Eindverslag OCR onderzoek

Hans van Dormolen

Mei 2008

Managementsamenvatting

In dit onderzoek is de bruikbaarheid van 35 mm krantenmicrofilms als bronmateriaal onderzocht voor een werkproces waarbij een digitale afgeleide het eindproduct is. Deze afgeleide is bestemd voor gebruik op internet. Een belangrijk technisch aspect voor gebruik van afgeleiden op internet is goede OCR-nauwkeurigheid. Daarom is de bruikbaarheid van 35 mm krantenmicrofilms getoetst door middel van het bepalen van de uiteindelijke OCR-nauwkeurigheid. Naast de uiteindelijke OCR-nauwkeurigheid verschaft dit onderzoek inzicht in de factoren die van invloed zijn op de OCR-nauwkeurigheid in de gehele workflow van krant naar microfilm naar digitale afgeleiden. Met dit onderzoek kan een realistische inschatting gemaakt worden van de uiteindelijke OCR-nauwkeurigheid bij het scannen van krantenmicrofilms.

Bij het beoordelen van de OCR-nauwkeurigheid in de hier bovengenoemde workflow spelen de volgende factoren een rol:

- Optische kwaliteit van de originele kranten
- Technische kwaliteit van de 35 mm microfilms
- Technische prestaties van de huidige microfilmscanners

Om deze drie factoren goed te kunnen beoordelen zijn een moderne krant en oude krant geselecteerd en vervolgens op vier nationaal en internationaal gangbare methodieken verfilmd. Vervolgens zijn microfilm scanners (rolfilmscanners) voor het digitaliseren gebruikt die algemeen in de microfilm wereld gebruikt worden.
Om inzicht te verschaffen in de verhouding van de OCR-nauwkeurigheid verkregen door scannen van microfilm en scannen van origineel zijn ook scans van de originelen kranten gemaakt. De relatie tussen de kwaliteit van de microfilmscanner en uiteindelijke OCR-nauwkeurigheid is inzichtelijk gemaakt door de microfilms ook te scannen met behulp van een kwalitatief hoogwaardige scanner die veel in de fotografie wordt gebruikt.

De eerste generatie microfilm, de camerafilm, is bestemd voor conservering. Tijdens het scannen van een microfilm kan de film krassen. Daarom wordt bij het scannen van een microfilm in de regel nooit de camerafilm gebruikt maar de tweede generatiefilm. Deze film kan een positieve of negatieve polariteit hebben. De polariteit heeft grote invloed op de OCR-nauwkeurigheid. Daarom zijn in dit onderzoek tweede generatie microfilms gemaakt met zowel een positieve als negatieve polariteit.

Uit dit onderzoek blijkt:

- Scannen van een tweede generatie microfilm met positieve polariteit geeft de beste OCR-nauwkeurigheid.
- Laagcontrast verfilming is beter dan hoogcontrast verfilming
- De technische prestaties van de huidige microfilmscanners zijn zeer slecht met betrekking tot het scannen van microfilms met een negatieve polariteit

Met de in dit onderzoek verkregen inzichten wordt er een 35 mm scan-ijkstrook met bijbehorende richtlijnen gemaakt.

In dit onderzoek zijn uitsluitend 35 mm microfilms gebruikt. Om een origineel op 16 mm te kunnen verfilmen, i.p.v. op 35 mm microfilm, moet het origineel ongeveer twee keer zoveel verkleind worden. In dit onderzoek zijn de originelen 21 keer verkleind. De uitkomst van dit

onderzoek is daarom ook toepasbaar op 16 mm microfilms waarvan de originelen niet meer dan 12 keer zijn verkleind.
Vervolg onderzoek naar scherpte overdracht specifiek voor 16 mm microfilms met een verkleining groter dan 12 keer zal in de toekomst, indien noodzakelijk, worden uitgevoerd.

Eindverslag OCR onderzoek
Hans van Dormolen
mei 2008

Inhoudsopgave

Inleiding

Inleiding

In dit onderzoek is de bruikbaarheid van krantenmicrofilms als bronmateriaal onderzocht voor een werkproces waarbij een digitale afgeleide het eindproduct is. Deze afgeleide is bestemd voor gebruik op internet. Bij het gebruik van afgeleiden op internet speelt het gemak waarmee afgeleiden doeltreffend doorzocht kunnen worden een grote rol. Met doeltreffend doorzoeken wordt hier effectief doorzoeken op woorden bedoeld. Voor het zoeken op woorden wordt gebruik gemaakt van optische tekenherkenning (optical character regognition, OCR). Daarom is de OCR-nauwkeurigheid naast beeldkwaliteit zoals o.a. scherpte en helderheid een belangrijk aspect van de kwaliteit van de digitale afgeleide.
Hoe goed en hoe geschikt zijn nu de krantenmicrofilms die gedurende de laatste 40 a 50 jaar gemaakt zijn door de KB en andere instellingen en leveranciers? En hoe verhoudt deze kwaliteit van de microfilms zich tot de OCR-nauwkeurigheid? Om een antwoord op deze vraag te kunnen geven is dit OCR-onderzoek in maart 2007 gestart.

