
Digitising surrogates: Scanning from Microfilm

IMPACT Case Study

IMPACT project

Astrid Verheusen, Koninklijke Bibliotheek
Hans van Dormolen, Koninklijke Bibliotheek
Lotte Wilms, Koninklijke Bibliotheek

Released under Creative Commons - Attribution-NonCommercial-ShareAlike v3 Unported
(International)

Table of Contents

Executive Summary	1
Description of the test	1
Methods and procedures used	2
Evaluation	4
Slow scanning versus production scanning	4
OCR accuracy in production scanning	4
Calibration of scanners	4
Detailed results	5
Conclusions	6
Further research	6
References	7
Links	7
Glossary	7
Addendum 1	8
Addendum 2	10

Digitising surrogates: Scanning from Microfilm – A Case Study

Executive Summary

In mass digitisation projects scanning from already available microfilms is often considered as a cost effective way to produce high volumes of digital images. The quality of digital images scanned from microfilm may be doubtful and depends on the optical quality of the original print, i.e. the clarity of the letters in contrast to their background, the technical quality of the microfilms, the technical accomplishments of modern microfilm scanners. This research shows that second generation microfilms with a positive polarity give the best OCR accuracy, that microfilm scanners in use today for mass digitisation yield very poor results when it comes to microfilms with a negative polarity and that low contrast microfilming can be expected to lead to higher OCR accuracy than high contrast microfilming.

Description of the test

In this test done by the Koninklijke Bibliotheek a comprehensive set of microfilms has been manufactured. These have been run through a number of scanners: scanners used typically in mass digitisation where throughput is very important (production scanning), but also a scanner that is slow to operate but gives

out a high quality image (slow scanning). In this way it was possible to compare how much quality is sacrificed for the sake of throughput. OCR accuracy of the results was then compared. The results were also matched against reference scans, the same pages scanned directly from the original source.

In order to relate the quality of the results to the quality of the originals, two different newspaper pages were used for this research, to represent both ends of the quality spectrum:

- 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. The original was 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 grey readable letter can be followed by a very bold, black letter. The background has discoloured evenly, from a yellowish tint in the centre to a light and darkish brown at the edges of the page. The original contains many grey tones.

Methods and procedures used

For OCR, ABBYY FineReader 8.0 Corporate Edition was used.

OCR accuracy was measured by counting the number of correct and incorrect characters on a page. When the OCR score was very low not all characters on the page were counted, the score was simply listed as '< 40%'.

All generally accepted methods of black and white microfilming on a 35 mm negative film were imitated in this research. Microfilming was carried out conforming to Metamorfoze Preservation Microfilming Guidelines. Microfilms used for digitising have generally been produced in two steps: a 'first generation master' film is produced by microfilming the original. First generation films are used for long term preservation. A 'second generation' is produced by making a copy from the first generation microfilm before it is stored for preservation. Second generation films are used to produce user copies ('third generation') on microfilm or microfiche and increasingly, for making digital copies by scanning. Second generation films can have either a positive or a negative polarity.

In order to cover the whole range of first generation films that are used to produce second generation films, a number of features have been taken into account:

- 1) First generation by high or by low contrast microfilming;
- 2) Density of first generation microfilming (in case of high contrast microfilming). This works out into the following types of first generation microfilms. They are indicated by a combination of the contrast used in microfilming (HC= high contrast; LC= low contrast) and D-max, a certain indication of maximum/minimum density used in quality assurance. D-max is calculated as D-max minus D-min.

	Contrast	Density	D-max
HC 1.04	High contrast	Density 0.70-1.00	1.04
HC 1.35	High contrast	Density 1.00-1.30	1.35
HC 1.62	High contrast	Density 1.30-1.60	1.62
LC 1.24	Low contrast	Density 1.00-1.20	1.24

Table 1 Types of first generation microfilms in use

All testing was done on second generation films. For the test four second generation films were produced from every type of first generation film. One with a negative polarisation and three with a positive polarisation: one over-exposed, one normal-exposed and one under-exposed¹. The reason only normal

¹ See Addendum 1

exposure was used for negative films is that there is almost no contrast change in the image when this film is duplicated. Microfilms with a positive polarisation are much more susceptible to over- and underexposure (see below).

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.

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, is 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 master film.

For scanning the microfilms, three microfilm scanners were used; Zeutschel OM 1200, Zeutschel OM1400 and Imacon Flextight 848. Zeutschel OM 1200 and OM 1400 represent production scanning. Both scanners produce comparable results and are not distinguished in the test results. Imacon Flextight 848 represents slow scanning. For scanning directly from the original, a Zeutschel OS 10000 was used.

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)². Optimal adjustment means that the scanner is adjusted in such a way that patch A, with an accurately defined D-max in the master 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.)

We have also tried to show the entire tonal scale on the grey 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 42%. 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.

