https://doi.org/10.2352/issn.2168-3204.2021.1.0.22
This work is licensed under the Creative Commons
Attribution-NonCommercial-ShareAlike 4.0 International License.

Evaluation of Text Legibility in Alternative Imaging Approaches to Microfiche Digitization

Hilda Deborah, Dipendra J. Mandal; Norwegian Colour and Visual Computing Laboratory, Department of Computer Science, NTNU–Norwegian University of Science and Technology; Gjøvik, Norway

Abstract

Microfiche was a common format used in microforms reproductions of documents, extensively used for archival storage before the move to digital formats. While contemporary documents are still available for digitization, others from older historical periods are no longer physically accessible for various reasons. In some cases, their microfiche copies are available, making microfiche digitization a must. However, a microfiche reader is not always available and, even then, it is a machine made for the purpose of reading and not for data collection. In this work, the performance two imaging devices are evaluated as alternatives to the traditional microfiche reader, by means of optical character recognition (OCR). Results show that this alternative surpasses the performance of a microfiche reader in terms of text legibility.

Introduction

In the recent decades, we have seen an increase in the digitization of historical manuscripts using not only high-end color cameras and scanners, but also using multispectral [1], [2] and hyperspectral [3] imaging. There are significant advantages in doing so, not only for more accurate documentation purposes, but also for more advanced tasks, e.g., recovering hidden information [4]–[6]. Despite the need and advantages of such a digitization, especially for historical documents, there are cases where a *rescanning* of an object is no longer possible. Access to many historical documents in library collections across the globe can be difficult due to the fragile condition of the object. There are even cases where manuscripts or fragments have been lost [7]. Fortunately, when the documents have been kept in collections or institutions, often their records or analog copies are available in microforms. In this work, we focus on a specific format of microforms, i.e., microfiche.

Prior to the advance of digital technologies, microforms were the only available way to archive and preserve large documents. It was quickly adopted by the cultural heritage sector to capture their collection for preservation, access, and distribution. Microfiche is plasticky flat film sheets commonly used for reproducing historical printed documents, e.g., books and newspapers, in an optically reduced size or microforms [8]. These are of various types, e.g., silver-halide, diazo, and vesicular, and are available with different reduction ratios and life expectancy up to 500 years. The microfiche may be negative, i.e., clear lettering on a dark background, or its opposite, i.e., positive microfiche.

An example of a microfiche is provided in **Figure 1** and, taking note of its physical dimension, we can see that a single microfiche contains multiple photos or pages. Due to its significant reduction ratio, a microfiche reader [9], [10] is required to be able to observe and read its content or pages. Today's commonly available consumer or phone cameras would rarely have the resolution required to read a microfiche. This poses two challenges. The first is that a microfiche reader might not be as available as it was before since the technology has largely been replaced by digital technologies. A microfiche reader is also an analog machine made for the purpose of reading and not for digitization or data collection purposes. Thus, despite providing a high resolution, the use of a microfiche reader for digitizing microfiches is very time consuming. A single page in a microfiche equals a single image capture, requiring manual adjustments or placements of the lens such that it points to the right page. For one microfiche alone, use **Figure 1** as an example, 60 image captures are needed. And when talking about a digitization effort, we have hundreds if not thousands of microfiches, making the use of a microfiche reader impractical and highly costly. Additionally, it is also important to ensure the quality of the digitized images to meet user objectives.

The aim of this work is to find alternative imaging technologies for the digitization of microfiches. Trading off resolution with accessibility and time constraints, we are comparing two different imaging setups for the task of microfiche digitization. Information obtained from a microfiche must be readable. Thus, we define legibility as the evaluation criteria, and it is to be assessed by means of Optical Character Recognition (OCR) [11]–[13]. By using OCR, we limit the legibility assessment to system performance and, therefore, excluding assessment by human observers.

