Scanning Preservation Microfilms: Key Issues

Hans van Dormolen; Metamorfoze, Koninklijke Bibliotheek, National Library of the Netherlands; The Hague, The Netherlands

Abstract

In my paper "Scanning Preservation Microfilms: Key Issues, I have presented an OCR-research (Optical Character Recognition) that was initiated by the KB in March 2007. The goal of this research is to test the use of microfilms of newspapers as source material for a work process in which a digital derivative is the end product. For use on the internet, the OCR-quality of the digital derivative is very important. For this research, we have studied the relationship between the technical quality of the microfilms, the technical quality of the microfilm scanners and the final OCR accuracy.

Originals

In every work process in which originals are transferred to another carrier, the quality of the originals plays an important role. In this particular process, we are dealing with newspapers. Newspapers can be discolored quite badly. Besides this, the quality of the print varies between different newspapers. Therefore, we have used two different newspaper pages for this research:

- A page from a modern newspaper, October 2006. This page represents high quality print. That is to say: a clear, black letter on a clear white background. This original is bitonal.
- A page from an old newspaper, September 1892. This page represents very low quality print. Low quality in this case means that on a single page, a very thin light gray readable letter can be followed by very bold, black letter. The background has discolored evenly, from a yellowish tint in the centre to a light and darkish brown at the edges of the page. The original contains many gray tones.

Scope of the test

All general accepted methods of black and white microfilming on a 35 mm negative film have been imitated in this research. Errors that may occur during filming and which will undoubtedly have a negative influence on the accuracy have not been researched here. These errors are: gutter shadow, skew, filming without the glass sheet which may cause a disturbing shadow on the page. These errors occur frequently in old microfilms, until ca. 1995. The adverse effects of "set through" ("set through" happens when the ink on the front of a page shows through the back of a page) on the OCR accuracy have not been tested specifically in this research. Set through is difficult to quantify. The percentages presented in this research must be seen as the maximum attainable. Depending on the film errors referred to above and on the "set through" the OCR accuracy will in reality be slightly lower.

For reference, we have also made scans of the original newspaper pages that have been used in this research. These scans were made in May 2007 with a Zeutschel scanner, the OS 10000.

The quality of these scans does not comply with the guidelines as described in the draft version of the Guidelines Preservation Imaging Metamorfoze. The technical quality of these scans was of a standard that was widely accepted throughout the Netherlands at that period (May 2007). But this level is now considered rather low. The technical deviations in the images however have only a very insignificant influence on the OCR accuracy researched for this research. The level of noise is rather high (The standard deviation is measured on a Q-13 and goes from 4.5 on patch A to 12.5 on patch 19). This might have an adverse effect on the OCR accuracy. On the other hand, the highlight gamma is rather high (around 1.4 for all the color channels). This might have a positive effect on the OCR accuracy. Some editing, such as sharpening or increasing of the contrast of the images in order to influence positively the OCR accuracy, has been left out deliberately. We have opted for a relatively cheap standard workflow. For the OCR we have used ABBYY FineReader 8.0 Corporate Edition software.

Microfilm scanners

The microfilms have been scanned with microfilm roll film scanners: the Zeutschel OM 1200 and OM 1400. The performances of these scanners can be compared with other brands of microfilm scanners. The performances of the tested scanners are referred to as production scanners in this report. In order to assess the technical performances of these production scanners and also to assess what we have to give up in terms of quality in favour of high production (bulk) and speed, we have also made reference scans of the microfilms used in this research. These reference scans give us an insight into what is technically possible for the microfilms made for this research in terms of data transmission. The reference scans were made with an Imacon Flextight 848. This Imacon scanner is absolutely not suitable for production scanning of microfilms. Scanning with this scanner is very time consuming. But the Imacon scanner does show what is technically possible with the microfilms produced for this research. The performances of this scanner are referred to as slowscan.

