Interaktiven sistem za kontinuirano učenje vizualnih konceptov *

Danijel Skočaj, Alen Vrečko, Matej Kristan, Barry Ridge, Gregor Berginc in Aleš Leonardis
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Tržaška 25, SI-1001 Ljubljana, Slovenija
danijel.skocaj@fri.uni-lj.si

Interactive system for continuous learning of visual concepts

We present an artificial cognitive system for learning visual concepts. It comprises of vision, communication and manipulation subsystems, which provide visual input, enable verbal and non-verbal communication with a tutor and allow interaction with a given scene. The main goal is to learn associations between automatically extracted visual features and words that describe the scene in an open-ended, continuous manner. In particular, we address the problem of cross-modal learning of visual properties and spatial relations and analyse several learning modes requiring different levels of tutor supervision.

1 Uvod

Za spoznavni sistem je zelo pomembna sposobnost neprestanega učenja in prilagajanja novim situacijam in izzivom, ki lahko nastanejo v realnem nenehno spreminjajočem se okolju. Takšno učenje je že po naravi večmodalno (ang. multi-modal), saj mora sistem uporabiti vse svoje zaznavne senzorje in spoznavne sposobnosti, da lahko zazna in razume okolje v katerem se nahaja in ustrezno osvežuje in nadgrajuje svoje znanje. V ta namen mora znati učinkovito komunicirati z ostalimi spoznavnimi sistemi (vključno s človekom), ki lahko bistveno poenostavijo in pohitrijo proces učenja, ter ga naredijo bolj robustnega in zanesljivega.

V tem članku predstavljamo umetni spoznavni sistem za interaktivno učenje vizualnih konceptov, ki upošteva prej omenjene zahteve. Temelji na nenehnem učenju in izmenjevanju uporabe že naučenih modelov ter osveževanjem in nadgrajevanjem-le teh ter dodajanjem novih. Sistem je sestavljen iz podsistemov za vizualno zaznavanje, komunikacijo in interakcijo, ki zagotavljajo vizualno informacijo, omogočajo verbalno in neverbalno komunikacijo s človekom ter interakcijo z okoljem. Takšen večmodalni interaktivni sistem omogoča učinkovito in uporabniku prijazno in naravno kontinuirano učenje[1].

Osredotočili se bomo na učenje vizualnih lastnosti predmetov (kot so barve in oblike) ter prostorskih relacij (kot sta 'levo od' ali 'blizu'). Glavni cilj je najti asociacije med besedami, ki opisujejo te koncepte in enostavnimi vizualnimi značilnicami pridobljenimi iz slik. Rešitev tega ti. *problema senzorske utemeljitve simbolov*[2] (ang. symbol grounding problem)[2, 10] je postavljena v okvir kontinuiranega učenja, ki poteka v interakciji med sistemom in človekom.

S podobnim problemom se je ukvarjalo že veliko raziskovalcev z različnih področij, od psihologije, lingvistike, umetne inteligence do računalniškega in spoznavnega vida. Naše delo je še najbolj sorodno pristopu D. Roya [6, 5], pri katerem se sistem uči besed in vizualnih atributov iz zvočnih in video posnetkov. S podobnim problemom se ukvarjata tudi L. Steels in F. Kaplan [8]. Delo, ki so ga predstavili Chella idr. [1] ima tudi za cilj zgraditi okvir za spoznavno učenje in povezovanje simbolov z nižjenivojskimi signali. S sistemskega vidika so soroden sistem zgradili M. Vincze idr. [9]. Še bolj kompleksen sistem za interaktivno inkrementalno učenje objektov implementiran na humanoidnem robotu pa so razvili v [4].

