# Zoek ‘wortel’!

## Inhoud video lastig te vinden

Metadateren van bewegende beelden, een gat in de markt voor de slimme catalogiseerder? Waarom niet? In de toekomst zijn we nog veel meer dan nu visueel georiënteerd, en de computer kan geen beelden ‘zien’. Of toch?

**Door Gerard Bierens**

Catalogiseren kun je automatiseren. De kans is groot dat je met zo’n openingszin flink wat collega’s tegen het zere been schopt. Jawel, catalogiseren is een vak, een specialisme waar je een behoorlijk aantal jaren voor gestudeerd moet hebben. En dan ben je er nog niet, want het vak echt leren doe je natuurlijk pas in de praktijk. Catalogiseerders hebben in de regel dan ook weinig op met al die web 2.0-types die doodleuk verkondigen dat wat hen betreft het publiek best zelf in staat is om goede trefwoorden toe te kennen aan de boeken die ze in de bibliotheekcatalogus hebben gevonden.

Toch is dit een ontwikkeling die – los van de catalogiseerders – in de bibliotheekwereld verder nauwelijks nog serieuze weerstand kent. Steeds meer openbare bibliotheken voegen de *social tagging*-module MyDiscoveries toe aan hun Aquabrowser, universiteitsbibliotheken verrijken de catalogus met trefwoorden uit LibraryThing, Delicious of andere bronnen. En wie op het web ziet hoe vrijwel ieder platform zijn gebruikers de vrijheid geeft om naar hartelust te taggen kan toch wel stellen dat de catalogiseerder in de strijd met de massa op een zo goed als onoverbrug-

bare achterstand is gekomen. Bovendien gelden op het web nu eenmaal andere kwaliteitsnormen, de *wisdom of the crowd* is lang niet perfect, maar de gemiddelde gebruiker kan best leven met een foutmarge van 20 procent.
Binnen de bibliotheekwereld is er heus nog wel waardering voor het specialistische werk van de catalogiseerder, maar omdat bibliotheken zelf steeds meer moeite moeten doen om hun meerwaarde aan te tonen, gaan ze zich steeds meer richten op meer in het oog springende activiteiten. En met de verwachting dat binnen een paar jaar de meeste content toch wel in digitale vorm beschikbaar zal zijn, is de traditionele bibliotheekcatalogus in te ruilen voor zoiets als een *full-text metadiscovery search engine.*

## STEEDS MINDER TEKST

Over de juiste benaming van een dergelijke tool zullen de experts het vast nooit eens worden, maar over de functionaliteiten wel. Deze systemen zullen steeds beter worden in het analyseren, kwalificeren en rubriceren van content, steeds beter ook in het interpreteren en vertalen van zoekvragen. Geheel geautomatiseerd, eerst nog op basis van frequentietellingen, thesaurustermen en zoekalgoritmen, later op basis van zoiets als *kunstmatige zoekintelligentie*. Voor de catalogiseerder valt in zo'n toekomstscenario nog weinig eer te behalen.
Wie de tijd neemt om toch even wat langer over het geschetste toekomstscenario na te denken, zal zich realiseren dat dit beeld niet helemaal correct is. In dit scenario wordt namelijk uitgegaan van een volledig tekst-georiënteerde omgeving, terwijl onze dagelijkse omgeving juist steeds minder tekst en steeds meer audiovisuele elementen bevat.
Het voorbeeld bij uitstek is natuurlijk het immens populaire YouTube. Om een idee te geven van de omvang: één persoon zou van geboorte tot hoogbejaarde leeftijd non-stop moeten blijven kijken om het materiaal dat op één dag geüpload is naar YouTube langs te zien komen. Maar ook andere platformen als iTunes, Flickr en Last.fm exploiteren gigantische audiovisuele collecties.
Ook in Nederland gaat het om grote getallen. Zo beheert het Instituut voor Beeld en Geluid een videocollectie van vele duizenden uren met uitzendingen van de publieke omroepen. Pas een relatief klein deel van die content is online beschikbaar op uitzendinggemist.nl, maar het platform voorziet duidelijk in een snel groeiende behoefte. In 2009 werden er bijvoorbeeld ruim 150 miljoen video's bekeken, 25 procent meer dan het jaar ervoor. En om even een andere trend aan te stippen, ook in 2009 werden meer dan 2,7 miljoen video's van uitzendinggemist.nl bekeken via mobiele devices. Indrukwekkende cijfers, maar dat terzijde.