First generation microfilm

All general accepted methods of black and white microfilming on a 35 mm negative film have been imitated in this research. The methods are:

- High contrast microfilming with average, high and with low density
- Low contrast microfilming

High contrast microfilming

High contrast microfilming has for a long time been a generally accepted and widespread method for microfilming. This way of microfilming is characterized by the relatively high

contrast of the mother negative. The gamma or gamma value (Gamma or gamma value is the contrast or contrast factor. The gamma simply way indicates the relation between the contrast of the original and the contrast of the image in the film (1)) of these first generation microfilms has an average of 3 (2).This means that the contrast in the mother negative is on average three times as high as the contrast in the original. This means that two thirds of the original gray tones are lost, which is a loss of 66.66%. Particularly in the high lights, which are the light gray parts (light gray letter), this loss of gray tones has important consequences. Because of the loss of gray tones, holes may appear in the light gray letters. And holes in letters result in a decreased OCR accuracy.

High contrast microfilming in this research is divided in three groups:

- Density 1.00 – 1.30, Abbreviation HC 1.35. The D-max (maximum density-minimum density) of patch A on the Kodak Gray Scale (Q-13) on the first generation microfilm is 1.35. At this density, calibration is possible at various stages of the production process. HC stands for high contrast.
- Density 1.30 – 1.60, Abbreviation HC 1.62. The D-max (maximum density-minimum density) of patch A on the Kodak Gray Scale (Q-13) on the first generation microfilm is 1.62. At this density, calibration is possible at various stages of the production process. HC stands for high contrast.
- Density 0.70 – 1.00. Abbreviation HC 1.04. The D-max (maximum density-minimum density) of patch A on the Kodak Gray Scale (Q-13) on the first generation microfilm is 1.04. At this density, calibration is possible at various stages of the production process. HC stands for high contrast.

Low contrast filming is a microfilming method developed by Metamorfoze between 1999–2006. The essence of low contrast filming is to retain as much as possible the gray tones in all generations microfilms. With the help of a gray scale the loss of gray tones in different generations is made clear. The first generation low contrast microfilms currently have a gamma of 1.5. The contrast within these films is therefore on average 1.5 times as high as in the original. All low contrast first generation microfilms have a density of 1.00 to 1.20. A density below 1.00 is considered underexposed. A density over 1.20 is considered overexposed. Newspapers have been microfilmed low contrast by Metamorfoze since 2006. Low contrast microfilming is formed by 1 group:

- Density 1.00 – 1.20, Abbreviation LC 1.24. The D-max (maximum density-minimum density) of patch A on the Kodak Gray Scale (Q-13) on the first generation microfilm is 1.24. At this density, calibration is possible at various stages of the production process. LC stands for low contrast.

Second generation microfilm

For this research we have scanned from a second generation microfilm. We have tested the usefulness of second generation microfilms with a negative as well as films with a positive polarity.

Second generation microfilm with a negative polarity

The microfilm (Kodak 2470 Intermediate) that was used here has a gamma of around 1. This means that there is no contrast change in the image when this film is duplicated. In other words: all information that is there on the first generation microfilm is retained in this second generation microfilm. Another advantage besides the gamma 1 is that it is easy to define correct exposure and development of this second generation microfilm in a guideline by defining the D-min (minimal density, base plus fog). The D-max of the first generation microfilm, however, decreases slightly in the second generation. This only applies to the density area over 1.00. Metamorfoze has been using this type of film as a second generation microfilm since 2005.

Second generation microfilm with a positive polarity

The microfilm (Agfa Copex) that has always been used for this purpose in the Netherlands has a reasonably high contrast, a gamma of around 2. The dynamic range of this film is rather restricted, 3 to 3.5 stop. The comparatively high contrast of this film, as well as the limited dynamic range, are disadvantageous aspects of this type of film. The direct consequences of these two aspects is that these films may alternately have the right exposure or be slightly overexposed or underexposed, depending on the exposure used for duplicating and the density of the mother film.

These three variants, correct exposure, slightly overexposed and slightly underexposed have been imitated in this research, see Table 4.