Figure 1. *An example of a positive microfiche, of physical dimension 105mm × 148mm and a reduction ratio of 24X or 24 times, which will be used in the assessment of text legibility experiment. This material comes from Ref. [14].*

Imaging Approaches

Access to a microfiche reader or any microform reader machine for that matter, is scarce even though the technology used to have a central role in archiving. The challenge of reading a microfiche is mainly related to its very reduced size. However, considering advances in optical devices in the past decade alone, it is highly likely that an alternative imaging solution is available. The first one to consider is a flatbed scanner, which nowadays are available with a resolution up to 6400 dpi. Then, if we also consider the availability of macro lenses coupled with a high-resolution camera, it might be sufficient to resolve the reduction ratio of a microfiche. Based on these considerations, we select a professional grade flatbed scanner and an in-house film scanning system that couples a monochrome camera and a macro lens as alternatives to a microform reader.

Microform reader

A microform reader Zeutschel delta plus was available to us through the local library in Gjøvik, Norway. This device was marketed for public or professional use in digitizing all formats of microfilm and photographic materials. Microfiche is also listed as a compatible input type. This device is said to support a *reduction ratio* of 7X-105X. Reduction ratio expresses the linear relationship between the size of a document and its photographically reduced format or *microimage* [14]. For example, if a 10 cm object has been reduced 10X, it means the microimage is of size 1 cm. Other specifications of the reader machine that is relevant for this study can be seen in **Table 1**. Despite the high reduction ratio support, we consider the operation ease of using this device for digitization purposes to be low. This is mainly due to the need to manually position the lens for every single page within the microfiche, see maximum fiche per scan factor in the table, i.e., $1/n$ with $n = 98$.

Flatbed scanner

A flatbed scanner used in this study is a professional grade scanner aimed for scanning films, i.e., Epson Perfection 4870 Photo. This scanner has up to 4800 dpi and provide an option to scan in transmissive or transparency mode, thus suitable for the purpose of microfiche scanning. The immediate advantage of its use is in time saving. Even though its throughput in **Table 1** is given in terms of seconds per line instead of per image, it still more efficient than a microfiche reader since it can scan two whole microfiches in one capture. This makes the operational ease high because of a significant reduce in time and efforts that are needed for the manual adjustments of apparatus and materials before each image capture.

In-house film scanner

An in-house LED-based multispectral film scanner with the main purpose of capturing various kinds of film colors in transmission mode [15]. This scanner couples a monochrome camera with a macro lens, see details in **Table 1**, and therefore suitable for microfiche scanning. Since microfiche materials in this study is black and white, we only take grayscale images with one light source instead of multispectral images with the full range of the LED lights. The light source used was 415.5 nanometer, chosen arbitrarily but kept constant throughout the acquisition of all images. The maximum scan area of this scanner is not only due to the field of view of the scanner, but also due to how the apparatus is built for capturing images in transmissive mode. It has a square hole of roughly the size of a 35 mm film and only objects smaller than that size can be captured. For the specific test microfiche used in this study, the apparatus allows capturing six pages within a single microfiche. Nevertheless, both the throughput speed and operational ease can still be considered as high.

Table 1. *Comparison of the specifications and characteristics of the three imaging setups evaluated in this study. Note that this summary is formulated within the specific context of reading 105 mm x 148 mm microfiches in a monochrome setup. The test microfiche has 24X reduction ratio and maximum $n = 98$ pages.*

Factors	Microform reader	Flatbed scanner	In-house film scanner
Model	Zeutschel delta plus	Epson Perfection 4870 Photo	QHY600 16BIT BSI, atx-i 100mm F2.8 FF MACRO
Compatible input types	Microfiche, microcards, 16/35 mm roll microfilm, photographic slides, negatives, 35 mm perforated films	A4 size document, transparencies, photos, 35 mm films, negatives, 4”x5” formats	35 mm photographs and motion picture films, small objects of different kinds
Max. scan area	35 x 47 mm	216 x 297 mm	35 x 40 mm
Max. fiche/ scan	$1/n$	$2n$	$6/n$
Effective pixels	10 MP	40,800 x 56,160 at 4800 dpi	9,576 x 6,388 (±60 MP)
Illumination	Custom-calibrated LED array	Cold cathode fluorescent lamp	Calibrated LEDs
Throughput speed	Medium (±0.3 sec/ image)	High (±0.027 sec/ line)	High (±0.4 sec/ image) *
Operation ease	Low	High	High

*The speed is calculated from specification of the camera given which was given as 2.5 fps for 16-bit output.