## FRAGMENT BOER ZOEKT VROUW

Interessanter is de vraag in hoeverre die overvloed aan audiovisueel materiaal eigenlijk *inhoudelijk* ontsloten is. Aan audiovisueel materiaal wordt nauwelijks beschrijvende metadata toegevoegd. Een titel, samenvatting en wat trefwoorden, en dat is het wel zo ongeveer. Nu is het vrij eenvoudig om bijvoorbeeld aflevering 4, jaargang 2009 van *Boer zoekt vrouw* via uitzendinggemist.nl terug te vinden, want verwacht mag worden dat zulke elementaire metadata bij de betreffende video wordt opgeslagen. Maar zodra de zoekvraag ingaat op de inhoud van de video wordt het al snel een *mission impossible*. Probeer bijvoorbeeld maar eens bij diezelfde aflevering 4 van *Boer zoekt vrouw* te zoeken naar het fragment waarin een wortel in beeld komt. Gegarandeerd zal zo'n zoekactie nul treffers opleveren, terwijl die wortel toch echt in de video voorkomt. Wie 'm wil vinden, zal toch echt de video aandachtig moeten bekijken. Zou het hier een geschreven tekst betreffen, dan zou een zoekmachine vrijwel direct verwijzen naar exact de juiste pagina en exact de juiste alinea waarin het woord 'wortel' staat geschreven. Hetzelfde probleem is aan de orde voor gesproken tekst. Probeer die ene leuke uitspraak van boer Wietse over tafelmanieren maar eens terug te vinden als je niet meer zeker weet in welke aflevering dat fragment zit. Daar is geen beginnen aan.
Audiovisueel materiaal is op dit moment dus maar heel beperkt toegankelijk, in essentie nog veel beperkter dan de wijze waarop materialen in bibliotheekcatalogi ontsloten worden. Je zou bijna denken dat hier een uitgelezen kans ligt voor catalogiseerders, wier toekomstperspectief zojuist toch al een flinke knauw heeft gekregen.
Metadateren van bewegende beelden, een gat in de markt voor de catalogiseerder? Waarom ook niet, het menselijk brein heeft immers geen enkele moeite om die wortel in bewegende videobeelden te kunnen onderscheiden. Sterker nog, we zien binnen dezelfde seconde ook dat de zon schijnt, dat het gras groen is en dat het stevig waait.
Voor een computerprogramma ligt dat anders, die 'ziet' feitelijk geen videobeelden, maar alleen een patroon van enen en nullen, bits en bytes. Om daarmee geautomatiseerd objecten in bewegende beelden te kunnen onderscheiden lijkt een welhaast onmogelijke opgave.
Maar helaas, het alternatief is echter net zo onmogelijk. Reken maar eens uit hoeveel tijd er nodig zou zijn om alle objecten in een video van één uur handmatig te annoteren. Precies, minimaal één uur, maar twee à drie is meer waarschijnlijk. Met de stortvloed aan audiovisueel materiaal is daar werkelijk geen beginnen aan, hoe spijtig ook voor het toekomstperspectief van de catalogiseerder.

## AL TIEN JAAR ONDERZOEK

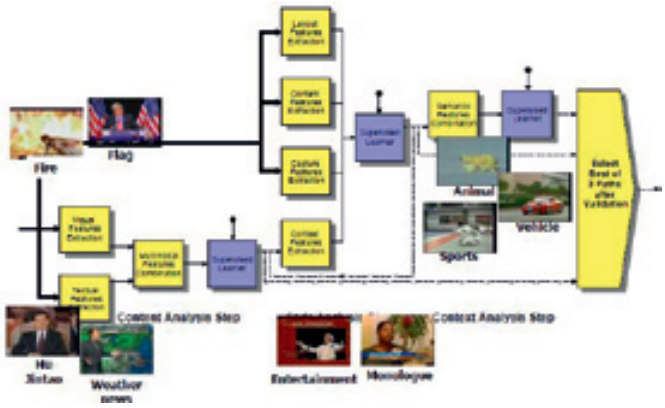
Aan de ene kant kampen we dus met onvoldoende menselijke capaciteit, aan de andere kant is een computer niet in staat om beelden te kunnen 'zien'. Daarmee lijkt de inhoudelijke ontsluiting van audiovisueel materiaal in een impasse te zijn beland. Maar impasses zijn er juist om te doorbreken, precies daarin zit voor wetenschappers vaak de uitdaging. Ook hier is dat het geval. Wetenschappers van de Universiteit van Amsterdam doen al tien jaar onderzoek naar kunstmatige beeld- en spraakherkenning. Een speciaal daarvoor opgericht lab, het Intelligente Systemen Lab Amsterdam (ISLA) houdt zich bezig met cognitieve beeldbewerking, visuele emotieherkenning, machinaal herkennen van voorwerpen en beeldzoekmachines. Vooral dat laatste trekt de aandacht, de semantische TRECVID