² See Addendum 2

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 Metamorfoze³ 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. In other words: it is difficult to judge what exactly happens with the weak grey 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 “show through” and when there are drawings with relevant information in the black parts.

Evaluation

Slow scanning versus production scanning

In our test results, the difference between slow scanning and production scanning of microfilms is not very large for modern newspapers. Based on our test results, it would seem that for modern material the disadvantages of slow scanning do not translate into a substantial higher quality. Only films with a negative polarity were tested, films with a positive polarity would require more research.

OCR accuracy in production scanning

In production scanning, low contrast microfilms produce overall the highest and most consistent OCR accuracy score. For old newspapers combined with high contrast microfilms OCR accuracy is very inconsistent, due to a combination of disadvantageous qualities of the microfilm in this workflow such as high contrast and a limited dynamic range. When scanning from high contrast microfilms it may be more advantageous to scan from original in those cases where the originals are relatively bad (e.g. show through) or that contain many dark areas that need to be preserved. Microfilms with a positive polarity produce a higher OCR accuracy score than microfilms with a negative polarity.

Calibration of scanners

Quality depends heavily on the scanners being used and the way they are adjusted optimally (calibrated) for each type of film. Calibration of microfilm scanners is a specialist task and should be left to specialists. Even so, the quality of microfilm scanners is often not sufficient for high quality scanning. In general one can say that the performance of microfilm scanners, when scanning microfilms with a negative polarity, is insufficient, while scanning microfilms with a positive polarity gives better results.

³ <http://www.metamorfoze.nl/en/methodiek/guidelines2006.pdf>

Detailed results

The following results can be seen as a guide on the quality one might expect when scanning from microfilm. Frequently occurring errors on microfilms such as gutter shadow, show through or skew will negatively affect the OCR accuracy further.

Modern newspaper	99.95%
Old newspaper	95.75%

Table 2: Scan of the original and OCR accuracy

	LC 1.24	HC 1.35	HC 1.62	HC 1.04
Modern newspaper	99.88%	Unknown*	94.34%	94.26%
Old newspaper	95.45%	95.35%	94.84%	81.54%

Table 3: Slow scanning. OCR accuracy - LC and HC with negative polarity

*Unknown: Preservation of all grey tones during slow scanning also has its disadvantages. Scanning is more difficult and time-consuming. While scanning 'HC 1.35 modern newspaper' the image has become too grey and could not be interpreted by the OCR engine.

	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%

Table 4: OCR accuracy - LC and HC with negative polarity

	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 - LC and HC with positive polarity and normal-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 - LC and HC with positive polarity and over-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%

Table 7: OCR accuracy - LC and HC with positive polarity and under-exposed

Conclusions

For both high and low contrast microfilms: scanning from a second generation film with positive polarity gives the gives a higher OCR accuracy score than a second generation with a negative polarity.

When scanning a film with positive polarity: low contrast films will give both a better OCR score than high contrast films and a more reliable workflow.

It seems necessary that the quality of the microfilm scanners have to be improved with regard to the scanning of microfilms with a negative polarity. As long as this is not the case second generation microfilms with positive polarity are best used for scanning.

Further research

Based on this study, the Koninklijke Bibliotheek has developed a microfilm target with a negative polarity for the calibration of microfilm scanners and a set of guidelines to use it. The calibration is build on correct sampling rate, sufficient sampling efficiency and limited geometric distortion for a specific used reduction ratio (this applies to any specific reduction ratio from 8:1 up to 21:1). The correct tonal capture is related to the microfilm. The correct tonal capture has to be checked by visually inspection of the scanned image of the microfilm.

As further research the Universal Test Target (UTT)⁴, a mounted target with size A-3, will be put on 35 mm and 16 mm microfilm to improve the microfilm scanning work flow. We expect that using UTT on microfilm will reduce the amount of time that is needed for quality assurance (research due to be completed in 2011).



⁴ <http://www.universaltesttarget.com/>

To improve the tonal capture of microfilm scanners different scanners from different vendors are being tested and the performance will be discussed with the vendors.

References

Dormolen, H. van (2008). Eindverslag OCR onderzoek. [research paper, available in Dutch only]

Dormolen, H. van (2009). Richtlijnen Scannen Preservation Microfilm Metamorfoze [guidelines microfilm preservation imaging, available in Dutch only]

Links

Dormolen, H. van (2006). "Metamorfoze preservation microfilming guidelines". Retrieved 18 January 2011 from: <http://www.metamorfoze.nl/en/methodiek/guidelines2006.pdf>

Dormolen, H. van (2006). "Metamorfoze preservation microfilming guidelines, blue prints and technical drawings". Retrieved 18 January 2011 from: <http://www.metamorfoze.nl/en/methodiek/guidelinesblueprints.pdf>

