements

To bring images to the standard InfyReader, you can use software products for working with images such as Adobe Photoshop.

InfyReader standard:

- 1) the image should be clean and not have noise.
- 2) The recommended density of dots per inch (DPI) image should be 400-600, but was found by experimentation that the threshold density dpi for the qualitative detection 300 begins.

1) Recommendations to PDF

- PDF on the Internet are often scanned as 200DPI or converted to program screen style. To try InfyReader images with a resolution of around 400-600 dpi for reliable quality.

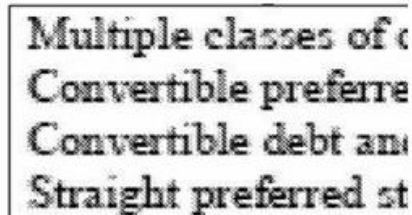


Figure 6 - Low DPI image

- Color or light and shadow images provide an unstable process. These images can cause a hang or other problems for the program. Newer versions after InftyReaderVer.2.6.3 can process such images to some extent.

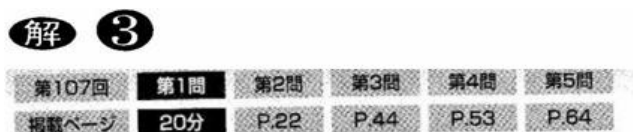


Figure 7 - is not the right color scheme

- Double-check the settings, and the original image.

If the result is clearly inconsistent, check the language settings in the Input.

- Check the original image. Inverted image can not be processed.

2) The types of images that are not appropriate for the recognition

- handwritten

- reverse black and white, gray tone.

- scratches, blurred, sliding, two letters stuck together.

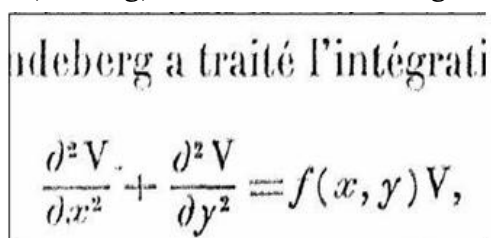


Figure 8 - Poor quality text

- Text typed with a typewriter has a lower recognition result.

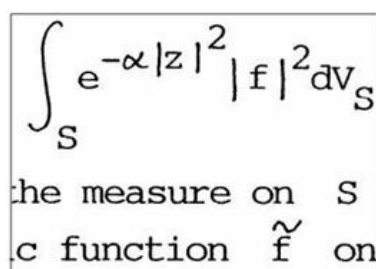


Figure 9 - The text was typed on a typewriter

If incorrect recognition is more than 10 percent of the page, it is recommended rescan, these cases require new settings and rescan.

Try black and white monochrome mode with a resolution of 400 dpi.

Common problems when scanning books:

- Distorted letters around the seam in the middle.
- Shadow in the background.
- Difference in book page heights

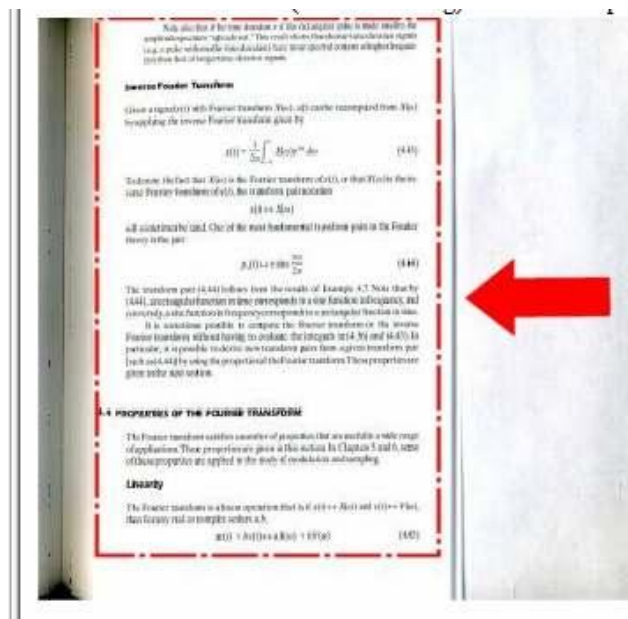


Figure 10 - The area of the page to the correct recognition

Solution: it is necessary to open the book and press it to the scanning glass in order to smooth the stitching, then repeat the scanning. If possible, divide the thick book on each page, trimming the pages along the stitching or trimming the book's spine. Split pages can be perfectly processed through the automatic document feeder.

- Shade the center or outside of the page.
- The presence of extraneous images on the page.
- Two columns are recognized from left to right.
- A 2-page spread should be divided into 2 different pages. Cut out each page from the image and paste into new pages.

