SCIENTIFIC PAPERS
OF THE UNIVERSITY OF PARDUBICE
Series A
Faculty of Chemical Technology
19 (2013)

INFLUENCE OF TYPES TONE VALUE ON CHARACTER FORMATION AND ACCURACY OF OPTICAL CHARACTER RECOGNITION

Igor KARLOVIĆ[1], Ivana TOMIĆ and Ivana JURIČ
Faculty of Technical Sciences,
The University of Novi Sad, RS–21000 Novi Sad

Received September 30, 2013

The characters, primary carriers of textual information, fulfil a significant role in print quality. Character formation is important not just for the human reader, but also to a machine based reading and optical character recognition (abbreviation OCR). In this paper we investigated the influence of the characters optical density on the accuracy of commercial and open source optical recognition performance. The intention was to investigate the possibility of toner saving and the thresholds for typeface sizes that will enable good readability. The test chart with three type sizes (24pt, 12pt and 6 pt) and two typefaces (Times New Roman as a serif and Arial as sans serif type) were printed by varying the optical density through tone values (from 10 % to 100 % by a 10 % increment). Test form was printed on 80 g m^{-2} uncoated office paper using Riso EZ 570E digital screen printing system. The prints were scanned using a calibrated HP Scanjet G3010. Scanning resolution was set to 300 spi, as a recommended value for this kind of documents.

[1] To whom correspondence should be addressed.

Character area and perimeter were characterized using image analysis method (open source ImageJ software). For optical character recognition we used several OCR software: OmniPage, Abbyy Fine Reader, ReadIris Pro, Presto Pro, CuneiForm. The results obtained from the study indicate that text tone value has an important role in OCR accuracy for both type faces.

Introduction

Optical character recognition (OCR) software is nowadays commonly used to digitalize printed content for archiving and also for electronic distribution through Internet and other digital channels. The fast books digitalization for large online libraries and archives require not just a good technical equipment, but technical knowledge and preferably quality printed and conserved samples. The OCR methods were developed with different technological solutions. Mori *et al*. [1] in their review paper described the history of the OCR technologies development and classified them into three generations. Fujisawa in his extensive overview [2] described the development path of the technology with some insight into possible future applications of the OCR technology. Within the recognition phase there are four main steps: Line Find, Character Segmentation, Feature Extraction and Classification [3]. Character Recognition techniques can be classified according to two criteria [4]: the way pre-processing is performed on the data and the type of the decision algorithm (where algorithm strategies heavily depend on the nature of the data to be recognized). Pre-processing techniques include three main categories: the use of global transforms (correlation, Fourier descriptors, etc.), local comparison (local densities, intersections with straight lines, variable masks, etc.) and geometrical or topological characteristics (strokes, loops, openings, diacritical marks, skeleton, etc.).

Influencing factors which determine the algorithms' quality are the typefaces and quality of printed characters. The best fonts for OCR are designed for machine reading and have uniform character spacing. It is also important that each character is designed to be unique, so that it cannot be confused with any other. Other important factors which can influence OCR performance are ink absorption and background noise due to material structure or process residue.

Materials and Methods

Test chart with randomly generated text was made with three type sizes 24pt, 12pt and 6pt. Typefaces were also varied — Times New Roman was used as serif and Arial as sans serif font. The text contained all specific Latin Serbian characters and numbers and was varied through optical density with different tone values (TV)

from 10 % to 100 %, by 10 % increment. Test chart was printed on a 80 g m^{-2} uncoated office paper using Riso EZ 570E digital screen printing system. This kind of a digital printer creates a master by means of tiny heat spots on a thermal plate burning voids (corresponding to image areas) in a master sheet. This master is then wrapped around a drum and ink is forced through the voids in the master. The paper runs flat through the machine while the drum rotates at high speed to create each image on the paper. Printing resolution was 300×600 dpi (the default value).

After printing, the printed sheets were scanned with HP Scanjet G3010 CCD based flatbed scanner, which was calibrated using the VueScan software. The scanning resolution was set to 300 spi, as a recommended value for this kind of documents [5]. After scanning, digitalized samples were processed through several OCR software: OmniPage, Abbyy Fine Reader, ReadIris Pro, Presto Pro, CuneiForm.

The accuracy of the recognition was assessed by reading the recognized text and hand correcting the mistakes. Mistakes were noted and counted on the level of characters. The most accurate way of determining the quality is hand correction. The text, which requires too much hand correcting, was not analyzed due to poor character recognition, and thus the sample was considered useless for OCR. The analysis of character formation through the area and perimeter was performed for the f character using open source ImageJ software.

