

Zoek- machine versus video- tsunami



Hoe meer videobeelden, des te ingewikkelder het vinden van dat ene bijzondere fragment. Hoog tijd dus voor een nieuwe generatie zoekmachines – en een heel andere weergave van de zoekresultaten.

DE MANIER WAAROP we zoeken op internet zou wel eens helemaal kunnen veranderen. Niet langer zouden we na het intikken van een zoekterm een lijst met resultaten zien. Denk eerder aan een scherm gevuld met beeldjes, waardoorheen je horizontaal, verticaal of diagonaal kunt scrollen, op zoek naar dat ene gewenste resultaat. Een oude belofte van de informati-

ca zou daarmee eindelijk binnen bereik komen: een geautomatiseerd systeem dat in video kan zoeken. Zo'n systeem is hard nodig, want het aantal videobeelden op internet neemt explosief toe. In 2006 bestond het gegevensverkeer nog maar voor 12 procent uit online video. Een jaar later was dat al 22 procent en eind dit jaar zal het naar verwachting 32 procent

zijn. Afgelopen zomer voorspelde internet-multinational Cisco Systems dat de totale hoeveelheid gegevensverkeer in 2012 verzesvoudigd zal zijn, waarvan liefst de helft bestaat uit online video.

Om nog maar te zwijgen van de al bestaande film- en videoarchieven. Alleen al in Nederland beheert het Nederlands Instituut voor Beeld en Geluid circa 700.000 uur aan beeldmateriaal, dat de komende jaren zal worden gedigitaliseerd. Voorts krijgen inlichtingendiensten en beveiligingsbedrijven steeds meer beeldmateriaal van beveiligingscamera's te verwerken en komt er op YouTube elke minuut maar liefst tien uur aan nieuwe beelden bij. Ofwel: de noodzaak om in al deze uren aan materiaal te zoeken wordt steeds groter.

Veruit de meest gebruikte manier om de juiste film of video te vinden is zoeken op trefwoord. Toch kunnen trefwoorden en korte beschrijvingen nooit alles beschrijven wat er op een video is te zien. Zoek op YouTube maar eens naar rode auto's (*red cars*): het resultaat is teleurstellend. Als een rode auto niet het hoofdonderwerp is in een filmpje, zal niemand ooit 'rood' of 'auto' als trefwoord opgeven.

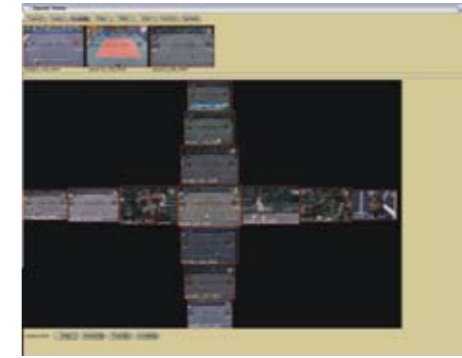
Dat moet beter kunnen. Geavanceerdere videozoeksystemen helpen de gebruiker een handje door zelf een beschrijving toe te voegen. Een daarvan is *Blinkx*, die met 26 miljoen uur aan materiaal de grootste diep geïndexeerde videozoekmachine ter wereld claimt te hebben. *Blinkx* gebruikt spraakherkenningstechnieken en zet gesproken

tekst automatisch om in geschreven tekst, waarin de gebruiker kan zoeken op trefwoord. Klinkt handig als je zoekt naar fragmenten waarin veel wordt gesproken, zoals in journaaluitzendingen. Maar het grootste nadeel van deze techniek is dat ze eigenlijk alleen maar goed werkt bij Engelse en niet bij de Nederlandstalige video's. Laat staan Arabische of Chinese.

Technieken om visueel te zoeken bestaan ook. De zoekopdracht bestaat dan uit een voorbeeldvideo, in plaats van een zoekwoord. De computer geeft vervolgens een video terug die veel lijkt op het voorbeeld. Maar ook dat werkt niet altijd. Anders dan computers letten mensen immers niet zozeer op visuele gelijkenissen, maar op semantische gelijkenissen, dus of ze in bete-

kenis op elkaar lijken. Zo komen bij het zoeken naar een trein ook vaak plaatjes van bruggen terug, omdat de bovenkant van een trein dezelfde ronding heeft als een brug. En dat is natuurlijk niet de bedoeling. Maar het begrijpen van beelden is voor een computer dan ook erg ingewikkeld. "We weten niet eens hoe het herkennen van beeld in onze eigen hersenen werkt", verklaart Cees Snoek, postdoc bij de informaticafaculteit van de Universiteit van Amsterdam. "Als ik tien mensen vraag om te vertellen wat ze precies op een beeld zien, krijg ik tien verschillende verhalen. Beelden zijn dus subjectief. Daarom kunnen we beeldherkenning niet zomaar modelleren met een computer." Het fundamentele probleem van het

- Het nieuwe browsen: de crossbrowser en de forkbrowser, essentieel om zoekresultaten in video goed te kunnen weergeven.



herkennen van plaatjes en het zoeken in videobeelden is dan ook het overbruggen van de zogeheten 'semantische kloof'. Deze kloof wordt door Snoek en z'n collega's beschreven als de discrepantie tussen enerzijds wat de computer kan bepalen, zoals kleur, vorm en textuur, en anderzijds de semantische interpretatie: de betekenis die mensen toekennen aan wat ze zien. Verspreid over de wereld zijn verschillende onderzoeksteams bezig om deze kloof zo goed mogelijk te overbruggen.

