Jeroen Vendrig en Marcel Worring

Interactief annoteren

Samen over het semantisch gat in multimedia

Nu breedband internet en sterke standaarden als MPEG-2 de technische belemmeringen voor de verspreiding en opslag van multimediagegevens wegnemen, wordt de vraag actueel hoe de juiste informatie in grote collecties kan worden gevonden. Jeroen Vendrig en Marcel Worring over het interactief annoteren van video's.

Multimediadata bestaan uit observaties van de echte wereld. Neem als simpel voorbeeld een foto van uw zoontje. De data bestaan uit een tweedimensionaal raster van getallen, maar voor u relevante informatie is niet expliciet aanwezig. Er is een interpretatieslag door uzelf voor nodig om te weten dat het om een kind gaat, en specifiek uw eigen kind.
Een computer die zich moet baseren op signaalanalysetechnieken kan deze slag niet maken. Het verschil tussen informatie die uit gegevens kan worden gehaald en de interpretatie die een mens aan diezelfde gegevens toekent, noemen we het 'semantisch gat'[1] (zie figuur 1 zie p. 24).
Het semantisch gat speelt bij multimedia-informatie een veel grotere rol dan bij bijvoorbeeld tekstuele informatie. De belangrijkste reden hiervoor is, dat bij de interpretatie van tekst een afgebakend en algemeen aanvaard vocabulaire kan worden gebruikt. Voor multimediagegevens is niets vergelijkbaars beschikbaar, het vocabulaire is oneindig groot. Slechts in specifieke domeinen, bijvoorbeeld medische beelden, is het mogelijk om de voorkomende informatie uitputtend te beschrijven.
Het gevolg is dat het zoeken in multimedia een hogere complexiteit kent. Automatische indexering van multimediadata is vanwege het semantisch gat in het algemeen niet succesvol. Systemen die hierop gebaseerd zijn, eisen dan ook een hoge mate van interactie met de eindgebruiker. Beter zou het zijn om ervoor te zorgen dat de index wel een foutloze beschrijving van de data is. Het is echter onmogelijk om de gigantische hoeveelheid gegevens waar het om gaat handmatig te verwerken. Door in een interactief proces rekenkracht en analysemogelijkheden van computers te combineren met de expertise van informatieprofessionals, is het toch mogelijk om grote hoeveelheden multimediagegevens efficiënt te indexeren en te ontsluiten.
In de huidige praktijk betekent interactieve indexering dat de informatieprofessional de resultaten van een computersysteem corrigeert. Wij zien interactie echter als een echt samenspel van mens en machine, die van begin tot eind gebruik maken van elkaars sterke punten. We geven hier als voorbeeld het interactief annoteren van video's.

Het primaire doel van het interactieve indexeringssysteem is het verbeteren van de zoekmogelijkheden van informatiesystemen in organisaties die voor hun bedrijfsvoering afhankelijk zijn van video-informatie. Het systeem is algemeen van opzet en niet voor een specifiek domein ontwikkeld. De huidige kenmerken die gebruikt worden zijn bij-

voorbeeld bruikbaar op nieuwsredacties om te kunnen zoeken naar persconferenties met bepaalde personen. Met andere features zou het systeem bijvoorbeeld eenvoudig gebruikt kunnen worden om in een trainingsvideo van de brandweer te kunnen zoeken op de shots waar de daadwerkelijke brand wordt getoond.

