HET BEELD, HET WOORD EN DE ALGORITMEN
MOGELIJKHEDEN EN ONMOGELIJKHEDEN VAN AUTOMATISCHE INDEXERING

Annemieke de Jong
Nederlands Instituut voor Beeld en Geluid

Oorspronkelijk verschenen in TMG, Tijdschrift voor Mediageschiedenis, jaargang 3, nummer 2, december 2000

INHOUD

HET BEELD, HET WOORD EN DE ALGORITMEN
MOGELIJKHEDEN EN ONMOGELIJKHEDEN VAN AUTOMATISCHE INDEXERING

Annemieke de Jong
Nederlands Instituut voor Beeld en Geluid

INLEIDING

Er wordt steeds meer software ontwikkeld die kenmerken en eigenschappen van gedigitaliseerd audiovisueel materiaal automatisch extraheert en 'beschrijft'. Voor iedereen die zich bezig houdt met beeld- en geluidscollecties - of het nu gaat om de invoerkant of om het raadplegen - lijkt deze ontwikkeling ongelofelijke perspectieven te bieden. Beheerders van grote audiovisuele collecties verheugen zich nu al op de grote bedrijfseconomische voordelen van het automatisch ontsluiten van beeld en geluid. Gebruikers hebben de garantie dat véél meer materiaal beschikbaar zal komen voor consultatie en hergebruik en dat zoeken makkelijker wordt. De indexeersystemen waaraan wordt gewerkt, lijken steeds 'slimmer' te worden. Ging het eerst om automatisch gegenereerde formele kenmerken van film en video, zoals *kleur* en *contrast*, nu zijn er al programma's die de president van Amerika herkennen, eigenhandig kunnen aangeven waar hij het over heeft en of hij spreekt op een binnen of een buiten gegeven persconferentie. Films en televisie-programma's hoeven straks ook niet meer altijd integraal bekeken te worden: ten behoeve van een snelle voorselectie kunnen de meest betekenisvolle delen uit een film of programma, als bewegende, visuele samenvatting mét audio, aan de gebruiker worden gepresenteerd.

Deze ontwikkeling zal ongetwijfeld invloed hebben op de manier waarop audiovisuele bronnen ons worden aangeboden. Maar wat kunnen we werkelijk van deze systemen verwachten? Wat kunnen ze wel en wat kunnen ze niet? Om deze vragen te kunnen beantwoorden gaat dit artikel allereerst in op de wijze waarop beelds- en geluidsmateriaal wordt geordend en beschreven om het toegankelijk te maken. Wat is er eigen aan het analyseren en indexeren van dit soort materiaal? De klassieke obstakels vanuit zowel de archiveerpraktijk als de filmanalyse worden in kaart gebracht. Het classificeren van de kenmerken en eigenschappen van audiovisuele media - noodzakelijk om deze te kunnen overbrengen naar het digitale domein - vormt hierbij het kernpunt. Vervolgens wordt de 'state of the art' van een aantal automatische indexeertechnieken besproken, waarvan sommige al als product op de markt zijn. Wat doen ze precies en hoe werkt dat? Twee voorbeelden van toepassingen op dit gebied worden uitgediept, met een sterk accent op de wensen en verwachtingen van toekomstige gebruikers. Aan het slot worden problemen en perspectieven van automatische indexering nog eens samengevat.

HET VIRTUELE ARCHIEF

Veel audiovisuele archieven zijn in rap tempo bezig hun collecties te digitaliseren. Dit geldt zeker voor archieven van grote omroepen en media producenten. Hun klanten 'gaan digitaal' en zij - als belangrijke toeleveranciers van materiaal - doen mee. Maar ook andere, kleinere en grotere beeld- en geluidscollecties worden steeds vaker in 'geëncodeerde' vorm aangeboden. Zij verschijnen als CD Rom of DVD op de markt of worden online, via het Internet, verspreid.

De toegangsdrempels tot audiovisuele collecties vallen langzaam weg. Wordt het traditionele archief nu nog ondergaan als een fysiek, moeizaam binnen te dringen gebouw met een beperkte, tastbare collectie films en banden, straks kan iedereen putten uit één groot virtueel archief dat zich zal uitstrekken over de grenzen van tijd en ruimte. In de digitale omgeving wordt het zoeken in toenemende mate uitgevoerd vanaf de eigen computer: thuis, op school, in de bibliotheek of in het studiecentrum. De menselijke bemiddeling tussen gebruiker en beeldmateriaal verdwijnt. Daarvoor in de plaats krijgt de gebruiker een *interface* aangeboden met bepaalde zoekstrategieën.

De classificatie en de trefwoorden die archivarissen en documentalisten bij de beelden hebben bedacht, zijn straks niet meer de enige mogelijke ingang tot het materiaal: vanuit een gepredefinieërde set parameters kunnen gebruikers zelf opgeven hoe ze het materiaal geïndexeerd en geselecteerd willen hebben. Het systeem presenteert vervolgens het resultaat. Via *streaming video*-applicaties

worden de gekozen fragmenten naar de desktop gestuurd, waar ze kunnen worden gedownload om er naar believen in te browsen, te plakken en te knippen. Deze technologie opent het perspectief van de oneindige, dynamische bron van beeld, geluid en tekst, waaruit gebruikers iedere keer weer hun eigen collectie kunnen samenstellen.

De grootscheepse digitalisering en het toenemend aantal gebruikers dat toegang heeft tot audiovisuele content dwingen – veel meer dan voorheen - doordachte methoden te ontwikkelen om het materiaal te beheren en op duidelijke en effectieve wijze te ontsluiten. Jain en Hampapur omschrijven deze uitdaging als volgt:

> Video (the term is used to refer to audiovisual temporal data) is by far one of the most powerful and expressive non-textual media because it is a streaming media (temporally extended) with high resolution and multiple channels. These properties of video make it a popular media for capturing and presenting information. At the same time these very properties of video, along with its massive storage requirements present technical challenges from the data management perspective. The challenges include storage of video on computer systems, realtime synchronized delivery of video and content-based retrieval. (Jain en Hampapur 1998:246)

Contentbased retrieval is het benaderen van de inhoud *op basis van deze inhoud zelf,* waarbij (intrinsieke) kenmerken van het beeld- en geluidsmateriaal zèlf zoekingangen vormen. Software moet in staat zijn deze kenmerken te herkennen, te indexeren en geordend aan te bieden. Om een computerprogramma dergelijke bewerkingen te laten uitvoeren, zijn regels of *algoritmen* nodig. De uitdaging waarop hierboven wordt gedoeld, schuilt kort gezegd in het vangen van vorm en inhoud van audiovisuele informatie in dit soort regels. Het blijkt dat de ordening en modellering van audiovisueel materiaal – noodzakelijk voor het ontwikkelen van mechanismen voor het 'content-based' zoeken - veel problemen oproepen. Deze problemen hangen samen met de specifieke kenmerken van audiovisuele documenten en vooral met het eenduidig classificeren van deze kenmerken. In dit opzicht is er overigens niets nieuws onder de zon. De obstakels die bestaan rond het manueel beschrijven van beeld en geluid komen in verhevigde mate terug als het gaat om het bedenken van rekenschema's voor het automatisch indexeren van dit soort materiaal.