1. Kranten

In elk werkproces waarbij originelen op een andere drager worden gezet, speelt de kwaliteit van de originelen een belangrijke rol. In dit onderzoek zijn kranten het bronmateriaal. Kranten kunnen erg verkleurd zijn, ook kan de drukkwaliteit bij kranten onderling erg afwijken. Daarom zijn voor dit onderzoek twee verschillende krantenpagina's gebruikt: Een pagina van de Volkskrant en een pagina van de Amersfoortsche Courant.
De pagina van de Volkskrant is van recente datum, oktober 2006. Deze pagina staat voor kwalitatief goed drukwerk. Dat wil zeggen: strakke zwarte letter op een helder witte ondergrond.
De pagina van de Amerfoortsche Courant is veel ouder, namelijk van september 1892. Deze pagina staat voor kwalitatief zeer slecht drukwerk. Slecht drukwerk betekent in dit geval dat de pagina een zeer dunne nauwelijks leesbare letter heeft afgewisseld met een dikke zwarte, letter. De ondergrond is egaal verkleurd naar lichtbruin, de randen zijn iets donkerder bruin. De dunne, nauwelijks leesbare letters staan tot in de bruine rand.

2. Omvang test

Alle gangbare methodieken[1] van zwart/wit verfilmen op 35 mm negatieffilm zijn in dit onderzoek nagebootst.

Fouten die op kunnen treden tijdens de verfilming en die ongetwijfeld een negatieve invloed zullen hebben op de nauwkeurigheid zijn in dit onderzoek niet onderzocht. Het gaat hierbij om de volgende fouten: schaduw in de kneep, originelen zeer scheef verfilmen, verfilmen zonder glasplaat waardoor hinderlijke schaduw op de pagina kan ontstaan. Deze fouten komen veelvuldig voor op oude microfilms tot ongeveer 1995. Invoer van het gebruik van een glasplaat bij verfilming is afhankelijk van de verfilmer. Globaal zal de invoer rond 1995 liggen. Schaduw in de kneep kwam ook in de beginjaren van Metamorfoze verfilming herhaaldelijk voor. De oorzaak van schaduw in de kneep is vaak dat de krant te strak is gebonden. Het advies om, na toestemming van de eigenaar van de originelen, een te strak gebonden origineel uit elkaar te halen en dan te verfilmen is pas sinds 2005 in de richtlijnen[2] opgenomen.

[1] Laagcontrast en hoogcontrast met verschillende densiteiten.
[2] Richtlijnen Preservation Microfilming Metamorfoze, 2005 en 2006, Punt 2.10: Andere gebreken.

Het nadelige effect van doorslag[3] op de OCR-nauwkeurigheid is in dit onderzoek niet specifiek getest. Doorslag laat zich moeilijk kwantificeren.

De in dit onderzoek gepresenteerde percentages moeten als maximaal haalbaar gezien worden. Afhankelijk van bovengenoemde verfilmfouten en doorslag zal de OCR-nauwkeurigheid in praktijk iets lager liggen.

Als referentiescans zijn er ook scans gemaakt van de originelen krantenpagina's die in dit onderzoek zijn gebruikt. Deze scans zijn in mei 2007 gemaakt door Karmac met een Zeutschel scanner, de OS 10000.
De kwaliteit van deze scans voldoet absoluut niet aan de richtlijnen zoals omschreven in de conceptversie Preservation Imaging Metamorfoze. De technische kwaliteit van deze scans was in mei 2007 op een voor die tijd voor Karmac, en algemeen binnen Nederland, geldend niveau. Dit niveau zien we nu als kwalitatief laag. De technische afwijkingen in het beeld zijn echter marginaal van invloed op de in dit onderzoek onderzochte OCR-nauwkeurigheid. De hoeveelheid ruis in de images is vrij hoog[4]. Dit is nadelig voor de OCR-nauwkeurigheid. Anderzijds is de highlight gamma[5] aan de hoge kant. Dit heeft een positief effect op de OCR-nauwkeurigheid.

Enige nabewerking, zoals verscherpen en/of contrast verhogen van de images om de OCR-nauwkeurigheid positief te beïnvloeden is bewust niet uit gevoerd. Gekozen is voor een relatief goedkope standaard workflow.