Naš sistem zasleduje podobne cilje kot zgoraj omenjena dela (in nekatera druga), vendar se bistveno razlikuje v tem, da je še posebej izpostavljeno kontinuirano učenje v sodelovanju s človekom. Le-ta igra v tem procesu pomembno vlogo, saj sistemu zagotavlja zelo zanesljivo informacijo o opazovanih prizorih. To informacijo lahko sistem poskuša pridobiti tudi sam, ne da bi s tem obremenjeval človeka, vendar je tako pridobljena informacija manj zanesljiva in lahko pripelje do poslabšanja že naučenih modelov. V tem članku bomo najprej na kratko opisali paradigmo kontinuiranega učenja, nato predstavili sam sistem, ki smo ga razvili. Nato bomo tudi predstavili uporabo sistema in analizirali več različnih načinov učenja, ki zahtevajo različne nivoje vpletenosti človeka.

* To delo so deloma podprli MVZT (Raziskovalni program *Računalniški vid P2-0214*) ter EU projekta *CoSy* (FP6-004250-IP) in *VISIONTRAIN* (MRTN-CT-2004-005439).

[1] Neprestano, nenehno, vseživljensko nadgrajevanje znanja.

[2] Povezovanje (lingvističnih) simbolov z sub-simboličnimi interpretacijami fizičnega sveta.

2 Paradigma kontinuiranega učenja

Predstavljen sistem omogoča različne načine kontinuiranega učenja, ki zahtevajo različne nivoje sodelovanja človeka. Pri **popolnoma usmerjanem** (krajše *PU*) načinu učenja vso informacijo sistemu zagotovi človek, ki mu nedvoumno in pravilno opiše prizor, tako da lahko sistem vedno osveži trenutno znanje s pravilnimi in popolnimi podatki. Pri **deloma nadzorovanem** (*DN*) načinu sistem najprej poskuša sam razpoznati prizor in vizualne koncepte, ki se pojavljajo. Če mu to upe dovolj zanesljivo, potem sam, brez pomoči človeka, osveži svoje znanje, sicer pa o pravilnosti povpraša človeka in osveži svoje znanje s tako dobljeno pravilno interpretacijo. Pri **popolnoma samostojnem** (*PS*) načinu učenja pa sistem vedno osveži trenutno znanje samostojno s svojo interpretacijo brez vpletanja človeka. Druga dva načina učenja (*DN* in *PS*) nadalje še razdelimo na **konzervativno** in **liberalno**, pri čemer je zaupanje v lastno sposobnost razpoznavanja večje pri slednjem, medtem, ko pri konzervativnem načinu sistem osveži svoje znanje samo, če je zelo prepričan v pravilnost razpoznave.

Seveda je v takšnem sistemu pomemben tudi sam učni algoritem, ki skrbi za osveževanje in nadgrajevanje znanja. V principu bi lahko uporabili katerikoli algoritem, ki ustreza pravilom inkrementalnosti[3]. Trenutna implementacija sistema uporablja algoritem, ki posamezen vizualni koncept asocira z vizualno značilnico, katere vrednosti so najbolj konsistentne pri vseh slikah predmetov, ki predstavljajo isti vizualni koncept. Podrobnejši opis tega algoritma, kot tudi celotnega okvirja kontinuiranega učenja, se nahaja v [7].

3 Integriran sistem

Integriran sistem, ki ga predstavljamo v tem članku, je sestavljen iz več podsistemov, ki skupaj tvorijo homogeno celoto in omogočajo implementacijo paradigme kontinuiranega učenja, ki smo jo na kratko orisali v prejšnjem poglavju.

Slika 1(a) prikazuje sistem na delu, medtem ko Slika 1(b) shematično ponazarja vse komponente sistema in povezave med njimi. V nadaljevanju bomo na kratko opisali vsako komponento posebej.

Upravitelj vizualnega učenja in razpoznavanja je centralni modul v sistemu. Neprestano spremlja dogajanje in sprejema zahteve za razpoznavanje oz. učenje, ki prihajajo bodisi s strani pogovornega podsistema (ko dialog sproži človek), bodisi s strani modula za usmerjanje pozornosti (ko iniciativo prevzame sistem sam). Glede na trenutno stanje sistema in vsebino zahteve nato sproži ustrezne akcije (segmentacija, razpoznavanje, učenje, postavljanje vprašanja, ipd.), ter nato ustrezno obdela rezultate teh akcij. Ta modul implementira različne načine učenja opisane v prejšnjem razdelku in koordinira celotno delovanje ter določa obnašanje sistema.

