Nuances in visual recognition

Gavves, E.

Publication date
2014

SAMENVATTING

Het onderwerp van deze dissertatie is het automatisch begrijpen van fijnmazige visuele nuances. We starten met rigide objecten zoals gebouwen en monumenten en behandelen generieke object categorieën om vervolgens te eindigen met fijnmazige diersoorten zoals vogels en honden. Alvorens terug te komen op de onderzoeksvragen uit de introductie, volgt hier een kort overzicht van de hoofdstukken.

In Hoofdstuk 2, behandelen we de incoherentie van zogenaamde visuele woorden. Een visueel woord is een afgevaardigde van een groep visueel gelijke stukken beeld. Andere onderzoekers wijzen zulke visuele woorden vaak toe op basis van gelijkheid in een beschrijvingsruimte. Wij onderscheiden ons door paren van afzonderlijke visuele woorden te groeperen die elk dezelfde visuele werkelijkheid beschrijven (visuele synoniemen). We richten ons op monumenten en gebouwen met een rigide geometrie. Deze geometrie maakt het mogelijk om dezelfde fysieke locaties te berekenen over verschillende aangezicht beelden. Een visueel synoniem is een ander visueel woord op dezelfde fysieke locatie. Een analyse van de visuele synoniemen maakt het mogelijk om 98% van de redundante woorden te verwijderen zonder de herkennings nauwkeurigheid geweld aan te doen.

In Hoofdstuk 3 beschrijven we de toepasbaarheid van visuele woorden voor het terugzoeken van bekende objecten in beelden. Andere onderzoekers tonen het voordeel aan van een groot visuele woordenboek, tot maar liefst 1 miljoen woorden. Wij richten ons op kleine, robuuste, woordenboeken die we verkrijgen door te concentreren op de meest informatieve visuele woorden. Een volledige combinatorische analyse van visuele woord combinaties is computationeel onmogelijk en daarom benaderen we het probleem met behulp van kruisentropy. We monsteren visuele woord combinaties die weinigvoorkomende, informatieve, visuele informatie beschrijven. In tegenstelling tot Hoofdstuk 2 maken we hier geen expliciet gebruik van rigide geometrie. Niettemin vinden wij woorden die zowel visueel als geometrisch betekenisvol zijn. Een evaluatie op bekende monumenten en op logos laat zien dat zelfs met een reductie van 95% van het visueel woordenboek de zoek nauwkeurigheid beter is dan andere methodes met een veel groter woordenboek.

In Hoofdstuk 4 zetten we uiteen wat de beperkende factoren zijn van snelle en effectieve beeld classificatie in relatie tot het aantal geanalyseerde beelden *en* de dimensionaliteit van de beeld representatie. Onder de aanname van een voortdurende groei van het aantal beschikbare beelden willen we in dit hoofdstuk de dimensionaliteit van de beeld representatie beperken. Hiervoor stellen we voor om gereduceerde lineaire kernels te gebruiken die maar een fractie van de representatie dimensionaliteit gebruiken. We formuleren dit probleem als een convex optimalisatie probleem dan efficiënt opgelost kan worden. In contrast tot andere reductie methodes is onze voorgestelde methode sneller en nauwkeuriger. We laten dit zien op zowel de Scenes en de PASCAL VOC 2007 data sets. Ook laten we zien dat een combinatie met de Fisher Vector en met niet-lineaire embeddingen tot hoge nauwkeurigheid leidt. Wederom laat een analyse van de gekozen kernel dimensies zien dat visueel betekenisvolle structuren worden gevonden zonder gebruik te maken van geannoteerde voorbeelden.