Als vervolg op dit onderzoek zou het interessant kunnen zijn om met behulp van de bestaande images de effectiviteit van relatief simpele nabewerkingen, zoals verscherpen, te testen.

Alle rapporten die zijn opgesteld tijdens dit onderzoek van zowel de verfilming, het dupliceren, als van het scannen en ijken zijn bij ondergetekende in te zien. Ook kunnen deze rapporten op aanvraag worden verstrekt.

3. Microfilmscanners
Voor het scannen van de microfilms zijn Zeutschel microfilmscanners gebruikt: de OM 1200 en OM 1400. De prestaties van deze scanners zijn vergelijkbaar en worden in dit verslag aangeduid met de term productiescannen. Om de technische prestaties van deze productie scanner goed te kunnen beoordelen en om tevens te kunnen beoordelen wat we aan kwaliteit inleveren ten gunste van hoge productie (bulk) en snelheid zijn ook referentiescans van de in dit onderzoek gemaakte microfilms gemaakt. Deze referentiescans geven inzicht in wat technisch mogelijk is met betrekking tot informatieoverdracht van de in dit onderzoek gemaakte microfilms. De referentiescans zijn gemaakt met behulp van een Imacon Flextight 848. Deze Imacon scanner is absoluut niet geschikt voor het productiescannen van microfilms. Scannen met deze scanner is zeer tijdrovend. Deze scanner laat wel zien wat technisch mogelijk is met de gemaakte microfilms. De prestaties van deze scanner worden aangeduid met de naam slowscan.

[3] Doorslag is de benaming voor het verschijnsel dat de inkt op de voorkant van de pagina doorschijnt op de achterkant van de pagina. Doorslag kan nadelig zijn voor de OCR nauwkeurigheid.
[4] Standaard deviatie is gemeten op Q-13 en loopt op van 4,5 (vak A) naar 12,5 (vak 19).
[5] Highlight gamma per kleurkanaal: vak A-1 = 1,3, 1,4, 1,6, vak 1-2 = 1,3, 1,3, 13.

4. Werkproces

Het werkproces waarbij gebruikt wordt gemaakt van een microfilm als bronmateriaal voor het vervaardigen van een digitalegebruikers kopie ziet er in het algemeen als volgt uit:

1. De originelen worden verfilmd. De camerafilm heeft een negatieve polariteit en wordt de eerste generatie film of het moedernegatief genoemd. Deze film wordt, na diverse controles, voor lange termijn opgeslagen.
2. Voordat de eerste generatie film voor lange termijn wordt opgeslagen moet er een duplicaatfilm (tweede generatie film) van de eerste generatie film gemaakt worden. Deze tweede generatie film kan een positieve of negatieve polariteit hebben, afhankelijk van de gebruikte richtlijnen. Deze tweede generatie film wordt, afhankelijk van de polariteit, duplicaatnegatief of zilverpositief genoemd. Van deze film wordt het gebruikersexemplaar gemaakt, door middel van dupliceren of, zoals in deze test, door scannen van deze film.
3. Het gebruikersexemplaar kan zowel een analoge als een digitale afgeleide zijn, afhankelijk van de voorkeur van de opdrachtgever. Om een analoge gebruikersfilm te kunnen maken moet de tweede generatie film gedupliceerd worden. Voor het maken van een digitalegebruikers kopie moet de tweede generatiefilm gescand worden.

Eerste generatie microfilm

Alle gangbare methodieken van zwart/wit verfilmen op 35 mm negatieffilm zijn, zoals al eerder vermeld, in dit onderzoek nagebootst. Hiermee wordt hoogcontrast verfilmen met een gemiddelde, hoge en lage densiteit bedoeld en laagcontrast verfilmen.

Hoogcontrast verfilmen is een algemeen geaccepteerde en wijdverspreide methodiek van verfilmen. Deze manier van verfilmen kenmerkt zich door het hoge contrast van het moedernegatief. De gamma[6] waarde van deze eerste-generatiefilms is gemiddeld 3. Dit betekent dat het contrast in het moedernegatief gemiddeld drie maal zo hoog is als het contrast in het origineel. Dit betekent dat er 2/3de deel van de oorspronkelijke grijstonen verloren gaat, dit is een verlies van 66,66%. Vooral in de hoge lichten, dit zijn de licht grijze partijen (lichtgrijze letter), heeft dit verlies aan grijstonen grote gevolgen. Door het verlies aan grijstonen kunnen er gaten in deze lichtgrijze letters vallen. Gaten in letters zorgen voor minder goede OCR-nauwkeurigheid.

