Nieuwheidsdetectie voor Activiteit Herkenning door Robots

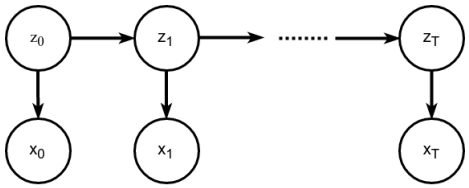
Thomas Moerland
T.M.Moerland@tudelft.nl

1 Introductie

Robots zullen in de komende jaren in toenemende mate ons dagelijks leven binnentreden. Een belangrijke toepassing zijn de thuis-robots, die onder andere een oplossing voor het toenemende ouderenzorg probleem kunnen worden. Echter, de vereiste interactie met mensen en het functioneren in een 'open omgeving' (in tegenstelling tot een relatief gecontroleerde productieomgeving) creëert meerdere uitdagingen.

Een van de vereiste taken voor thuis-robots is het kunnen herkennen van menselijke activiteit. De robot heeft een camera die een continue stroom van hoog-dimensionale RGB(-D) beelden produceert. Het doel van activiteit herkenning is voorspellen (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 voor dit soort hoog-dimensionale voorspel 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 in dit geval onder andere uitdagend vanwege de dimensionaliteit van de input data 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 creëert echter nog een extra uitdaging. Stel dat het is gelukt om bovenstaande activiteit classifier te leren, dan zullen we nog 1) nooit alle menselijke activiteiten in de trainingsset hebben, en 2) in verschillende robot toepassingen 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/nieuwe klasse, en is tevens een eerste vereiste voor een zelf-lerend (zichzelf uitbreidend) systeem. In dit artikel bestuderen we nieuwheidsdetectie in de context van activiteitsherkenning.



Figuur 1. Hidden Markov Model.

2 Activiteit Herkenning

We beschrijven nu eerst een activiteit classifier gebaseerd op het Hidden Markov Model (HMM) [2]. De trainingsdata bestaat uit een set gelabelde video's $\{X_i, s_i\}_{i=1}^N$ voor video klasse labels $s \in \mathcal{S} = \{1, 2, .., M\}$. Elke video X_i van lengte T_i bestaat uit een serie observatie vectoren $x_{i,t}$ voor frame-index t. In de praktijk gebruiken we nog enkele voorbewerkingsstappen om de frame feature-vector $x_{i,t}$ te krijgen (zie [1] voor details).

Om de temporele correlatie tussen de observatie vector te vatten nemen we een HMM structuur aan, een grafisch model zoals weergegeven in Figuur 2. We nemen aan 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 'kern houdingen' $W = \{w_1, w_2, ...w_K\}$. Iedere kern houding omschrijft een bepaalde positie van het lichaam in de observatie vector. Het transitiemodel tussen de tijdsstappen is te beschrijven als een transitiematrix A van grootte $K \times K$, waarbij $A_{k,l} = P(z_t = l|z_{t-1} = k)$. De relatie tussen de latente variabelen en de observatie vector is gegeven door een diagonale Gaussiaans 'emissie verdeling' per klasse: $P(x_t|z_t = k) = \mathcal{N}(\cdot|\mu_k, \Sigma_k)$. Het voorgaande definieert een generatief model van een video.

Om onderscheid te maken tussen video klassen trainen we een aparte HMM per video klasse. In praktijk laten we alleen de transitie modellen verschillen, met A^s het transitiemodel voor klasse s. Een activiteit wordt daarbij als een serie van lichaamsposities gemodelleerd, waarbij voor verschillende activiteiten ook verschillende transities te verwachten zijn (welke we schatten uit de trainingsdata). De emissie verdelingen worden gedeeld tussen de klassen (het is aannemelijk dat bepaalde kern houdingen voorkomen in meerdere activiteit klassen, 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 klasse-specifieke transitie modellen, en $\mu = \{\mu_1, ..\mu_K\}$ en $\Sigma = \{\Sigma_1, ..\Sigma_K\}$ de emissie modellen per kern houding.

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) [3], 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-step 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:

$$\begin{aligned}\hat{s} = \arg\max_{s \in \mathcal{S}} P(s|X) &= \arg\max_{s \in \mathcal{S}} \frac{p(X|s)p(s)}{p(X)} \\ &= \arg\max_{s \in \mathcal{S}} p(X|s) \end{aligned} \tag{1}$$

In de laatste versimpeling nemen we de prior over de klassen, $p(s)$, uniform aan, en kunnen we de marginale video kans, $p(X)$, negeren omdat het de argmax niet beïnvloed.