Analysis of OCR Accuracy

There are few suggestions for OCR accuracy measurement method. Tanner [6] describes the method for evaluation of the accuracy and check-up of the feasibility of the OCR system, similarly like in [7]. For the purpose of this paper, we decided to base an accuracy on the effort needed to correct the OCR-ed document. The number of the correcting operations in the text on the character level is counted as a mistake (character insertion, deleting, replacing). The number of the mistakes compared to the total number of characters in the text is defined in mistake percentage. The accuracy of the particular software is determined when the mistake percentage is subtracted from 100 %.

We followed the guidelines described by Holley in Ref. [8] where a scale for good, average and bad OCR is defined as follows:

- Good OCR accuracy = 98-99% accurate recognition (1-2% OCR inaccurate recognition)
- Average OCR accuracy = 90-98% accurate recognition (2-10% OCR inaccurate recognition)
- Bad OCR accuracy = below 90% accurate recognition (more than 10% in-accurate recognition)

Results

The results of the OCR software testing is presented in terms of accuracy of recognition. Situations where the OCR software was unable to make a recognition or there was too many mistakes were omitted. The results for the 24pt size Times New Roman typeface analyzed by all OCR software are presented in Fig. 1.

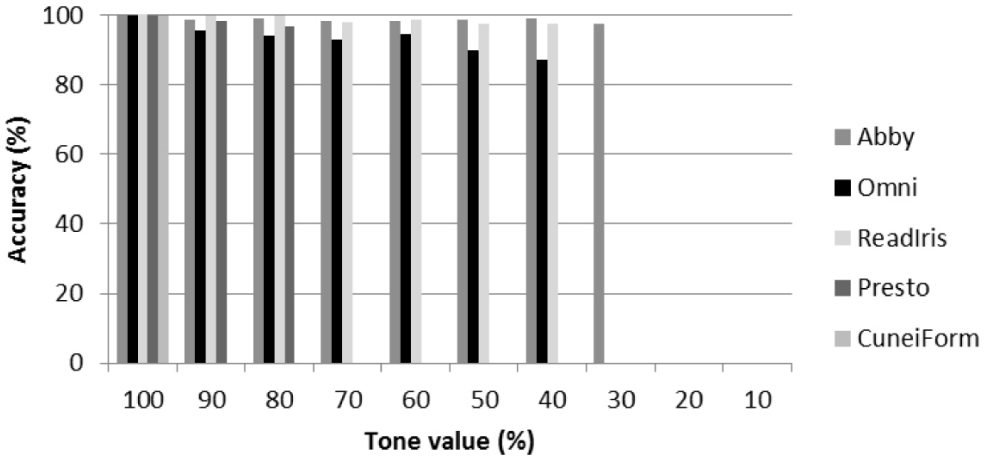

Fig. 1 The results for OCR accuracy with different tone values (Times New Roman 24pt)

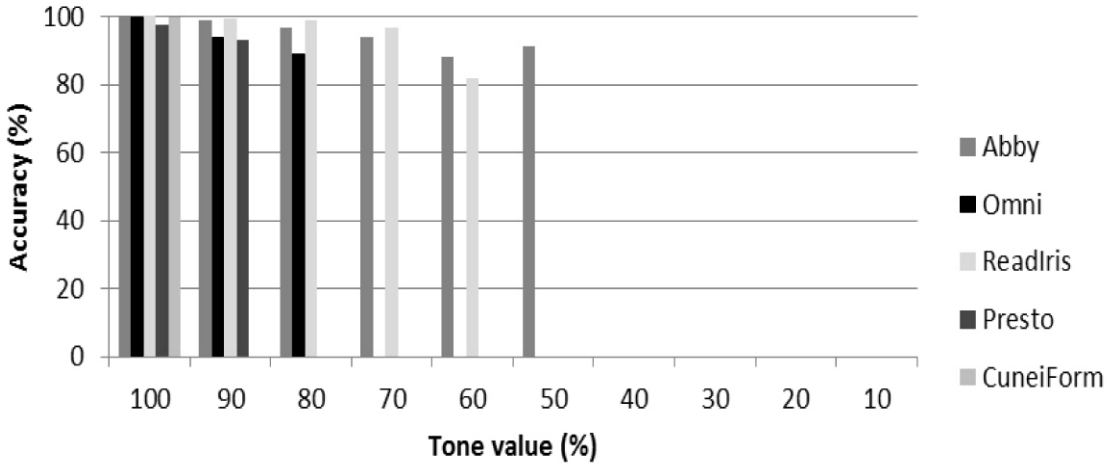

Fig. 2 The results for OCR accuracy with different tone values (Times New Roman 12pt)

As we can observe from Fig. 1, all software had maximum accuracy at 100% tone value (at the largest optical density). On lower TV values the best results are performed by Abby Fine Reader and ReadIris Pro. They maintain good OCR accuracy up to 60 % TV (Abby 98.11 % and ReadIris Pro 98.03 %). Abby Fine Reader maintaines this degree of accuracy until the TV of 40 % while the rest of software (OmniPage Pro, Presto Pro and CuneiForm) falls below the good OCR threshold on higher TVs 90 %-80 % tone values (averagely accurate) and after 70-60 % TV accuracy, based on suggested scale, can be categorized as bad. The 30

% TV at this type size is the threshold where the text in unreadable by any software. Figure 2 contains results of OCR accuracy for Times New Roman, size 12 pt.