Hoogcontrast verfilming is in dit onderzoek onderverdeeld in drie groepen:

1. Gemiddelde densiteit[7], 1.00 – 1.30. Deze groep wordt in het onderzoek aangeduid met de afkorting HC 1.35[8].
2. Hoge densiteit, 1.30 – 1.60. Deze groep wordt in het onderzoek aangeduid met de afkorting HC 1.62[9].

---

[6] Gamma of gammawaarde is het contrast of de contrastfactor. De gamma geeft op eenvoudige wijze de verhouding aan tussen de contrastomvang van het onderwerp en die van het negatief (uit: Foto Techniek door P. Charpentier).
De gammawaarde is tangus van het lineaire gebied van de zogenaamde “S”curve. De gamma kan eenvoudig worden berekend met behulp van een Kodak Gray Scale. Zie punt 2.4 Kodak Gray Scale, *Richtlijnen Preservation Microfilming Metamorfoze, versie III, 2006.*
http://www.metamorfoze.nl/publicaties/richtlijnen/richtlijnenfeb06.pdf

[7] Opzichtdensiteit of densiteit is het logaritme van de opaciteit. Opaciteit is het tegenovergestelde van transparantie. Opaciteit is het opvallendlicht gedeeld door het gereflecteerde licht. Bijvoorbeeld: Opvallendlicht is 1 of 100%. Het gereflecteerde licht is 0,5 of 50%. De opaciteit is 1/0,5 = 2, Log 2 = 0.30. De densiteit of opzichtdensiteit = 0.30.

[8] De D-max (maximale densiteit-minimale densiteit) van vak A op de Kodak Gray Scale (Q-13) is 1.35. Op deze densiteit kan in diverse stappen in het productie proces (conversie slagen) geijkt worden. HC staat voor hoogcontrast.

3. Lage densiteit, 0.70 – 1.00. Deze groep wordt in het onderzoek aangeduid met de afkorting HC 1.04[10].

Deze indeling komt overeen met de hoogcontrast verfilming zoals die vanaf 1960 tot heden door de KB en andere instellingen nationaal en internationaal is uitgevoerd.

Laagcontrast verfilming is een methodiek van verfilmen die door Metamorfoze is ontwikkeld in de periode 1999 – 2006. Vanaf 2003 is laagcontrast verfilming omschreven in diverse microfilm richtlijnen van Metamorfoze. De essentie van laagcontrast verfilming is het zoveel mogelijk behouden van grijstonen in alle generaties microfilms. Met behulp van een grijstrap wordt het verlies aan grijstonen in diverse generaties inzichtelijk gemaakt. De eerste generatie laagcontrast films hebben tegenwoordig gemiddeld een gamma van 1,5. Het contrast in de films is dus gemiddeld 1,5 keer zo hoog als in het origineel. In de beginjaren van laagcontrast verfilming was het contrast van de films gemiddeld gamma 2. De densiteit is sinds 1999 niet gewijzigd. Alle films hebben een densiteit van 1.00 tot 1.20. Een densiteit onder de 1.00 wordt gezien als onderbelichting. Een densiteit boven de 1.20 wordt gezien als overbelichting. Kranten zijn tot 2006 altijd hoogcontrast verfilmd.

Laagcontrast verfilming wordt in dit onderzoek aangeduid met de afkorting:
LC 1.24[11]

Tweede generatie microfilm
Deze tweede generatie film wordt, afhankelijk van de polariteit, zilverpositief of duplicaatnegatief genoemd.

Zilverpositief. De reden voor de keuze van een positieve polariteit voor de tweede generatie film heeft te maken met het gebruik van een “slijtvaste” film met positieve polariteit in de leeszaal, de diazofilm. Deze diazofilm draait de polariteit niet om bij het dupliceren. Positief blijft positief. Deze film heeft een hoog contrast, gamma groter dan 2. Vanwege het hoge contrast van deze film is sinds 2005 het gebruik van deze film afgeraden. De zilverpositieffilm heeft ook een gamma van rond de 2. De kopieeromvang van deze film is vrij beperkt, 3 stoppen, densiteit 0.90, of te wel 1:8. De moederfilm heeft veelal een contrastomvang groter dan 3 stoppen namelijk 4 tot 5 stoppen, 1:16 tot 1:32. Een altijd geldende standaard voor het maken van een zilverpositieffilm is niet eenvoudig te maken en daarom is er dan ook nooit een standaard in richtlijnen opgenomen. Met als gevolg dat zilverpositieffilms afwisselend goed, iets over -en iets onderbelicht kunnen zijn, afhankelijk van de dekking van de moederfilm. Deze drie varianten zijn in het onderzoek nagebootst. De uitslagen van deze drie varianten zijn bij elkaar opgeteld en vervolgens door 3 gedeeld. Om zo een gemiddelde te kunnen vaststellen.