Het doel van Hoofdstuk 5 is automatisch fijnmazige visuele classificatie zonder menselijke interactie. Andere onderzoekers gebruiken object onderdelen waarvoor ze specifieke detectoren bouwen. In tegenstelling tot specifieke detectoren, stellen wij voor om objecten in hun geheel ruwweg uit te lijnen op basis van de globale vorm. Deze uitlijning maakt het mogelijk om correspondenties te vinden naar gelabelde onderdelen in een set voorbeelden met supervisie, of blind maar consistent zonder supervisie. We beargumenteren dat voor het herkennen van sub-classes verdeling-kenmerken belangrijk zijn, zoals de kleuren Fisher Vector. Deze kunnen de locale visuele omgeving beter beschrijven dan traditionele orientatie histogrammen zoals HOG. Onze methode vangt subtiele lokale verschillen tussen sub-classes, en is tegelijkertijd robuust tegen fouten in de uitlijning. We evalueren onze aanpak op de CUB-2011 en de Stanford honden dataset, die uit 200 en 120 soorten vogels en honden die visueel alleen door een expert te scheiden zijn. In onze experimenten bestuderen we het voordeel van de kleuren Fisher Vector, de invloed van de uitlijning, en de invloed van object segmentatie op fijnmazige classificatie. Een ruwe uitlijning mengt natuurlijk samen met standaard object-hypothese algoritmes, wat resulteert in een volledig automatisch systeem zonder enige menselijke interactie. Onze resultaten verbeteren alle methodes tot nu toe op beiden data sets. Een analyse van de lokale uitlijning laat zien welke details het belangrijkst bleek voor elke fijnmazige categorie.

In Hoofdstuk 6 mikken we op segmentatie en classificatie van objecten. We stellen *codemaps* voor, die een gezamenlijk formulatie mogelijk maken van de classificatie score en de lokale regio in een beeld. We verkrijgen codemaps door de codering, groepering en classificatie stappen te herordenen. We geven een theoretische generalisatie van de wiskundige eisen aan een decomposie van beeld codering en classificatie. Waar bestaande lineaire decomposities de nadruk alleen op de efficiëntie leggen, heeft onze methode drie nieuwe bijdrages. Als een eerste bijdrage introduceren we ℓ_2-normalisatie voor arbitraire beeld regio's, die snel genoeg is voor semantische segmentatie met be hulp van Fisher codemaps. De tweede bijdrage buit de onafhankelijkheid van de codemaps tot het aantal classifiers uit waardoor we localiteit kunnen introduceren in het leren van attributen. De derde bijdrage is een vaste netwerk structuur, zoals ook wordt gedaan in object classificatie. We stellen voor om kernel groepering met codemaps te doen, waardoor niet-linearitetien kunnen worden geëncapsuleerd in een recursieve aanpak resulterende in een Feed-Forward architectuur. De resultaten tonen aan dat ℓ_2-normalisatie voor Fisher codemaps de huidige resultaten verbeterd op de PASCAL VOC 2011 en VOC 2012 datasets. Verder, door lokaliteit te introduceren verbeteren we de huidige resultaten met attributen op de uitdagende Caltech-UCSD 200-2011 Vogel dataset. Als laatste tonen we aan dat de Feed-Forward architectuur van de codemaps de classificatie nauwkeurigheid verhoogd zonder additionele kosten in een object classificatie taak op de PASCAL VOC 2007 dataset. We concluderen dat codemaps accuraat en efficiënt zijn waardoor we lokaliteit met sterke beeld representaties kunnen uitbuiten voor een divers aantal taken.

Het doel van deze dissertatie was een studie van fijne nuances die kunnen leiden tot een beter visueel begrip. Verscheidene aspecten van dit probleem zijn onderzocht, elk draagt bij tot een stuk van de puzzel.