Woestijn Sinds vorig jaar demonstreren al deze onderzoekers elkaar hun systemen op de Videolympics. Een soort Olympische Spelen voor videozoekmachines, waarbij meedoen belangrijker is dan winnen. Wat al deze innovatieve systemen in ieder geval met elkaar gemeen hebben, is een soort woordenboek met semantische begrippen. Dat kan een object zijn zoals een trein of een brug, maar ook een bepaalde stijl van regisseren. Zo is een interview te herkennen aan een rechthoekig tekstblokje dat onderin beeld verschijnt en aan het gegeven dat de mensen op het beeld weinig bewegen.

De meeste videozoeksystemen kunnen nu zo'n honderd tot vijfhonderd verschillende begrippen van elkaar onderscheiden. Variërend van vliegtuig tot voetbal, van cartoon tot woestijn en van explosie tot voedsel. Dit is overigens nog steeds weinig vergeleken bij het aantal begrippen dat een mens kent, dus is er nog een lange weg te gaan. Want hoe groter het vocabulaire van een zoekstelsel, des te succesvoller het systeem.

Het denken in dergelijke concepten is nog een vrij jonge wetenschap, gegrondvest in de jaren negentig van de vorige eeuw. Toen leerde men computers in videobeelden te herkennen of er een gezicht op stond en waar dat gezicht zich in het beeld bevond. "We konden kijken of de kleur van een pixel overeenkwam met huidskleur en bepalen hoeveel huidskleurpixels aan elkaar vast zaten. Als al die pixels bij elkaar dan de vorm van een ellips hadden, konden we concluderen: dit is een gezicht. Maar we konden niet herkennen wie er op stond,"

nuanceert Cees Snoek.

En volgens hem kan dat nog steeds niet, tenzij er van elke persoon meerdere videobeelden bestaan. Een commercieel initiatief op internet dat een poging doet om gezichten te herkennen is *Viewdle* van persbureau Reuters. Het gaat dan niet om de gezichten van beroemdheden die vaak in films of op televisie komen. Ook het gezicht van Jan Peter Balkenende wordt door de technologie herkend. Gewone stervelingen zijn echter onvindbaar op *Viewdle*.

"Als een computer niet genoeg voorbeelden heeft van degene die hij moet herkennen, dan zal hij hem of haar nooit kunnen vinden", zegt Snoek. "Hoe meer voorbeelden, des te beter het systeem semantisch leert denken."

Vandaar dat zoeken in video, nogal tegen het gezond verstand in, wat gemakkelijker is dan zoeken in stilstaand beeld. "Omdat een video uit meerdere beelden bestaat, is het zoeken binnen video makkelijker. Je hebt simpelweg meer voorbeelden."

Woordenboek Toch is het niet alleen een kwestie van veel voorbeelden, maar ook van de computer laten 'kijken' naar de juiste, zo onderscheidend mogelijke, kenmerken. Bij de Universiteit van Amsterdam concentreren Snoek en zijn collega's zich met name op dit probleem. De belangrijkste kenmerken zijn kleur, textuur en vorm. Alle andere kenmerken, zoals randen, hoeken en punten, zijn daarvan afgeleid. "Zo is een Nederlandse koe gemakkelijk te herkennen aan een wit met zwart kleurenpatroon en een lijf met vier poten", legt de UvA-onderzoeker uit. "Een hond daarentegen is bijna niet te herkennen, omdat er geen prototype hond bestaat. Daarvoor kunnen we dus moeilijk goede kenmerken verzinnen."

Toch is 'hond' wel degelijk een van de begrippen die het systeem in zijn woordenboek heeft staan. "Een hond is vaak het centrale onderwerp van een video en staat daarom vaak in het midden van een scherm," denkt Rong Yan, ontwikkelaar bij IBM Research. "Bovendien worden honden vaak op de arm gedragen." Maar Snoek ziet dat anders: "Honden komen juist vaak toe-

vallig voorbij lopen. Alleen als je er een documentaire over maakt, komt zo'n beest centraal in beeld."

Handig zou het zijn als geautomatiseerde zoeksystemen ook meer gebruik konden maken van de bewegingen op video. Daarmee zou een trein bijvoorbeeld wel degelijk gemakkelijk van een brug zijn te onderscheiden zijn, en een vliegtuig van een vogel. "Computers kunnen dat helaas nog niet goed," zegt Snoek. En dat is vooral een kwestie van rekenkracht: "Deel van het probleem is dat de hoeveelheid gegevens enorm de pan uitgroeit als je het tijdsaspect wilt meenemen. Er is veel meer onderzoek nodig om bewegingskenmerken efficiënt uit te kunnen rekenen en ze zodanig weer te geven dat je er handig mee kunt verder rekenen."