Scanning microfilms

Before the second generation microfilms were scanned, the scanner was adjusted optimally (calibrated) for each type of film using patch A of the Kodak Gray Scale on the microfilms (HC 1.35, HC 1.62, HC 1.04, LC 1.24). Optimal adjustment means that the scanner is adjusted in such a way that patch A, with an accurately defined D-max in the mother negative is translated consistently around pixel value 242. Besides this, we have tried to translate the size of the step between patch A and patch 1 as realistic as possible. (LC 1.24. The D-max of patch A in the second generation negative film is a density of 1.10. We translate this value to white, to a pixel value of around 242. Patch 1 in the second generation negative film has a density of 0.93. This is a density difference of 0.17 points. In an optical model a density difference of 0.17 points equals a pixel value difference of 39 points. Now in Photoshop, using the eyedropper tool (5x5 pixels), we measure merely 3 points difference. The difference measured here divided by the theoretical difference, 3/39, is 0.076. See the calculation of the highlight gamma (3)) We have also tried to show the entire tonal scale on the gray level from D-max to D-min. While scanning the microfilms with negative polarity it turned out that only very limited adjustment was possible to make with the tested microfilm scanners. A gamma adjustment (contrast adjustment) for optimal scanning of the microfilms with a negative polarity cannot, or at any rate can only very limitedly be made. This deficiency renders the microfilm scanners incapable to register correctly the contrast transitions in the density area of about 1.10 to 0.60, between patch A and patch 3, on the Kodak Gray Scale. Of the size of the step between patch A and patch 1, only 7.6% remains. The calculation of this percentage is based on the density difference in the high lights of an optical model with a positive polarity. When scanning a

film with a negative polarity the highlights are located in the dark parts. The difference in pixel values in the dark parts is always smaller than in the highlights. A density difference of 0.17 in the dark parts (optical density 1.78 – 1.95) results in a difference in pixel values of 7 points (with monitor gamma 2.2). In percentages, the contrast transition is 3/7, or 4.2%. This is also a very poor contrast transfer. All the more so because in this calculation we assume a D-max defined as 1.95. On the negative film, however, the D-max is only 1.10.

In general we can say that the tonal capture performance of the tested microfilm scanners, when scanning microfilms with a negative polarity, is insufficient. The direct result of this insufficient tonal capture performance is digital files with a low OCR-accuracy.

The tonal capture performance of the reference scanner, the Imacon Flextight 848, is, after calibration, acceptable. The difference between patch A to 1 on microfilm LC 1.24 neg. is conveyed by 31 pixel values. This is a contrast transfer of 79% (Pixel value patch A is 242, pixel value patch 1 is 211. The difference is 31. Highlight gamma is 31/39 is 0.79), which is acceptable. In the Guidelines Preservation Imaging Metamorfoze a highlight gamma of 0.8 to 1.08 (80% - 108%) is given as tolerance value. Correct tonal capture performance guarantees high OCR-accuracy.

The tonal capture performance of the tested microfilm scanners, the Zeutschel OM 1200 and 1400, is hard to express in figures when scanning microfilms with a positive polarity. This is partly due to the fact that the dynamic range of the positive microfilm is limited. The difference between patch A and 1 is generally hardly visible on a film with positive polarity. In pixel values this difference is therefore nil. It does turn out, however, after visual inspection, that no or hardly any information is lost on the film. With other words: it is difficult to judge what exactly happens with the weak gray tones of the letters. In film LC 1.24 pos, the difference between patch A and patch 2, after scanning is 54 points. The highlight gamma between patch A and 2 is 1.17, the contrast transfer is 117%. However, this does not mean very much, as it is not clear what information is lost between patch A and 1. In general, we can say that the contrast transfer between a film with positive polarity and its digital derivatives is in harmony. This means that the differences in pixel values in the highlights are high and in the dark parts low. Because of the combination of the limited dynamic range of the film with positive polarity and the limited capacity of the microfilm to transfer tonal information, blacks will fuse easier. This can cause difficulties if the information in the black parts is relevant, such as in the combination of text and "set through" and when there are drawings with relevant information in the black parts.

OCR accuracy

In order to determine the OCR accuracy we counted the characters that were rendered correctly and those that were rendered incorrectly for a certain page. After that, we calculated the accuracy percentage using the total number of characters on the same page, see Table 1-4. We often had to cease counting the characters rendered correctly and incorrectly in the texts of the scanned pages. In such cases we did not deem it worthwhile to keep count for these texts as the amount of characters that were rendered incorrectly was very high. The OCR accuracy of these scans is very low, although we do not know exactly how low. We do however know that accuracy percentage is below 40% and sometimes even much lower.