Duplicaatnegatief. Sinds 2005 wordt een negatieffilm als tweede generatie film gebruikt. Voor deze film is in die tijd gekozen omdat de gamma van deze film, mits goed belicht en ontwikkeld, in het lineaire gebied 1 is. Dit betekent dat er geen contrast verandering plaats vindt in het beeld met dupliceren op deze film. Met andere woorden: Alle informatie die aanwezig is in de eerste generatie microfilm blijft behouden in deze tweede generatie microfilm. De D-max in de eerste generatie microfilm loopt iets terug in de tweede generatie.

---

[9] De D-max van vak A op de Kodak Gray Scale is 1.62. Op deze densiteit kan in diverse stappen in het productie proces (conversie slagen) geijkt worden.
[10] De D-max van vak A op de Kodak Gray Scale is 1.04. Op deze densiteit kan in diverse stappen in het productie proces (conversie slagen) geijkt worden.
[11] De D-max (maximale densiteit) van vak A op de Kodak Gray Scale (Q-13) is 1.24. Op deze waarde kan in diverse stappen in het productie proces (conversie slagen) geijkt worden. LC staat voor laagcontrast.

Dit geldt enkel voor het densiteitsgebied boven de 1.00. Een ander voordeel naast gamma 1 is dat goed belichten en ontwikkelen van deze tweede generatie microfilm zich gemakkelijk laat omschrijven in een richtlijn door vastlegging van de D-min[12]. Zie Richtlijnen Preservation Microfilming Metamorfoze, versie III, punt 2.2 film type en generatie.

Overzicht van de gemaakte films en uitgevoerde beoordelingen

	Moeder negatief	Duplicaat negatief	Zilverpositief
Volkskrant	4 films	4	12
Amersfoortsche Courant	4 films	4	12
Totaal	8 films	8 films	24 films
Beoordeeld			
Slowscan		8 films	geen
Productiescan		8 films	24 films

5. Scannen microfilms

Voordat de tweede generatie microfilms zijn gescand is de scanner optimaal afgesteld (geijkt) per filmsoort[13] met behulp van de D-max op vak A van de Kodak Gray Scale. Optimaal afstellen is de scanner zo af stellen dat de aangegeven D-max consequent wordt vertaald rond pixel waarde 242[14]. Daarnaast is geprobeerd om de overgang tussen vak A en 1 zo realistisch mogelijk te vertalen[15]. Ook is geprobeerd om de gehele toonschaal op de grijstrap van D-max tot D-min te laten zien. Bij het scannen van de negatieffilms is gebleken dat de geteste microfilmscanners zeer beperkt zijn in te stellen. Een gamma-instelling (contrastinstelling) voor het optimaal scannen van negatieffilms is niet te maken. Dit gebrek maakt de microfilmscanners incapabel om de contrastovergangen in het densiteits gebied van grofweg 1.10 tot 0.60, vak A op de Kodak Gray Scale tot vak 3, correct te registreren. Van de contrastovergang tussen vak A en vak 1[16] blijft nog maar 7,6% over. De berekening van dit percentage is gebaseerd op densiteitsverschil in de hoge lichten van een opzichtmodel met een positieve polariteit. Bij het scannen van een film met negatieve polariteit bevinden de hogelichten zich in de donkere partijen. Het verschil in pixelwaarden in de donkere partijen is altijd geringer dan in de hogelichten. Een densiteitverschil van 0.17 in de donkere partijen (opzichtdensiteit 1.78 – 1.95) resulteert in een pixelwaarden verschil van 7 punten (bij monitor gamma 2.2). In percentages is de contrastovergang dan 3/7 is 42%. Dit is ook een slechte contrastoverdracht. Ook al omdat we in deze berekening uitgaan van een D-max gedefinieerd als 1.95. Op de negatieffilm echter is de D-max slechts 1.10.

[12] D-min, minimale densiteit, grondsluier.
[13] Film soorten en bij behorende ijkstroken HC 1.35, HC 1.62, HC 1.04, LC 1.24.
[14] Pixelwaarde 242 komt overeen met opzichtdensiteit 0.05 vertaald naar een 8 bit ruimte met monitor gamma 2.2. Formule: pixelwaarde =255(reflectiewaarde^1/monitor gamma).
[15] LC 1.24. De D-max van vak A in de tweede generatie negatiefilm is een densiteit van 1.10. We vertalen deze waarde naar wit, naar een pixelwaarde die ligt rond de 242. Vak 1 in de tweede generatie negatieffilm heeft een densiteit van 0.93. Dit is een densiteits verschil van 0.17 punten. In een opzicht model staat een densiteits verschil van 0.17 punten gelijk aan een pixelwaarde verschil van 39 punten. Nu meten we in Photoshop met pipet (5 x5 pixels) slechts 3 punten verschil. Het nu gemeten verschil gedeeld door het theoretische verschil, 3/39 is 0,076.
Zie berekening *highlight gamma*, punt 2.7.1. *Richtlijnen Preservation Imaging Metamorfoze*.
[16] Densiteit op de film van vak A en 1 is respectievelijk 1.10 en 0.93