METADATA

Indexeren is het benaderbaar maken van aan een document toegekende of ontleende termen, zodat daarop gezocht kan worden. Fulltext retrieval systemen indexeren automatisch ieder woord van een willekeurige digitale tekst, zodat deze woorden kunnen fungeren als ingang tot het document. Audiovisueel materiaal is opgebouwd uit informatie die wordt overgebracht door beeld en geluid. In tegenstelling tot tekstuele documenten is het niet goed mogelijk dat een digitaal audiovisueel document *zijn eigen index* vormt. Dit komt omdat dit materiaal niet is opgebouwd uit concrete symbolische eenheden (zoals woorden), die kunnen worden gebruikt als directe toegang tot de inhoud. En er bestaat geen algemeen geaccepteerde inhoudsopgave van audiovisuele data, zoals er binnen de audiovisuele communicatie ook niets vergelijkbaar is met het functionele systeem van natuurlijke taal. Bovendien zijn audiovisuele documenten samengesteld uit sequentiële en temporale objecten, wat een snel en synthetisch begrip van de inhoud bijzonder lastig maakt. Om een film of televisieprogramma in zijn geheel te kunnen bevatten is de gebruiker te allen tijde afhankelijk van de tijdlijn van het document, i.c. het ritme dat is bepaald door de maker: de film zal doorgaans helemaal moeten worden bekeken.

Om audiovisuele documenten te kunnen benaderen is het nodig een beschrijving van het document te gebruiken. Deze beschrijvingen kunnen bij wijze van shotlist in tijd gelijk oplopen (*time-alligned*) met het beeld- en geluidsmateriaal. De tekst fungeert als een vervanging of representatie van de audiovisuele content waarop computerbewerkingen kunnen worden uitgevoerd en er wordt een index op het document gemanipuleerd en geannoteerd die door de gebruiker kan worden doorzocht. Wanneer eenmaal een relevant element van de representatie is geselecteerd, kan het stukje corresponderende *content* (het daadwerkelijke fragment) op de computer worden afgespeeld.

Deze representatie kan worden gezien als *metadata*: data over data, in dit geval informatie over de digitale content. Metadata zijn alle data die worden gebruikt om digitale video en audio zodanig te

ordenen dat *contentbased retrieval* van de inhoud wordt vergemakkelijkt. Deze informatie kan handmatig worden ingevoerd, zoals de door een documentalist gemaakte inhoudsamenvatting en de aan een fragment toegekende trefwoorden. Metadata kunnen ook automatisch worden gegenereerd tijdens het productieproces, zoals bijvoorbeeld een tijdcode en een productie-datum. Steeds meer soorten metadata kunnen automatisch aan het audiovisuele document zèlf worden ontleend.

Wanneer deze metadata worden gebruikt als toegang tot audiovisuele documenten, moet de vastlegging worden geformaliseerd en gestandaardiseerd. Om coherentie en consistentie aan te brengen in de geautomatiseerde verwerking, zal er een standaardmodel moeten komen waarin de vorm en de inhoud en de onderlinge relaties van de metadata exact vastliggen. Op basis van zo'n standaardmodel kan digitale informatie door allerlei verschillende systemen binnen een netwerk worden begrepen en uitgewisseld. De leer van de tekens, de semiotiek ligt aan zo'n model ten grondslag. Metadata kent de volgende semiotische onderdelen:

- *Het semantische aspect*: wordt onder elk element hetzelfde verstaan, wordt dezelfde inhoudelijke betekenis ontleend aan de tekens, de objecten, de eigenschappen en de relaties?
- *Het syntactische aspect* ofwel de wijze waarop een metadata-element digitaal gecodeerd is en dus de regels volgens welke men tekens mag gebruiken. De syntax definieert de naam en waarden van een element, geeft bijvoorbeeld het voorschrift dat de datum wordt weergegeven als '*jjmmdd'*.
- *De structuur.* De structuur detailleert de mechanismen voor het groeperen van de metadata en de relaties daartussen. Ook wordt de betekenis van de relaties gedefinieerd.
- *De praktijk.* Bedoeld worden de praktische richtlijnen bij de daadwerkelijke specificatie en creatie van metadata. Welke inwerking heeft een bepaald signaal op de gebruiker in de praktijk?
  (De Jong, 2000)

Over syntax kunnen goede afspraken worden gemaakt. Syntax is geen technisch hoogstandje maar vereist wel een hoog niveau aan consensus en standaardisatie. Protocollen zijn immers puntsgewijs opgesteld: systemen kunnen met elkaar communiceren of ze kunnen dat niet.
Structuur omvat het raamwerk waarmee verschillende syntaxen, relaties en semantische elementen kunnen worden ondersteund. De structuur kan informatie verschaffen over de koppelingen van elementen in andere bestanden en catalogi (*hoe* heet *wat* in *welk* schema?) en kan meerdere talen ondersteunen Er zijn al modellen in ontwikkeling die sommige eigenschappen van film en video hebben gedefinieerd en hun classificatie hebben vastgelegd. Een voorbeeld is de ontologie van MPEG-7, een internationale standaard voor de multimedia content.

Als er eenmaal zo'n model is gemaakt, blijft de belangrijkste vraag *wat* er precies gerepresenteerd moet worden. De semantiek, die ervoor moet zorgen dat de samenstellende delen van een model daadwerkelijk *informatie* vormen, is verreweg het lastigste aspect van de metadata. En zonder goed gedefinieerde semantiek geen syntax, geen structuur en dus geen representatie. Bij het ontwikkelen van modellen voor audiovisuele representaties wordt dan ook goed gekeken wordt naar conventies die gelden binnen de audiovisuele archivering waar op dit gebied veel ervaring en expertise bestaat.

BESCHRIJVINGSREGELS VOOR DOCUMENTALISTEN

Het definiëren van de content van een beeld teneinde zijn representatie vast te stellen is een zeer gecompliceerde taak. Documentalisten en archivarissen van audiovisuele collecties weten daar alles van. Het uiteenrafelen, benoemen en koppelen van de verschillende elementen en betekenissen in radio, televisie en film behoort tot hun dagelijks werk. Zij moeten daarbij bewerkstelligen dat *hetzelfde* programma of programmafragment uiteindelijk kan worden benaderd vanuit het *verschillende* zoekgedrag van allerlei mogelijke gebruikers, zoals programmamakers, wetenschappers, beeldresearchers en het algemene publiek. Dit eist nogal wat van hun vernuft en vaardigheden.

Veel regelgeving voor het beschrijven van audiovisuele documenten lijkt in de kern op elkaar. Naast formele kenmerken dient een inhoudskarakteristiek te worden gemaakt die antwoord geeft op de vragen *wie, wat, waar, wanneer, waarover, hoe* en *waarom*. Deelonderwerpen moeten ook worden benoemd. Als richtlijnen gelden hierbij de *seuil de finesse*, de drempel van detail en *seuil de pertinence*: de drempel van relevantie (Film en Beeldbandarchief NOB, 1993). Drempel van detail

betekent dat niet wordt benoemd wat onmiskenbaar een integraal deel is van een groter geheel. Als een landschap te zien is, wordt 'landschappen' geïndexeerd, maar niet alle afzonderlijke bomen, struiken en vogels. Als er echter een helling met wijnstokken is gefilmd, wordt wijnstokken weer wel geïndexeerd, wanneer deze beelden mogelijk interessant kunnen zijn voor een gebruiker. De drempel van relevantie heeft betrekking op het indexeren van betekenisvolle gehelen. Om benoemd te worden, moet het betreffende geheel zich op de een of andere manier onderscheiden, d.w.z. dat het zich, duidelijk identificeerbaar, op de voorgrond plaatst. Een object verdient ook vermelding als de aanwezigheid ongewoon is, een bijzonder fenomeen binnen het beeld als totaal. In de regel gaat het in het laatste geval om a-typische zaken.