Experimental Setup

The flowchart of assessing text legibility of the three imaging setups for the context of microfiche digitization in this study can be seen in **Figure 2**. Microfiche materials will be captured using the different devices, resulting in grayscale digital images. Note that despite the ability of these devices to capture color or multispectral images, it is unnecessary for the purpose of this experiment.

Depending on the experiment, a post-processing of the digital image might be carried out to remove noise and smooth an image by means of a median filter. Then, the images will be passed on to an optical character recognition (OCR) engine, which in this case is the open-source Tesseract-OCR. An OCR engine takes an image as input and return texts it can read from the input image. By comparing this recovered text with its corresponding ground truth, a text similarity measure using Levenshtein edit distance [16] will be calculated.

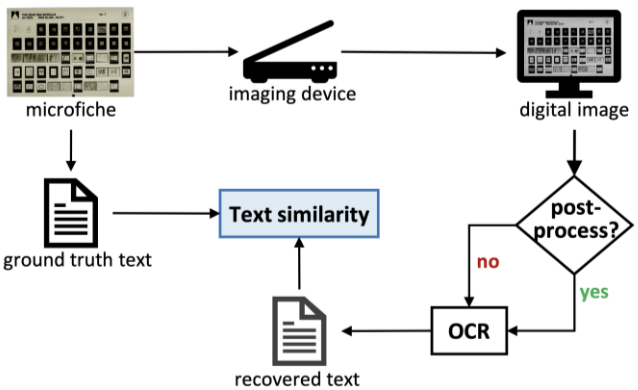
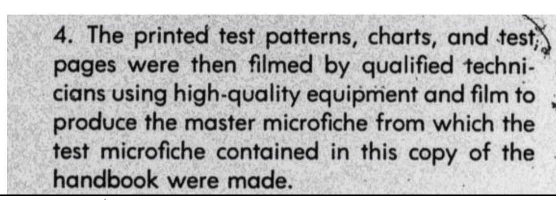


***Figure 2.** Experiment flowchart of the assessment of text legibility by means of Optical Character Recognition (OCR). Three imaging devices are compared, i.e., a microform reader, a flatbed scanner, and an in-house film scanner.*

Microfiche materials and their processing

The test microfiche used in this experiment is one of the microfiches provided by a handbook for evaluating microfiche readers [14]. It is the positive microfiche with 24X size reduction and, therefore, allowing a single microfiche to contain a maximum of 98 images. The microfiche itself contains of only 60 images, see **Figure 1**, composed of the microimages of all pages in the handbook. Considering their relevance for text legibility assessment using OCR, only 16 pages are used. Fourteen pages used in the experiment are written in two-columns page. This poses a necessity to split the image of a page into its individual column to avoid confusion in the order of reading by OCR. Consequently, the ground truth text is also made following such order. A subset of a column and its corresponding ground truth text can be seen in **Figure 3**.


4. The printed test patterns, charts, and test pages were then filmed by qualified technicians using high-quality equipment and film to produce the master microfiche from which the test microfiche contained in this copy of the handbook were made.

```
4. The printed test patterns, charts, and test
pages were then filmed by qualified techni-
cians using high-quality equipment and film to
produce the master microfiche from which the
test microfiche contained in this copy of the
handbook were made.
```

***Figure 3.** A subset image of a microfiche-column and its ground truth text.*

Tesseract-OCR

An optical character recognition (OCR) engine allows the conversion of digital images of typed, handwritten, or printed text into machine-encoded text, by recognizing a character at a time. It is particularly useful for automatizing a data entry process from printed records. In addition to working with images of documents, it can also be used to recognize text in a photograph of a scene. Tesseract-OCR is an open-source OCR engine that has been trained not only to detect single characters, but also optimized to recognize the shapes of letters for better recognition in case of blurred images. Furthermore, it also uses dictionary to improve text accuracy at the character segmentation step [17]. Our use of Tesseract-OCR is done through the Python wrapper `pytesseract`, and it returns the extracted text which will then be compared to the ground truth text.

