Hoe de computer een glimlach kan waarnemen

Facial Action Coding System

De ontwikkeling van een geautomatiseerd systeem dat menselijke gezichtsuitdrukkingen kan herkennen en interpreteren is een grote uitdaging. Aan de Technische Universiteit Delft wordt gewerkt aan een systeem voor het automatisch herkennen van de bewegingen van gezichtsspieren in video-opnamen van gezichten en een redenatie-systeem waarmee het mogelijk is gezichtsuitdrukkingen te classificeren in emotiecategorieën.

Maja Pantic

Het menselijk gezicht geeft signalen die essentieel zijn voor de communicatie met andere mensen in ons sociale leven. Het gezicht omvat het spraakapparaat en wordt gebruikt om andere mensen te herkennen, om het gesprek te sturen door iemand aan te kijken of te knikken en om te interpreteren wat er is gezegd door middel van liplezen. Het is onze directe en natuurlijke manier van communiceren en helpt ons te begrijpen wat iemands gemoedstoestand en intenties zijn op basis van zijn gelaatsuitdrukking (Keltner & Ekman, 2000).

Het automatisch analyseren van gezichtsuitdrukkingen zou zeer nuttig zijn voor allerlei verschillende disciplines, zoals beveiliging, geneeskunde en onderwijs. Op het gebied van beveiliging spelen gelaatsuitdrukkingen een doorslaggevende rol bij het bewijzen van iemands geloofwaardigheid of helpen ze juist het tegendeel te bewijzen. In de geneeskunde zijn gelaatsuitdrukkingen het middel om waar te nemen wanneer welke specifieke mentale processen plaatsvinden. In het onderwijs kan de leraar uit de gezichtsuitdrukking van de leerling afleiden of de aangeboden leerstof dient te worden aangepast. Wanneer het gaat om interfaces tussen mensen en computers (pc's, robots, machines), zijn gelaatsuitdrukkingen een mogelijkheid aan de machine te communiceren wat ervan verwacht en gevraagd wordt. Registratie van waar de gebruiker naar kijkt (*gaze tracking*) kan doeltreffend worden gebruikt om computergebruikers te bevrijden van het traditionele toetsenbord en de muis. Daarnaast kunnen bepaalde gezichtssignalen (zoals een knipoog) worden geassocieerd met bepaalde commando's (bijvoorbeeld een klik met de muis) en zo een alternatief bieden voor de traditionele muis- en toetsenbordcommando's. Het menselijk vermogen om emoties te kunnen aflezen van iemands gezichtsuitdrukking vormt de basis voor *facial affect processing*, wat kan leiden tot het ontwikkelen van interfaces met emotionele communicatie en uiteindelijk tot een flexibeler, beter aanpasbare en meer natuurlijke interactie tussen mensen en machines.

De ontwikkeling van een geautomatiseerd systeem dat menselijke gezichtsuitdrukkingen kan herkennen en interpreteren is echter nogal moeilijk. Aan de Technische Universiteit Delft wordt een systeem ontwikkeld voor het automatisch

Samenvatting

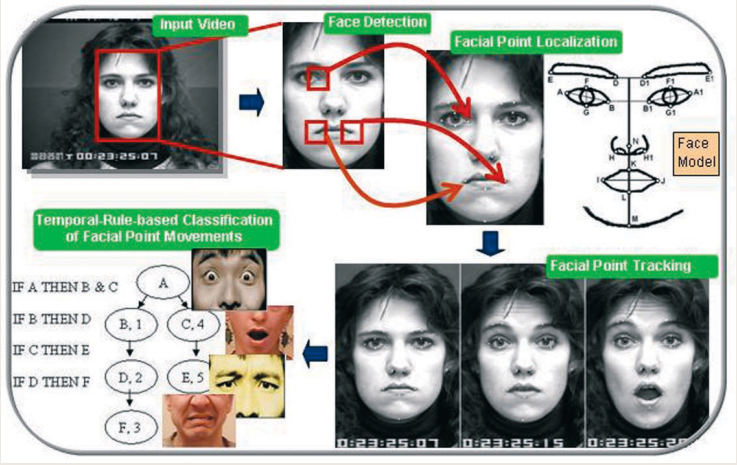
Aan de TU Delft wordt een systeem ontwikkeld voor het automatisch herkennen van de bewegingen van gezichtsspieren (*action units*, au's) in video-opnamen van het gezicht, en een redenatiesysteem waarmee gezichtsuitdrukkingen worden geclassificeerd in emotiecategorieën die van de gebruiker geleerd zijn. Er is meer onderzoek nodig om het gehele scala van menselijke gezichtsuitdrukkingen automatisch te coderen.