Video strežnik skrbi za zajemanje slik z barvno kamero. Slike opremi s časovno oznako in jih da na voljo ostalim komponentam sistema.

Modul za usmerjanje pozornosti spremlja dogajanje in detektira spremembe v prizoru. O spremembi nato obvesti modul za upravljanje in mu poda tudi regijo, kjer je do sprememb prišlo.

Modul za segmentacijo predmetov skrbi za segmentacijo predmetov od ozadja. Ker je kamera statična, se najprej nauči model ozadja, ki ga nato uporablja za segmentacijo.

Modul za izločanje značilnic nato prejme detektirano področje zanimanja (ang. ROI) s pripadajočo segmentacijsko masko in iz nje izloči značilnice, ki se potem uporabljajo pri razpoznavanju in učenju. V principu bi lahko sistem delal s katerikolimi značilnicami; trenutno je implementirano izločaje nekaterih enostavnih značilnic pridobljenih iz videza in oblike predmeta ter razdalj med predmeti.

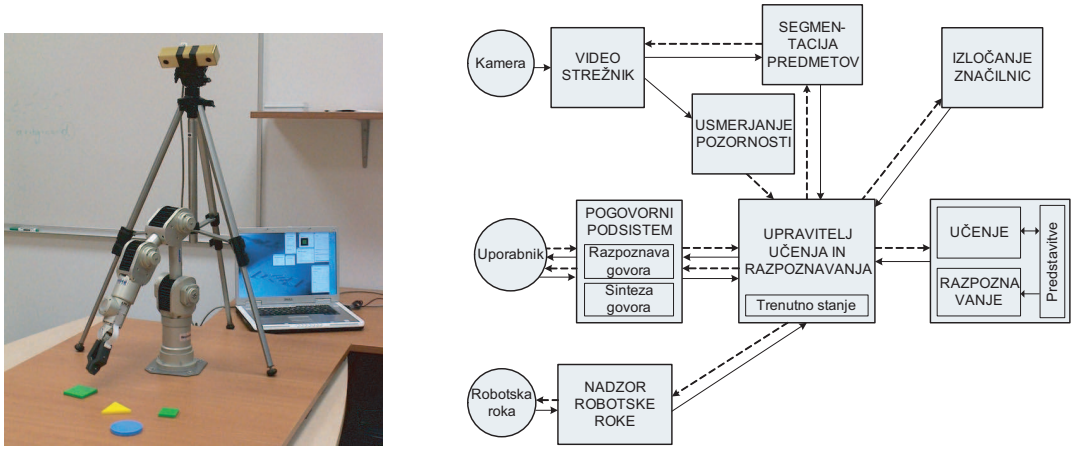
Modul za razpoznavanje in učenje je implementacija učnega algoritma. Njegova naloga je generiranje, uporaba in vzdrževanje predstavitev konceptov. Na osnovi izločenih značilnic poskusi razpoznati vizualne koncepte ter nadgraditi njihove predstavitve, oz. takšne predstavitve dodati, ko naleti na nov koncept.

Pogovorni podsistem skrbi za komunikacijo s človekom - procesira uporabnikove izraze in iz njih generira simbolične predstavitve in obratno, iz simbolov, ki jih dobi od modula za upravljanje, generira izraze v naravnem jeziku. Za razpoznavanje govora uporablja Spinx 4 standardno distribucijo. Ker pa razpoznavanje govora včasih ni zanesljivo, imamo kot rezervno rešitev na voljo tudi vnos preko tipkovnice in miške. Odgovori sistema se v naravnem jeziku uporabniku predvajajo z uporabo Mary TTS sistema.

