Zoeken in multimedia collecties

Thijs Westerveld

Abstract

Information retrieval behoeft vandaag de dag nauwelijks een introductie. Dankzij webzoekmachines als Google en Yahoo! is bijna iedereen wel bekend met het zoeken van informatie in grote collecties. Maar hoe zit het met beeldmateriaal? Is dat net zo eenvoudig te vinden op basis van enkele zoektermen? En, kan een zoekvraag ook uit visuele kenmerken bestaan?

Multimedia documenten bevatten een schat aan informatie. Naast de beeld- en geluidsinformatie in het document zelf is er vaak allerlei afgeleide informatie en achtergrondinformatie te vinden. Neem bijvoorbeeld een bioscoopfilm, daar is flink wat metadata over beschikbaar: de regisseur, de scenarioschrijver, de belangrijkste personages en hun vertolkers en ga zo maar door, de aftiteling van een film kan behoorlijk lang zijn. Daarnaast is ook een behoorlijke hoeveelheid afgeleide data voorstelbaar. In het audio signaal zou bijvoorbeeld geclassificeerd kunnen worden wanneer er gesproken wordt, wanneer het stil is en op welke momenten er muziek te horen is. Op de gesproken delen kunnen we spraakherkenningstechnologie loslaten om een transcript te krijgen van wat er gezegd wordt en bij de muziek is vaak extra metadata aanwezig die vertelt wat de titel van het stuk is en wie componist en uitvoerende zijn. Ook in het beeldsignaal zit veel informatie. Om te beginnen kan het worden opgeknipt in kleinere coherente eenheden als scènes en shots en voor elk van deze fragmenten kan een representatief frame gekozen worden. De belangrijkste visuele eigenschappen in dat frame kunnen vervolgens worden beschreven in een model dat gebruikt kan worden als representatie van het fragment en wellicht kunnen we een aantal generieke concepten herkennen in deze representatieve frames—is het bijvoorbeeld een binnen- of buitenopname en is er een gezicht te zien. Door de grote hoeveelheid informatie aanwezig in multimedia documenten zou het zoeken in dergelijke collecties eenvoudiger kunnen zijn dan het zoeken in collecties met louter tekstuele informatie, er zijn immers zo veel wegen die naar de relevante informatie kunnen leiden. De praktijk is echter niet altijd even rooskleurig. Hieronder bespreek ik de mogelijkheden en beperkingen van hedendaagse technieken voor het doorzoeken van multimedia collecties.

Tekst als zoekingang

Een van oudsher toegepaste methode voor de ontsluiting van multimedia collecties is via het handmatig toekennen van tekstuele ingangen aan de te ontsluiten documenten. Ook vandaag de dag is deze ontsluitingsmanier nog op vele plekken succesvol. Grote foto- en videoarchieven zoals Corbis, Getty en Beeld en Geluid werken bijvoorbeeld op deze manier. Het is echter evident dat het handmatig annoteren van foto- en videomateriaal een tijdrovend proces is, om nog maar te zwijgen van de hoeveelheden training en standaardisatie die nodig zijn om annotaties consistent te krijgen.

Een alternatief waar recentelijk veel naar gekeken wordt is ontsluiting op basis van toevallig aanwezige tekst. Multimedia documenten komen vaak met een flinke hoeveelheid gerelateerde tekstuele informatie. Zo komen tv-uitzendingen en bioscoopfilms vaak met spraaktranscripties en metadata, en heeft beeldmateriaal op het web heeft in het algemeen een tekstuele context in de vorm van bestandsnamen, labels en bijschriften. Deze geassocieerde tekst geeft vaak een goede indicatie van wat er in het beeld te zien is. Het is niet voor niks dat deze tekstuele data door beeld zoekmachines als die van Google of AltaVista gebruikt wordt om beeldmateriaal te ontsluiten. Voor veel zoekvragen levert dat goede resultaten, maar het gaat ook vaak mis.