Video Search Engine heeft bijvoorbeeld in 2009 een aantal internationale competities op dit gebied gewonnen.
Belangrijk onderdeel is de *semantic pathfinder algorithm*. Hoewel het gaat om extreem complexe en futuristische technologie die voor een leek moeilijk lijkt te doorgronden, zijn de basisprincipes niet heel ingewikkeld. Een visualisatie[1]:


1. *Choose concept*: een willekeurig videobeeld wordt als concept vastgelegd, in dit geval een boot.
2. *Label and divide images*: de boot wordt aan de hand van een test- en trainingsset voorzien van het meest waarschijnlijke label 'boat'.
3. *Build model from training set*: aan de hand van vergelijkbare materialen in de trainingsset wordt er een kleurpatroon gekoppeld aan het model.
4. *Test model on test set*: het model wordt vergeleken met de inhoud van de test set, het percentage geeft aan in welke mate er een match is met het model.
5. *Find best path through multiple models*: meerdere modellen worden inhoudelijk vergeleken op basis van de daarin aanwezige content. Het resultaat met het hoogste scoringspercentage bepaalt de uiteindelijke rubricering.

Het baanbrekende werk van ISLA krijgt dus veel waardering in wetenschappelijke kring, maar blijft ook daarbuiten niet onopgemerkt. Dat komt vooral ook door de strategische samenwerkingsverbanden die het ISLA aangaat. Bijvoorbeeld met het eerder genoemde Nederlands Instituut voor Beeld en Geluid, maar ook met Human Media Interaction (spraakherkenningexperts van de Universiteit Twente), het CWI en de NPS. Dr. Theo Gevers, ISLA-onderzoeker bij de Universiteit van Amsterdam, zegt er het volgende over: "Audiovisuele annotatie is inderdaad een belangrijk nieuw en uitdagend onderdeel van archivering. Het Nederlands Instituut voor Beeld en Geluid is daarmee bezig en wij bieden hen softwareoplossingen aan. Niet alleen voor Beeld en Geluid belangrijk maar ook voor AT5, BCC en andere organisaties".
Samen met Beeld en Geluid en andere partners heeft ISLA onlangs de website Hollands Glorie op Pinkpop[2] gelanceerd. De binnen ISLA ontwikkelde beeldherkenningssoftware MediaMill[3] wordt hier gebruikt om concepten in concertvideo's te detecteren. Er zijn voor dit project meer dan twee miljoen videoframes geanalyseerd, computers met een enorme rekenkracht hadden daar nog altijd 3600 uur voor nodig.
De bedoeling is dat wanneer er bijvoorbeeld een gitarist in beeld is, dit wordt opgemerkt en vastgelegd door de beeldherkenningssoftware. Komt de drummer in beeld, dan wordt het label 'drummer' vastgelegd in de tijdlijn. Ook zijn de backstage-interviews met artiesten geïndexeerd met spraakherkenningssoftware. De resultaten daarvan worden aangeboden in een aanklikbare trefwoordenwolk. De beeld- en spraakherkenningstechnologie achter Hollands Glorie op Pinkpop ziet er indrukwekkend uit, maar werkt zeker nog niet foutloos. Het gaat de goede kant op, maar de technologie is zeker nog niet op het punt waarop alle objecten in beeld foutloos herkend en correct worden geannoteerd. Om bij het Pinkpop-voorbeeld te blijven, de toepassing is op dit moment redelijk in staat om automatisch een zanger te herkennen. Het vaststellen dat het gaat om de zanger Herman Brood is vele malen complexer.
Gevers beaamt dat: "Automatische herkenning van bijvoorbeeld gezichten *in the wild* is een moeilijk probleem. Dat zie ik niet zo snel gebeuren. Ook analyse van menselijk gedrag staat nog, qua nauwkeurigheid, in de kinderschoenen, maar men is er wel sterk mee bezig".


*Beeld- en spaakherkenning in actie op Hollands Glorie op Pinkpop. Giel Beelen interviewt Moke.*




*De architectuur achter de semantic pathfinder in beeld.*

Bijzonder bij dit experiment is dat er ruimte is voor gebruikersparticipatie. Bezoekers van de site kunnen helpen om de toepassing te verbeteren, simpelweg door een Pinkpop-optreden te bekijken en daarbij af en toe aan te geven of de computer het wel bij het rechte eind had. Eigenlijk een hele prettige, ongedwongen manier van annoteren. Al is het in een wat meer bescheiden controlefunctie, maar wie van muziek houdt zal het zeker geen straf vinden om hier veel tijd door te brengen. Wie zich daartoe geroepen voelt, moet overigens wel snel zijn. Vanwege auteursrechtelijke aspecten mag hollandsglorieoppinkpop.nl niet langer dan drie maanden online blijven.