herkennen van de bewegingen van gezichtsspieren (*action units*, au's) in video-opnamen van gezichten. Daarnaast wordt gewerkt aan een op een casus gebaseerd redenatiesysteem waarmee het mogelijk is gezichtsuitdrukkingen te classificeren (gecodeerd in au's) in emotiecategorieën zoals die van de gebruiker geleerd zijn.

Waarneming van gezichtsuitdrukkingen

De meeste methoden voor het automatisch analyseren van gelaatsuitdrukkingen proberen een beperkte combinatie van prototypische emotionele gelaatsuitdrukkingen te herkennen, bijvoorbeeld angst, verdriet, walging, blijdschap, boosheid en verrassing (voor een volledig overzicht van het reeds gedane onderzoek, zie Pantic & Rothkrantz, 2003). Deze gewoonte vloeit voort uit onderzoek gedaan door Darwin en meer recentelijk Ekman (Keltner & Ekman, 2000), die meenden dat basisemoties vergezeld gaan van bijbehorende prototypische gelaatsuitdrukkingen. In het leven van alledag echter komen zulke prototypische gelaatsuitdrukkingen relatief zelden voor; emoties worden vaker getoond in de vorm van subtiele veranderingen in een of een paar discrete gelaatstrekken, zoals het optrekken van de wenkbrauwen bij verbazing. Om zulke subtiele menselijke emoties waar te nemen en meer algemeen om informatie, die gelaatsuitdrukkingen verschaffen, bruikbaar te maken voor de verschillende eerdergenoemde toepassingen, is de automatische herkenning van de bewegingen van gezichtsspieren, zoals de bewegingseenheden ofwel *action units* (au's) van het FACS-systeem (Ekman & Friesen, 1978), noodzakelijk. Het *Facial Action Coding System* (FACS) is ontwikkeld opdat menselijke waarnemers de veranderingen in gelaatsuitdrukking kunnen beschrijven aan de hand van waarneembare bewegingen van de gezichtsspieren (au's). FACS biedt regels voor het visueel waarnemen van 44 verschillende au's en hun verloop (*onset*, *apex* en *offset*, oftewel opgaand, hoogtepunt, neergaand) in een opeenvolgende serie beelden van een gezicht. Met deze regels kan een menselijke waarnemer een getoonde gezichtsuitdrukking ontleden tot op de specifieke au's waaruit de gezichtsuitdrukking is opgebouwd.

Er zijn weinig methoden bekend voor wat betreft de automatische herkenning van au's met behulp van beelden van gezichten (voor een gedetailleerd overzicht van het voltooide onderzoek, zie Pantic (2005)). In tegenstelling tot deze methoden, die vooral ingaan op het probleem van het ruimtelijk modelleren van gezichtsuitdrukkingen, gaat de methode die wij voorstellen tevens in op het probleem van het modelleren van de temporele aspecten van gezichtsuitdrukkingen. Met andere woorden, onze methode is bij uitstek geschikt voor het coderen van temporele activatiepatronen (onset, apex, offset) van au's zoals zichtbaar in video-opnamen van gezichtsuitdrukkingen. Figuur 1 geeft een overzicht van de methode zoals wij die toepassen. Hierbij wordt een opeenvolgende serie beelden van een gezicht ingevoerd in vier stappen: *face detection* (gezichtsdetectie), *facial fiducial points detection* (detectie van de referentiepunten op het gezicht), *point tracking* (volgen van de punten) en *AU coding* (au-codering).