The results presented in Fig. 2 indicate that again all OCR solutionsshow good recognition accuracy at 100 % tone value. At 90 % TV, only Abby (with 98.8 %) and ReadIris Pro (99.2 %) remain in the good accuracy range, while OmniPage with 89.12 % and Presto (93 %) fall in the average scale grade. The CuneiForm had a large number of mistakes already at this tone value. With declining tone values only the aforementioned Abby and ReadIris Pro maintained an average value accuracy till the 70 % tone values. After that, text was recognized at 60 and 50 % tone value but with poor accuracy. The best accuracy of 99.9 and 99.65 was obtained for 24pt and 12pt text at 100% tone value (highest optical density). The accuracy stays stable for the 24pt up to the 30 % tone value of characters while for the 12pt the recognition threshold is around 50 % of tone coverage. At the smallest typeface size of 6pt only at 100 % tone value, the software Abby Fine Reader and Omnipage were able to recognize the text with average and poor quality.

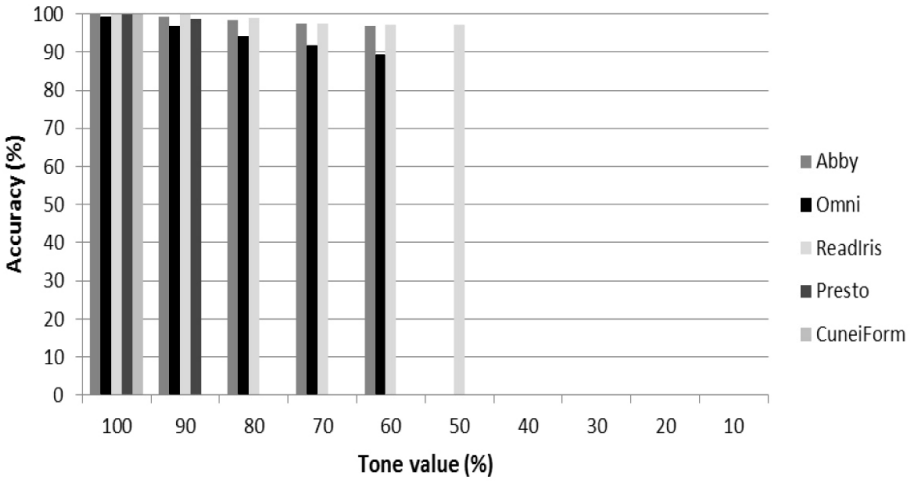

Fig. 3 The results for OCR accuracy with different tone values (Arial 24pt)

The next test of the OCR software was done using a standard sans serif typeface Arial, also in three font sizes of 24pt, 12pt and 6pt. The results of the OCR accuracy are presented in Figs 3 and 4.

As we can see from Fig. 3, the recognition accuracy of all OCR software for sans serif Arial font (24pt type size) was 100 at 100 % tone value, while with degrading tone values the accuracy also degrades. At 90 % tone value Abby Fine Reader had a good quality read out of 99.4 %, OmniPage an average of 97, ReadIris Pro a good quality of 100 % while Presto and CuneiForm both had good recognition accuracy of 99.4 %. With lower density and tone values the accuracy can be graded as average for Abby and Omni (below 98 %), while ReadIris still proved good OCR results with 99% accuracy and Presto and CuneiForm were

unable to make a quality recognition. At 70 % and 60 % tone values almost all OCR software fall into bad accuracy rating with values lower than 98 %. Again a margin around 60 % and 50 % tone value proved to be a final threshold line where the software were able more or less to recognize the characters. For the 12pt size Arial the results are presented in Fig. 4.

The results of the recognition accuracy of 12pt Arial showed that all software had 100% accuracy at the highest tone and optical density values. At lower tone values all software except OmniPage pro and CuneiForm remain in the good accuracy rank with values above 98 %. At 80 and 70 % tone value Abby Fine Reader and Presto remain in the higher values of average accuracy while OmniPage and CuneiForm showed much weaker text recognition accuracy. At even lower tone values the accuracy is additionally reduced. Threshold for each software is at around 50 % and specifically 40 % TV for Abby Fine Reader. At 6pt Arial the results are similar to the Times New Roman values where only Abby Fine Reader and Omnipage Pro had successfully recognized characters at 100 % and 90 % tone values with average and bad accuracy (below 98 %) and all other OCR software were unable to make a successful recognition.