Algemeen kan gesteld worden dat de contrastoverdracht bij het scannen van een microfilm met negatieve polariteit met behulp van de geteste microfilmscanners onvoldoende is. Het directe gevolg van deze onvolledige contrastoverdracht is digitale bestanden met slechte OCR-nauwkeurigheid.

De contrastoverdracht van de referentie scanner, de Imacon Flextight 848, is, na ijking, acceptabel. De contrastovergang van vak A en 1 op film *LC 1.24 neg* wordt weergegeven met 31 pixelwaarden. Dit is een contrastoverdracht van 79%[17]. Dit is acceptabel. In de Richtlijnen Preservation Imaging Metamorfoze wordt een highlight gamma van 0,8 tot 1,08 (80% - 108%) als tolerantie waarde gegeven. Correcte contrastoverdracht staat borg voor goede OCR-nauwkeurigheid.

De contrastoverdracht van het scannen van een microfilm met positieve polariteit met de geteste microfilmscanners, de Zeutschel OM 1200 en 1400, is moeilijk in een cijfer uit te drukken. Dit komt deels door dat de kopieeromvang van de positieffilm beperkt is. Het verschil tussen vak A en 1 is veelal nauwelijks zichtbaar op een film met positieve polariteit. In pixelwaarden is dit verschil dan ook nihil. Toch blijkt, na visuele inspectie, dat er op de film geen of nauwelijks informatie verloren is gegaan. Met andere woorden: Het is moeilijk in te schatten wat er precies met de zwakke grijstonen van de letters gebeurt. Het verschil tussen vak A en vak 2 is bij film *LC 1.24 pos* na scannen 54 punten. De highlight gamma tussen vak A en 2 is dan 1,17 de contrastoverdracht is dan 117%. Dit zegt niet zoveel omdat niet duidelijk is wat er precies verloren is gegaan aan informatie tussen vak A en 1. Algemeen kan wel gezegd worden dat de contrastoverdracht tussen een film met positieve polariteit en digitale afgeleide in harmonie is. Hiermee wordt bedoeld dat de verschillen in pixelwaarden in de hogelichten groot zijn en in de donkere partijen klein zijn. De doortekening in de donkere parijen in de digitale afgeleide is wel een stuk geringer dan op de positieffilm. Dit kan problemen geven indien informatie in de donkere partijen relevant is, zoals in de combinatie tekst en doorslag en bij tekeningen met informatie in de donkere partijen. Dit kan ook problemen veroorzaken bij vet gedrukte letters, letters kunnen dichtlopen. Het gevolg hiervan is minder goede OCR nauwkeurigheid.

6. OCR-nauwkeurigheid

Bij het vaststellen van de OCR-nauwkeurigheid zijn de goed en fout weergegeven karakters per pagina geteld. Vervolgens is met behulp van het totale aantal karakters van dezelfde pagina het nauwkeurigheids percentage berekend, zie Tabel 1 t/m 4. Het tellen van de goed en fout weergegeven karakters in de scans van de pagina van de oude krant is regelmatig gestaakt. Verdere telling werd bij deze scans niet meer zinvol geacht omdat het aantal fout weergegeven karakters zeer hoog was. De OCR-nauwkeurigheid van deze scans ligt zeer laag. Hoe laag precies weten we niet. We weten wel dat het nauwkeurigheids percentage onder de 40% ligt en soms nog lager. In dit onderzoek is OCR pakket Abbyy FineReader 8.0 Corporate Edition gebruikt.

Tabel 1. **Scannen van origineel. Scanner Zeutschel OS 10000**

	OCR-nauwkeurigheid
Volkskrant	99,95 %
Amersfoortsche Courant	95,75 %

[17] Pixelwaarde vak A is 242, pixelwaarde vak 1 is 211. Het verschil is 31. Highlight gamma is 31/39 is 0,79.

