NIEUWHEIDSDETECTIE VOOR ACTIVITEITHERKENNING DOOR ROBOTS

Scène uit de film *Chappie* (2015) van Neill Blomkamp

Thomas Moerland

Robots zullen de komende jaren in toenemende mate ons dagelijks leven binnentreden. Zo lijken robots onder andere een veelbelovende rol te kunnen gaan spelen bij het verzorgen van ouderen. Echter, daar waar robots in de industrie in relatief gecontroleerde productie-omgevingen werken, zorgt het alledaagse leven voor een enorme variatie aan situaties. De robot moet functioneren in een zogeheten 'open' omgeving met veel verschillende woonkamers en de inherent onzekere interactie met mensen. Zo'n open omgeving stelt ons voor meerdere uitdagingen, waarvan we er in dit artikel een behandelen.

Menselijke activiteit

Een van de vereiste taken voor thuisrobots is het kunnen herkennen van menselijke activiteit. De robot heeft een camera die een continue stroom van hoog-dimensionale RGB-D (diepte)-beelden produceert. Het doel van activiteitherkenning is om te detecteren (classificeren) of een persoon voor de camera op dit moment aan het eten is, leest, of misschien zojuist gevallen is en er alarm geslagen moet worden. De gebruikelijke benadering van dit type problemen is via *machine learning*. We verzamelen een trainingsset van gelabelde voorbeeld-video's, en leren daarmee een (geparametriseerde) *predictor/classifier* die voorspellingen kan maken op nieuwe, ongelabelde data. Deze taak is onder meer uitdagend vanwege de dimensionaliteit van de inputdata en door de sterke temporele correlatie (een activiteit wordt vaak gedefinieerd door een tijdsafhankelijke serie van menselijke posities, en is zelden aan een los beeld te herkennen).

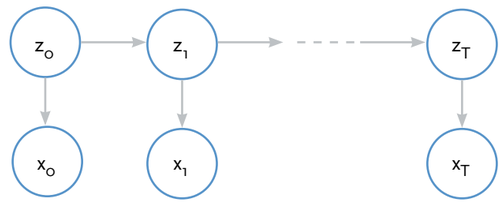
De 'open' omgeving van de thuissituatie schept echter nog een extra uitdaging. Stel dat het is gelukt om de bovengenoemde *classifier* te leren om nieuwe activiteiten te classificeren, dan nog zullen we 1. nooit alle menselijke activiteiten in de trainingsset hebben kunnen onderbrengen, en 2. bij verschillende robottoepassingen verschillende sets van activiteiten observeren. Een classifier die een activiteit uit een onbekende klasse toch toewijst aan de beste passende klasse uit de trainingsset zit per definitief fout. Dit probleem staat bekend als 'nieuwheidsdetectie', ofwel de detectie van data uit een onbekende en nieuwe klasse, en is tevens een eerste vereiste voor een zelflerend (zichzelf uitbreidend) systeem. In dit artikel bestuderen we nieuwheidsdetectie in de context van activiteitherkenning.

Activiteitherkenning

We beschrijven nu eerst een activiteit-classifier gebaseerd op het Hidden Markov Model (Baum & Petrie, 1966). De trainingsdata bestaan uit een set gelabelde video's $\{X_i,s_i\}_{i=1}^N$ voor videoklasselabels $s \in S = \{1,2,..,M\}$. Elke video X_i van lengte T_i bestaat uit een serie observatievectoren $x_{i,t}$ voor frame-index t. In de praktijk gebruiken we nog enkele voorbewerkingsstappen om de frame feature vector $x_{i,t}$ te krijgen (voor details zie Moerland, Chandarr, Rudinac & Jonker, 2016).