Een aardige illustratie van de problemen die optreden bij tekst gebaseerd zoeken naar beeldmateriaal is het online spelletje Guess-the-Google. De opgave daar is te raden welke zoekterm gebruikt is om de getoonde foto's te vinden. Al snel blijkt dan dat tekst op dit niveau niet altijd onderscheidend is, zo is een mix van huisdieren, ongedierte en computertoebehoren het resultaat van de query *mouse* en levert een zoektocht naar *gates* zowel foto's van hekken als van de topman van Microsoft.

Een tweede probleem bij tekst gebaseerd zoeken is dat niet alle beeldmateriaal vergezeld wordt van tekstuele beschrijvingen. Een goed voorbeeld van deze categorie is de groeiende hoeveelheid privé-materiaal afkomstig van digitale foto- en videocamera's. Dit materiaal komt vaak zonder enige context. Het enige aanknopingspunt is dan de bestandsnaam en ook die levert vaak geen nuttige informatie—zoekt u bij Google image search bijvoorbeeld maar eens naar F1050038.jpg.

Visuele kenmerken

Het mag duidelijk zijn dat tekst niet in alle gevallen de geëigende ingang is om informatie te vinden of terug te vinden. Aan alternatieven wordt hard gewerkt. Een systeem dat automatisch een beschrijving toevoegt aan willekeurige ongelabelde foto's uit persoonlijke archieven is nog ver weg, maar voor specifieke taken wordt beeldanalyse al succesvol toegepast. Een bekend voorbeeld is persoonsidentificatie op basis van biometrische kenmerken: vingerafdrukken en hand- en irisscans worden automatisch geanalyseerd en op basis van de gelijkenis met soortgelijke afbeeldingen in een database wordt bepaald of iemand al dan niet toegang wordt verschaft tot sportschool of luchthaven. Bij deze identificatie taken hebben we de omstandigheden redelijk onder controle. Zaken als de positie van de iris in het beeld en de lichtval zijn min of meer bekend, waardoor identificatie relatief eenvoudig is. Het wordt anders wanneer gezichten herkend moeten worden vanaf videobeelden afkomstig van een bewakingscamera, maar ook daarmee wordt in de praktijk geëxperimenteerd, bijvoorbeeld om notoire winkeldieven te herkennen. Op andere plekken worden bewakingscamerabeelden geanalyseerd om verdachte situaties te detecteren, zoals in een parkeergarage waar normaalgesproken niet veel meer te zien dan auto's die van en naar een parkeerplek rijden en mensen die af en aan lopen. Patronen van deze normale situatie kunnen worden gemodelleerd of geleerd, zodat bij afwijkende patronen—zeg, wanneer iemand te veel draalt—alarm kan worden geslagen.

In alle bovengenoemde voorbeelden hebben we te maken met uitermate specifieke taken en eigenlijk is dat het enige wat vandaag de dag praktisch haalbaar is qua beeldanalyse. Willekeurige visuele zoektaken in heterogene collecties als het web behoren vandaag de dag nog tot het domein van de onderzoeker. Een veel bestudeerde methode, geïpireerd op de successen bij de specifieke taken, is het zoeken aan de hand van van te voren gedefinieerde, afgebakende concepten. Bij deze benadering worden speciale systemen ontwikkeld voor het detecteren van relatief algemene concepten. De achterliggende gedachte is dat een combinatie van dergelijke algemene concepten kan leiden tot het vinden van dat specifieke beeld waar de gebruiker naar zoekt. Zo worden bijvoorbeeld detectoren gebouwd voor water, lucht, gras, personen, gebouwen en auto's. Een zoektocht naar een zonnige dag in het park kan dan beantwoord worden door te zoeken naar beelden waarop personen, gras en lucht te zien zijn, maar geen gebouwen en auto's. In het verleden werden detectoren wel gebouwd door het opstellen van regels, zo is het bijvoorbeeld waarschijnlijk dat er een persoon te zien is als het aantal huidskleurige pixels in een beeld groot genoeg is. Tegenwoordig worden de kenmerken voor een bepaald concept veelal automatisch geleerd op basis van een groot aantal gelabelde voorbeelden. Beide aanpakken zijn nogal bewerkelijk en bovendien zijn niet alle concepten even eenvoudig te detecteren. Dit betekent dat we niet kunnen verwachten dat dergelijke detectoren voor een brede set van concepten beschikbaar zullen komen, en dus dat lang niet alle zoekvragen op deze manier te beantwoorden zijn—hoe zoek je bijvoorbeeld naar beelden van een kop koffie wanneer de enige beschikbare concepten de genoemde zes zijn?