Belangrijk beschrijvingsonderdeel vormt het herkennen en beschrijven van *stockshots,* sequenties van beelden die hergebruikt kunnen worden in een andere context. Een exterieurshot van een besneeuwd Witte Huis in het NOS- journaal kan een bruikbaar stockshot zijn, een tussenshot van twee zwervers in een documentaire over architectuur kan worden voorzien van het trefwoord 'daklozen', beelden van een zonsondergang kunnen worden geïndexeerd als 'sfeershots' etc. Veel audiovisuele bedrijfsarchieven zijn speciaal gericht op hergebruik van bestaand materiaal in nieuwe producties. Hier is het belangrijk om deze shots en fragmenten apart toegankelijk te maken zodat ze een soort beeldbank vormen naast de op onderwerp gerangschikte documenten.

In de context van een enkele beschrijving zijn dus meerdere semantische lagen zichtbaar. Voor een deel zijn deze betekenissen ontleend aan het document zelf, voor een ander deel worden ze toegekend vanuit een betekenis die buiten het materiaal ligt. Om in semiotische termen te blijven: de *conceptual signs* geven aan waar het document over gaat, maar omvatten ook de subtielere, metaforische kenmerken van audiovisuele communicatie: waar gaat het programma *echt* over. Wat op het scherm *zichtbaar* is, valt onder de *physical signs*, die op hun beurt weer kunnen worden onderverdeeld: wat is er te zien, waar staat dat (eventueel) voor en wat, van hetgeen te zien is, kan worden hergebruikt in een andere
context?

Door deze gelaagdheid is het zeer lastig film en video consistent te identificeren, te analyseren en te interpreteren in termen van concept of betekenis. Regels voor het inhoudelijk ontsluiten van audiovisueel materiaal kunnen dan ook niet meer zijn dan globale richtlijnen. In feite is het voor een documentalist onmogelijk om alle mogelijke betekenissen, interpretaties en benaderingen die verborgen zitten in een enkel shot of sequentie, in ogenschouw te nemen. Doorgaans ontstaat -op basis van intersubjectiviteit tussen documentalisten- uiteindelijk een soort *best practice*, die vooral wordt gestuurd vanuit de vraag van gebruikers. De gebruikers immers, geven vanuit hun eigen (formele en inhoudelijke) criteria voortdurend aan welke shots en fragmenten ze wel, en welke ze niet geschikt achten voor hun doel. Deze *best practice* is er niet voor de eeuwigheid. De vragen die aan het materiaal worden gesteld zijn steeds weer anders en als gevolg daarvan wisselt ook de overeenstemming tussen documentalisten.

FILMANALYSE

Naast de archiveerpraktijk biedt ook de wetenschappelijke filmanalyse mogelijk aanknopingspunten voor een systematische rubricering van betekenislagen in film en video, nodig om deze media het digitale domein in te loodsen. Het onderscheid dat de semioticus Daniel Chandler's maakt tussen 'digitale' en 'analoge' tekens mag illustratief heten maar is overigens nauwelijks bemoedigend.

> Analogical signs, such as visual images and gestures, involve graded relationships on a continuum. They can signify infinite subtleties which seems beyond words (…) One cannot specify the number of different smiles and laughs available in one person's repertoire. Digital signs, on the other hand, involve discrete units such as words and numeral and depend on the categorization of what is signified (..).
> (Chandler,1994)

Alleen al op het niveau van shots onderscheidt de filmanalyse doorgaans drie systemen waaruit betekenis kan worden gehaald.

- Symboliek: de conventionele, afgesproken symbolische betekenis van een beeld of het shot.

- Iconiciteit: de gelijkenis van een beeld of shot met een gegeven persoon, situatie of handeling in de realiteit.
- Indexicaliteit: de directe verwijzing van een beeld of shot naar zaken in de realiteit. (Chandler 1994):

De Deense filmanalyticus Karsten Fledelius voegt daar nog een vierde bron aan toe n.l. de plaats van een fragment of shot binnen het geheel: betekenis door positie. Welke betekenis dominant is, hangt volgens Fledelius deels af van het algemene referentiesysteem in de communicatiesituatie en deels van de actuele context (Fledelius,1979: 283). Chandler gaat ook in op het belang van deze context. Hij stelt dat de interpretatie van een individueel shot afhangt van zowel paradigmatische analyse (de wellicht onbewuste vergelijking met het gebruik van afwisselende shots) als de syntagmatische analyse (de vergelijkingen met voorafgaande en navolgende shots.) Hetzelfde shot, gebruikt in een andere sequentie van shots, kan een geheel andere betekenis krijgen, aldus Chandler. Zo zal een tussenshot van een uitdagend rokende puber in een documentaire over longziekten een totaal ander effect hebben dan exact hetzelfde shot als onderdeel van een vlot gemonteerd lifestyle magazine voor jongeren.

Het is duidelijk dat een audiovisueel document niet alleen geduid moet worden vanuit het visuele deel maar ook vanuit het geluidsspoor. Geluid is op te delen in taal, muziek en omgevingsgeluid, terwijl ook de mixage (de wisselende balans tussen de verschillende soorten geluid) hiertoe wordt gerekend. De geluidsbron kan daarbij in beeld te zien zijn, zich *offscreen* bevinden of noch in beeld, noch in de fictionele ruimte aanwezig zijn:
het is dan 'extra-diëgetisch' (van Loo, 1989). Fledelius schaart sommig geluid, waaronder muziek, onder de zgn. *modal elements*: tekens die geen specifieke betekenis hebben voor de feitelijke inhoud van het document, maar worden gebruikt om de onderliggende boodschap te versterken. Hieronder kunnen ook de camerahoek en het ritme van de montage vallen.

Er zijn veel semotici die vinden dat in film en televisie iconen het sterkst aanwezig zijn. Op het eerste gezicht lijken iconische tekens inderdaad de dominante vorm. Sommige tekens in de filmtaal zijn echter behoorlijk arbitrair: betekent een *dissolve* (het langzaam vervagen van het beeld) bijvoorbeeld echt in alle gevallen dat een scène uit iemands herinnering gaat volgen? Of een teken symbolisch is, of iconisch of indexicaal hangt grotendeels af van de manier waarop het teken wordt gebruikt. Typische voorbeelden kunnen hier misleiden. Voor de ene gebruiker valt een object in de ene, voor de ander in de andere categorie. Hetzelfde teken kan in de ene context iconisch worden gebruikt, terwijl het in een ander verband een symbolische betekenis heeft: de opname van één bepaalde vrouw kan een hele categorie vrouwen vertegenwoordigen, maar kan ook staan voor een specifieke vrouw. Tekens in film en video kunnen volgens Chandler niet worden geclassificeerd in termen van iconisch, symbolisch of indexicaal *zonder* referentie aan het doel van de gebruikers van het materiaal in een bepaalde context: wanneer we spreken over een icoon, een index of een symbool dan hebben we het niet over objectieve kwaliteiten van het teken zelf, maar over hoe een gebruiker het teken ervaart.

Een film of programma *als geheel* wordt geanalyseerd vanuit twee invalshoeken: wat is de feitelijke content en welke impliciete of expliciete normen en waarden worden overgebracht?
Voortbordurend op de syntagma-indeling van de semioticus Christian Metz (1972) onderscheidt Fledelius een aantal betekenisniveau's: het micronivau (het *paradigmatic plane*) dat enkele tekens of een combinatie van enkele tekens omvat en het medioniveau van de subsyntagmas en syntagmas. Deze niveau's worden gebruikt om de feitelijke content van de film te bepalen. Het macroniveau (het *syntagmatic plane*) omvat de bovenliggende betekenis-eenheden (de zgn. super- mega- en gigasyntagmas).Voor de ideologische analyse van een document moet worden geput uit de tekens op alle niveaus.