Dormolen, H. Van (2007). "Metamorfoze preservation imaging guidelines". Retrieved 18 January 2010 from: <http://www.metamorfoze.nl/en/methodiek/guidelinespijune07.pdf>

Imacon Flextight 848. Retrieved 18 January 2011 from: http://www.hasselbladusa.com/media/af870cd4-d074-4eff-b81c-2cfd18fb6cac-Flextight_848_English.pdf

Universal Test Target. Retrieved 18 January 2011 from: <http://www.universaltesttarget.com/>

Zeutschel OM 1200. Retrieved 18 January 2011 from: http://www.zeutschel.com/products/microfilm_scanner_om1200.html

Zeutschel OM 1400. Retrieved 18 January 2011 from: http://www.zeutschel.com/products/color_scanner_os14000_a0.html

Glossary

Contrast microfilming

High contrast microfilming:

Highly adopted method of microfilming from the original (first generation film). The gamma value (contrast factor) of these films has an average of 3. This means the contrast in the master negative is three times as high as the contrast in the original, resulting in a loss of 66.67% grey tones. Light grey areas especially suffer from this.

High contrast microfilms are divided into three groups according to density:

1. Average density, 1.00-1.30
2. High density; 1.30-1.60
3. Low density, 0.70-1.00

Low contrast microfilming:

This method of microfilming from the original (first generation film) has been developed by the Koninklijke Bibliotheek in their Metamorfoze preservation program in the period 1999-2006. Since 2003 Metamorfoze guidelines based on this method have been available.

The essence of low contrast microfilming is to preserve as many grey tones as possible. The gamma value (contrast factor) of these films have an average of 1.5 (for earliest films: 2). All films have a density of 1.00-1.20.

D-max	The measure of the greatest, or maximum, density of silver or dye image attained by a microfilm in a given sample. There are two ways to calculate D-max : D-max minus D-min (D-max – D-min) or D-max plus D-min (D-max + D-min), where D-min is base plus fog.
Density	The range of tones that can be captured by an imaging device. Optical density is measured on a scale of 0 for white to 4 for black
Kodak Gray Scale	The KODAK Grey Scale is a quality control device that helps compare tone values of reflection copy with its reproduction. Also helps find the correct exposure and processing conditions. Balances negatives and positives in a colour reproduction process and plots tone reproduction curves.
Microfilm (generations)	First generation microfilm: A first generation film is produced by microfilming from the original. First generation films are used for long term preservation and to produce second generation films. Second generation microfilm: A ‘second generation’ is produced by making a copy from a first generation microfilm before it is stored for preservation. Second generation films are used to produce user copies (‘third generation’) on microfilm or microfiche. And, increasingly, for making digital copies by scanning. Second generation films can have either a positive or a negative polarity.
OCR accuracy	As most commonly used, the term OCR accuracy refers to the number of correctly recognised characters/words in relation to the total number of characters/words in a document. OCR accuracy is assessed by comparing OCR results with a document’s ground truth.
Optical Character Recognition (OCR)	Optical character recognition is the mechanical or electronic translation of images of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text (Wikipedia).
Quality assurance	Quality assurance refers to planned and systematic production processes that provide confidence in a product’s suitability for its intended purpose. (Wikipedia [http://en.wikipedia.org/wiki/Quality_assurance])

Addendum 1

Overexposed	Normally exposed	Underexposed
-------------	------------------	--------------