Een andere veel bestudeerde zoekmethode is die van query-by-example, of zoeken aan de hand van een voorbeeld. In deze situatie heeft een gebruiker al een voorbeeld plaatje of fragment van wat gezocht wordt, maar om een of andere reden is dit voorbeeld niet goed genoeg om direct te gebruiken; misschien is het beeld al te vaak gebruikt, ontbreken de benodigde gebruiksrechten of is de belichting niet goed. Het zou mooi zijn als we een zoeksysteem dan kunnen vragen om beelden met soortgelijke kleuren,

texturen en compositie. Ook op dit gebied wordt veel onderzoek verricht. Voor het berekenen van de gelijkenis tussen twee beelden worden doorgaans eerst visuele kenmerken van een plaatje geëxtraheerd. Vaak zijn dit zogenaamde low-level features, kenmerken die via eenvoudige algoritmes op basis van de pixel waardes berekend kunnen worden. Denk hierbij bijvoorbeeld aan het aantal pixels van een bepaalde kleur of het aantal scherpe kleurovergangen (edges) in een bepaalde richting. Wanneer veel kenmerken van een plaatje berekend zijn, kan een plaatje gerepresenteerd worden als een punt in een hoogdimensionale ruimte waarin elke dimensie overeenkomt met één van de geëxtraheerde kenmerken. De gelijkenis tussen twee beelden kan vervolgens worden bepaald aan de hand van de afstand tussen de punten in de kenmerkenruimte. Een variant hierop is het bouwen van een kansmodel voor elk van de beelden in de collectie. Zo'n model is een statistische beschrijving van de belangrijkste kleuren, texturen en compositionele eigenschappen in het beeld. Zoeken in de collectie gebeurt dan door voor elk van de modellen te berekenen wat de kans is dat het voorbeeldplaatje gegenereerd kan worden uit dat model. De beelden behorend bij de meest waarschijnlijke modellen worden vervolgens aan de gebruiker getoond.

Uitvoerige evaluaties van multimedia retrieval zoeksystemen, met name bij TRECVID, de internationale multimedia retrieval benchmark, hebben laten zien dat zoeken met visuele kenmerken nog maar beperkt succesvol is; voor de meeste vragen geeft een tekst gebaseerde aanpak toch nog steeds betere resultaten. Een kanttekening die we daarbij moeten maken is dat het hier in het bijzonder gaat om het zoeken in nieuws video's. Het is mogelijk dat tekstueel zoeken minder succesvol is in andere domeinen. Daarnaast zijn er zoals gezegd collecties waarvoor geen tekstuele informatie aanwezig is. Toch zijn er ook in de TRECVID evaluaties al een aantal zoekvragen waarvoor visuele informatie wel degelijk een nuttige bijdrage levert. Het gaat dan vooral om de gevallen waarbij de zoekvraag overeenkomt met een gebouwde detector, of daar dicht in de buurt zit. Zo lukt het bijvoorbeeld wel om fragmenten van Clinton voor een Amerikaanse vlag te vinden als we al een systeem hebben dat Clinton herkent. Een andere categorie van zoekvragen waarbij visuele kenmerken helpen is die waarbij de te vinden fragmenten een hoge mate van overeenkomst vertonen met het gegeven voorbeeld fragment. Dat is onder andere zo bij veel sportfragmenten; honkbal velden lijken bijvoorbeeld nogal op elkaar: stukken gras met wat zand erin, witte mannen erop en een menigte er omheen. Bij meer algemene zoekvragen als fragmenten met een kopje koffie is die overeenkomst tussen de verschillende fragmenten veel kleiner: een koffiekopje kan in allerlei verschillende situaties voorkomen. Wanneer een van die situaties overigens een commercial blijkt te zijn, dan is het wel weer eenvoudig om de herhalingen van die commercial op verschillende plekken in de collectie te lokaliseren en zo verschillende kopieën van hetzelfde fragment te vinden.