Solution: Using the image software, cut out the required area and insert a new file. Use the “eErase” tool.

- Wrong direction.

Solution: use image software to rotate until the top of the document rises.



For the correct recognition result, first check the scanned image. It is recommended to erase blemishes, unreadable letters, patterns, header and footer, or any additional information using image software and rescan for distortion.

- Incorrect recognition factors, such as shadow, blur, or unreadable information, should be removed prior to digitization.

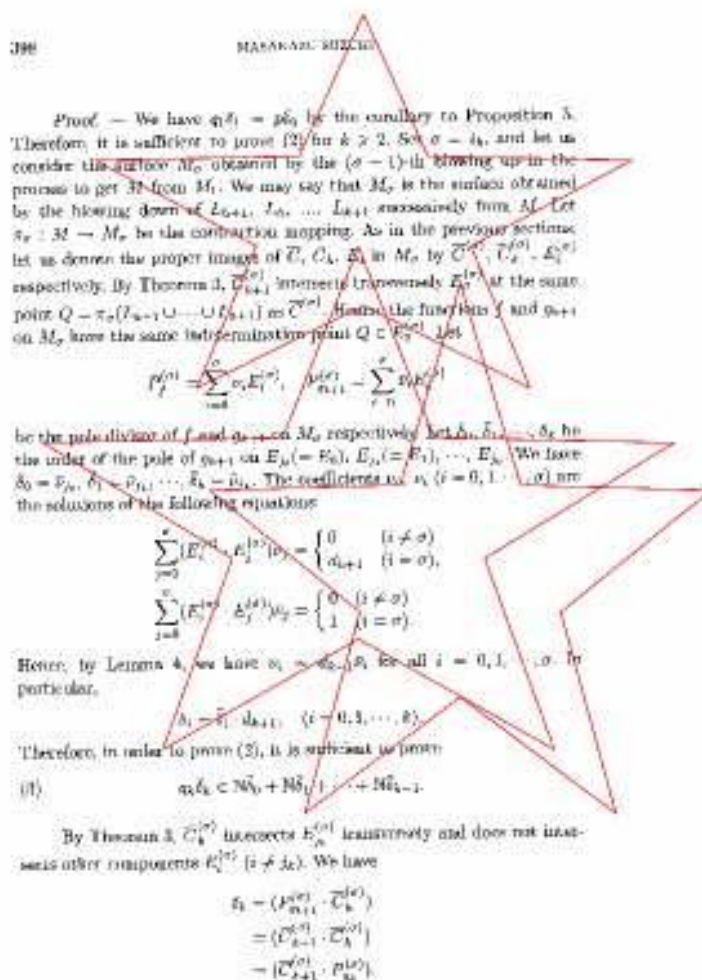


Figure 11 - the presence of extraneous image

### Conclutions

The main results of the study:

- 1) The algorithm used in InftyReader allows you to quickly and accurately identify formulas contained in the image.
- 2) The recognition process consists of four phases: analysis of the page layout, recognition of formulas and text, structural analysis of mathematical expressions and manual correction.
- 3) Testing of the program was carried out, which confirmed its operability (no more than 10% of erroneous recognition).

4) There were identified requirements and image format for high-quality recognition.

The practical value of the study: In the course of the work, the program InftyReader was analyzed on the question of the possibility of working with images containing formulas and its performance was confirmed. Designed guidelines for the format and image quality.

Future prospects: continue research to improve the quality of identification of documentation containing mathematical formulas. It seems appropriate to use this program in the recognition of paper articles and their subsequent testing for uniqueness (antiplagiat).

#### REFERENCES

1. Research Project on Mathematical Information Processing Mathematical Document Recognition and Analysis, Accessibility of Scientific Documents URL: <http://Inftyproject.org>

2. sAccessNet Nonprofit organization - Science Accessibility Net A support network of the accessibility of scientific information for visually impaired people URL: <https://www.sciaccess.net/en/>

3. Simone Marinai, Hiromichi Fujisawa (Eds.), Machine Learning in Document Analysis and Recognition: Studies in Computational Intelligence ISSN 1860-949X // Kokubunji-shi, Tokyo 2008. pp. 1-21

4. M. Suzuki, F. Tamari, R. Fukuda, S. Uchida and T. Kanahori. INFTY – An Integrated OCR System for Mathematical Documents // Proc. DocEng, 2003.

5. Suzuki M., Tamari F., Fukuda R., Uchida S., Kanahori T. / Infty -an integrated OCR system documents. In Proceedings of ACM Symposium on Document Engineering 2003, pp. 95-104 (2003).

6. InftyReader user manual. Inftyhelp.exe ver.3.1 (C). Copyright 2000-2018: Masakazu Suzuki (Kyushu University) / Non Profit Organization Science Accessibility Net.