Figuur 1. **Overzicht van methode voor detectie van gezichtsspierbewegingen in video-opnamen van gezichtsuitdrukkingen**

Figuur 2. **Resultaten van tracking van punten op het gezicht in inputreeks gezichtsbeelden**

Voor het bepalen van de oppervlakte van het gezicht in het eerste frame van een ingevoerde gezichtsvideo gebruiken we een commercieel verkrijgbare real-time gezichtsdetector. Het waargenomen gezichtsoppervlak wordt vervolgens onderverdeeld in twintig relevante *regions of interest* (roi's), die elk corresponderen met een bepaald punt op het gezicht dat dient te worden waargenomen. Dit is mogelijk door gebruik te maken van een combinatie van heuristische technieken gebaseerd op de analyse van de verticale en horizontale beeldhistogrammen. De toegepaste detectiemethode voor het herkennen van gezichtskenmerkende punten maakt gebruik van individuele gelaatstreklocatietemplates voor het detecteren van punten in de relevante roi. Deze gelaatstrekmodellen bestaan uit GentleBoost-templates van 13 bij 13 pixels en zijn opgebouwd uit zowel grijsniveauwaarden als Gabor-*wavelet*-karakteristieken. In de trainingsfase worden de gezichtsmodellen aangeleerd door gebruik te maken van een representatieve set positieve en negatieve voorbeelden, waarbij de positieve voorbeelden bestaan uit 'beeldplekken' gericht op een bepaald karakteristiek punt in het gezicht en de negatieve voorbeelden uit 'beeldplekken' die willekeurig geplaatst zijn dicht bij dezelfde gelaatstrek. In de testfase wordt iedere roi eerst gefilterd door dezelfde set Gabor-filters die gebruikt zijn in de trainingsfase (in totaal worden 48 Gabor-filters gebruikt). Vervolgens wordt voor een bepaald punt op het gezicht een inputvenster van 13 bij 13 pixels (*sliding window*) pixel voor pixel over 49 beelden van de relevante roi geschoven (grijswaarde plus 48 Gabor-filterbeelden). Voor elke stand van het sliding window produceert een GentleBoost-*classifier* een respons die de gelijkenis aangeeft tussen het 49-dimensionale beeld van het sliding window en de aangeleerde punten op het gezichtsmodel. Na het scannen van de gehele roi onthult de positie met de hoogste respons wat het betreffende punt op het gezicht is.
Nadat zo twintig referentiepunten zijn gelokaliseerd in het eerste frame van de inputreeks gezichtsbeelden, worden rond elk van de gezichtspunten een aantal kleurentemplates gedefinieerd. Vervolgens traceren we elke kleurentemplate gedurende de rest van de beeldenreeks (zie figuur 2) met behulp van het deeltjesfilter, een algoritme dat werd geïntroduceerd door Pitt en Shepard (1999). Het filteren van deeltjes is inmiddels een toonaangevend traceringsparadigma vanwege het vermogen om succesvol om te gaan met storing, occlusie en ruis. Om dit aan te passen aan het probleem van op kleur gebaseerde templatetracering hebben we een observatiemodel opgesteld op basis van een op kleur gebaseerde afstand tussen de doelkleurentemplate zoals gedefinieerd voor het eerste frame van de inputreeks en de kleurentemplate van het huidige frame. Mogelijke problemen in verband met schaduwen voorkomen we door de globale verandering in intensiteit te compenseren.