Levenshtein edit distance

We have seen an example of an image input for the OCR and its corresponding ground truth text in **Figure 3**. The accuracy of the text returned by the OCR, however, will vary depending on the quality of the input image. This further means that the accuracy depends on the quality of the imaging device. Using the last two lines from the image in **Figure 3** as an example, below are the texts returned by OCR for the exact image:

```
test microfiche contained in this oaPy of the
handbook were made. :
```

Comparing the above text to its ground truth in **Figure 3**, two mistakes can be spotted. The word `copy` is recognized as `oaPy` and there is also an extra colon (`:`). In computational linguistics, edit distance is used to quantify the difference between two texts by calculating the minimum number of *operations* required to transform one string to another. Different edit distances consider different operations in its calculation, e.g., deletion or substitution. Levenshtein edit distance (ED) [16] is chosen since it considers deletion, insertion, and substitution. This enables comparing two strings of different lengths unlike, e.g., Hamming distance [18]. Calculating the difference of the above text and its ground truth using ED, we obtain the score of 5. Since ED is a distance function, smaller value means higher text similarity, therefore indicating a better imaging setup for the task at hand. In addition to the standard ED, we will also use cumulative or aggregate ED to allow better comprehension of the overall performance of an imaging device.

Results and Discussion

The results of legibility assessment of the three imaging setups for the use of microfiche digitization can be observed in **Figure 4**. In this graph, four entries are provided since two different dpi are evaluated for the flatbed scanner, i.e., at 4800 (Flatbed 4k) and 2400 (Flatbed 2k) dpi. By a quick observation, we can see that for certain microfiches, the use of a flatbed scanner with a 2400 dpi is insufficient to resolve the text from the microimage. Interestingly, it can also be seen that at 4800 dpi, the flatbed scanner almost always outperforms the microform reader. To have another point of view of the performance, see **Figure 5**, where ED is plotted in a cumulative manner along the x-axis. A cumulative ED at any point of the x-axis is a sum of all EDs from the previous points. Here, it becomes clearer that the flatbed scanner at 4800 dpi outperforms the others and the in-house film scanner in general performs better than the microform reader.

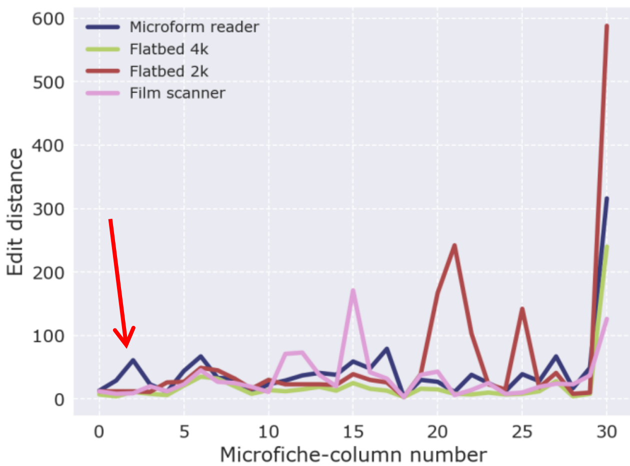


Figure 4. *Levenshtein edit distance (ED) of the compared imaging setups, computed against each document's ground truth text. Two different dpi are evaluated for the flatbed scanner, i.e., Flatbed 4k and 2k.*

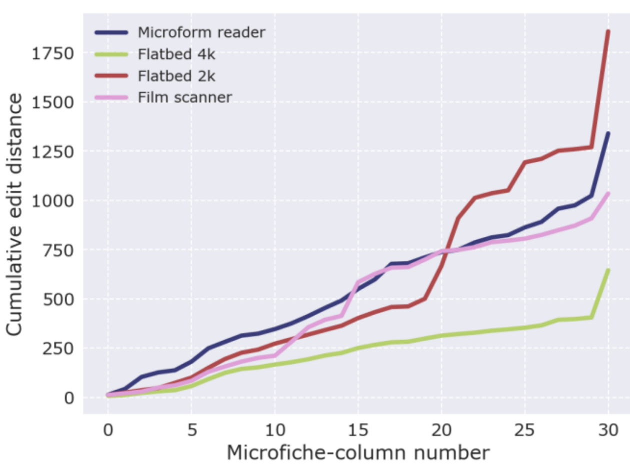


Figure 5. *ED of the compared imaging setups, shown in a cumulative manner along the x-axis. It shows that the flatbed scanner at 4800 dpi is the best performing one at providing legible texts as evaluated by an OCR and that, interestingly, the microform reader is not clearly superior from the rest.*