Semioticus Stefan Herrmann benadrukt vooral de manier waarop de verschillende tekens worden gecombineerd tot codes, die door de kijker moeten worden ontcijferd om er betekenis aan te verlenen. De tekens die in televisieprogramma's worden gebruikt, kunnen worden verdeeld in sociale, technische en representatiecodes. Sociale codes verschaffen informatie over bijvoorbeeld de sociale status van een karakter, zijn beroep of opleiding. Hiervoor worden vaak stereotype tekens gebruikt die geen verdere uitleg vereisen en de kijker aangeven wat men kan verwachten (voorbeeld: zwarte kleding impliceert een 'vals' personage.) Technische codes, zoals de cameravoering, de montage en het gebruik van geluid, beïnvloeden de sfeer en de stemming en maken ook duidelijk of men kijkt naar een documentaire, een dramaproductie, een quiz of een nieuwsbulletin. Subtieler werken zaken als

belichting en camerahoek. Deze codes voegen effectief betekenissen toe aan het beeld en kunnen een gefilmd karakter bijvoorbeeld kwetsbaar of juist machtig laten overkomen. Het ontcijferen van representatiecodes is een kwestie van ervaring. Zelfs al kunnen we de taal van een programma niet verstaan, dan is het ons toch direct duidelijk of we zitten te kijken naar een nieuwsuitzending, een show of een soapserie. Jarenlang televisiekijken heeft de kijker getraind in het feilloos herkennen van de conventies van de verschillende genres. Herrmann geeft echter duidelijk aan dat de interpretatie van al deze codes geenszins statisch is en altijd zal afhangen van de overeenstemming tussen kijkers op een bepaald moment. Televisiekijken is in die zin geen 'taal' die voor eens en voor altijd te leren is. (Herrmann, 2000)

HET AUTOMATISCH INDEXEREN VAN BEELD EN GELUID

Het predefiniëren van classificaties en schema's voor alle semantische tekens en betekenissen in audiovisueel materiaal kan zowel vanuit de praktijk als vanuit de theorie op vele manieren benaderd worden. Het proces kent vele dimensies en valkuilen en is in ieder geval uiterst complex, zo blijkt. Dit belooft niet veel goeds voor het genereren van standaardmodellen voor het automatische indexeren. Als *mensen* al grote problemen hebben met het eenduidig classificeren van beeld- en geluidsopnamen, hoe kunnen hier dan vaste computerprotocollen voor worden opgesteld ? De researchers van IBM onderkennen het probleem:

> ´Fundamentally, the description of content that human operators can identify and associate with the video (mostly keywords describing semantic content) is very different from and very difficult for, algorithmic operations--, i.e machine interpretation of video semantics. In addition, no one may ever develop an algorithm to interpret all of the semantics from the video automatically and extract meaningful annotation automatically for every video sequence´. (Bolle e.a.,1998)

Jain en Hampapur beamen ook dat het proces van extraheren van de semantische content van film en video feitelijk niet automatisch kan worden uitgevoerd. De hiervoor benodigde contextuele kennis kan niet in indexeersystemen worden ingebracht. Naast de semantische laag bestaat een gedigitaliseerd programma echter uit 'audiovisuele content' en 'informatie content'. Voor het extraheren van deze contentsoorten is geheel andere kennis nodig, aldus Jain en Hampapur. De audiovisuele inhoud toont wat er te zien en te horen is. Afhankelijk van hoe de film is gemaakt, kan dezelfde semantische content worden weergegeven door meerdere audiovisuele representaties. De audiovisuele content is in de eerste plaats georiënteerd op auditieve en visuele waarneming en vereist geen begrip van de informatie. Deze content kan met behulp van audio-, spraak- en beeldherkenning automatisch worden geïndexeerd. IBM spreekt overigens niet over semantische en audiovisuele content maar over hoge en lage niveaus van semantische informatie (de zgn. *high level features* en *low level cues*). Deze indeling lijkt meer recht te doen aan het feit dat audiovisuele inhoud evenzeer semantische informatie kan verschaffen. De MPEG-7 standaardiseringsgroep heeft het over verschillende niveaus van *abstractie* die een beschrijving kan vertonen. De automatische beschrijving van vorm en camerabeweging resulteert in een laag abstractieniveau. Door menselijke interventie kan een beschrijving een hoog abstractieniveau krijgen.

Op dit moment zijn er meerdere systemen op de markt die de audiovisuele content automatisch kunnen extraheren. De meeste van deze systemen richten zich op één toepassing en indexeren of beeld, of geluid of tekst. Met uitzondering van enkele producten behoren gecombineerde extractie-technieken veelal nog tot het domein van de research.

Beeldherkenning
De nadruk bij de temporale, ruimtelijke video-informatie ligt op de segmentatie in groepen frames, op basis van allerlei automatisch geïndexeerde kenmerken. Voor elk groep of soort metadata worden algoritmen ontwikkeld. Hoewel er steeds gesproken wordt over *contentbased* indexeren is er ook zoiets als *contentfree* indexering. Dit proces vindt plaats zonder rigoureuze analyse van de inhoud maar genereert statististische gegevens over bijvoorbeeld kleurverdeling, textuur en de vormopbouw in bepaalde videofragmenten. Deze techniek is inmiddels heel robuust en kan voor heel accurate resultaten zorgen.

Scène-wisselingen in het document worden meestal gebruikt om de visuele inhoud op te delen. Soms verlopen deze overgangen zo ongemerkt dat ze bijna niet statistisch kunnen worden gegenereerd. Daarom worden ook andere overgangen geïndexeerd, zoals *dissolves, fades, wipes* en *blends*. Uitgangspunt blijft echter de verandering van de visuele inhoud. De indirecte, statistische informatie die vanuit het beeld zelf wordt gedestilleerd, dient voornamelijk om op stilstaand beeld te kunnen zoeken (*query by image*) en is nuttig voor het snel vergelijken en categoriseren van enkele shots. Deze mogelijkheid kan ook vragen verfijnen, die in de eerste plaats zijn gesteld vanuit de inhoud. Voorbeelden zijn zoekacties naar 'soortgelijke scènes' en 'vergelijkbare gebouwen'. De techniek is nog niet in alle toepassingen even betrouwbaar, wat soms leidt tot onverwachte zoekresultaten.

Een andere vorm van visuele segmentatie is gebaseerd op de detectie van camerabewegingen zoals *zoom-in* en *zoom-out*, *panning* en *forward* camerabeweging. De computer volgt al interpreterend de intensiteit van de camera en maakt berekeningen voor gelijksoortige camerabewegingen door de hele film heen. Zo kan duidelijk worden dat statische interviewsequenties veel identieke videoframes bevatten en dat de aanwezigheid van fraaie, visuele effecten in een fragment niettemin weinig 'beschrijvende' informatie oplevert. Niet alleen de camera-activiteit kan worden geïndexeerd, ook beweging van door de camera opgenomen *objecten* kan worden gedetecteerd. Waar camerabewegingen in een soort *flow* door de hele film voorkomen, zijn objectbewegingen voorbehouden aan specifieke regio's binnen een bepaald beeld.