Tabel 2. **Slowscan laagcontrast en hoogcontrast 2e generatie met negatieve polariteit.**

	LC 1.24 neg	HC 1.35 neg	HC 1.62 neg	HC 1.04 neg
Moderne krant	99,88 %	onbekend	94,34 %	94,26 %
Oude Krant	95,45 %	95,35%	94,84 %	81,54 %

Onbekend: Het scannen met behoud van alle grijstonen heeft ook nadelen. Het scannen is moeilijker en kost meer tijd. Bij het scannen van negatief "HC 1.35 moderne krant" is de afgeleide te grijs geworden. Het OCR pakket, Abbyy FineReader 8.0 Corporate Edition, kon hier niets mee. In een vervolgonderzoek wordt dit negatief opnieuw gescand en beoordeeld.

Tabel 3. **Productiescan laagcontrast en hoogcontrast 2e generatie met negatieve polariteit.**

	LC 1.24 neg.	HC 1.35 neg	HC 1.62 neg	HC 1.04 neg
Moderne krant	**93,11 %**	**96,69 %**	**93,71 %**	**97,44 %**
Oude krant	**< 40%**	**< 40%**	**< 40%**	**< 40%**

In deze tabel is de beperking van de microfilmscanner duidelijk waarneembaar. Alle opnamen van de oude krant scoren zeer slecht. In punt 5. *Scannen microfilms* wordt deze slechte score technisch onderbouwd. Opvallend is de hoge nauwkeurigheidsscore van de combinatie moderne krant en hoogcontrast verfilming.

Tabel 4. **Productiescan LC en HC 2e generatie met positieve polariteit.**

	LC 1.24 pos		HC 1.35 pos		HC 1.62 pos		HC 1.04 pos	
Moderne krant	Norm. belicht	**99,65%**	Norm. belicht	**99,72%**	Norm. belicht	**98,30%**	Norm. belicht	**99,54%**
	Over belicht	**99,49%**	Over belicht	**99,51%**	Over belicht	**99,07%**	Over belicht	**98,92%**
	Onder belicht	**99,47%**	Onder belicht	**99,73%**	Onder belicht	**97,71%**	Onder belicht	**99,76%**
Oude krant	Norm. belicht	**95,39%**	Norm. belicht	**93,97%**	Norm. belicht	**94,42%**	Norm. belicht	**88,06%**
	Over belicht	**94,27%**	Over belicht	**93,82%**	Over belicht	**< 40%**	Over belicht	**< 40%**
	Onder belicht	**94,22%**	Onder belicht	**< 40%**	Onder belicht	**< 40%**	Onder belicht	**92,68%**

Uit tabel 3 en 4 blijkt

- Algemeen geldt dat het productiescannen van een film met positieve polariteit een betere OCR nauwkeurigheidsscore oplevert dan het productiescannen van een film met negatieve polariteit.
- Laagcontrast verfilming behaalt de hoogste nauwkeurigheidsscore.
- Laagcontrast verfilming levert een beduidend stabielere en betrouwbaardere workflow op dan hoogcontrast verfilming.
- De combinatie oude krant en hoogcontrast verfilming levert een onbetrouwbare OCR nauwkeurigheidsscore op. Dit is te wijten aan de combinatie van nadelige filmeigenschappen zoals hoogcontrast (eerste generatie film) en beperkte kopieeromvang (tweede generatie film). Zie voor verdere technische uitleg punt 5.0 *Scannen microfilms*.

Het is zeer moeilijk om een stabiele en betrouwbare workflow op te bouwen met een tweede generatie film met een beperkte kopieeromvang en een gamma van 2, zoals de gebruikte film met positieve polariteit. Zover ik weet bestaat er geen tweede generatie microfilm met positieve polariteit en betere film eigenschappen, zoals een iets grotere kopieeromvang dan 3 stoppen en iets lagere gamma dan 2. Het lijkt mij daarom absoluut noodzakelijk dat de kwaliteit van de microfilmscanners snel wordt verbeterd met betrekking tot het scannen van microfilms met een negatieve polariteit.

7. Conclusie
Voor zowel laag- als hoogcontrast microfilms geldt: Scannen van een 2e generatie film met een positieve polariteit geeft een betere OCR-nauwkeurigheid dan scannen van een 2e generatie film met een negatieve polariteit.

Voor het scannen van een tweede generatie film met positieve polariteit geldt:

- Laagcontrast verfilming geeft een betere OCR-nauwkeurigheid dan hoogcontrast verfilming.
- Laagcontrast verfilming levert een beduidend stabielere en betrouwbaardere workflow op dan hoogcontrast verfilming.

8. Advies
Zolang de scankwaliteit van de geteste microscanners niet is verbeterd, is het zinvol om uitsluitend te scannen van een tweede generatie microfilm met positieve polariteit.