To obtain a more thorough understanding of why and when a certain device is a better choice, an observation of the images is needed. Microfiche-column 3/5-1 is the one resulting in the first peak in the microform reader plot in **Figure 4**, as pointed by the red arrow. A subset area of that image can be observed in **Figure 6**. Upon a visual observation, both the contrast and sharpness of the flatbed scanner images in **Figure 6**(b)-(c) are significantly reduced compared to the one in **Figure 6**(a). The legibility score as measured by ED is, however, conversely related. Despite seemingly having a lower image quality, the flatbed scanner image at 4800 dpi has a lower ED of 11 compared to that of the microform reader with ED of 61. Even the image at 2400 dpi provides a better ED of only 12. The visual similarity of the flatbed scanner and film scanner images, those that provide low EDs, are the smoothness of the background. On the other, despite sharp letters, the microform reader image is noisy and granular in its background content. With the hypothesis that the background content tampers with the text recognition of the OCR, a smoothing filter can be used to improve the general performance of the microform reader and the film scanner.

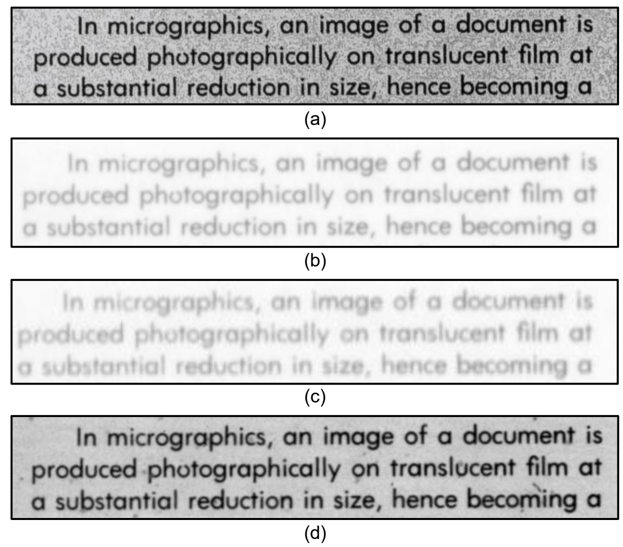


Figure 6. *A subset area of microfiche-column 3/5-1 obtained from the compared (a) microfiche reader, (b) flatbed scanner at 4800 dpi and (c) 2400 dpi, and (d) in-house film scanner. The obtained ED scores for the microfiche-column for the respective devices are 61, 11, 12, and 9.*

After applying median filters of varying kernel size, the performance of the in-house film scanner is measured and plotted in **Figure 7**. We can see that the use of kernel size 3 improves the performance, albeit insignificantly. However, with kernel size 5, OCR struggles at recognizing the text in the images. The impact of smoothing on the images can also be observed through examples in **Figure 8**. Here, it becomes clear that smoothing with kernel size 5 blurs the individual letters unlike in size 3 where they are still sharp. Smoothing is also applied to the images from the microform reader and the performance can be observed in **Figure 9**. Note that here we choose to visualize it in terms of cumulative ED for ease of reading and clarity purposes. In the figure, it can be observed that smoothing increases the legibility of the text, although only up to kernel size 11. When reaching size 13, the legibility performance starts to decrease again as shown by MF-13 in the plot. The impact of median filters at these cutoff sizes to the images are shown in an example in **Figure 10**. Compared to the original image in **Figure 6**(a), both median filtered images show less granular artefacts in the background, allowing improvements in the legibility aspect. Nevertheless, the improvement is only possible when the text itself is not blurred, which is the difference that can be observed between the two images in **Figure 10**.