Context

Het beantwoorden van een visuele zoekvraag in een generieke collectie is duidelijk geen opgelost probleem. Gelukkig staan zoekvragen vaak niet op zichzelf en ook over de documenten in de collectie is vaak nog wel wat meer bekend. De eerder besproken geassocieerde tekst is slechts één vorm van metadata. Ook andere contextuele informatie kan ons dichter bij de gezochte informatie brengen. Voor televisie archieven zijn bijvoorbeeld dag en tijdstip van uitzending interessante gegevens, net zoals het tijdstip van een fragment binnen een uitzending. Zo is het bijvoorbeeld bekend dat sportprogramma's voornamelijk laat op de avond en in het weekend te zien zijn en dat eventueel sportnieuws in een journaal vaak tegen het eind van de uitzending te vinden is. Dit is informatie die van pas kan komen wanneer er gezocht wordt naar dergelijke fragmenten. Zo kunnen we uitzendingen van Den Haag vandaag veilig negeren wanneer we op zoek zijn naar beeldmateriaal van de Tour de France. Dergelijke patronen kunnen expliciet gemaakt worden voor sommige categorieën van zoekvragen (sport, politiek, weer), maar ze kunnen ook automatisch ontdekt worden op basis van interactie met gebruikers. Wanneer een gebruiker na een initiële zoekvraag aangeeft welke resultaten interessant zijn, dan is het mogelijk patronen van contextuele kenmerken te herkennen en in te springen op de gebruikersvoorkeuren. Op deze manier maken we gebruik van de overeenkomsten van contextuele informatie tussen documenten om zoekvragen beter te kunnen beantwoorden. Ook informatie van eerdere gebruikers kan nuttig zijn en als een soort aanbevelingsservice werken, zoals bekend van webwinkels als Amazon ("people who bought this also bought..."). Wanneer de documenten die gebruiker A relevant vindt voor een groot

deel overlappen met de beoordelingen van gebruiker B, dan kunnen we daar op inspringen door nog meer beeldmateriaal te laten zien dat door gebruiker B gewaardeerd werd.

Combineren

Vooralsnog lijkt multimedia retrieval op basis van visuele kenmerken vooral praktisch toepasbaar voor duidelijk afgebakende taken. Voor algemene, willekeurige zoekvragen worden hier en daar wat onderzoekssuccessen geboekt, maar de internetzoekmachines branden hun vingers er nog niet aan. Om dergelijk zoekvragen te kunnen beantwoorden zullen alle zeilen moeten worden bijgezet. Dit artikel begon met een beschrijving van de vele wegen die naar relevante informatie kunnen leiden in multimedia collecties. Elk van de hierboven beschreven technieken bewandelt slechts een van die wegen en is daarmee niet in alle situaties bruikbaar. Voor het beantwoorden van generieke zoekvragen is het nodig de verschillende technieken en informatiestromen te combineren. Elk van de stromen levert informatie met betrekking tot de mogelijke locatie van relevante informatie. Stel iemand is binnen de film Lost in translation op zoek naar het fragment waar Charlotte over een Porsche spreekt. De verschillende informatiestromen van de film zullen daar verschillende antwoorden geven. Van de metadata weten we dat Charlotte een van de hoofdkarakters is in deze film, dus ze zal van begin tot eind te zien zijn. Audioanalyse vertelt ons wanneer gesproken wordt; misschien is het Charlotte die spreekt. Spraakherkenning zal op verschillende plekken de term Charlotte herkent hebben en op andere plekken de term Porche. Verder zijn er vast een aantal representatieve frames die erg op het voorbeeldplaatje van Charlotte lijken en zijn er op andere plekken misschien gezichten gedetecteerd.