We begonnen met de vraag *is de machinale representatie van fysiek identieke elementen constant?* Onze bevindingen ondersteunen de conclusie dat huidige beeld representaties met grote visuele woordenboeken aardig onstabiel zijn, vooral als men probeert zich te concentreren op de fijnmazige nuances; wanneer te diep wordt gegaan zullen nuances die corresponderen met hetzelfde fysieke element op een gegeven moment te verschillend worden en uiteindelijk anders worden behandeld. Door het gebruik van geometrie worden gedetecteerde nuances in een stabiele manier beschreven, maar ze zijn niet noodzakelijkerwijs optimaal als een beeld representatie. Dus, we stellen onze tweede vraag: *welke set van nuances zijn het belangrijkst voor het vinden van objecten?*. We vinden dat het behandelen van object zoeken vanuit een belangrijkheids-monsterings oogpunt we die weinige visuele nuances kunnen ontdekken die specifiek accuraat zijn voor bepaalde objecten. Aangemoedigd door de resultaten onderzochten we de vraag *welke set van nuances zijn het belangrijkst voor het classificeren van objecten?* We vertaalde het probleem in termen van een geregulariseerde kleinste kwadraten methode. We vonden dat de nuances niet alleen maar onderscheiden tussen categorieën zijn, maar dat ze ook kunnen uitleggen wat de bijdrage is van minder onderscheidende nuances. Voor onze volgende vraag behandelen we *welke visuele nuances onderscheiden fijnmazige object categorieën?*, waar het ontdekken van de belangrijke nuances van essentieel belang is en niet alleen maar behulpzaam voor een succesvolle classificatie. Onze bevindingen ondersteunen het idee dat fijnmazige object categorieën meestal dezelfde vorm en pose delen, zoals in algemene zin teruggevonden in de gemeenschappelijke bovenliggende categorie. Het uitbuiten van hun gemeenschappelijke ruimtelijke karakteristieken, fijnmazige objecten kunnen onderzocht worden door hun lokale details gegeven door corresponderende object regio's. We verkregen een enorm accurate herkenningsmachine, in een probleem dat zelfs voor menselijke (niet-experts) als moeilijk wordt beschouwd. Als laatste ondervroegen we *kunnen we de interpretatie van beeld nuances opdelen in lokale, herbruikbare componenten die een exacte reconstructie op regio niveau toestaan*. De resultaten laten zien dat dit mogelijk is, wat ons toestaat om sterke, moderne beeld representaties te gebruiken voor lokaal zoeken. Dit leidt tot excellente semantische segmentaties en object classificatie resultaten.

Voor een richting van toekomstig onderzoek noten we dat mensen makkelijk ongeveer 30.000 basis categorieën kunnen herkennen [12]. Experts kunnen veel, veel meer fijnmazige sub-categorieën herkennen. Voor zoveel categorieën zullen een aantal erg veel op elkaar lijken tot een punt waar onderscheidende nuances subtiel zijn. Het oogsten van voldoende voorbeeld gegevens om zulke subtiele nuances te kunnen ontdekken zal vrij moeilijk zijn. Dit is vooral waar als we ze verder willen raffineren in relatie tot hun locatie in het beeld.

Om het makkelijker te maken om zoveel categorieën te leren, zou het een mogelijkheid kunnen zijn om automatische kennis-overdracht methodes [98] te gebruiken, waar men mogelijkerwijs a-priori ervaringen kan gebruiken om de machine in toekomstige richtingen te leiden. Een andere mogelijkheid biedt zich aan in nul-voorbeelden leren [70] waar men een representatieve tussenlaag gebruikt voor het herkennen. In [90]

was ontdekt dat het uitbuiten van niet-visuele patronen, zoals object-naar-object relaties, een interessante richting is voor nul-voorbeelden leren. Verder, als fijnmazige herkenning het unieke object benaderd, in [128] wordt bewijs gepresenteerd dat in instanties waar maar een enkel beeld beschikbaar is, de lokaliteit van belang is. Als laatste, als recente sprongen in diep-lerende architecturen hebben ontsloten, is het leren van beeld representaties van data de ultieme manier is om de visuele voorkomens van fijne nuances te representeren. Al met al, de computer visie is verzekerd om meer flexibele manieren te ontdekken om visuele nuances te leren die een betrouwbare en praktische herkenning van zulk een extensief pallet van categorieën. We sluiten deze dissertatie af door onze fundamentele vraag nogmaals te bezoeken: *wat zijn de fijne details, de nuances, die we moeten ontdekken om een visueel complex beeld te begrijpen?*. Er is geen universeel antwoord, het hangt simpelweg af van waar we naar op zoek zijn. In feite, om de juiste nuance te vinden moeten we ze zien voor wat ze zijn en niet voor wat we willen dat ze zijn. Om bijvoorbeeld de Penrose driehoek van de kaft te begrijpen, zoals de achterkant van de kaft illustreert, hoeven we de driehoek alleen maar een beetje te draaien om het van een ander perspectief te bekijken.