Op basis van de veranderingen in de positie van de referentiepunten meten we veranderingen in gezichtsuitdrukking. Veranderingen in de positie van de referentiepunten worden eerst omgezet in een reeks middenniveauparameters voor de herkenning van au's. We hebben twee parameters gedefinieerd: *up/down*(P) en *inc/dec*(PP'). De parameter *up/down*$(P) = y(P_{t1}) - y(P_t)$ beschrijft opwaartse (*up*) en neerwaartse (*down*) bewegingen van punt P en de parameter *inc/dec*$(PP') = PP'_{t1} - PP'_t$ beschrijft de *increase* (toename) of *decrease* (afname) van de afstand tussen de punten P en P'.
Op basis van de temporele consistentie van middenniveauparameters codeert een op regels gebaseerde methode temporele segmenten (*onset, apex,*

offset) van 27 au's zoals die alleen of in combinatie voorkomen in de inputgezichtsvideo's. Om bijvoorbeeld de tijdelijke segmenten van AU4 te herkennen, waarbij de wenkbrauwen dichter bij elkaar worden getrokken, gebruiken we de volgende temporele regels (zie figuur 1 voor het gezichtsmodel en de betekenis van de punten D en D1):

ALS $([inc/dec(\mathrm{DD1})]_t > [inc/dec(\mathrm{DD1})]_{t-1} + \varepsilon)$
EN $inc/dec(\mathrm{DD1}) > \varepsilon$ DAN ***AU4-onset***
ALS $|\,[inc/dec(\mathrm{DD1})]_t - [inc/dec(\mathrm{DD1})]_{t-1}\,| \leq \varepsilon$
EN $inc/dec(\mathrm{DD1}) > \varepsilon$ DAN ***AU4-apex***
ALS $([inc/dec(\mathrm{DD1})]_t < [inc/dec(\mathrm{DD1})]_{t-1} - \varepsilon)$
EN $inc/dec(\mathrm{DD1}) > \varepsilon$ DAN ***AU4-offset***

Bij het uittesten op de Cohn-Kanade Facial Expression Database (Kanade, Cohn & Tian, 2000) en onze eigen MMI Facial Expression Database (Pantic e.a., 2005) behaalde de voorgestelde methode een herkenningsgraad van 90 procent bij het waarnemen van 27 au's die elk op zich of in combinatie voorkwamen in een inputreeks van beelden van gezichten.

Interpretatie van gezichtsuitdrukkingen

Vrijwel alle systemen voor het automatisch analyseren van gezichtsaffecten trachten een kleine groep universele emoties of basisemoties te herkennen (Pantic & Rothkrantz, 2003). Echter, pure uitdrukkingen van basisemoties komen zelden voor; mensen tonen meestal een mengsel van emoties (Keltner & Ekman, 2000). Daarom is de classificatie van menselijke non-verbale affectieve feedback naar een enkele basisemotiecategorie niet realistisch. Evenzo kunnen niet alle non-verbale affectieve signalen worden geclassificeerd als een combinatie van de basisemotiecategorieën. Denk bijvoorbeeld aan frustratie, twijfel of verveling. Bovendien is aangetoond dat het begrip van een bepaald emotielabel en de manieren waarop de daarbij behorende affectieve toestand wordt uitgedrukt, kunnen verschillen van cultuur tot cultuur en zelfs van persoon tot persoon (Russell & Fernandez-Dols, 1997). Daarom zijn pragmatische keuzes noodzakelijk (keuzes aangepast aan het gebruikersprofiel) voor wat betreft de selectie van de affectieve toestanden die worden herkend door een automatische analysator van menselijke affectieve feedback.
Daarom hebben wij een systeem van diagnostisch redeneren ontwikkeld waarbij au's worden geclassificeerd in emotiecategorieën zoals die van de gebruiker geleerd zijn. De case die wordt toegepast, betreft een dynamisch, in toenemende mate auto-organiserend *event-content*-addresseerbaar geheugen dat het mogelijk maakt feiten te achterhalen en events te evalueren op basis van de door de gebruiker ingegeven voorkeuren en de generalisaties zoals die ontstonden na eerdere input. Elk event (case) bestaat uit een of meer micro-events, die elk bestaan uit een serie au's. Micro-events die gezamenlijk tot doel hebben een specifieke affectieve toestand te communiceren, zijn gegroepeerd binnen een en hetzelfde dynamische geheugenblok. Met andere woorden, elk geheugenblok vertegenwoordigt een specifieke emotiecategorie en omvat alle micro-events waaraan de gebruiker het emotielabel in kwestie heeft toegewezen. De indexen behorend bij elk dynamisch geheugenblok zijn opgebouwd uit individuele au's en combinaties van au's die het meest karakteristiek zijn voor die bepaalde emotiecategorie. Ten slotte zijn alle micro-events van elk dynamisch geheugenblok hiërarchisch geordend in overeenstemming met hun typische karakter: hoe vaker een bepaald micro-event voorkwam, des te hoger is de hiërarchische positie binnen dat bepaalde blok. De initiële kwaliteit van het dynamische geheugen wordt bereikt door de gebruiker te vragen een interpretatielabel (emotielabel) te associëren met een reeks van 40 typische gezichtsuitdrukkingen (micro-events die onlosmakelijk verbonden zijn met emoties volgens Russel & Fernandez-Dols (1997)). De classificatie van de au's zoals waarneembaar in een inputbeeld van een gezicht in de emotiecategorieën zoals die geleerd zijn van de gebruiker, wordt vervolgens bereikt door middel van diagnostisch redeneren over de inhoud van het dynamische geheugen. Om een nieuw probleem met betrekking tot classificatie van een reeks input-au's in door de gebruiker gedefinieerde interpretatiecategorieën op te lossen, worden de volgende stappen genomen:

1. Doorzoek het dynamische geheugen op gelijksoortige cases, roep deze opnieuw op en interpreteer de input-au-reeks gebruikmakend van de interpretaties zoals ingegeven door de opgeroepen cases.
2. Is de gebruiker tevreden met de interpretatie, sla de case dan op in het dynamische geheugen. Is dat niet het geval, pas het geheugen dan aan volgens de door de gebruiker geleverde feedback over de interpretatie die hij associeert met de inputgelaatsuitdrukking.

De toegepaste oproep- en aanpassingsalgoritmes maken gebruik van een voorselectie van cases die

is gebaseerd op de groepsgewijze rangschikking van het dynamisch geheugen, de indexerende structuur van het geheugen en de hiërarchische rangschikking van cases binnen de groepen/blokken al naargelang hun typische karakter.
Er zijn twee validatieonderzoeken uitgevoerd op een prototypesysteem. De vraag waarop het eerste validatieonderzoek zich richtte was: hoe acceptabel zijn de interpretaties die het systeem oplevert nadat het heeft geleerd zes basisemoties te herkennen (namelijk angst, verdriet, walging, blijdschap, boosheid en verrassing)? De vraag die tijdens het tweede validatieonderzoek werd gesteld was: hoe acceptabel zijn de interpretaties die het systeem oplevert nadat het is getraind om een willekeurig aantal door de gebruiker gedefinieerde interpretatiecategorieën te herkennen? In het eerste geval werd een menselijke FACS-codeerder gevraagd het systeem te trainen. In het tweede geval werd een amateurexpert, zonder formele opleiding in het herkennen van emotiesignalen, gevraagd het systeem te trainen. Dezelfde expert die werd gebruikt om het systeem op te leiden werd tevens gebruikt om de prestaties te evalueren, dat wil zeggen om te beoordelen hoe acceptabel de interpretaties die het systeem gaf waren. Wat betreft basisemoties was de expert het in 100 procent van de testcases eens met de interpretaties die het systeem opleverde. Wat betreft de door de gebruiker gedefinieerde interpretatiecategorieën was de amateurexpert het in 83 procent van de testcases geheel eens met de interpretaties en in 14 procent van de testcases was de expert het eens met de meeste maar niet met alle interpretatielabels die het systeem aan de relevante cases toekende.