Voor scannen van microfilms waarbij de eerste generatie microfilms hoogcontrast zijn geldt: Zit er in de originelen veel essentiële informatie in de donkere partijen, met andere woorden: zijn de originelen overwegend slecht, is er veel doordruk of is er spraken van afbeeldingen met veel informatie in de donkere partijen die behouden moet blijven, dan is het zinvoller om de originelen te scannen.

9. Vervolg
Met de verschillende bedrijven (o.a. Leaf en GMS) wordt onderzocht of het mogelijk is om een betere scanner te ontwikkelen voor het scannen van microfilms met negatieve polariteit.

Microfilmscanners en filmscanners van andere fabrikanten zullen in de toekomst worden getest op contrastoverdracht met behulp van de in dit onderzoek gebruikte scanijkstrook met positieve en negatieve polariteit. Een voorwaarde voor het testen van filmscanners is dat automatische doorvoer van 35 mm film redelijkerwijs gefaciliteerd kan worden.

Met de in dit onderzoek verkregen inzichten wordt er een 35 mm scan-ijkstrook met bijbehorende richtlijnen gemaakt. Deze scan-ijkstrook wordt in twee polariteiten, positief en negatief, uitgevoerd voor zowel hoogcontrast als laagcontrast microfilms. Indien noodzakelijk wordt de scan-ijkstrook ook in 16 mm uitgevoerd.

10. Toepassing en gebruik onderzoeksresultaten
Voor alle 35 mm microfilms die er in de loop der jaren gemaakt zijn is het mogelijk om met behulp van de verkregen uitslagen in dit onderzoek een realistische inschatting te maken van de maximaal haalbare OCR-nauwkeurigheid. Zoals eerder opgemerkt is de OCR-nauwkeurigheid afhankelijk van een aantal factoren. Al deze factoren moeten beoordeeld worden en pas dan kan een realistische inschatting worden gemaakt. De factoren die beoordeeld moeten worden zijn:

- Optische kwaliteit originelen (met behulp van visuele inspectie tweede generatie microfilm)
- Verfilmfouten (met behulp van visuele inspectie tweede generatie microfilm)
- Polariteit tweede generatie microfilm (met behulp van visuele inspectie tweede generatie microfilm / database)
- Hoogcontrast / laagcontrast eerste generatie microfilm (kranten zijn tot 2006 hoogcontrast verfilmd. Bij twijfel overleg met de kwaliteitmanagers)
- Densiteit eerste generatie hoogcontrast microfilm (met behulp van de kwaliteitsmanagers)

11. Microfiches en 16 mm microfilms

In het OCR onderzoek is de relatie tonale weergave in verschillende generaties en OCR nauwkeurigheid onderzocht. Uitsluitend zijn 35 mm films gebruikt met verkleiningsfactor 21. De in dit onderzoek geconstateerde problematiek met betrekking tot de tonale weergave zal voor 16 mm films en microfiches vergelijkbaar zijn. Maar de verkleiningsfactor en de grootte van de afbeelding speelt altijd een rol bij informatie overdracht, zoals bij het scannen van een microfilm. Het beeld op een 16 mm is grofweg de helft zo groot als het beeld op een 35 mm film. Volgens de richtlijnen microverfilming van Metamorfoze mochten tot 2006 boeken op 16 mm worden verfilmd. Een regel voor het verfilmen op 16 mm is wel dat de buitenmaat van de boeken (opengeslagen) kleiner moet zijn dan A4. De reductiefactor die dan gebruikt werd is 12. Dit komt overeen met reductiefactor 21 voor 35 mm. Daarom kunnen we nu, vooruit lopend op een specifiek 16 mm scherpte scantest, voorzichtig stellen dat voor het scannen van 16 mm film geldt: Indien de verkleiningsfactor niet groter is dan 12 zijn de in dit onderzoek verkregen uitslagen van toepassing.
Dit geldt niet voor microfiches. Veelal, niet altijd, zijn microfiches een afgeleide van 16 mm moedernegatieven. Het maken van deze afgeleiden wordt aangeduid met de term: conversie. De methodieken die voor deze conversieslag worden gebruikt kunnen per leverancier verschillen. Het scherpteverlies dat optreedt tijdens deze conversieslag verschilt per methodiek, maar kan aanzienlijk zijn. De verkleiningsfactor waarmee de 16 mm beelden op een microfiche gezet worden kunnen veelal aangepast worden. Algemeen geldt dat microfiches ongeschikt zijn om van te scannen.

In de 16 mm workflow van Metamorfoze werden aanvankelijk geen tweede generatie rolfilms gemaakt. Er werd een “moederfiche”gemaakt. Dit “moederfiche”werd gedupliceerd en het duplicaatfiche was het gebruikers exemplaar.

Hans van Dormolen
Mei, 2008.