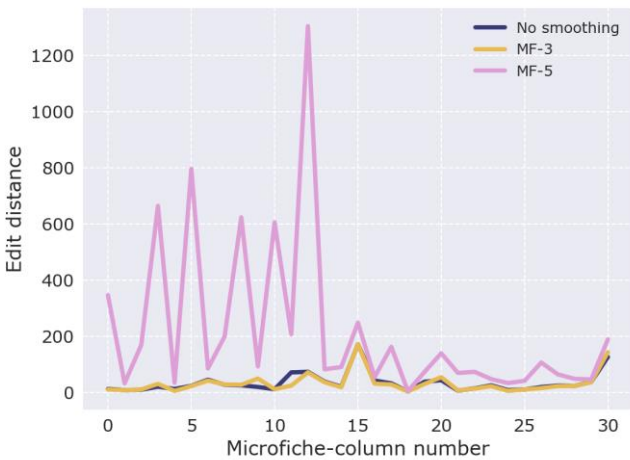
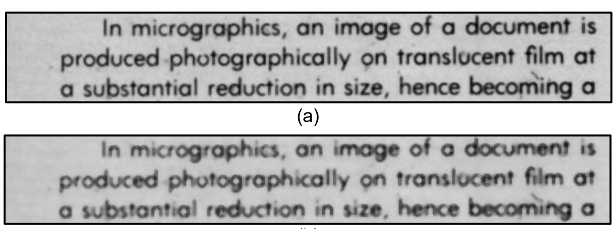


Figure 7. ED obtained by the in-house film scanner, at varying level of smoothing using median filters. MF-x in the figure legend means a median filter of kernel size x has been applied to the microfiche-column images.


In micrographics, an image of a document is produced photographically on translucent film at a substantial reduction in size, hence becoming a

(a)

In micrographics, an image of a document is produced photographically on translucent film at a substantial reduction in size, hence becoming a

(b)

Figure 8. Subset area of microfiche-column 3/5-1 obtained from the in-house film scanner, and after applying a median filter of kernel size (a) 3 and (b) 5. The former gives the highest legibility, while the latter the lowest.

Finally, taking the best combination of with or without smoothing, the performance of each imaging device in terms of text legibility can be observed through **Figure 11**. The flatbed scanner at 4800 dpi is still the best performing one. If we recall the initial result without smoothing as a post-processing step shown in **Figure 4** and **Figure 5**, we can now see that the performance of the microform reader has significantly improved when combined with a median filter of kernel size 11. Its result approximates the flatbed scanner at 4800 dpi. The improvement made by applying a median filter of size 3 to the in-house film scanner images, however, is insignificant. This makes its performance comes at the third place, after the flatbed scanner at 4800 dpi and the microform reader. Nevertheless, it is still a better choice for when the available flatbed scanner only provides a resolution of up to 2400 dpi.

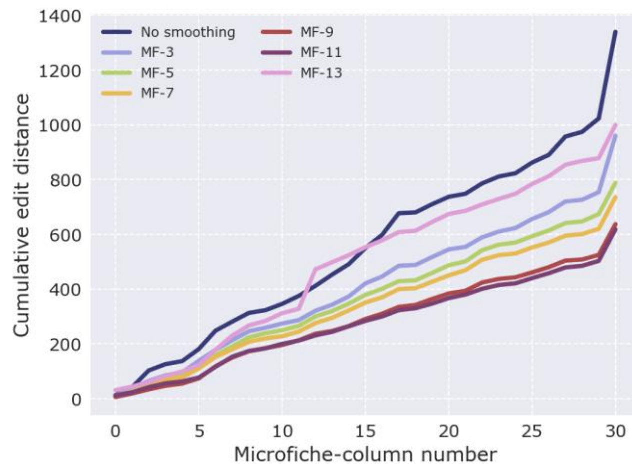
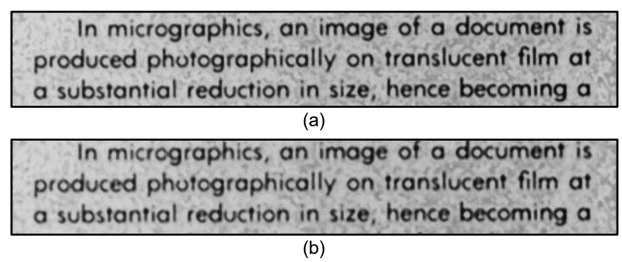


Figure 9. Cumulative ED obtained by the microform reader, after applying a median filter of kernel size x. The best performance is provided by applying a median filter of size 11, as shown by MF-11.