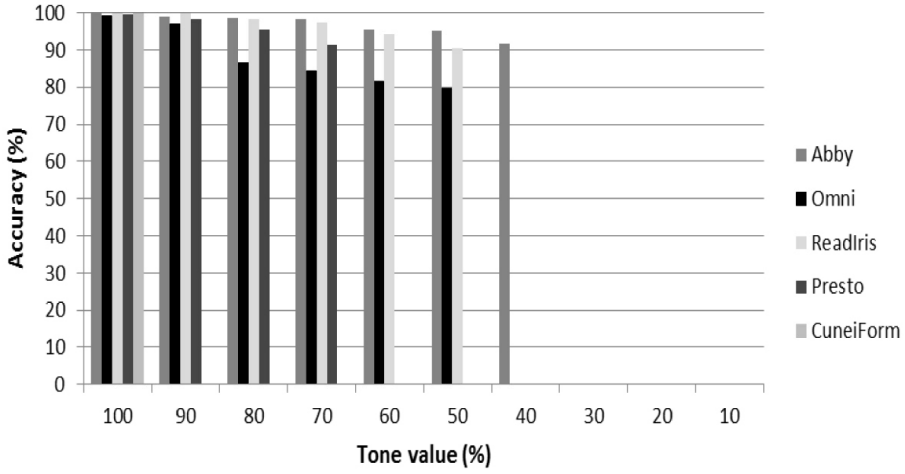

Fig. 4 The results for OCR accuracy with different tone values (Arial 12pt)

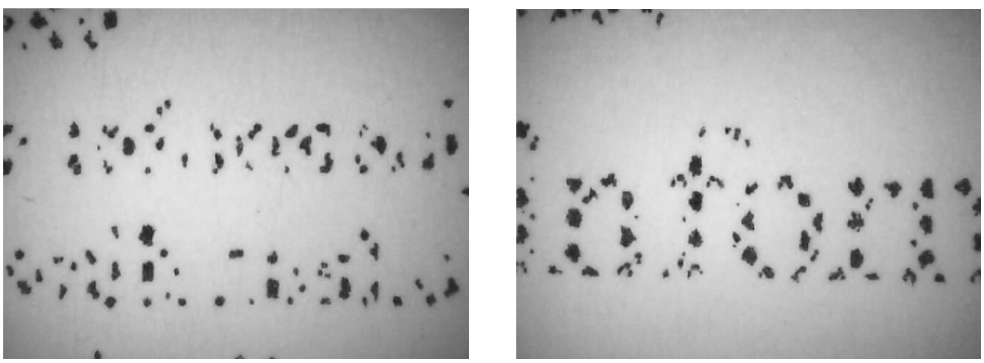

Fig. 5 Times New Roman 6pt and 12pt size at 30% tone value

To further explore the deformation of the characters on different tone values, images were captured with Sibress PIT USB camera. The images of the Times New Roman typeface, on 30% tone values of 6pt and 12pt typeface are presented in Fig. 5, and Arial 6pt,12pt and 24pt on 10 % tone values are presented in Fig. 6.

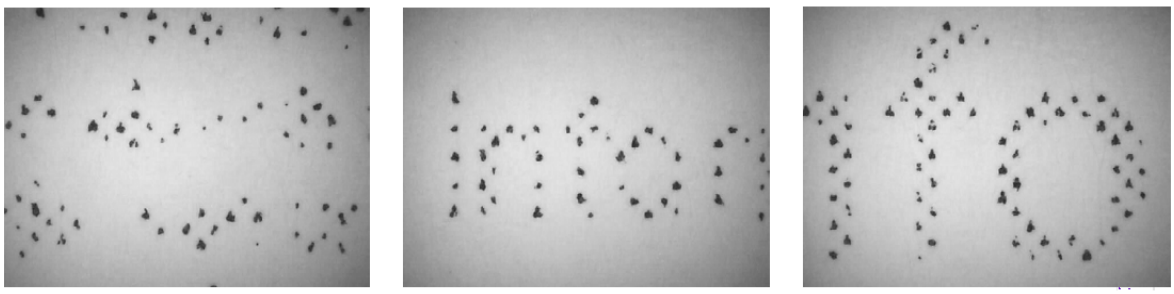

Fig. 6 Images of 6, 12 and 24pt Arial at 10 % tone value

Area defines the quantity of toner which forms a character. To establish a connection between optical density and area, area of the f character was measured with ImageJ software. In the case the letter was fragmented, we took into account all its parts by summarizing its areas. The f character was chosen as a closed but fairly complex character which would help evaluate the accuracy of the investigated softwares. The presumption was that lower toner density on the halftoned types will have an impact on the accuracy of the tested OCR software. The results of the character area measurements are presented in Figs 7 and 8.