Domein en kenmerken

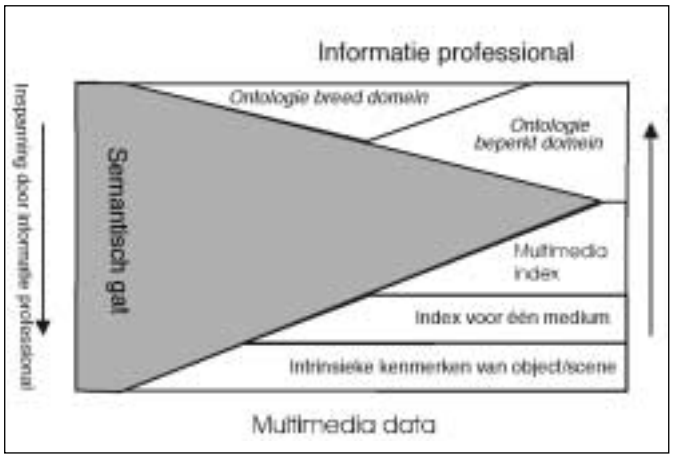
Een computer kan multimediadata alleen analyseren als die zijn getransformeerd tot één of meerdere kenmerken. Een voorbeeld van een kenmerk is een kleurenhistogram dat aangeeft welke kleuren in welke mate voorkomen in een beeld. Het succes van de analyse hangt af van de uitdrukkingskracht van de kenmerken, wat weer is gerelateerd aan het domein waarop het wordt toegepast.
Bij het gebruik van kenmerken moet rekening worden gehouden met diverse variaties. Geobserveerde objecten zoals stoelen, auto's en mensen, kunnen veranderen, bijvoorbeeld doordat ze roteren, of de opnameomstandigheid verandert bijvoorbeeld door verandering van het licht. Als u op zoek bent naar foto's van een bepaalde locatie, kiest u voor een kenmerk dat invariant is onder de lichtintensiteit. Zoekt u daarentegen alle foto's die 's morgens zijn genomen, dan kiest u een kenmerk dat wel gevoelig is voor de lichtintensiteit. Kortom, voor elke toepassing is een specifieke verzameling kenmerken nodig, afhankelijk van het domein en het probleem.
Hoewel de domeinkennis in handen is van de informatieprofessionals, kan niet van hen worden verlangd dat zij iets van multimediakenmerken begrijpen. In een beperkt domein kunnen informatieprofessionals en signaalverwerkingsexperts samenwerken om tot een ontologie te komen met daaraan gekoppelde multimediakenmerken. Voor bredere domeinen is de complexiteit van het probleem echter te hoog. Wij stellen daarom voor om de informatieprofessionals zodanig met computers te laten communiceren dat zij alleen hun domeinkennis hoeven te uiten en dat computers daarop vervolgens stap voor stap steeds beter kunnen inspelen.

Case: interactief annoteren van personen

Een voorbeeld van een interactieve video-indexeringapplicatie is ons *i-Notation*-systeem.[2] Deze case laat zien hoe de namen van spelers in een film kunnen worden geannoteerd. Het doel van het systeem is om met zo min mogelijk moeite aan te geven wie op elk moment in de film in beeld is. Algemener geformuleerd willen we de semantische kennis van de informatieprofessional over wat in een video zichtbaar is, overdragen aan de computer, zodat andere delen van de video automatisch kunnen worden geannoteerd.
Het mooiste systeem om dit probleem op te lossen zou een gezichtsherkenningssysteem zijn. In de praktijk is deze oplossing voor films niet economisch, afgezien van de technische haalbaarheid. Voor de unimodale aanpak (alleen beeldinformatie) is een flink aantal voorbeelden nodig om zo'n systeem te trainen, voor ieder personage. Voor hoofdrolspelers, die gemiddeld in meer dan honderd shots voorkomen, is dat misschien nog wel haalbaar. Maar voor de minder belangrijke acteurs is het verzamelen van de voorbeelden meer werk dan het bereiken van het einddoel van de applicatie zelf.
De multimodale aanpak (naast beeld ook tekst en geluid) heeft minder voorbeelden nodig. De aanpak is gebruikt in het domein van nieuwsuitzendingen,[3] waar de verschillende modaliteiten semantisch synchroon lopen. Dat wil zeggen dat als Gerrit Zalm in beeld is, er een tekstbalk met zijn naam wordt getoond, en de nieuwslezer ook nog eens zijn naam noemt. In films is dit niet het geval. Sterker nog, als iemand een naam noemt, is het juist waarschijnlijk dat de desbetreffende persoon niet in beeld is, omdat hij of zij slechts onderwerp van gesprek is. Zelfs als de genoemde gezichtsherkenningssystemen nagenoeg foutloos zouden werken, zouden ze niet geschikt zijn voor de annotatie van filmpersonages.