Digitising surrogates:
Scanning from Microfilm

Negative HC 1.05	N.A.	<p>zullen verschijnen, de heeren Klomp en Jongensma protesteerden tegen de leiding en het beleid van het bestuur, de laatste noemde het zelfs een schamige manier, maar echt sociaal-democratisch, om het debat te smoren. Deze beide heeren vertrokken met enkele anderen onder hilariteit, waarna de heer Aukes zijn debat voortzette totdat een politieagent aankondigde dat het sluitingsuur was aangebroken en de voorzitter de vergadering sloot. Op het debat komen we morgen terug.</p> <p>DE ATHLETISCHE WEDSTRIJDEN. De uitgeleefde medailles voor de Zaterdag te houden wedstrijden op het Alemaria terrein zijn heden avond te bezichtigen in de etalage van den heer P. de Frenne, Langestraat. Morgenavond komen we neg even op de wedstrijden terug.</p> <p>DRANKWET. Had de makelaar J. P. Wagenaar te Alkmaar, eenigen tijd geleden succes, zoo schrijft men ons, doordat door zijne bemiddeling, de vergunning tot verkoop van sterke drank in het klein, op grond van art. 26 der Drankwet (maatschappelijk verkeer) in het persoon „Diligentia” te Alkmaar, werd overgeschreeven, thans heeft hij wederom succes behaald met een dergelijk geval, namelijk ten behoeve van den heer C. Kos, Damrak te Amsterdam, wiens vader ten vorigen jare was overleden, en in u toch de vergunning ten zijner name wordt overgeschreeven.</p>	N.A.
Negative HC 1.35	N.A.	<p>zullen verschijnen, de heeren Klomp en Jongensma protesteerden tegen de leiding en het beleid van het bestuur, de laatste noemde het zelfs een schamige manier, maar echt sociaal-democratisch, om het debat te smoren. Deze beide heeren vertrokken met enkele anderen onder hilariteit, waarna de heer Aukes zijn debat voortzette totdat een politieagent aankondigde dat het sluitingsuur was aangebroken en de voorzitter de vergadering sloot. Op het debat komen we morgen terug.</p> <p>DE ATHLETISCHE WEDSTRIJDEN. De uitgeleefde medailles voor de Zaterdag te houden wedstrijden op het Alemaria terrein zijn heden avond te bezichtigen in de etalage van den heer P. de Frenne, Langestraat. Morgenavond komen we neg even op de wedstrijden terug.</p> <p>DRANKWET. Had de makelaar J. P. Wagenaar te Alkmaar, eenigen tijd geleden succes, zoo schrijft men ons, doordat door zijne bemiddeling, de vergunning tot verkoop van sterke drank in het klein, op grond van art. 26 der Drankwet (maatschappelijk verkeer) in het persoon „Diligentia” te Alkmaar, werd overgeschreeven, thans heeft hij wederom succes behaald met een dergelijk geval, namelijk ten behoeve van den heer C. Kos, Damrak te Amsterdam, wiens vader ten vorigen jare was overleden, en in u toch de vergunning ten zijner name wordt overgeschreeven.</p>	N.A.
Negative HC 1.62	N.A.	<p>zullen verschijnen, de heeren Klomp en Jongensma protesteerden tegen de leiding en het beleid van het bestuur, de laatste noemde het zelfs een schamige manier, maar echt sociaal-democratisch, om het debat te smoren. Deze beide heeren vertrokken met enkele anderen onder hilariteit, waarna de heer Aukes zijn debat voortzette totdat een politieagent aankondigde dat het sluitingsuur was aangebroken en de voorzitter de vergadering sloot. Op het debat komen we morgen terug.</p> <p>DE ATHLETISCHE WEDSTRIJDEN. De uitgeleefde medailles voor de Zaterdag te houden wedstrijden op het Alemaria terrein zijn heden avond te bezichtigen in de etalage van den heer P. de Frenne, Langestraat. Morgenavond komen we neg even op de wedstrijden terug.</p> <p>DRANKWET. Had de makelaar J. P. Wagenaar te Alkmaar, eenigen tijd geleden succes, zoo schrijft men ons, doordat door zijne bemiddeling, de vergunning tot verkoop van sterke drank in het klein, op grond van art. 26 der Drankwet (maatschappelijk verkeer) in het persoon „Diligentia” te Alkmaar, werd overgeschreeven, thans heeft hij wederom succes behaald met een dergelijk geval, namelijk ten behoeve van den heer C. Kos, Damrak te Amsterdam, wiens vader ten vorigen jare was overleden, en in u toch de vergunning ten zijner name wordt overgeschreeven.</p>	N.A.
Negative LC 1.24	N.A.	<p>zullen verschijnen, de heeren Klomp en Jongensma protesteerden tegen de leiding en het beleid van het bestuur, de laatste noemde het zelfs een schamige manier, maar echt sociaal-democratisch, om het debat te smoren. Deze beide heeren vertrokken met enkele anderen onder hilariteit, waarna de heer Aukes zijn debat voortzette totdat een politieagent aankondigde dat het sluitingsuur was aangebroken en de voorzitter de vergadering sloot. Op het debat komen we morgen terug.</p> <p>DE ATHLETISCHE WEDSTRIJDEN. De uitgeleefde medailles voor de Zaterdag te houden wedstrijden op het Alemaria terrein zijn heden avond te bezichtigen in de etalage van den heer P. de Frenne, Langestraat. Morgenavond komen we neg even op de wedstrijden terug.</p> <p>DRANKWET. Had de makelaar J. P. Wagenaar te Alkmaar, eenigen tijd geleden succes, zoo schrijft men ons, doordat door zijne bemiddeling, de vergunning tot verkoop van sterke drank in het klein, op grond van art. 26 der Drankwet (maatschappelijk verkeer) in het persoon „Diligentia” te Alkmaar, werd overgeschreeven, thans heeft hij wederom succes behaald met een dergelijk geval, namelijk ten behoeve van den heer C. Kos, Damrak te Amsterdam, wiens vader ten vorigen jare was overleden, en in u toch de vergunning ten zijner name wordt overgeschreeven.</p>	N.A.

Addendum 2



Scanoutput from
2e generation microfilm
with negative polarity