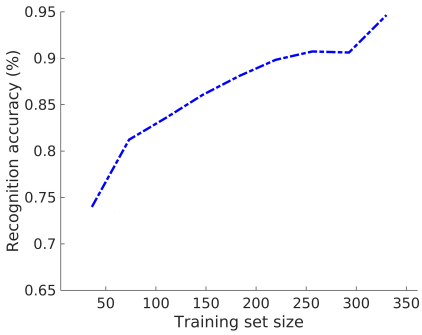
3 Nieuwheidsdetectie

We gaan nu verder met de nieuwheidsdetectie methodologie, welke 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 blijkt op deze manier onderscheid te maken tussen een daadwerkelijk nieuwe klasse en een ruizige video van een bekende klasse.

Een theoretisch intuïtivere 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 model ruimte beschrijft, inclusief alle onbekende klassen. Probleem is echter dat we deze klassen juist niet geobserveerd hadden. Daarom introduceren we nu 'achtergrond modellen'. Achtergrond modellen zijn nieuwe HMM'en welke dezelfde trainingsset, kernhoudingen en emissie modellen als de modellen uit Sectie 2 gebruiken. We zullen nu echter een nieuwe (mogelijke klasse-specifieke) achtergrond transitiematrix $A^{\star}_{(s)}$ trainen.

We beschrijven drie typen achtergrond modellen. 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 achtergrond model is het 'anti-model'. Deze is klasse-specifiek (elke klasse heeft zijn eigen anti-model), en wordt getraind op alle data die *niet* bij die klasse hoort, herwogen naar de kans om verkeerd herkend te worden als de bewuste klasse. Daarmee zijn ze bedoelt om de regio rondom de ware kans regio van een klasse te modelleren, wat voor betere separatie zou kunnen zorgen. Als derde proberen we het 'vlakke model', welke de transitiematrix uniform initialiseert.



Figuur 2. Leercurve als functie van de grootte van de trainingsset. Voor dit plaatje maken we een random split, waardoor de sensitiviteit iets lager is dan zoals beschreven in de tekst.

Uiteindelijk wordt een test video gedecodeerd onder de herkenningsmodellen (Sectie 2) en onder de achtergrond modellen (Sectie 3) 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 bepaalt op de trainingsset. Als $\log p(s|X) \geq \tau$ dan identificeren we de video als 'bekend' en gaan we verder met standaard classificatie (Sectie 2), en als $\log p(s|X) \leq \tau$ dan identificeren we de video als onbekend/nieuw.

4 Experimenten

We valideren onze methode met de publiek beschikbare 'Microsoft Research Action (MSRA) 3D' dataset, waaruit we 15 actie klassen selecteren (onder andere zwaaien, gooien, klappen, etc.) met in totaal 366 video's. All resultaten zijn gemiddeldes over 3 herhalingen van een 3-voudige kruisvalidatie. Eerst testen we de sensitiviteit (accuraatheid) van de HMM classifier (Sectie 2) in de standaard training/test set-up zonder onbekende klassen. 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 (moeilijkere) classificatie taak, 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 achtergrond modellen 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

	Ware label	
Voorspelde label	**Bekend**	**Nieuw**
Bekend (correct)	78%	22%
Bekend (foutief)	1%	-
Nieuw	21%	78%

Figuur 3. Test set resultaten voor gewogen combinatie van vlakke en anti-modellen.

typen fouten kunnen maken. Allereerst op het nieuwheid niveau, 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 achtergrond modellen gebruiken, hetgeen aangeeft dat ze ook bruikbaar zijn om robuustheid te vergroten in gesloten set problemen (hetgeen dan wel ten koste gaat van een aantal gevallen waarin we weigeren te classificeren).

Voor een gedetailleerdere discussie van resultaten zie [1].

5 Conclusie

Dit artikel beschreef een nieuwe methode om de robuustheid van actie herkenning in open omgevingen te vergroten. Onze nieuwheidsdetectie procedure gebaseerd op achtergrond modellen 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 activiteit herkenning) 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 model parameters) van aard, vinden in toenemende mate hun toepassing in de machine learning, en zouden ook op het hier beschreven probleem van toepassing kunnen zijn.

[1] Moerland T.M., Chandarr A., Rudinac M. and Jonker P.P (2016). Knowing What You Dont Know: Novelty Detection for Action Recognition in Personal Robots. VISIGRAPP Vol. 4: VISAPP, 317-327.
[2] Baum, L. E., and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. The annals of mathematical statistics, 37(6), 1554-1563.
[3] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood

from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.

Thomas Moerland (1988) 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. Voor bovenstaand werk ontving hij de Vvs+OR Jan Hemelrijk prijs 2017.