In dit voorbeeld, het annoteren van filmpersonages, is een volautomatisch systeem dus nog verre van realistisch. Daarom hebben we het in de doelstelling over 'zo min mogelijk moeite': het systeem moet inspelen op de informatie die een gebruiker aandraagt, zodat de gebruiker

Een informatieprofessional kan het i-Notation-systeem, een interactieve video-indexeringapplicatie, gebruiken voor het annoteren van shots in films.
Hier afgebeeld zijn stills van de film 'Shakespeare in love' (bron: www.imdb.com)



Figuur 1. Het semantisch gat tussen menselijke interpretatie en binaire multimediagegevens

minder energie hoeft te steken in hetzelfde eindresultaat. Een voorbeeld is het laten bevestigen van de hypothese van het systeem (zoals 'dit is Robert De Niro') door de informatieprofessional. Het verschil met zelf de naam intikken lijkt misschien klein, maar op een video van twee uur, scheelt dit al veel werk.

Voor de analyse van film gebruiken we een multimodale techiek op basis van drie gegevensbronnen: beeld, geluid en tekst. Geluid is de tekst die de acteurs uitspreken. Voor het gemak gaan we ervan uit dat geluid beschikbaar is in de vorm van ondertiteling in de oorspronkelijke taal. Tekst is een script dat is gebruikt voor het maken van de film. Het script kan afwijken van de echte film, maar vertoont in grote lijnen overeenkomsten. De herkende spraak en het script vullen elkaar goed aan. Dankzij de herkende spraak weten we precies op welk moment iets wordt gezegd, maar we weten nog niet wie het zegt. Het script geeft wel aan wie wat zegt, maar is niet gerelateerd aan de tijd in de video. Door de uitgesproken tekst te vergelijken met het script, weten we in de meeste gevallen wie wat zegt op welk moment. Omdat dit nog weinig zegt over wie in beeld is, is meer werk nodig om namen aan het videobeeld te koppelen.

Similariteitscores

Een interactieve sessie met het i-Notation-systeem werkt als volgt. Eerst laat het systeem een aantal videoshots zien, gevisualiseerd door middel van een representatief beeld. Vervolgens annoteert de informatieprofessional een shot met de namen van de mensen die in het beeld voorkomen en daarna selecteert hij alle shots waarvoor die annotatie geldt. Vervolgens toont het systeem een volgende reeks videoshots, waarbij in het ideale geval alle shots dezelfde annotatie kunnen gebruiken. De informatieprofessional hoeft dit alleen nog maar te verifiëren en op het knopje 'OK' te drukken (figuur 2).
De hamvraag is natuurlijk hoe het systeem de juiste shots kan laten zien. Het systeem toont een ranglijst van shots die het meest op de gezochte annotatie X lijken. Hiervoor berekent i-Notation voor elk shot een score op basis van de drie gegevensbronnen en kennis die de informatieprofessional invoert. De eindscore is een combinatie van vijf similariteitscores die gecombineerd een eindscore opleveren.
De eerste twee similariteitfuncties gebruiken alleen beeldinformatie om nieuwe shots te vergelijken met reeds geannoteerde shots. Het gaat dan niet om een precieze vergelijking van gezichten. Dat is technisch en economisch nog niet haalbaar, zoals we eerder hebben geconstateerd. We maken in plaats daarvan gebruik van structuur in films. Films zijn geen chaotische verzamelingen van beelden, maar meestal een visueel verhaal met een doordachte ordening. Een film bestaat uit diverse scènes waarvan de shots in elke scène bepaalde karakteristieken delen. Een voorbeeld is dat alle shots in een scène binnen zijn opgenomen, of nog gedetailleerder dat een bepaalde achtergrond telkens terugkomt, zoals een bloemetjesbehang.
Uiteraard komen er meerdere achtergronden binnen een scène voor, maar meestal kunnen personen een op een aan een van die achtergronden worden gekoppeld. Op het moment dat voor een van de shots in de scene de annotatie bekend is, kan die kennis worden gebruikt om de inhoud van andere shots te analyseren. Dat resulteert in de '*visuele overeenkomsten*' score: de mate waarin de visuele inhoud lijkt op die shots met annotatie X.
De '*visuele verschillen*'-score gebruikt dezelfde informatie als de eerste score, maar met een tegengesteld doel. De score is te vergelijken met een zwarte lijst. Voor shot A kan bekend zijn dat het niet annotatie X heeft, omdat A wel aan de informatieprofessional is getoond toen hij of zij voor X koos, maar niet werd geselecteerd. Shots die op A lijken, hebben dan waarschijnlijk ook niet annotatie X. De score plaatst deze shots lager op de ranglijst.
Ook de '*temporele nabijheid*'-score maakt gebruik van structuur in films, maar dan onafhankelijk van de inhoud. Het gaat ervan uit dat het waarschijnlijk is dat iemand die op een bepaald moment in de film verschijnt ook een paar shots eerder of later voorkomt.
De '*menselijke aanwezigheid*'-score vergelijkt shots met een mogelijke annotatie door gezichten te tellen. De gezichten worden gedetecteerd in shots en het aantal gezichten wordt vergeleken met het aantal namen in de annotatie. Let wel dat detecteren niet hetzelfde is als herkennen. We weten alleen dat er een gezicht in een beeld is, maar niet wie, en zelfs niet of het een man of een vrouw is. Deze score helpt vooral om onderscheid te maken tussen shots waarin wel en shots waarin geen mensen voorkomen.
De '*naam overeenkomst*'-score maakt alleen van tekst gebruik. We kijken of in een tijdvenster rond het shot de personen in annotatie X aan het woord zijn. Dat ze praten betekent niet dat ze op dat moment in beeld zijn, maar het vergroot wel de waarschijnlijkheid dat ze rond die tijd een keer in beeld komen.