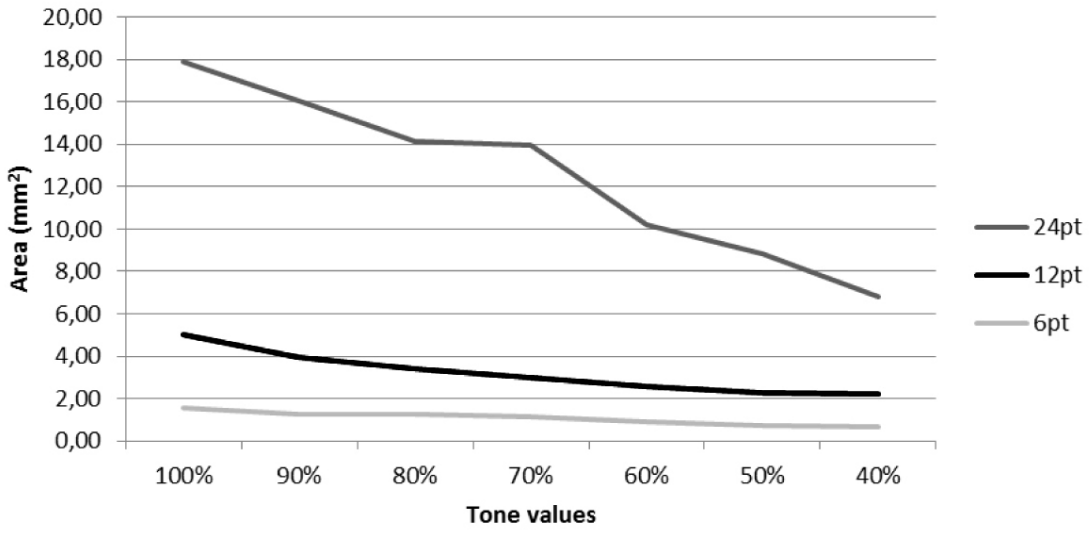

Fig. 7 Area of the letter f printed with Times New Roman on different tone values

The results for both typefaces show a fairly linear declining of area values with declining of tone values applied to the halftoned typefaces. To establish a connection between area of the character and OCR accuracy, we plotted the values