Modul za nadzor robotske roke skrbi za interakcijo z okoljem. Nadzoruje robotsko roko Neuronics Katana Arm 6M180, ki premore pet prostorskih stopenj in prijemalo. V glavnem se robotska roka uporablja v primerih, ko je v prizoru več predmetov in je potrebno pokazati na enega izmed njih, da se razreši dvoumnost pri referenciranju predmetov.

Za komunikacijo med moduli sistem uporablja komunikacijsko orodje BALT [3], ki sloni na arhitekturi CORBA (Common Object Request Broker Architecture). Komunikacija med moduli poteka popolnoma v ozadju preko TCP/IP protokola, tako da je lahko sistem porazdeljen preko več računalnikov.

[3]Trenutne predstavitve lahko učinkovito osveži z na novo pridobljeno informacijo, pri čemer nima dostopa do predhodno obravnavanih učnih podatkov, temveč le do njihovih predstavitev.



(a) (b)

Slika 1: (a) Sistem na delu. (b) Shematični prikaz sistema.

4 Eksperimentalni rezultati

Večino eksperimentov smo opravili s ploščicami različnih barv, velikosti in oblik, ki smo jih uporabljali pri pouku matematike v osnovni šoli (Slika. 2(a)). Ta domena je primerna za analizo takšnega sistema, ker so lastnosti predmetov zelo raznovrstne in jasno določene.

Pri učenju in razpoznavanju vizualnih lastnosti predmetov poskuša sistem najti asociacije med desetimi vrednostmi vizualnih lastnosti[4] in šestimi avtomatsko pridobljenimi vizualnimi značilnicami[5]. Pri učenju in razpoznavanju prostorskih relacij, pa sistem obravnava enajst relacij[6] ter pet značilnic, ki temeljijo na položajih središč predmetov[7].

Predstavljeni sistem je sestavljen tako, da čim bolj olajša komunikacijo med človekom in sistemom. Slika 2(c) prikazuje izseke iz dialoga pri učenju in razpoznavanju vizualnih lastnosti prej omenjenih ploščic. Na začetku usmerja potek učenja uporabnik, čez nekaj časa pa prevzame iniciativo sistem in sam po potrebi postavlja vprašanje človeku.

Na podoben način poteka tudi razpoznavanje prostorskih relacij. V tem primeru sta na prizoru dva predmeta, za katera sistem poskuša ugotoviti v kakšnih relacijah sta in na osnovi tega opisati prizor. Slika 2(b) prikazuje primer avtomatskega opisa prizora, ki je bil narejen na podlagi modelov zgrajenih v procesu kontinuiranega učenja prostorskih relacij.

Za bolj obsežno analizo in evaluacijo predlaganih načinov učenja pa je takšno interaktivno delo neprimerno. Zato smo posneli večje število prizorov ter jih ročno označili. Nato smo lahko odgovore uporabnika umetno generirali ter tako simulirali različne načine učenja.