Bij IBM Research concentreert het onderzoek zich intussen vooral op machinelere. Rong Yan is een van de Amerikaanse ontwikkelaars van iMARS, een afkorting voor het IBM Multimedia Analysis and Retrieval System. "In het kort is machinelere het gebruiken van statistische methoden om bepaalde kenmerken te leren", vat Yan samen. De Amerikaanse vlag herken je vooral aan de patronen – sterren en strepen – en een voetbalwedstrijd aan een groen vlak met mensen.

Het idee achter iMARS is hetzelfde als dat achter het Nederlandse systeem: de onderzoekers verzamelen een grote hoeveelheid voorbeeldvideo's, en kennen daaraan met de hand een semantische betekenis toe. Dus vliegtuig bij vliegtuig, hond bij hond en vlag bij vlag. Omdat al deze concepten op verschillende manieren in beeld gebracht kunnen worden, leert het systeem vervolgens zelf algoritmen om alleen unieke elementen te bewaren, zodat soortgelijke patronen herkend worden in andere videobeelden. Inderdaad, een koe heeft vier poten, maar die zijn niet altijd zichtbaar. Een koe moet ook herkenbaar zijn als hij ligt te slapen in de wei. Om die reden zal een wit met zwart gevlekt patroon soms een beter kenmerk zijn dan de vier poten. Statistische algoritmen maken daarin een afweging. ►



Het verschil met het Nederlandse systeem is volgens Snoek dat iMARS iets sneller leert en dus minder voorbeelden nodig heeft. "Uiteindelijk proberen we allebei om het systeem op een zo efficiënt mogelijke manier zo veel mogelijk aan te leren. Nu kost het nog heel veel rekenkracht om video's te analyseren, maar straks zou het ook mogelijk moeten zijn met je mobiele telefoon."

Bolvormig Maar hoe moeten mensen straks met zulke superslimme zoeksystemen omgaan? We zijn eraan gewend dat zoekresultaten onder elkaar op het scherm worden getoond. Bij MediaMill, het systeem van de Universiteit van Amsterdam, komt daar een extra dimensie bij. Daar worden resultaten niet alleen verticaal, maar ook horizontaal getoond. In de vorm van een kruis dus. De 'crossbrowser', noemt Ork de Rooij dat. Hij promoveert aan de Universiteit van Amsterdam op de gebruikersinterface van videozoeksystemen. Op de verticale as staan de zoekresultaten zoals die in steeds verschillende video's voorkomen. En op de horizontale as staat de tijd, omdat de auto

die je zoekt ook meerdere keren binnen dezelfde film of televisie-uitzending kan voorkomen. Deze twee dimensies worden bolvormig op het scherm geprojecteerd om de navigatie makkelijker te maken. De Rooij ontwikkelt ook andere manieren om de zoekresultaten te laten zien. Zoals de 'forkbrowser', die met twee extra diagonale assen inderdaad een beetje lijkt op een vork. De extra assen zijn eigenlijk bedoeld om het feit dat videozoekmachines zich nog wel eens in interpretatie willen vergissen te compenseren: ze bieden zoekresultaten die bijvoorbeeld qua kleur of semantiek erg lijken op het gevonden plaatje. Als je alleen rode auto's zoekt, verschijnen er op de ene diagonale as – die voor kleur – andere rode objecten, terwijl de andere diagonaal as juist alleen auto's met dezelfde vorm laat zien, zonder dat die per se dezelfde kleur hebben. Met dit soort gecombineerde browsers is het al bijna mogelijk om mensen te herkennen, bijvoorbeeld vergeten voetballers of terroristen. Cees Snoek: "Mensen persoonlijk herkennen is nog een wetenschappelijke brug te ver. In een wereld met zes miljard mensen is de variatie tussen mensen

namelijk heel klein. Je hebt dan wel heel veel gezichtskenmerken en heel veel voorbeelden van al die personen nodig." Maar op kleinere schaal werkt het al wel, zegt Snoek. "Zoals bij mensen die in hetzelfde gebouw werken en allemaal op dezelfde manier gefotografeerd worden. Dan is de kans op fouten een stuk kleiner." En verder wordt de software steeds beter in het herkennen van allerlei markante zaken. "We kunnen bijvoorbeeld doelpunten detecteren. En we weten of op beelden mensen wel of niet aanwezig zijn, of ze een rugzak dragen en of ze een baard hebben." Hoewel dergelijke technieken nog niet perfect werken, kan door de manier van presenteren al snel veel irrelevant beeldmateriaal worden weggefilterd. "Met de crossbrowser of de forkbrowser kan iemand daarna uit de gevonden doelpunten alleen nog maar die van Ruud Krol zoeken, of uit de bewakingscamera's alleen die shots bekijken die mensen met rugzakken bevatten," denkt Snoek. Straks leven we in een wereld waarbij video-beelden en films helemaal geen trefwoorden meer nodig hebben om toch gevonden te worden. Welke geheimen zouden die beelden dan onthullen? ●