Kerntechniek in de automatische indexering is het segmenteren van video op basis van de presentie van een bepaald object of een combinatie van objecten. Menselijke content is in de context van beeldherkenning bijzonder belangrijk. Gezichtsherkenning is een van de bekendste toepassingen. De hiervoor benodigde algoritmes zijn bijzonder ingewikkeld en hangen af van strenge voorwaarden: een bepaalde belichting bij de opname, er mag niets 'in de weg staan' en het gezicht mag niet al te ver zijwaarts gedraaid zijn. Er wordt ook gewerkt aan het herkennen van menselijke objecten die optreden in een specifieke omgeving. Zo bestaan er al systemen die settings 'beschrijven': vindt de handeling binnen of buiten plaats, gaat het om een studiodecor of een bosrijke buitenomgeving etc. Er wordt ook gesleuteld aan herkenningsystemen voor complexe objecten zoals dieren en meer 'rigide' zaken zoals auto's en vliegtuigen. Op het gebied van systemen die alleen *bepaalde objecten* herkennen zijn inmiddels al heel redelijke resultaten geboekt. In het algemeen hebben deze systemen een hogere precisie dan systemen die erop uit zijn algoritmen in te bouwen die *alle objecten* aankunnen.

Tekst en graphics
Omdat zij de kijker veel feitelijke informatie willen laten absorberen in een korte tijdspanne gebruiken veel nieuwsprogramma's tekst in beeld. Deze tekst komt niet voor in het geluidsspoor. Het gaat hier om zgn. 'printed characters' zoals onderschriften en tekstblokjes, maar ook om bijvoorbeeld opschriften en uithangborden die in de opname zèlf voorkomen.
Ook tekstvormen die als signaal met het beeld en geluid worden meegestuurd, zoals ondertitels, teletekst en grafische elementen als logo's en statistieken vallen in deze categorie. Een wijziging of verschuiving van deze elementen in het beeld is meestal een indicatie van het veranderen van de semantische content. Nadat de teksten en grafische vormen door het systeem zijn gedetecteerd, worden zij met een Optical Character Recognition toepassing (OCR) bewerkt en als doorzoekbare full-text toegevoegd. Het is mogelijk om deze teksten te koppelen aan thesauri of andere gecontroleerde woordsystemen, zodat zij ook via deze weg kunnen worden benaderd.

Spraak
Harde geluiden in een film of televisieprogramma impliceren doorgaans een verhoging van de emoties. Deze (tijdelijke) veranderingen in het geluidsspoor kunnen gemakkelijk gemeten worden. Dat geldt ook voor stiltes en pauzes. Dit soort momenten kunnen iets zeggen over de semantische inhoud, n.l. dat het gaat om een relatief onbelangrijke of juist belangrijke sequentie. Sprekerwisselingen, het timen van audio en achtergrondmuziek en veranderingen in de inhoud van het gesproken woord kunnen helpen bij het segmenteren. Audio-segmentatie is nodig om het gesproken woord te onderscheiden van de andere geluiden en de woorden om te kunnen zetten in tekst, de zgn *transcripten*. Met behulp van taaltechnologische analyses worden de semantisch meest 'belangrijke' woorden eruit gefilterd. Omdat geïndexeerde, natuurlijke taal de zoekmogelijkheden enorm kan vergroten wordt momenteel gewerkt aan spreker-onafhankelijke spraakherkenners die op basis van een 'onbeperkte' vocabulaire automatisch de audioband transcriberen Een taalbegrijpend systeem analyseert en organiseert het transcript en slaat het op in een full-text informatie-retrieval systeem. Via dezelfde tekstdatabase kan zowel gezocht worden op individuele videosegmenten die beantwoorden

aan een vraag op basis van woorden in de geluidsband, als op handmatig toegevoegde trefwoorden en annotaties.

Structuur
Als de video eenmaal is gesegmenteerd is het zaak deze informatie te gebruiken voor het interpreteren van langere fragmenten en zo mogelijk voor het complete programma. Dit kan niet worden bereikt door het verwerken van data die toebehoort aan enkele shots. Een shot veschaft op zichzelf weinig semantische informatie en is feitelijk niet meer dan een bouwsteen, vergelijkbaar met een woord in een zin. Voor een analyse van de structuur is minstens zo belangrijk wat er *tussen* de shots gebeurt. Immers, het is de relatie of juist het ontbreken van de relatie tussen shots die de meeste zeggingskracht heeft.

Het 'verhaal' in een film of televisieprogramma kan bestaan uit een groep shots, die doorloopt in de tijd. In een audiovisueel document komen vaak meerdere verhaaleenheden voor. De logisch-chronologische opeenvolging van de gebeurtenissen in een film wordt *fabel* genoemd. Bij het *sujet* gaat het om de manier waarop de gebeurtenissen in een audiovisuele product aan ons worden gepresenteerd. Het sujet gaat op een geheel eigen wijze om met de interne logica en chronologie: het verhaal kan van de ene locatie naar de andere springen, of zich keer op keer verplaatsen naar een andere tijd. De continuïteit in tijd in een audiovisueel document is dan ook veel minder veelzeggend dan de continuïteit in betekenis. Belangrijke vraag is steeds: gaat hetzelfde verhaal door in de volgende verhaalunit?
Het blootleggen van de structuur van een film of video gebeurt door groepen shots met dezelfde 'betekenis' te formeren en samen te brengen. Het systeem identificeert daartoe ook alle shotwissels op punten waar een discontinuiteit in betekenis zichtbaar is. Door deze analyse kan inzicht ontstaan in de relatie tussen fabel en sujet en kan semantische betekenis worden verleend aan de sequenties.

Het destilleren van structuur is volgens de researchers van IBM een bijzonder belangrijk onderzoeksdomein, waarbij het combineren met informatie uit de geluidsband onschatbare voordelen gaat bieden.

> This is perhaps the most challenging aspect of automatic video annotation, finding the underlying discontinuities of meaning, or equivalently establishing from shot to shot where there is continuity of meaning if the tv programma goes to a commercial break or if the anchorperson of the news changes to a different news item. (Bolle e.a. 1998)

Automatische *genre-detectie* van een film of televisieprogramma is mogelijk door het aantal en de volgorde van de verhaaleenheden (de *visual narratives*) en de shots te gebruiken als indicatie van het type programma. De structuur van veel televisieprogramma's, zoals documentaires, nieuwsbulletins, quizzen en sportuitzendingen is erop gericht onderliggende narraties en boodschappen over te brengen. Vaak gaat het om vaste, vooraf gedefinieerde programma-formats waarin de shots of groepen shots op een bepaalde manier en in een vaste volgorde zijn gerangschikt. Veel kort op elkaar volgende shots wijzen meestal op een reportage uit de nieuws- en actualiteitensfeer; weinig en lang uitgesponnen shots en camerabewegingen horen veelal bij programma's met een meer beschouwend karakter. Door het automatisch analyseren van de visuele content kunnen deze temporale kenmerken worden herkend en kan geassocieerde informatie worden geëxtraheerd.