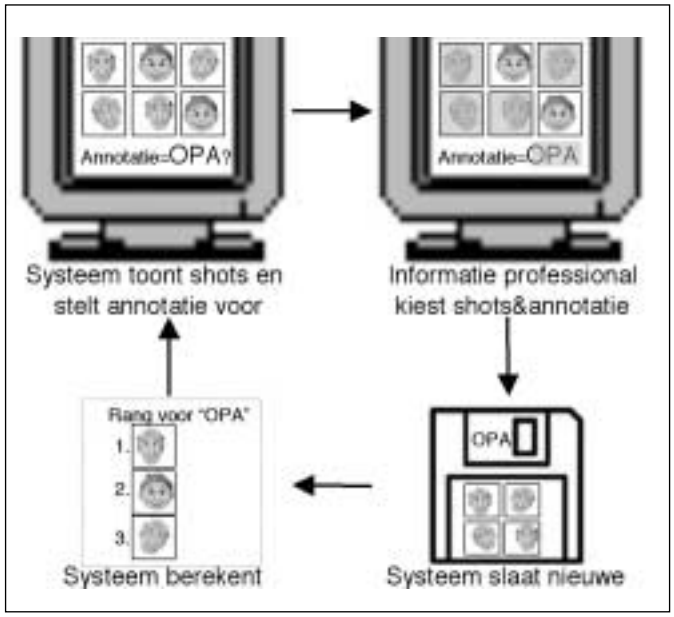
Winst

Evaluatie toont aan dat deze interactieve, adaptieve aanpak vele uren werk per film bespaart. We hebben de triviale sequentiële aanpak (de shots in de oorspronkelijke volgorde annoteren), onze adaptieve methode, en een theoretisch optimum afgezet tegen het simpelweg een voor een annoteren van shots. Voor een romantische komedie als 'Shakespeare in Love' leidt dit bij benadering tot de volgende werktijden. Als de informatieprofessional de sequentiële

aanpak volgt, is hij twaalf uur bezig. Met i-Notation is deze tijd tot acht uur ingekort. Het theoretisch optimum, een perfect systeem, houdt de informatieprofessional nog altijd vier uur bezig. Dit is omdat hij toch nog moet controleren of de voorgestelde annotaties juist zijn, en omdat hij erachter moet komen wat de namen van de spelers zijn.
De geboekte winst komt op diverse manieren terug bij eindgebruikers. De kosten voor het annotatieproces gaan bijvoorbeeld omlaag. Omdat veel annotatiewerk in handen is van non-profit organisaties met een vast budget, betekent dit dat er nu veel meer multimediamateriaal kan worden ontsloten op basis van semantische kenmerken. Ook is de kwaliteit van de annotaties hoger, omdat ze meer consistent zijn. Dit komt doordat het materiaal gegroepeerd in een kort tijdsbestek wordt geannoteerd.