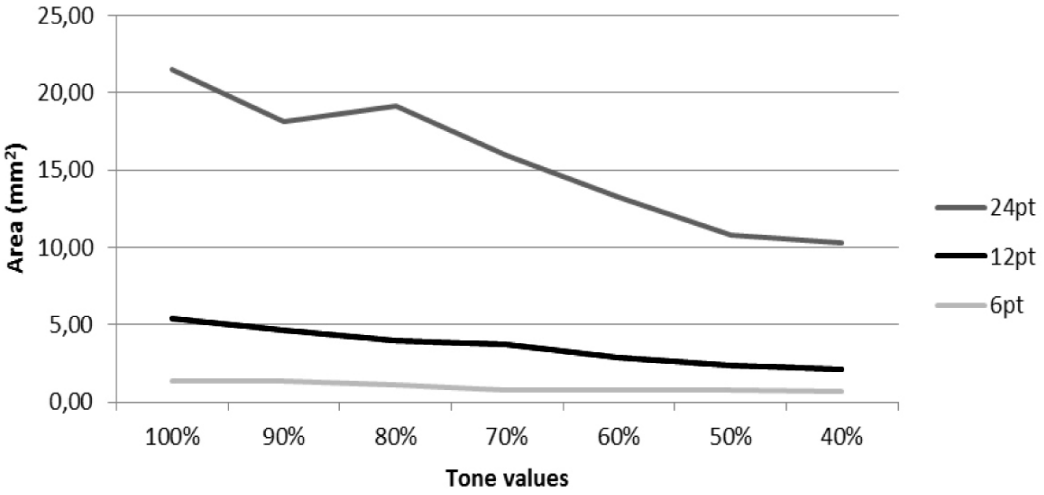

Fig. 8 Area of the letter f printed with Arial on different tone values

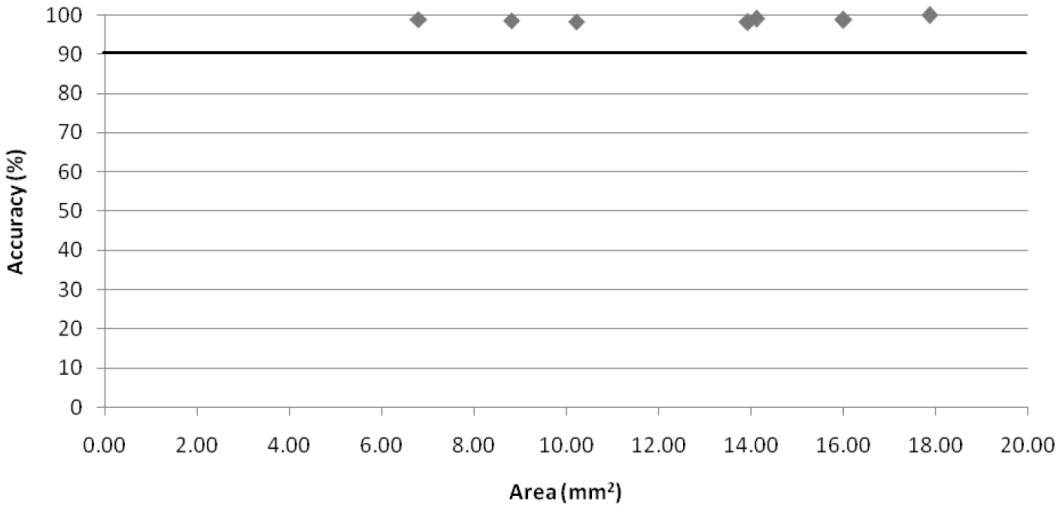

Fig. 9 The comparison of f character (24pt) area and Abby Fine Reader OCR accuracy at different tone values

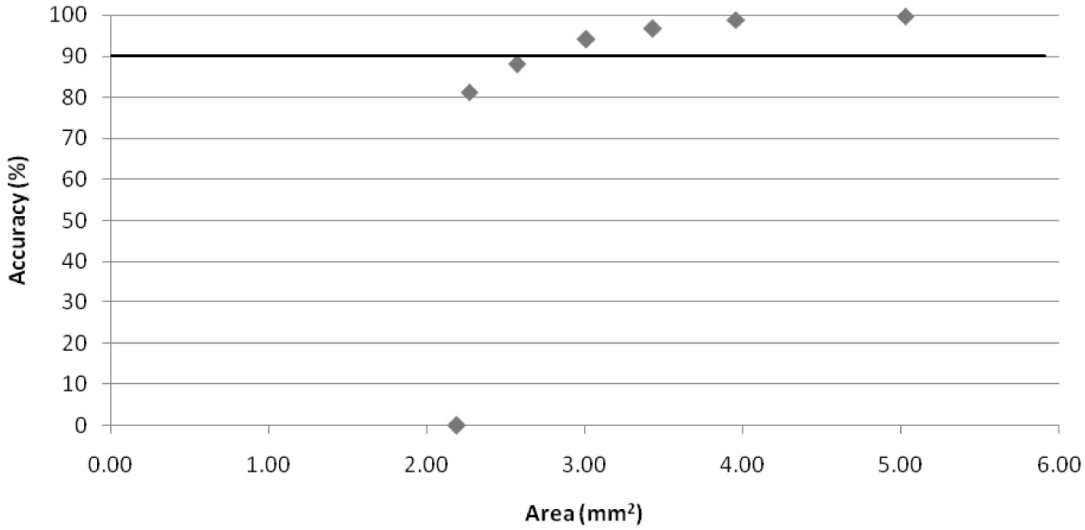

Fig. 10 The comparison of f character (12pt) area and Abby Fine Reader OCR accuracy at different tone values

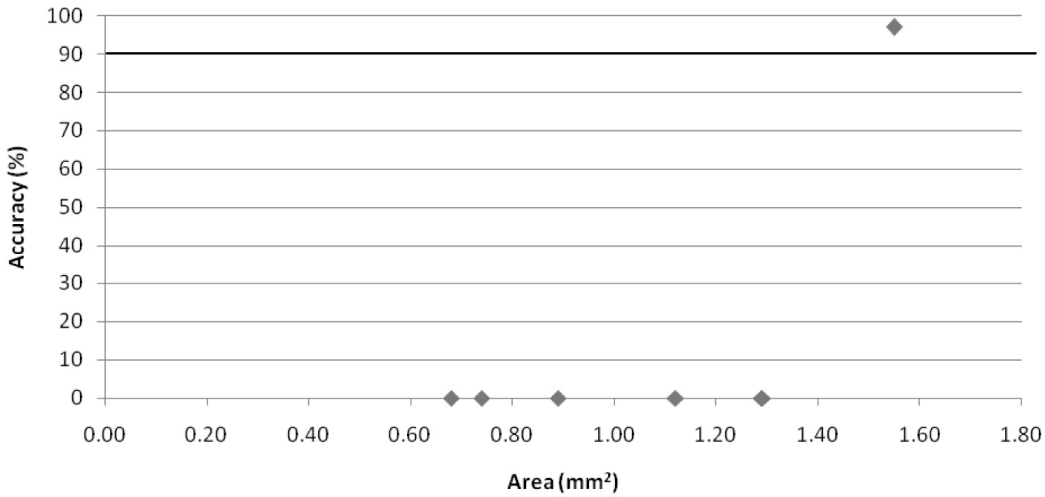

Fig. 11 The comparison of f character (6pt) area and Abby Fine Reader OCR accuracy at different tone values

of the accuracy results of Abby Fine Reader OCR software as it was the most accurate and consistent through most of the printed tone values. The results for the serif Times New Roman font are presented in Fig. 9, while the results for the sans serif Arial font are presented in Fig. 10. We plotted a line in the figures which denoted the average (useful) and bad OCR accuracy trehsold (90 % of accuracy).