Presentatie van de content
Via de interface kan de gebruiker kiezen hoe hij de geselecteerde content op zijn scherm wil krijgen. De presentatie kan bestaan uit tekstuele beschrijvingen, uit een weergave van audiofragmenten of uit beeldmateriaal in de vorm van *icons* of *thumbnails*, *imicons* (bewegende plaatjes) of fragmenten. Een combinatie is ook mogelijk. Systemen die videosegmentatie toepassen, plaatsen het zoekresultaat doorgaans in de vorm van een *storyboard* op het scherm. Hierbij wordt een serie keyframes gegenereerd, die zijn geselecteerd aan de hand van een bepaald paradigma, bijv. 'kleur' of 'shotwissels'
[zie figuur 5]

De metadata die bij de fragmenten hoort kunnen samen met een tijdcode onder het storyboard 'meelopen'. Door op een keyframe te klikken worden de bewegende beelden opgestart en kan de context van een segment worden vergroot. Een toepassing die in een enkel systeem al werkt zijn de zgn. *views*. De gebruiker kan hierbij zijn zoekresultaten in een bepaald verband laten presenteren,

bijvoorbeeld in een chronologisch of geografisch verband. Er kan ook een zgn. *personality view* over het zoekresultaat worden heen gelegd, dat inzicht verschaft in publieke persoonlijkheden die voorkomen in de gevonden documenten. Een combinatie van *cross-documents views* kan nog weer andere verbanden blootleggen: vanuit de gevonden informatie genereert de computer een grafisch overzicht dat de gebeurtenissen op een tijdlijn plaatst, aangeeft waar ter wereld ze plaatsvonden en laat zien welke VIPS in de film of video optreden.

Het verspreiden van audiovisuele content via niet-breedbandige netwerken vormt in veel gevallen (nog) een obstakel. Ook al om deze reden zijn visuele samenvattingen, die een snel inzicht kunnen verschaffen in de relevantie van een bepaald document, een belangrijk research onderwerp. Een *visual summary* is een automatisch gegenereerde weergave in (bewegend) beeld en geluid van de semantisch meest belangrijke momenten van een programma. Middels het zgn. *skimmen* kan de essentie van de visuele inhoud in een fractie van de normale speelduur worden overgebracht. Gecombineerde taal- en beeldherkennings- technieken genereren hierbij een synopsis van het origineel. Voor het audiodeel worden de gebruikte woorden in de getranscribeerde audiotrack 'gewogen' om de meest relevante trefwoorden te selecteren en deze als doorzoekbare index op te nemen. Ook andere clues, zoals overgangen tussen sprekers en onderwerpen - vaak te herkennen aan korte pauzes in het geluidssignaal - worden gedetecteerd. Het videodeel wordt gescreend en geanalyseerd op scène-wisselingen en overgangen, relevante objecten en beweging. Wanneer de video aldus is gesegmenteerd wordt statistisch het relatieve belang van elk stukje content uitgerekend. Levendige (en dus veelzeggende, informatieve) scènes worden eruit gelicht d.m.v. optische analyse van camerabeweging en de aanwezigheid van bewegende objecten. Gezichten, tekst en grafische elementen worden geïdentificeerd en gebruikt als basis voor de samenvatting. De uiteindelijke representatie van het document kan door de gebruiker zelf worden ingesteld en indien gewenst variëren in grootte en soort content. Onderzoek wijst uit dat er al bruikbare visuele samenvattingen gemaakt kunnen worden met een compressieratio van 6:1 en 20:1. (Wactlar, 1996)

WAT VINDT DE GEBRUIKER: TWEE ONDERZOEKEN

In een aantal gevallen zijn toepassingen voor automatische indexering tot stand gekomen in nauwe samenwerking met potentiële 'klanten', die niet alleen testcollecties aanleverden maar ook eisen en wensen naar voren schoven. Dergelijke samenwerkingsverbanden tussen ontwikkelaars en eindgebruikers zijn belangrijk om een ongecontroleerde invoering van technologie te vermijden. Twee voorbeelden zijn de producten VICAR (Video Indexing, Classification Annotation Retrieval) en ECHO (European Chronicles Online) Het gaat hier om twee systemen van een geheel andere orde. Het een, VICAR, is een zgn. *plugin*: een softwareapplicatie die aan andere systemen kan worden toegevoegd. ECHO biedt een compleet, 'volautomatisch' digitaal filmarchief.

VICAR
VICAR is ontwikkeld als Europees project. Er werd aan meegewerkt door een aantal technische en academische partners uit Duitsland, Zweden Oostenrijk en Nederland. Grote beeldarchieven uit deze landen fungeerden als contentproviders en brachten hun ervaring en kennis in.VICAR indexeert alleen beeld. Alle semantische informatie moet achteraf aan de 'beschrijving' worden toegevoegd. De software werkt op basis van de principes *query by image* (bijv. gezichtsherkenning) en *query by example*, gebaseerd op beeld-eigenschappen als vorm, kleur, contrastverdeling en textuur. Het systeem maakt voor gedigitaliseerd beeldmateriaal een index aan voor elk shot. Deze index wordt gekoppeld aan de tijdcode van de originele videoband. Daarna worden de indexen ingedeeld in een aantal standaardklassen zoals VIPS, settings (interieur, bos, stad, berglandschap), objecten (bomen, auto's) en camerabewegingen. De gebruiker kan uit de serie keyframes die het zoekresultaat vormt de video vanuit de browser starten of het beeldmateriaal ophalen vanaf een opslagplaats. Met behulp van een VICAR plug-in kan de zoekvraag steeds verder worden verfijnd: heeft men 'Helmut Kohl' gevonden in een interieurshot, dan kunnen via *query by image* ook nog eens alleen die shots worden geselecteerd, waarin de politicus gekleed is in een blauw colbert. VICAR biedt vooral de mogelijkheid tot associatief zoeken via het beeld zelf, waarbij de vorm het leidend principe is, niet de inhoud.

Een aantal ervaren gebruikers van de beeldcollecties van het Nederlands Instituut voor Beeld en Geluid, waaronder programmamakers en journaalredacteuren, heeft zich uitgesproken over VICAR. Het gaat om een klein, beschrijvend onderzoek, dat niettemin een indicatie geeft van de reacties op deze nieuwe technologie. Zo blijken keyframes en storyboard als representatie van het beeldmateriaal

een regelrecht succes. Vanuit een dergelijke weergave is praktisch en gemakkelijk een snelle selectie te maken, zo is de algemene mening. Sommige geïnterviewden vinden het onverantwoord om keyframes te extraheren op basis van andere criteria dan shotwisselingen, met als argument dat 'de representatie van een programma in tact moet blijven'. De manier waarop keyframes kunnen worden gegenereerd zal samenhangen met het programmagenre, zo verwacht men. Keyframes zouden kunnen fungeren als visuele geheugensteun van items en beelden die in het eigen geheugen zitten opgeslagen, aldus de meeste programmamakers. Ook kunnen door het gebruik van keyframes sommige semantische niveaus worden onderscheiden: een storyboard toont nl. alleen wat er *te zien* is. Een zoekresultaat dat tot stand kwam vanuit een tekstuele beschrijving waarin de niveaus 'waar gaat het over' en 'wat zie ik' noodgedwongen samen voorkomen, kan zo worden gepreciseerd.