## CAMERA'S OP ELKE HOEK

Die steeds betere technologie op het gebied van beeld- en spraakherkenning levert ook steeds meer toepassingsmogelijkheden op. Aan sommige ervan raken we heel makkelijk gewend. Een bestemming dicteren aan TomTom of een sms'je niet typen maar inspreken, daar kijken we niet eens meer van op. Maar tegelijkertijd komen ook futuristische scenario's zoals in *Minority Report* in zicht. In die film worden burgers 24 uur per dag gemonitord door intelligente volgsystemen. Op basis van in camera's ingebouwde irisscanners weet niet alleen *big brother* precies waar je bent, ook de reclamebillboards weten dat je op het punt staat om te passeren en stemmen precies op het juiste moment hun reclameboodschappen af op jou persoonlijke consumptieprofiel.
Dat lijkt nog vrij onschuldig, hooguit hinderlijk. En om iedereen gerust te stellen, te zijner tijd zal er heus ook wel een *opt out*-mogelijkheid worden aangeboden.
Het echte gevaar zal echter niet de beeldherkenningstechnologie zelf zijn, maar de wijze waarop bedrijven en overheden de technologie gaan inzetten. De verleiding is nu al groot om, zoals in Groot-Brittanie al gebeurt, camera's te hangen op elke hoek van iedere straat. Het is niet ondenkbaar dat ook onze overheid op enig moment bereid zal zijn om privacy-aspecten terzijde te schuiven, bijvoorbeeld terwille van criminaliteitspreventie of terrorismebestrijding. De verleiding om kunstmatige intelligentie en beeldherkenningstechnologie te koppelen aan allerlei andere informatiesystemen zal bij veiligheidsdiensten ongetwijfeld groot zijn. Alsof dat nog niet genoeg is, welk gevaar lopen we als geavanceerde beeldherkenningstechnologie in handen van echte *bad guys* valt?
Nogmaals Gevers: "Ik doe geen onderzoek naar veiligheid, agressie detectie, surveillance, et cetera, maar helaas is het wel mogelijk om de technieken daarvoor (beperkt) te gebruiken. Wat betreft het aantal camera's en inbreuk op privacy, dat zijn dingen waar ik zeer tegen ben. Dat is een groot maatschappelijk probleem waar je boeken vol over kunt schrijven...".

## GROTE BELANGSTELLING

Grote marktpartijen in de zoekmachinebusiness, zoals Google, Yahoo! en Microsoft moeten welhaast met grote belangstelling de ontwikkelingen bij ISLA en andere pioneers op dit gebied volgen. Gevers legt uit: "Er is grote interesse. We winnen de ene na de andere competitie. We werken samen met grote bedrijven, maar sommige bedrijven zoals Google zijn niet te bereiken. Andere bedrijven zijn zeer open, zoals Yahoo! met hun onderzoeksresultaten en richting. Aangezien we papers schrijven, zijn de technieken bekend en is het voor een groot bedrijf makkelijk om dat zelf te implementeren. Dat is ook sneller dan onze software te kopen. Die moet dan weer op maat worden gesneden voor het product, voor een ander platform, in een andere programmeertaal".
Het lijkt er sterk op dat we in de toekomst nog veel meer dan nu visueel georiënteerd zullen zijn. Het inhoudelijk ontsluiten van audiovisuele materialen is zelfs met toepassing van kunstmatige intelligentie een klus waarbij menselijke expertise onontbeerlijk is. De slimme catalogiseerder laat zich vanaf nu dus beter multimedia-archivaris noemen. Maar wat betekent beeld- en spraakherkenning in algemene zin voor de toekomst van de informatievoorziening? Het is de vraag of bibliotheken en hun bibliotheekzoeksystemen straks wel zijn voorbereid op deze toch rigoureuze omwenteling. Een ding is zeker; als de waarde van het geschreven woord gaat devalueren, neemt het belang van intelligente videozoeksystemen toe. **dib**

**Gerard Bierens** ontwikkelaar van de digitale mediatheek van Fontys Hogescholen en bovendien nauw betrokken bij de ontwikkeling van de HBO-Kennisbank. Zijn weblog is te vinden op www.gerardbierens.nl.

### Noten

[1] Deze visualisatie komt nog beter tot zijn recht in geanimeerde vorm, zie daarvoor deze link: www.science.uva.nl/research/mediamill/demo.
[2] www.hollandsglorieoppinkpop.nl
[3] www.mediamill.nl
[4] http://alecgoh.files.wordpress.com/2009/11/minority20report20lb20l.jpg of http://oseb79.free.fr/images/Cinema/Minority%20report%2002.JPG

### Nog meer relevante url's

- www.science.uva.nl/research/isla
- staff.science.uva.nl/~gevers
- staff.science.uva.nl/~cgmsnoek
- hmi.ewi.utwente.nl
- player.omroep.nl/?aflid=10659231