In micrographics, an image of a document is produced photographically on translucent film at a substantial reduction in size, hence becoming a

(a)

In micrographics, an image of a document is produced photographically on translucent film at a substantial reduction in size, hence becoming a

(b)

Figure 10. Subset area of microfiche-column 3/5-1 obtained from the microform reader, and after applying a median filter of kernel size (a) 11 and (b) 13. The former gives the highest legibility, while with the latter the legibility performance starts to decrease.

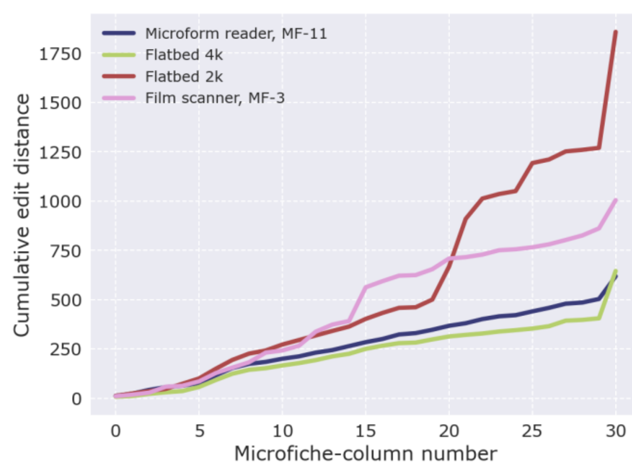


Figure 11. ED of the compared imaging setups, shown in a cumulative manner. Both microform reader and film scanner are combined with smoothing by means of median filters as a post-processing step.

Conclusion

In this study we have proposed a criterion for evaluating the quality of imaging devices for the task of microfiche digitization, i.e., text legibility. As an evaluation protocol, we have also proposed the use of OCR for an automatic recognition of the text in the images and Levenshtein edit distance as the metric. Three imaging devices have been compared, i.e., a microform reader, a flatbed scanner, and an in-house film scanner coupling a monochrome camera and a macro lens. As a conclusion, the flatbed scanner with 4800 dpi has been found to be the most suitable imaging device providing the highest quality of computer-legible texts.

This work has been motivated by our own research activities in the cultural heritage domain where, often, we do not have access to the physical objects for their rescanning using advanced imaging technologies. While in this study we have only assessed the legibility criteria, microfiche materials in our research are not only composed of written texts. There are also photographs and handwritten texts that will be unrecognizable by an off-the-shelf OCR. As a future work, we will develop more complete assessment protocols, considering other objective quality aspects as well as incorporating subjective evaluations by human observers.

Acknowledgment

Authors would like to thank Torleif Elgvin, Gregory High, and Gjøvik Biblioteket for providing access to the various materials and devices used in the formulation and execution of this study. This work is supported by The Lying Pen of Scribes–Manuscript Forgeries, Digital Imaging, and Critical Provenance Research funded by the Research Council of Norway (projectnr. 275293) and CHANGE-ITN project funded by EU Horizon 2020 (Marie Skłodowska-Curie, grant agreement No. 813789).

References

[1] R. L. (Jr.) Easton and W. Noël, "The Multispectral Imaging of the Archimedes Palimpsest," *Gaz. Livre Médiév.*, vol. 45, no. 1, pp. 39–49, 2004, doi: 10.3406/galim.2004.1646.

[2] B. Stegmann, "Collaborative Manuscript Production and The Case of Reykjabók: Paleographical and Multispectral Analysis," in *New Studies in the Manuscript Tradition of Njáls saga*, Medieval Institute Publications, 2018, pp. 29–54.