Op de mogelijkheid van zoekacties op basis van *query by image* werd door deze gebruikers verdeeld gereageerd. Men verwacht wel dat tekstueel zoeken naar materiaal *in combinatie* met keyframes in het algemeen minder ruis zal opleveren. Ook kunnen specifieke stockshots sneller worden achterhaald, zoals bijvoorbeeld de luchtopname van een voetbalstadion, beelden van een opgesloten dieren, een bergtop in de sneeuw etc. In deze gevallen zal *query by image* veel beter werken dan een beschrijving in tekst want: '(…) je ziet altijd veel meer dan in woorden te vatten is', zoals een van de gebruikers het uitdrukt. *Query by image* is nuttig als men associatief wil zoeken naar beelden waarbij *de vorm* het belangrijkst is, bijvoorbeeld in de oriëntatiefase bij het maken van nieuwe programma's, bij het zoeken naar beeld-elementen (eendeneieren, honden) en naar bepaalde esthetische kenmerken. Het 'op koppen' kunnen zoeken vinden de geïnterviewden bijzonder handig voor het traceren van de stemmingen waarin VIPS zich bevinden (Wim Kok die glimlacht, koningin Beatrix met een bezorgde frons etc.). Eén gebruiker wijst erop dat *query by image* wel degelijk ook semantische informatie kan geven, bijvoorbeeld doordat meteen zichtbaar is hoeveel personen in beeld staan. Van *query by image* is de algemene verwachting dat beelden zullen kunnen worden benaderd die anders niet geïndexeerd zouden worden. Overigens waren deze gebruikers lang niet enthousiast over alle foefjes van VICAR: het automatisch indexeren van camerabewegingen vond men bijvoorbeeld 'te ver gaan'. (Oomen, 2000)

ECHO
Het gebruikersonderzoek dat is verricht in het kader van het te ontwikkelen digitale European Chronicles OnLine (ECHO) archief toont overeenkomstige antwoorden. ECHO, eveneens een Europese project met technische, academische en archiefpartners uit Nederland, Frankrijk, Italie en Zwiterland, ontwikkelt een Internet-infrastructuur die toegang verschaft tot belangrijke digitale Europese filmcollecties. ECHO combineert zo'n beetje alle technieken op het gebied van beeld-, spraak- en tekst-indexering, met de bedoeling om vanuit het beeld en het geluid zelf zoveel mogelijke semantische informatie te genereren. De belangrijkste kenmerken van het systeem zijn: (semi)-automatische acquisitie van de metadata, multilinguale spraakherkenning (Italiaans, Nederlands, Duits, Engels en Frans), de mogelijkheid om in meerdere talen zoekvragen te stellen (*cross-language retrieval*), automatische transcripties van het geluidsspoor en visuele samenvattingen. Er wordt in eerste instantie toegewerkt naar een pilot, bestaande uit een beeldcollectie van belangrijke 20ste eeuwse thema's, belicht vanuit zowel nationale als Europese invalshoeken.

Een totaal van zestig potentiële gebruikers heeft, voorafgaand aan de bouw van dit systeem, aangegeven wat de wensen en verwachtingen zijn. De groep Franse, Italiaanse, Nederlandse en Zwitserse geïnterviewden bestond uit (media) historici, gebruikers uit de educatieve sector, programmamakers, documentalisten uit de audiovisuele productie- en archiefwereld en producenten van nieuwe media-producten. Ook dit onderzoek was beschrijvend van aard. Het ingeschatte, toekomstig gebruik van ECHO binnen deze groepen was: 24% research historisch filmmateriaal, 20% algemeen raadplegen van tekst en beelden, 19 % hergebruik van materiaal in nieuwe producties, 6 % invoeren van gegevens en 3% hergebruik t.b.v. educatieve doeleinden.

De ondervraagde documentalisten hebben, zoals verwacht mag worden, veel reserve tegenover automatisch gegenereerde metadata. Men vindt dat er in alle gevallen menselijke interventie mogelijk moet blijven, of het nu gaat om automatische trefwoordextractie, om scènedetectie of om het transcriberen van het commentaar. Bij alle toekomstige ECHO-gebruikers bestaat er een sterk geloof in de bruikbaarheid van transcripten, in het bijzonder voor hergebruik van de content en wetenschappelijke analyse van het document. Ook moet de transcriptie te allen tijde manueel geannoteerd kunnen worden. Men vindt belangrijk dat het systeem automatisch VIPS en objecten herkent. Metadata die de omgeving en 'setting' aangeeft (een bos, een persconferentie) is wenselijk,

maar wordt niet noodzakelijk geacht. Hetzelfde geldt voor automatisch geëxtraheerde camerabewegingen, waarbij *zoom-in* en *zoom-out* weer belangrijker worden gevonden dan een *pan*, en deze weer meer zeggingskracht heeft dan het enkele feit dat de camera zich van links naar rechts beweegt. De 'esthetische' kenmerken zijn voor iets meer dan de helft (60%) van de ondervraagden belangrijk. Als volgorde wordt aangegeven: kleur, belichting, framing en tot slot helderheid en contrast. Metadata over beweging vindt slechts 22% van alle ondervraagden interessant.

Query by image wordt door de documentalisten wenselijk geacht, maar ook niet meer dan dat. De filmhistorici onder de gebruikers zouden deze methode juist uitgebreid willen zien met algoritmen die kenmerken detecteren ten behoeve van het analyseren van vormen, stijlen en voor de algemene herkenning van filmtaal. De visuele samenvatting moet de chronologie van de film behouden, maar moet ook de semantische inhoud van de film weergeven. Deze *skim* zou zich moeten concentreren op de stemming van de film en op de tekst die erin voorkomt en vooral menselijke objecten moeten bevatten. Voor de keyframes geldt min of meer hetzelfde, zij het dat hier nog belangrijker is dat uit de keyframes meteen de 'mood' van een film of programma opstijgt. Zowel vanuit het gezichtspunt van de wetenschappelijk onderzoekers als van de kant van de hergebruikers bestaat grote interesse voor de mogelijkheid van automatische *views*. Chronologische en geografische overzichten vindt respectievelijk 78% en 52 % belangrijk, in *personality views* is 57 % geïnteresseerd en grafische overzichten van gebeurtenissen worden door 66 % van de ondervraagden gewenst. (ECHO User Requirement Report, 2000)

PERSPECTIEVEN EN PROBLEMEN

De onderzoeken naar de toepassing van automatische indexering leveren een wisselend beeld op. De toekomstige gebruikers zijn enthousiast, maar tonen ook scepsis. Een systeem dat automatisch indexeert wordt niet zonder meer 'vertrouwd', menselijk ingrijpen moet in de meeste gevallen mogelijk blijven. Uit de twee onderzoeken blijkt ook dat niet alle geëxtraheerde informatie - zoals sommige camera-acties en de detectie van beweging - even gewild is. Ook *query by image*, het zoeken met het beeld zelf, wordt verdeeld ontvangen. Het is natuurlijk mogelijk dat gebruikers zich nog niet precies voor kunnen stellen hoe bepaalde mogelijkheden te benutten, maar het ligt niet voor de hand dat het aanbod zonder meer de vraag zal creëren, zoals de leveranciers van deze systemen graag voorspellen.

Zal automatisch indexering de kostbare, tijdverslindende handmatige beschrijving van beeld en geluid overbodig maken? Hier past een genuanceerd antwoord. Ongetwijfeld zullen snelheid en efficiëntie in de beschrijving van documenten worden verhoogd, zeker waar het gaat om de ontsluiting van formele kenmerken. Véél meer materiaal zal na en ook tijdens de productie snel beschikbaar komen en in ieder geval op een minimaal ontsluitingsniveau zijn terug te vinden. Dat geldt voor iedere audiovisuele productieomgeving, waar voortdurend grote hoeveelheden uitgezonden en ruw materiaal circuleren. Een niet te onderschatten verbetering, zoals nu al blijkt bij organisaties als CNN waar met deze toepassingen wordt gewerkt. Het indexeren van sommige stockshots, beelden waarvan het begrip van de context niet belangrijk is, zal wellicht voor een groot deel automatisch kunnen geschieden. Ook kan de software zonder gevaar eigenhandig detecteren of het materiaal al dan niet 'schoon' is (i.c. of het geen logo's bevat, of verslaggevers die in beeld staan) en dus geschikt is om te worden hergebruikt.