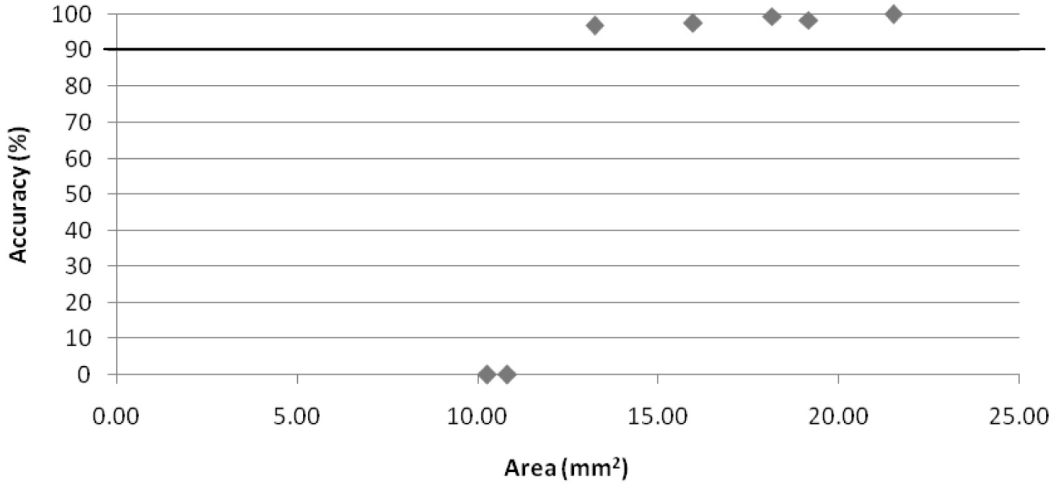

Fig. 12 The comparison of f character (24pt) area and Abby Fine Reader OCR accuracy at different tone values

As seen from Fig. 9 the accuracy for Abby Fine Reader with a 24 pt typeface is best at the largest area and then it shows very small variations in a small range of good and average accuracy OCR with the declining area value of around 6 mm^2. At 12 pt size the recognition accuracy of the Times New Roman text is declining with the area size congestion, and where the values between 2 and 3 mm^2 are the threshold area for even the bad accuracy value. At 6 pt only the 1.60 mm^2 area which represents the 100 % tone value text showed readable results. All characters with smaller tone values and area values where undetectable for the

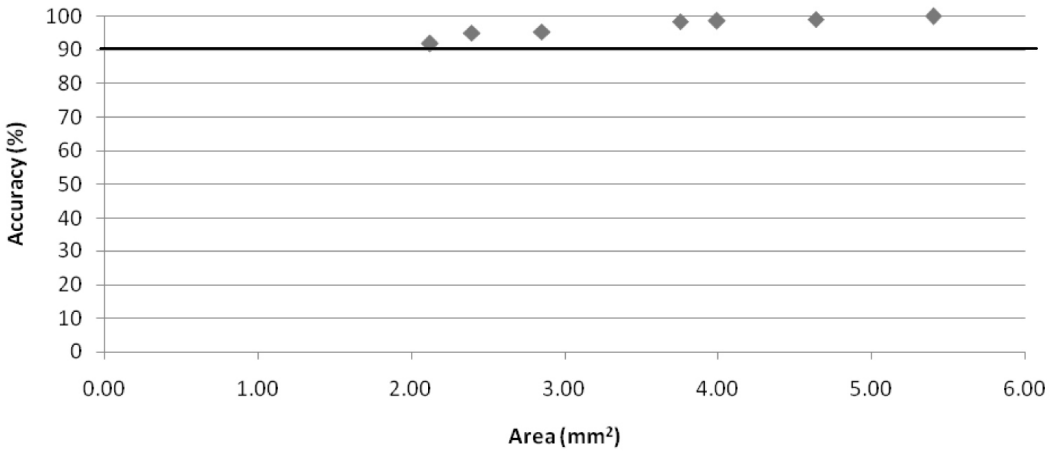

Fig. 13 The comparison of f character (12pt) area and Abby Fine Reader OCR accuracy at different tone values