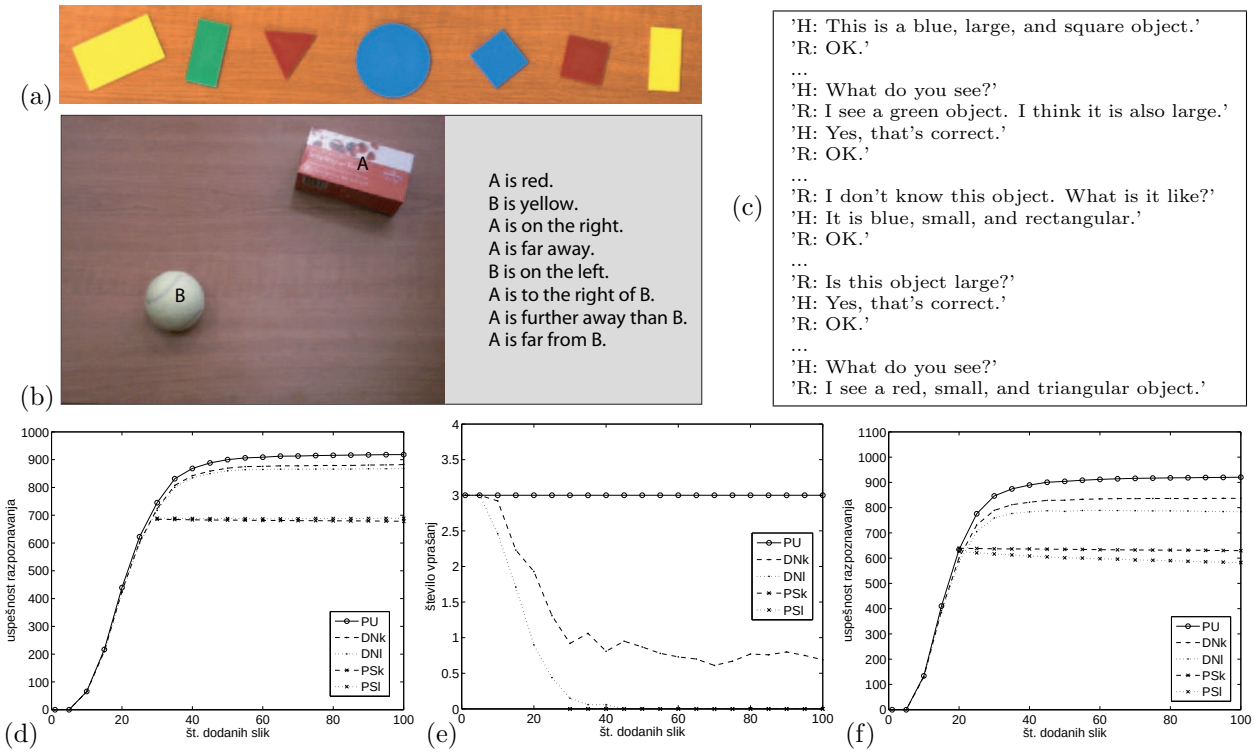
Najprej smo testirali rezultate kontinuiranega učenja vizualnih lastnosti. Na vsakem koraku učenja smo dodali po eno učno sliko ter poskušali osvežiti modele v skladu z različnimi učnimi strategijami. Tako osvežene modele smo nato uporabili za razpoznavanje vizualnih lastnosti na vseh testnih slikah. Rezultate smo ovrednotili z *uspešnostjo razpoznavanja* (ang. recognition score) [7]. Iz rezultatov, ki so predstavljeni na Sliki 2(d), je razvidno, da se uspešnost razpoznavanja izboljšuje skozi čas, ko sistem obdeluje nove slike in ustrezno temu osvežuje modele konceptov. Po pričakovanjih najboljše rezultate zagotavlja *popolnoma usmerjano učenje* (PU). Zato pa je št. vprašanj (oz. podatkov, ki jih uporabnik posreduje sistemu) v tem primeru zelo visoko (Slika 2(e)). Le malce slabše rezultate vrne konzervativno *deloma nadzorovano učenje* (DNk), le da je v tem primeru število postavljenih vprašanj precej manjše, saj sistem vprašuje uporabnika le, ko ni popolnoma prepričan v pravilnost svoje razpoznave. V primeru *liberalnega deloma nadzorovanega učenja* (DNl) je postavljenih vprašanj še manj, saj sistem vprašuje samo, če je zelo negotov glede pravilnosti svoje razpoznave, rezultati pa so le za odtenek slabši. Iz rezultatov pa je tudi razvidno, da *popolnoma samostojno učenje* (PSk ter PSl), pri katerem uporabnik ne sodeluje, ne bistveno izboljša začetnih modelov (ki so bili naučeni na popolnoma usmerjan način). Podobno metodologijo ovrednotenja sistema smo uporabili tudi v primeru učenja in razpoznavanja prostorskih relacij in tudi rezultati, ki so predstavljeni na Sliki 2(f), so zelo podobni.

[4]Štiri barve (rdeča, zelena, modra, rumena), dve velikosti (majhna, velika) ter štiri oblike (kvadratna, krožna, trikotna, pravokotna).