Table 1: Scan of the original and OCR accuracy

Modern newspaper	99.95%
Old newspaper	95.75%

Table 2: OCR accuracy Slowscan LC and HC with neg. polarity

	LC 1.24	HC 1.35	HC 1.62	HC 1.04
Modern newspaper	99.88%	Un-known	94.34%	94.26%
Old newspaper	95.45%	95.35%	94.84%	81.54%

Unknown: Scanning while retaining all gray tones does have its drawbacks as well. The scanning process is much more difficult and time consuming. When the negative HC 1.35 modern newspaper was scanned, the derivative images became too gray. The OCR package, AbbYY FineReader 8.0 Corporate Edition, was unable to cope with it. In a future research, this negative will be scanned and assessed again.

Table 3: OCR accuracy Production scan LC and HC with neg. polarity

	LC 1.24	HC 1.35	HC 1.62	HC 1.04
Modern newspaper	93.11%	96.69%	93.71%	97.44%
Old newspaper	< 40%	< 40%	< 40%	< 40%

In this table the limitations of the microfilm scanner are very obvious. Every exposure of the old newspaper has scored badly. However, the much higher accuracy of the high contrast microfilm for exposures of the bitonal original, the modern newspaper, is quite remarkable.

Table 4: OCR accuracy Production scan LC and HC with positive polarity and normal-exposed

	LC 1.24	HC 1.35	HC 1.62	HC 1.04
Modern newspaper	99.65%	99.72%	98.30%	99.54%
Old newspaper	95.39%	93.97%	94.42%	88.06%

Table 5: OCR accuracy Production scan LC and HC with positive polarity and over-exposed

	LC 1.24	HC 1.35	HC 1.62	HC 1.04
Modern newspaper	99.49%	99.51%	99.07%	98.92%

Old newspaper	94.27%	93.82%	< 40%	< 40%

Table 6: OCR accuracy Production scan LC and HC with positive polarity and under-exposed

	LC 1.24	HC 1.35	HC 1.62	HC 1.04
Modern newspaper	99.47%	99.73%	97.71%	99.76%
Old newspaper	94.22%	< 40%	< 40%	92.68%

Conclusion

From the production scan OCR accuracy tables we can conclude that low contrast filming gets the highest accuracy score and is also the most reliable. It also shows that the production scans made from a film with positive polarity are better than production scans made from a film with negative polarity. The low accuracy of the combination of high contrast microfilm and slightly over or under-developed second generation microfilms can be ascribed to the combination of disadvantageous qualities of the microfilm in this workflow such as high contrast and a limited dynamic range.

It is very difficult to build a stable and dependable workflow with a second generation microfilm with a limited dynamic range and a gamma of 2, such as the film with positive polarity used here. There is no second generation microfilm with positive polarity and better (film) qualities, such as a slightly bigger dynamic range than 3 stops and a gamma slightly lower than 2. To me it therefore seems absolutely necessary that the quality of the microfilm scanners will soon be improved with regard to the tonal capture performances of scanning microfilms with a negative polarity. As long as this is not the case we will use second generation microfilms with positive polarity for scanning.

References

[1] P. Charpentier, Foto Techniek.

[2] The gamma value is the tangent of the linear zone of the so called "S"curve. The gamma can easily be calculated using a Kodak Gray Scale. See par. 2.4 Kodak Gray Scale, Richtlijnen Preservation Microfilming Metamorfoze, versie III, 2006. http://www.metamorfoze.nl/publicaties/richtlijnen/richtlijnenfeb06.pdf

[3] Highlight gamma, par. 2.7.1. Guidelines Preservation Imaging Metamorfoze. Metamorfoze Preservation Imaging Guidelines http://www.metamorfoze.nl/publicaties/richtlijnen/richtlijnen.htm

Author Biography

After working for 15 years as a professional photographer Hans van Dormolen started in 1999 a career as quality manager for the National Library of the Netherlands. He is responsible for the technical quality of preservation substitutes, analog and digital. He is the author of the Metamorfoze Preservation Microfilming Guidelines and co author of the Metamorfoze Preservation Imaging Guidelines.