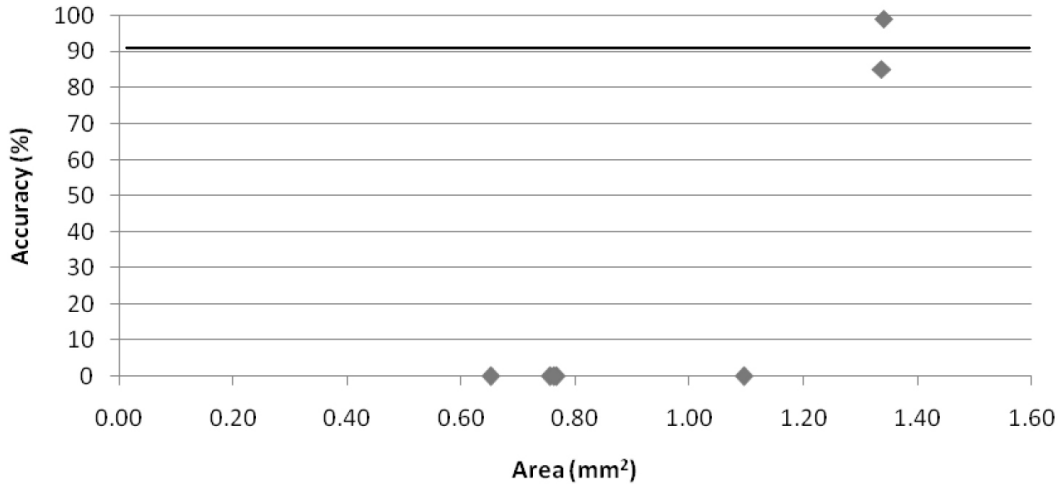

Fig. 14 The comparison of f character (6pt) area and Abby Fine Reader OCR accuracy at different tone values

software. For the Arial sans serif types the results are presented in Figs from 12 to 14.

As we can see in Fig. 12, for Arial typeface text with 24pt size the accuracy is fairly good for the character size above 15mm^2 which represents the text printed with 70 % tone value. After that value, the accuracy is drastically deteriorated. At 12pt size text, the accuracy is plummeting with the smaller area size of the characters, where the threshold is around 2 mm^2. For the 6pt size text values around 1.4 mm^2 were recognizable, and at smaller area and tone values there was no possibility for good OCR readout.

Conclusion

The results obtained from the study indicate that the tone value has an important role in OCR accuracy both for serif and sans serif typefaces. As it was expected, larger text sizes of 24pt and 12pt was recognized much better than 6pt size text. The results show that small typefaces quickly deteriorated in terms of shape due to smaller areas which form the character and this mainly influences the fact that it cannot be accurately recognized. There are also variations in the tested software where Abby Fine Reader, ReadIris and OmniPage Pro showed good results while the open source CuneiForm had fairly low results in all segments. The OCR algorithms differ in their quality and operating accuracy, and hence, it is concluded that the OCR software also constitutes an important role in the text recognition process. These results can be very useful in determining the threshold values of font sizes and tone values which could save toner usage while still be readabl. The most efficient was the Abby Fine Reader where the threshold was at about 50 % of TV (approximately half of the optical density of the full tone characters). Other software had lower accuracies and the saving is much lower due to narrow toner optimization (100-90 % TV). These results enable good optimization values where good quality printed text can be obtained with lower price, and which could be later still be useful for digital archiving.

Acknowledgment

This work was supported by the Serbian Ministry of Science and Technological Development, Grant No.: 35027 "The development of software model for improvement of knowledge and production in graphic arts industry".

References

[1] Mori S., Suen C.Y., Yamamoto K.: Historical review of OCR research and development, IEEE Proceedings, special issue on OCR, 1992, pp. 1029-1057.
[2] Fujisawa H.: Pattern Recog. **41**, 2435 (2008).
[3] Breithaupt M.: Improving OCR and ICR Accuracy Through Expert Voting, White Paper, 2006, Available from: http://www.csisoft.com/applications/OCE%20Intellidact%20Whitepaper.pdf (Accessed on 25.3.2013).
[4] Faure C. , Lecolinet E.: Written Language Input, Chapter 2 from Survey of the state of the Art in human language technology, Cole R. Ed., Cambridge University Press, 1997.
[5] Abby: OCR - Optimal Image Resolution, 2013, Available from: http://www.abbyy-developers.com/en:tech:insideocr:images_resolution_size

(Accessed on 25.3.2013).
[6] Tanner S.: Deciding whether Optical Character Recognition is feasible, 2004, Available from: http://www.odl.ox.ac.uk/papers/OCRFeasibility_final.pdf (Accessed on 8.4.2013).
[7] Atiz Innovation: Scanning for OCR Text Conversion, 2013, Available from www.atiz.com/resources/Scanning-for-OCR-Conversion.pdf (Accessed on 10.4.2013).
[8] Holley R.: How Good Can It Get? Analyzing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitization Programs, D-Lib Magazine, **15**, 2009.