The analysis of the graphical representation of mathematical formulas for extending descriptions of physical effects

D S Tolokin¹, D M Korobkin¹, S A Fomenkov¹ and S G Kolesnikov¹

¹ Volgograd State Technical University, 28, Lenina ave., Volgograd, 400005, Russia
E-mail: tolo1.dima@gmail.com

Abstract. The aim of the study is the software development of a methodology for analyzing graphical representations of mathematical formulas to expand the descriptions of scientific and technical effects. To achieve this goal, an analysis of existing systems for recognizing graphic representations of mathematical formulas was carried out, an algorithm for editing a PE database with integrated physical formulas and an algorithm for recognizing formulas and explanations were developed. Based on the developed methodology, an automated system is implemented in software. Thus, a technique for analyzing graphical representations of mathematical formulas has been developed to expand the descriptions of scientific and technical effects and an automated system created on its basis.

1. Introduction
In the database of physical effects (PE) [1], there are more than 1400 descriptions. These descriptions often lack formulas describing the physical processes of the effect.

One of the most updated sources of information on physical topics is the patent array, in particular, RosPatent, as well as journals of physical topics, abstracts, and dissertation texts containing descriptions of mathematical models.

Let us show an example of a graphical representation of the mathematical formula for PE in the patent [2].

A fragment of the patent text containing a formula, description of the formula, and descriptions of physical quantities from the formula, respectively:

“Calibrated magnetic measurements, are calculated by the following formula:

\[ \mathbf{m}_{\text{mag}} = (\mathbf{D}_{\text{mag}})^{-1} (\mathbf{m}_{\text{soft}} - \mathbf{b}_{\text{mag}}) \]

where \( \mathbf{D}_{\text{mag}} \) is the magnetic displacement state matrix corresponding to nine displacements from the soft iron; \( \mathbf{m}_{\text{soft}} \) is the magnetic displacement state vector corresponding to displacements from solid iron; and \( \mathbf{b}_{\text{mag}} \) represents three-dimensional magnetic measurements …”

The html structure:

<p>… According to one example, calibrated magnetic measurements, <img src = "https://img.findpatent.ru/img_data/1345/13457063.jpg" width = "56" height = "40" alt = "" /> , calculated by the following formula: </p>
where $D_{mag}$ is the magnetic displacement state matrix corresponding to nine displacements from soft iron; $\mathbf{b}_{mag}^B$ represents a state vector magnetic displacement corresponding to displacements from solid iron; and \( \tilde{\mathbf{m}}^B \) is three-dimensional magnetic measurements...

It is required to extract the formula in LaTeX format, as well as explanations to it in the form of definitions of physical quantities (can be presented both in text form and in graphical representation):

\[
\text{Calibrated magnetic measurements, } \mathbf{m}_{\text{cal}}^B, \text{ are calculated using the following formula:}
\begin{align*}
\mathbf{m}_{\text{cal}}^B &= (\mathbf{D}_{mag})^{-1} (\tilde{\mathbf{m}}^B) - \mathbf{b}_{mag}^B
\end{align*}
\]

where $\mathbf{D}_{mag}$ is the magnetic displacement state matrix corresponding to nine displacements from the soft iron; $\mathbf{b}_{mag}^B$ is the magnetic displacement state vector corresponding to displacements from solid iron; and \( \tilde{\mathbf{m}}^B \) represents 3D magnetic measurements...

2. Materials and methods

Thus, the task of automating the expansion of the description of the physical effect by replenishing the topic “Mathematical model of PE” is urgent.

The purpose of the work is to implement software development of a method for analyzing graphical representations of mathematical formulas to expand the descriptions of physical effects.

The analysis of existing systems for the recognition of graphic representations [3] of mathematical formulas was carried out (Table 1).

| Software                        | a | b | c | d | e |
|---------------------------------|---|---|---|---|---|
| Tesseract OCR                   | - | 0 | 5 | 1 | - |
| InftyReader                     | + | 2 | 6 | 5 | + |
| Mathematical Expression Recognition | + | 2 | 1 | 1 | + |
| MathType                        | + | 2 | 2 | 2 | + |

where a - the accuracy of recognition of mathematical formulas consists of the following estimates: “+” – “the accuracy is sufficient for applied use”, “-” – “the accuracy is insufficient for applied use”; b - API convenience consists of the following ratings: 2 – “API is convenient”, 1 – “API is inconvenient”, 0 – “API is absent”; c - the number of input image formats; d - the number of formats of recognized formulas; e - opportunity to output recognition results in LaTeX format "contains estimates: “+” – “Output in LaTeX format is possible” and “-” – “Output in LaTeX format is not possible”.

Based on the comparison of recognition systems, it turned out that the most convenient for solving the problem of extracting data from physical formulas from patent texts is the MathType [4] system in connection with a convenient API in combination with input image formats and output formats of recognized formulas that are practical in the authors' opinion.

As a result, an algorithm for editing a PE database with integrated physical formulas was developed (Figure 1) and an algorithm for recognizing formulas and explanations to them (Figure 2).
Figure 1. Algorithm for editing PE DB with integrated physical formulas.

Figure 2. Recognition algorithm for formulas and explanations to them.
Based on the developed methodology, the software was developed, which provides the following functions:

- recognition of graphical representations of physical formulas from patent texts in LaTeX format;
- extraction of explanations to the recognized formula from patent texts (description of physical quantities used in the formula);
- entering physical formulas into descriptions of physical effects in the database in LaTeX format.

Software is implemented in the languages Python (version 3.6) and Javascript. To recognize formulas from graphic images, the Mathpix OCR API is used. The BeautifulSoup library is used to find an explanation of the formula recognized in the patent. The request library is used to query the patent web pages.

The architecture is shown in Figure 3.

---

### 3. Results and discussion

The developed software displays (Figures 4,5):

- information on the results of extracting formulas from patent texts;
- the output of extracted formulas with explanations of physical quantities;
- descriptions of physical effects with the topics “Mathematical model” filled in with the extracted formulas.
4. Conclusion

The theoretical value of this work lies in the developed methodology for analyzing graphical representations of mathematical formulas for expanding the descriptions of scientific and technical effects and the software created on its basis.

5. Acknowledgments

The reported study was funded by RFBR (research projects 18-07-01086), RFBR and Administration of the Volgograd region (projects 19-47-340007, 19-41-340016).

References

[1] Vayngolts N A, Korobkin D M, Fomankov S A, Kolesnikov S G and Kravets 2019 Method of Chemical Formulas Comparison *MCCSIS* 151-158

[2] Sotak M, Dunik I and Kana Z 2018 Coarse determination for the hybrid navigation solution based on magnetic-calibrated measurements *Patent RU* 2673504

[3] Vereschak G, Korobkin D, Fomankov S and Kolesnikov S 2019 Identification of Descriptions of Scientific-Technical Effects in Patent Documents *I* 196-204.

[4] MathPix https://mathpix.com/ocr

[5] Beautiful Soup https://pypi.org/project/beautifulsoup4