Om de temporele correlatie tussen de observatie vector te vatten gebruiken we de structuur van het Hidden Markov Model (HMM), een grafisch model zoals weergegeven in figuur 1. We gaan ervan uit dat de transities tussen de tijdsstappen te modelleren zijn op een discrete, latente laag. De latente laag variabele $z_{i,t}$ neemt waardes aan in een set van 'kernhoudingen' $W = \{w_1,w_2,...w_K\}$. Iedere kernhouding omschrijft een bepaalde positie van het lichaam in de observatievector. Het transitiemodel tussen de tijdsstappen is te beschrijven als een transitiematrix A van grootte K x K, waarbij $A_{k,l} = P(z_t = l \mid z_{t-1} = k)$. De relatie tussen de latente variabelen en de observatievector is gegeven door een diagonale Gaussiaans 'emissie verdeling' per klasse: $P(x_t|z_t=k) = N(.|\mu_k,\Sigma_k)$. Het voorgaande definieert een generatief model van een video.

Om onderscheid te maken tussen videoklassen trainen we een aparte HMM per videoklasse. In praktijk laten we alleen de transitiemodellen verschillen, met A^s het transitiemodel voor klasse s. Een activiteit wordt



Figuur 1. Hidden Markov Model

daarbij als een serie van lichaamsposities gemodelleerd, waarbij voor verschillende activiteiten ook verschillende transities te verwachten zijn (welke we schatten uit de trainingsdata). De emissieverdelingen worden gedeeld tussen de klassen (het is aannemelijk dat bepaalde kernhoudingen voorkomen in meerdere activiteitklassen, en door deze gezamenlijk te schatten winnen we statistische efficiëntie). De volledige HMM parameters zijn daarmee gegeven door $\phi = \{\theta, \mu, \Sigma\}$, met $\theta = \{A^1,.. A^M\}$ de klassespecifieke transitiemodellen, en $\mu = \{\mu_1,..\mu_K\}$ en $\Sigma = \{\Sigma_1,..\Sigma_K\}$ de emissiemodellen per kernhouding.

De parameters van bovenstaand model zijn te schatten met maximum-likelihood. Het enige probleem is dat we de latente variabelen z niet geobserveerd hebben. We kunnen dit oplossen door toepassing van Expectation-Maximization (EM) (Dempster, Laird, & Rubin, 1977), waarbij we in de E-stap de parameters ϕ fixeren en de conditionele verwachting van alle $z_{i,t}$ variabelen berekenen, en in de M-stap deze z variabelen fixeren en maximaliseren ten opzichte van ϕ. Dit herhalen we tot convergentie van de parameters.

Voor een nieuwe video kunnen we nu de kans onder elke klasse s berekenen: $p(X|s)$. Voor classificatie gebruiken we vervolgens de maximum-a-posteriori (MAP) regel:

$$\hat{s} = \arg\max_{s \in \mathcal{S}} P(s|X) = \arg\max_{s \in \mathcal{S}} \frac{p(X|s)p(s)}{p(X)} = \arg\max_{s \in \mathcal{S}} p(X|s) \quad (1)$$

In de laatste versimpeling nemen we de prior over de klassen, $p(s)$, uniform aan, en kunnen we de marginale videokans, $p(X)$, negeren omdat het de arg max niet beïnvloedt.

Nieuwheidsdetectie

We gaan nu verder met de nieuwheidsdetectie-methodologie, die verder bouwt op de MAP-beslissingsregel uit de vorige paragraaf. Een voor de hand liggend idee voor nieuwheidsdetectie is om de kans van de video gegeven de klasse, $P(X|s)$, 'af te kappen'. Als de geobserveerde video een lage kans heeft onder alle klassen dan is het waarschijnlijk dat we een nieuwe klasse hebben getroffen. Dit blijkt echter in de praktijk niet optimaal te werken, voornamelijk omdat het moeilijk is om op deze manier onderscheid te maken tussen een daadwerkelijk nieuwe klasse en een ruizige video van een bekende klasse.