Het combineren van al deze informatie is niet triviaal. Ten eerste heeft elke techniek z'n eigen scorebereik. De scores kunnen natuurlijk genormaliseerd worden, zodat bijvoorbeeld alle scores tussen 0 en 1 liggen, maar dan nog kunnen de scoreverdelingen voor de verschillende informatiekanalen erg verschillen waardoor onbewust de ene informatiebron zwaarder weegt dan de andere. Dat is niet wat we willen. Soms willen we alles even zwaar wegen, maar in andere gevallen willen we juist bewust invloed uitoefenen op de mate waarin een bepaalde modaliteit wordt meegenomen. Vaak hebben we meer vertrouwen in de tekst gebasseerde kanalen dan in de visuele kanalen. Dat kan worden uitgedrukt worden door een gewicht toe te kennen aan elk van de informatiestromen. Maar hoe bepaal je deze gewichten? En zijn ze voor alle zoekvragen hetzelfde? Een derde probleem is de mis-alignment tussen de verschillende kanalen. Bij zoeken in videomateriaal, zullen fragmenten die op basis van de verschillende informatiestromen gevonden worden zelden synchroon lopen. Zo komt het vaak voor dat een enkel continu spraakfragment meerdere shots bestrijkt en dat meerdere personen spreken in hetzelfde shot. Dit betekent dat het gemakkelijk kan gebeuren dat het spraakkanaal vindt dat er relevante informatie te vinden is van seconde 40 tot 48, terwijl de visuele informatie ons vertelt dat de secondes 44 to 50 wellicht interessant zijn. Wat moet er in een dergelijk geval getoond worden aan de gebruiker? Een van de fragmenten, de doorsnede of de vereniging?

Wanneer alle informatiestromen in een generiek probabilistisch raamwerk worden gemodelleerd[1], verdwijnt een deel van deze problemen. Alle modaliteiten kennen dan kansen toe aan plaatjes of videofragmenten. Als we op zoek zijn naar beeldmateriaal dat op verschillende modaliteiten overeenkomt met de zoekvraag, dan kunnen we simpelweg gebruik maken van de gecombineerde kansverdeling. Ook het alignment probleem wordt opgelost; het combineren van kansdichtheden over tijd levert een kansfunctie die voor elk moment in een video aangeeft wat de waarschijnlijkheid is daar relevante informatie te vinden is (Fig. ???). Met een beetje geluk brengt dit ons in het Lost in translation voorbeeld bij het fragment waarin Scarlett Johansson tegen Bill Murray zegt: "You're probably just having a midlife crisis. Did you buy a Porche yet?".

References

[1] T. Westerveld *Using generative probabilistic models for multimedia retrieval.* PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2004.

SPIEGLE

Veel van de hier besproken technieken worden (verder) ontwikkeld in het nationale Bsik onderzoeksproject MultimediaN. In een van de subprojecten wordt gewerkt aan SPIEGLE een systeem dat op basis van gegevens over collectie, zoektaak en gebruikersgroep een zoekmachine voor een specifieke setting genereert. De door SPIEGLE gegenereerde systemen worden getoetst met behulp van een breed scala aan testbedden op het gebied van multimedia en tekst retrieval. Daarnaast wordt nauw samen gewerkt met twee specifieke gebruikersgroepen, te weten Van Dale lexicografie en Beeld en Geluid, het nationaal audiovisueel archief. Met name de laatste groep is van belang voor de evaluatie van de verschillende multimedia zoekmethoden in de praktijk. De archieven van Beeld en Geluid kennen vele gebruikersgroepen, variërend van programmamakers en onderzoekers tot onderwijsinstellingen en het algemene publiek. Elk van deze groepen heeft zijn eigen wensen en zoekpatronen. Met SPIEGLE wordt het eenvoudiger de verschillende groepen te bedienen. Ook kan het handmatige annotatieproces dat nu de basis vormt voor de ontsluiting van de collecties verlicht worden door zoekfunctionaliteit al tijdens dit proces beschikbaar te maken. Daarnaast kunnen zoektechnieken bijdragen aan de ontsluiting van delen waarvoor geen handmatige annotatie beschikbaar is.