[3] I. M. Cortea, L. Ghervase, L. Ratoiu, and R. Rădvan, "Application of Spectroscopic and Hyperspectral Imaging Techniques for Rapid and Nondestructive Investigation of Jewish Ritual Parchment," *Front. Mater.*, vol. 7, p. 601339, 2020, doi: 10.3389/fmats.2020.601339.

[4] A. Tournié *et al.*, "Ancient Greek text concealed on the back of unrolled papyrus revealed through shortwave-infrared hyperspectral imaging," *Sci. Adv.*, vol. 5, no. 10, 2019.

[5] L. Snijders, T. Zaman, and D. Howell, "Using Hyperspectral Imaging to Reveal a Hidden Precolonial Mesoamerican Codex," *J. Archaeol. Sci. Rep.*, vol. 9, pp. 143–149, 2016.

[6] A. Tonazzini *et al.*, "Analytical and mathematical methods for revealing hidden details in ancient manuscripts and paintings: A review," *J. Adv. Res.*, vol. 17, pp. 31–42, 2019.

[7] E. Tigchelaar, "On the Unidentified Fragments of 'DJD' XXXIII and PAM 43.680: A New Manuscript of '4QNarrative and Poetic Composition', and Fragments of '4Q13', '4Q269', '4Q525' and '4QSb'(?)," *Rev. Qumrân*, vol. 21, no. 3 (83), pp. 477–485, 2004.

[8] W. De Haas, "The Microfiche," *Am. Doc.*, vol. 9, no. 2, pp. 99–106, 1958, doi: 10.1002/asi.5090090204.

[9] C. Frost, "Microfiche readers," *Computer*, vol. 5, no. 3, pp. 58–60, 1971, doi: 10.1109/C-M.1972.216919.

[10] F. L. Keeler and W. A. Rizzo, "An evaluation of microfiche reader types for use with programmed instruction," Naval Training Equipment Center, TAEG-R-35, 1976.

[11] H. Deborah, H. M. Manurung, and A. M. Arymurthy, "Objective Criteria for Typewritten Old Document Image Enhancement and Restoration," *Proceeding Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS*, 2010.

[12] J. Owen, "Using Optical Character Recognition to Identify Legibility of Non-western Languages," California Polytechnic State University, 2011.

[13] M. R. Gupta, N. P. Jacobson, and E. K. Garcia, "OCR binarization and image pre-processing for searching historical documents," *Pattern Recognit.*, vol. 40, no. 2, pp. 389–397, 2007, doi: 10.1016/j.patcog.2006.04.043.

[14] W. R. Hawken, *Evaluating Microfiche Readers: A Handbook for Librarians*. Council on Library Resources, 1975.

[15] G. Trumpy, S. George, J. Y. Hardeberg, and B. Flueckiger, "Multispectral capture of film colors with LEDs," in *Colour Photography and Film: Sharing knowledge of analysis, preservation, conservation, migration of analogue and digital materials*, 2021, p. 28.

[16] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Sov. Phys. Dokl.*, vol. 10, p. 707, 1966.

[17] R. Smith, "An Overview of the Tesseract OCR Engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Sep. 2007, vol. 2, pp. 629–633.

[18] R. W. Hamming, "Error detecting and error correcting codes," *Bell Syst. Tech. J.*, vol. 29, no. 2, pp. 147–160, 1950.

Author Biography

Hilda Deborah received her BSc from the University of Indonesia (2010), MSc from the Erasmus Mundus Color in Informatics and Media Technology (2013), and PhD from NTNU and University of Poitiers (2016). She was a Marie Curie Postdoctoral Fellow at the University of Iceland and NTNU from 2018-2020. She is currently Senior Researcher at NTNU, with interests in hyperspectral imaging and fundamental image processing, with applications in the cultural heritage and remote sensing domains.

Dipendra Jee Mandal received his BE (2010) and MS (2017) degree in Electrical and Electronics Engineering from Kathmandu University, Nepal. He was an exchange student under Erasmus Mundus at University Lumiere Lyon 2, France (2017-2018). He is currently a Ph.D. student at NTNU under Marie Curie fellowship (CHANGE-ITN), his research field is quality assessment for cultural heritage digitization using hyperspectral imaging.