[5]Tri značilnice na podlagi videza (mediana vrednosti barvnega odtenka (ang. hue), nasičenosti barve ter intenzitete), ter tri značilnice oblike (površina, obseg in kompaktnost segmentirane regije).

[6]Šest binarnih relacij med dvema predmetoma (levo od, desno od, bližje kot, dlje kot, blizu, daleč od), ter pet unarnih relacij, ki opišejo položaj predmeta v prizoru (na levi, na sredini, na desni, blizu, daleč).

[7]Absolutne koordinate (x, y), razlika med koordinatami (dx, dy) ter razdalja med središči (d).



Slika 2: (a) Primeri predmetov. (b) Avtomatsko zgenerirani opis prizora. (c) Primer dialoga. (d) Uspešnost razpoznavanja ter (e) št. vprašanj med učenjem vizualnih lastnosti. (f) Uspešnost razpoznavanja prost. relacij.

5 Zaključek

V članku smo predstavili umetni spoznavni sistem, ki udejanja paradigmo kontinuiranega učenja na interaktiven, uporabniku prijazen način. Sestavljen je iz podsistemov za vizualno zaznavanje, komunikacijo in interakcijo. Predstavili smo rezultate njegovega delovanja in analizirali različne načine kontinuiranega učenja, ki zahtevajo različne nivoje vpletenosti uporabnika.

Rezultati kažejo, da je kontinuirano učenje učinkovito, saj se rezultati skozi čas, ko sistem opazuje nove in nove prizore, izboljšujejo. Po pričakovanju je naučen model boljši, če je nadzor uporabnika večji, vendar zelo zadovoljive rezultate dosežemo tudi ob visoki avtonomiji sistema, pri čemer je količina podatkov, ki jih mora uporabnik priskrbeti sistemu, bistveno manjša.

Nekatere komponente sistema so sicer enostavne, vendar je celoten sistem konceptualno močan in predstavlja dobro osnovo za nadaljnje nadgradnje in razširitve. Tako izboljšujemo učni algoritem, obogatili bomo nabor značilnic ter konceptov. Robotsko roko bomo uporabili tudi za učenje funkcionalnih lastnosti predmetov. Večino komponent sistema bomo tudi povezali tudi z novimi komponentami (za načrtovanje, sklepanje, ipd.) s ciljem zgraditi zmogljiv, inteligenten, čim bolj avtonomen, uporabniku prijazen, prilagodljiv in nenehno razvijajoč se spoznavni sistem.

Literatura

[1] Chella, A., Frixione, M., Gaglio, S.: A cognitive architecture for artificial vision. Artificial Intelligence **89**(1–2) (1997) 73–111

[2] Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena **42** (1990) 335–346

[3] Hawes, N.: BALT & CAAT: Middleware for cognitive robotics. TR CSR-07-1, Univ. of Birmingham (2007)

[4] Kirstein, S., Wersing, H., Körner, E. Rapid online learning of objects in a biologically motivated recognition architecture. In DAGM 2005 (2005) 301–308

[5] Roy, D.K., Pentland, A.P.: Learning words from sights and sounds: a computational model. Cognitive Science **26**(1) (2002) 113–146

[6] Roy, D.K.: Learning visually-grounded words and syntax for a scene description task. Computer Speech and Language **16(3)** (2002) 353–385

[7] Skočaj, D., Ridge, B., Leonardis, A.: On different modes of continuous learning of visual properties. In: ERK 2006 (2006) 105–108

[8] Steels, L., Kaplan,F. AIBO's first words. The social learning of language and meaning. Evolution of Communication **4(1)** (2001) 3–32

[9] Vincze, M., Ponweiser, W., Zillich, M.: Contextual coordination in a cognitive vision system for symbolic activity interpretation. In: ICVS 2006. (2006) 12

[10] Vogt, P.: The physical symbol grounding problem. Cognitive Systems Research **3**(3) (2002) 429–457