Voor archieven die tot nu toe niet zijn ontsloten, is vergaande annotatie van de videotheek nu een realistische optie. De tijd die moet worden geïnvesteerd in de indexering kan sneller worden terugverdiend door het sneller en beter zoeken op basis van sleutelwoorden en onderwerpen. Interactieve annotatie legt ook de basis voor verdere automatische indexering. Met behulp van de ingevoerde sleutelwoorden en rubricering kan eenvoudig gebruik worden gemaakt van andere informatiebronnen. In het filmvoorbeeld ligt het voor de hand om het systeem te koppelen aan de schat aan gestructureerde informatie die op internet aanwezig is. Ook is het mogelijk om via extra informatiebronnen verbanden te leggen tussen verschillende video's, vergelijkbaar met MediaMill's automatische indexering van Journaal-uitzendingen. Dit systeem kan videosegmenten rubriceren zonder dat expliciete vermelding van de rubriek in de beschikbare metadata nodig is. Deze aanpak zal ook in andere domeinen tot interessante, bruikbare toepassingen leiden.

Enkele principes van het i-Notation-systeem kunnen ook op beelden worden toegepast. Het moet dan wel om fotorapportages gaan, dat wil zeggen dat er een relatie tussen de beelden moet zijn. Omdat we om logistieke en andere praktische redenen een video reduceren tot een aantal beelden die representatief zijn voor de video, is een implementatie van i-Notation voor fotorapportages eenvoudig.



Figuur 2. Het interactieve annotatieproces in i-Notation

Interactie

Door interactie als uitgangspunt te nemen bij het ontwerpen van multimedia-informatiesystemen, kunnen informatieprofessionals sneller specifieke informatie ontsluiten. Interactie is de sleutel om het onbegrip tussen de manier waarop computers visuele data benaderen en de interpretatie die mensen aan visuele inhoud geven in goede banen te leiden.
Het hier beschreven i-Notation-systeem overbrugt het semantisch gat op een wijze die eenvoudig is voor de informatieprofessional. Hij of zij hoeft niets te weten over hetgeen zich onder de motorkap bevindt. Hoe het systeem visuele similariteit berekent, blijft met opzet verborgen. De kenmerken zijn bedacht door de systeemontwikkelaars en worden niet gecommuniceerd naar de informatieprofessionals die een film annoteren. Zij hoeven slechts aan te geven of de resultaten van het systeem kloppen.

In een andere applicatie hebben we wel geprobeerd om de beweegredenen van een systeem naar de informatieprofessional te communiceren. Daarbij wordt duidelijk dat zelfs als mens en computer het erover eens zijn dat twee beelden min of meer hetzelfde zijn, de computer er even goed nog flink naast kan zitten omdat het om de verkeerde redenen is. In zulke gevallen moet de informatieprofessional dus niet alleen de hypothese van een systeem beoordelen, maar ook nog eens controleren of dat om de juiste reden is. Dit vergt een flinke inspanning van de informatieprofessional, die slechts op lange termijn vruchten afwerpt.

Door interactie niet te zien als een lapmiddel om een eindresultaat te verbeteren, maar als startpunt voor het oplossen van een probleem, kunnen we met minder werk hetzelfde resultaat bereiken. Hoe we de sprong over het semantisch gat precies wagen, hangt af van de inspanning en expertise die we bereid zijn erin te stoppen.

Noten

1. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta en R. Jain, 'Content based image retrieval at the end of the early years', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, December, 2000.
2. J. Vendrig en M. Worring, 'Interactive adaptive movie annotation', *IEEE Multimedia*, July-September, 2003.
3. S. Satoh, Y. Nakamura en T. Kanade, 'Name-It: Naming and Detecting Faces in News Videos', *IEEE Multimedia*, January-March, 1999.

Jeroen Vendrig is onderzoeker/projectleider en Marcel Worring is universitair hoofd docent bij MediaMill en Informatica Instituut Universiteit van Amsterdam.