In het algemeen zal automatische indexering de mogelijkheden om archiefmateriaal te beschrijven vergroten. De technologie komt daarbij *niet zozeer in de plaats* van manuele invoer maar *voegt er wat aan toe*, n.l. die kenmerken die (bijvoorbeeld vanwege de tijdrovendheid) al lang niet meer handmatig werden geïndexeerd. De tekens die tezamen Stefan Herrmann's technische en representatiecodes vormen, zullen voor een deel automatisch te genereren zijn en veel van Fledelius' *modal elements* kunnen efficiënt uit een document worden gefilterd en geordend worden aangeboden. Compleet 'nieuwe' eigenschappen worden ook automatisch terugvindbaar, onder andere zaken die niet of alleen maar heel moeilijk in woorden te vatten zijn, zoals het 'ritme' van een tenniswedstrijd. De vraag rijst echter in hoeverre gebruikers van audiovisuele collecties hier behoefte aan zullen hebben.

De applicaties gaan ongetwijfeld betere zoekresultaten gaan opleveren, met een hogere precisie. Gebruikers krijgen toegang via objectieve observatie en identificatie van beeldkwaliteiten en (bepaalde) kenmerken van de inhoud. Hierdoor zal het benaderen van beeld- en geluidsmateriaal niet of veel minder worden gehinderd door subjectieve interpretatie vooraf. Filmanalytisch onderzoek kan

van deze mogelijkheden enorm profiteren: de automatische, gedetailleerde segmentatie van audiovisueel materiaal op basis van genoemde eigenschappen en kwaliteiten, kan de efficiënte, objectieve en systematische bestudering ten zeerste bevorderen.

Het zal duidelijk zijn dat de mogelijkheden en onmogelijkheden van automatische indexering niet kunnen worden teruggebracht tot een soort vorm-inhoud tegenstelling, waarbij vormkenmerken door de computer, en de semantische informatie door mensen wordt ingevoerd. Daarvoor is de relatie vorm en inhoud van film en televisie te complex. Zoals de zaken er nu voor staan zal de computer én (intelligente) hulp kunnen verlenen bij de manuele annotatie én (beperkt) kunnen interpreteren. Semantische informatie op de hoogste niveaus moet in alle gevallen nog handmatig worden toegevoegd. Slechts wanneer beeld en geluid dezelfde boodschap overbrengen, kan deze informatie rechtstreeks worden gegenereerd. Beelden van de ontmoeting van Barak, Clinton en Arafat in Camp David, begeleid door een commentaarstem die een duidelijke toelichting geeft, kunnen heel wel automatisch beschreven worden.

Genredetectie lijkt vooral perspectieven te bieden voor televisieprogramma's met eenvoudig herkenbare formats, zoals nieuwsprogramma's en uitzendingen met een vaste opbouw (shows, quizzen e.d.) Programma's met een 'onvoorspelbaar' verloop en subtiele, gelaagde verwijzingen kunnen nauwelijks automatisch worden geïndexeerd. Het lijkt onbegonnen werk om passende algoritmen te bedenken voor een documentaire van Hans Keller, waarin trage rijers langs Amsterdamse grachtenhuizen fungeren als associatieve illustratie bij een in het commentaar voorgedragen gedicht van een kampslachtoffer. En wat te doen bij satirische programma's en persiflages? Hoe vlot en precies de huidige software ook segmenteert, transcribeert en analyseert, een betekenisvol geheel zal er niet uitrollen.

Naast deze toepassingen blijft er ook behoefte aan in de beschrijving opgenomen *beoordelingen* van audiovisuele documenten. Waardeoordelen zijn nuttig om, temidden van de digitale overvloed, beelden en geluiden met een bijzonder gebruiks- dan wel cultuurhistorisch belang, te oormerken. Deze oordelen zullen per periode wisselen en zijn zeer waarschijnlijk nimmer in rekenschema's om te zetten. Tijdgebondenheid zal de automatische indexeertools zelf overigens ook parten spelen. Modellen die zijn ontwikkeld voor de televisiebeelden van nu kunnen niet straffeloos worden toegepast op historisch materiaal. Film- en opnameconventies verschillen bijna per decennium en ook *wat* er gefilmd wordt (objecten, automerken, gezichten, gebouwen) is steeds weer anders. De huidige spraak- en taalmodellen zijn niet bij machte materiaal uit eerdere periodes accuraat te indexeren. Een woord als 'excellentie' -bekende aanspreekvorm in de jaren '50- wordt door een moderne toepassing niet herkent, terwijl het veel semantische informatie bevat. Materiaal van vóór 1931 kan vanzelfsprekend alleen middels het beeld worden geïndexeerd. Hier zullen altijd beperkingen blijven bestaan ten aanzien van automatische extractie van de inhoud, ofschoon de aanwezigheid van tussentitels enige compensatie zou kunnen bieden.

Om alle soorten materiaal uit ons film- en televisieverleden te kunnen indexeren, zullen de regels en instructies voor de indexeringsoftware nog aanzienlijk moeten worden uitgebreid en verfijnd. Beeld-, spraak- en tekstmodules zullen getraind moeten worden in het herkennen van de specifieke vorm en inhoud van audiovisuele bronnen door de decennia heen. Voordeel bij rekenschema's voor historisch materiaal is dat ze in principe maar één keer bedacht hoeven te worden.Voor veel van het audiovisuele materiaal dat in de toekomst nog geproduceerd gaat worden, zullen steeds weer nieuwe algoritmen moeten worden ontwikkeld.

LITERATUUR

Bolle, R.M. e.a. (1998) Video Query : Research Directions [www.document]
URL http://www. research.ibm.com/journal/rd/422/bolle.text [010700]

Chandler, Daniel (1994) Semiotics for Beginners [www.document]
URL http://users.aber.ak.uk/dgc/semiotic.html [070700]

ECHO Project Summary (Pisa, 1999) IST-1999-11994

ECHO User Requirement Report Deliverable 1.2.1 (Hilversum, 2000) IST-1999-11994

Fledelius, Karsten (Kopenhagen, 1979) Considerations about content analysis of audiovisuals/History and the audiovisual media: Studies in History, film and Society I

Herrmann, Stefan (2000), Do we 'learn' to read television like a kind of 'language'? [www.document]
URL http://www.aber.ac.uk/educationb/Undgrad/ED10510/sfh0901.html [150700]

Jong, Annemieke de (Hilversum, 2000), Metadata in de audiovisuele productieomgeving, NAA Werkuitgave

Loo, Arjo van (Utrecht 1989), Handleiding Filmanalyse

Smith, Michael S/ Mozenter, Robert (1998) Emerging Technology for contentbased Acces of Digital Video [www.document] URL http://www.mediasite.net/info/wp3_etcba.htm [010700]

MPEG-7 Context and Objectives [www.document]
URL http://www.cselt.stet..it/ufv/leonardo/mpeg [010799)

Oomen, Johan (Hilversum, 2000) Evaluatie NAA pilot

Putten, Peter van de (1999), Content based Video Search Engines becomes a Reality

Regels voor de Beschrijving van Audiovisuele Documenten (1993, Film en Beeldbandarchief NOB

Regelwerk Gemeenschappelijke Thesaurus Audiovisuele Archieven (NAA/ Filmmuseum, 2000)

Jain.R/ A. Hampapur (1998, McGraw-Hill), Video Data Management Systems: Metadata and architecture/. Multimedia Data Management.

VICAR Final Evaluation Report, Projectnr. 24916 Doc.nr.VICAR-T4.6-SWR-xxx05.01.2000

VICAR website http://ppc210.joanneum.ac.at/vicar [010700]

Wactlar, Howard (1996) Intelligent Acces to Digital Video: Informedia project [www.document] URL http://www.cs.cmu.edu/afs/cs/usr/hdw/www/HDW_IEEE96.html [010700)