Conclusie

Onze benadering van de waarneming van au's vormt op twee manieren een uitbreiding van de stand van zaken op dit gebied, namelijk door de temporele weergave van gezichtsuitdrukkingen en het aantal gehanteerde au's (27 au's in totaal). De geautomatiseerde systemen voor au-detectie op basis van gezichtsvideo's die tot dusver zijn gerapporteerd, gaan namelijk vooral in op het probleem dat ruimtelijke weergave van gezichtsuitdrukkingen met zich meebrengt en kunnen – in het gunstigste geval – 16 tot 18 au's (uit een totaal van 44 au's) herkennen. Onze methode verbetert tevens andere aspecten rond geautomatiseerde au-detectie in vergelijking met eerdere onderzoeken. De prestaties van de methode worden niet beïnvloed door occlusies zoals brillen en gezichtsbehaaring zolang deze de benodigde referentiepunten op het gezicht niet geheel verbergen. De methode werkt bovendien goed onder alle lichtomstandigheden en is minder gevoelig voor veranderingen in lichtintensiteit. Gegeven het feit dat eerder geïntroduceerde gezichtsuitdrukkingsanalysatoren slechts in staat zijn gezichtsuitdrukkingen te classificeren in een van de zes basisemotiecategorieën, betekent onze methode voor het automatisch interpreteren van gezichtsuitdrukkingen een uitbreiding van de technologie op dit gebied doordat het hiermee mogelijk wordt gezichtsuitdrukkingen te interpreteren op een aan de gebruiker aan te passen wijze.
Onze methode is echter nog steeds niet in staat het gehele scala van gezichtsuitdrukkingen (dat wil zeggen alle 44 au's zoals gedefinieerd in FACS) te herkennen. Bovendien gaat de methode uit van invoerdata bestaande uit geïsoleerde beelden van gezichten die al dan niet van tevoren zijn opgedeeld en slechts een enkel temporeel patroon vertonen (onset, apex, offset) van een uitdrukking die begint en eindigt met een neutrale toestand. In werkelijkheid bestaat een dergelijke verdeling niet; het gedrag van het menselijk gezicht is veel complexer en overgangen van de ene naar de andere (emotionele) gelaatsuitdrukking worden niet noodzakelijkerwijs afgewisseld met een neutrale toestand. Daarom kunnen onze gezichtsuitdrukkingsanalysatoren niet omgaan met spontaan vertoonde gezichtsuitdrukkingen. Er is meer onderzoek nodig willen we het gehele scala van menselijke (spontane en geposeerde) gezichtsuitdrukkingen automatisch kunnen coderen.

Literatuur

Ekman, P. & W.V. Friesen (1978). *FACS Manual*. Palo Alto: Consulting Psychologist Press.

Kanade, T., J. Cohn & Y. Tian, (2000). Comprehensive database for facial expression analysis. *Proc. Int'l Conf. Face and Gesture Recognition (FGR'00)*, pp. 46-53.

Keltner, D. & P. Ekman, (2000). Facial Expression of emotion. *Handbook of Emotions*. New York: Guilford Press, pp. 236-249.

Pantic, M. (2005). Face for Interface. *The Encyclopedia of Multimedia Technology and Networking*. Hershey: Idea Group Publishing, vol. 1, pp. 308-314.

Pantic, M. e.a. (2005). Web-based database for facial expression analysis. *Proc. Int'l Conf. Multimedia and Expo (ICME'05)*,pp. 317-321.

Pantic, M. & L.J.M. Rothkrantz (2003). Toward an Affect-Sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE*, vol. 91, pp. 1370-1390.

Pitt, M.K.; & N. Shephard (1999). Filtering via simulation: auxiliary particle filtering, *J. Amer. Stat. Assoc.*, vol. 94, pp. 590-599.

Russell, J. & J. Fernandez-Dols (1997). *The psychology of facial expression*, Cambridge: Cambridge University Press.

Link

mmi.tudelft.nl/~maja

Maja Pantic
is werkzaam bij het Computing Department van het Imperial College in Londen. E-mail: m.pantic@imperial.ac.uk.