Een theoretisch meer intuïtieve maat om af te kappen is de kans van de klasse gegeven de video, $p(s|X)$, zoals gegeven in (1). Hiervoor kunnen we de prior over de klassen $P(s)$ nog steeds uniform aannemen, maar hebben we een schatter van de marginale video kans $p(X)$ nodig. Voor een nieuwe klasse zal de $p(X)$ term relatief groot zijn, terwijl deze voor een ruizige video laag zal uitvallen.

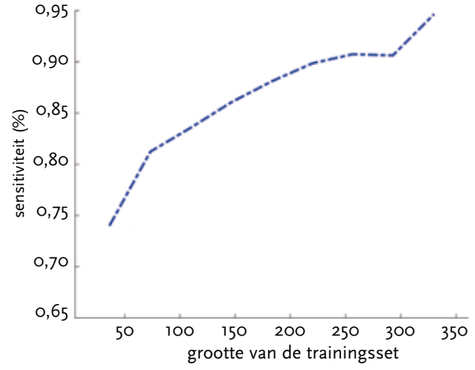
De marginale video kans is gegeven door $P(X) = \sum_G P(X|G)P(G)$, waarbij G de volledige modelruimte beschrijft, inclusief alle onbekende klassen. Probleem is echter dat we deze klassen juist niet geobserveerd hadden. Daarom introduceren we nu 'achtergrondmodellen'. Achtergrondmodellen zijn nieuwe HMM'en welke dezelfde trainingsset, kernhoudingen en emissiemodellen als de modellen uit de vorige paragraaf gebruiken. We zullen nu echter een nieuwe – mogelijke klassespecifieke – achtergrond transitiematrix $A^*_{(s)}$ trainen.

We beschrijven drie typen achtergrondmodellen. De eerste, het 'vul model', schat een nieuwe, generieke transitiematrix op alle data tezamen. Deze moet grofweg alle mogelijke menselijke transities (bewegingen) vatten, en is daarmee een goede kandidaat om $P(X)$ te benaderen. Het tweede type achtergrondmodel is het 'antimodel'. Dit is klasse-specifiek (elke klasse heeft zijn eigen antimodel), en wordt getraind op alle data die *niet* bij die klasse horen, herwogen naar de kans om verkeerd herkend te worden als de bewuste klasse. Daarmee zijn ze bedoeld om de regio rondom de ware kansregio van een klasse te modelleren, wat voor betere separatie zou kunnen zorgen. Als derde proberen we het 'vlakke model' dat de transitiematrix uniform initialiseert.

Uiteindelijk wordt een testvideo gedecodeerd onder de herkenningsmodellen uit de vorige paragraaf en onder de achtergrondmodellen in deze paragraaf om de (log) kans $log\ p(s|X) = log\ p(X|s) - log\ p(X)$ te bepalen. Op deze statistiek wordt vervolgens een afkapwaarde τ bepaald op de trainingsset. Als $\log p(s|X) \geq \tau$ dan identificeren we de video als 'bekend' en gaan we verder met standaardclassificatie, en als $\log p(s|X) < \tau$ dan identificeren we de video als onbekend of nieuw.

Experimenten

We valideren onze methode met de publiek beschikbare Microsoft Research Action (MSRA) 3D-dataset, waaruit we 15 actieklassen selecteren (onder andere zwaaien, gooien, klappen, etc.) met in totaal 366 video's. Alle resultaten zijn gemiddeldes over 3 herhalingen van een drievoudige kruisvalidatie. Eerst testen we de sensitiviteit (accuraatheid) van de HMM-classifier in de standaard training/test set-up zonder onbekende klassen.



Figuur 2. Leercurve als functie van de grootte van de trainingsset; door de random split is de sensitiviteit iets lager dan beschreven in de tekst.

De leercurve voor verschillende dataset-groottes is te zien in figuur 2. In de optimale setting haalt het systeem ongeveer 96% classificatie-accuraatheid, wat rond de *state of the art* voor deze dataset ligt.

Voor nieuwheidsdetectie maken we een dubbele dataset-split, waarbij we eerst gerandomiseerd 3 klassen afscheiden als 'onbekend' en de video's van deze klassen niet gebruiken voor het trainen van de HMM'en. Op deze (moeilijker) classificatietaak, waarbij in de test set zowel bekende als onbekende klassen voorkomen, haalt het systeem zonder achtergrondmodel 71% sensitiviteit (juist toegewezen bekende klassen) en 57% specificiteit (juist gedetecteerde nieuwe klassen). De achtergrondmodellen verbeteren deze scores allemaal. De beste resultaten worden bereikt door een half-half gewogen combinatie van vlakke en anti-modellen, met een sensitiviteit/specificiteit van 78% / 78%.

Tabel 1 geeft dit resultaat nog gedetailleerder weer. We zien hoe we drie typen fouten kunnen maken. Allereerst op het nieuwheidsniveau, waar we bekend als nieuw (21%) of nieuw als bekend (22%) kunnen toewijzen. Echter, we kunnen ook een 'putatieve fout' maken, waarbij een video correct als bekend wordt geïdentificeerd, maar vervolgens aan de verkeerde klasse wordt toegewezen. Deze laatste fout komt maar zeer zelden voor (1%) als we achtergrondmodellen gebruiken, hetgeen aangeeft dat ze ook bruikbaar zijn om de robuustheid te vergroten in gesloten-setproblemen. Dat gaat dan wel ten koste van een aantal gevallen waarbij we weigeren te classificeren. (Voor meer gedetailleerde resultaten zie Moerland, Chandarr, Rudinac & Jonker, 2016).

VOORSPELDE LABEL	WARE LABEL	
	bekend	nieuw
bekend (correct)	78%	22%
bekend (foutief	1%	–
nieuw	21%	78%

Tabel 1. Test set resultaten voor gewogen combinatie van vlakke en anti-modellen

Conclusie

Dit artikel beschrijft een nieuwe methode om de robuustheid van actieherkenning in open omgevingen te vergroten. Onze nieuwheidsdetectie procedure gebaseerd op achtergrondmodellen is in staat tot 78% van de nieuwe video's te filteren voordat ze foutief geclassificeerd worden. De methodiek is tevens geschikt voor andere velden (dan de hier beschreven activiteitherkenning) waar HMM classificatie van toepassing is.

In breder perspectief is onzekerheidskwantificering, van oudsher het domein van de statistiek, ook een cruciaal probleem binnen robotica en machine learning. Alternatieve methoden om onzekerheid te kwantificeren, zowel frequentistisch (e.g., de niet-parameterische bootstrap) als Bayesiaans (inferentie op modelparameters) van aard, vinden in toenemende mate toepassing binnen hoog-dimensionale machine-learning-problemen, en zouden ook op het hier beschreven probleem van toepassing kunnen zijn.

Literatuur

Moerland T.M., Chandarr A., Rudinac M., & Jonker P.P (2016). Knowing What You Don't Know: Novelty Detection for Action Recognition in Personal Robots. In *Proceedings of VISIGRAPP 2016, Vol. 4: VISAPP* (pp. 317–327).

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics, 37*(6), 1554–1563.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society.* Series B (Methodological), *39*(1), 1–38.

Thomas Moerland studeerde af in wiskunde (summa cum laude) en geneeskunde (cum laude) aan de Universiteit Leiden. Hij is momenteel als promovendus verbonden aan de afdeling Kunstmatige Intelligentie van de Technische Universiteit Delft. Zijn onderzoek focust op sequentiële beslissingsproblemen, in het bijzonder op het snijpunt van *reinforcement learning* en Bayesiaanse inferentie in (diepe) neurale netwerken. Hij ontving de VvS+OR Jan Hemelrijkprijs 2017.
E-mail: T.M.Moerland@tudelft.nl