



Univerza v Ljubljani  
Fakulteta za elektrotehniko

PHD THESIS

---

## Related work

---

*Author:*

Gregor Koporec, mag. inž.  
el.

*Menthor:*

prof. dr. Janez Perš

May 18, 2018

# Contents

<b>1</b>	<b>Podobna dela</b>	<b>2</b>
1.1	Primerjava med ljudmi in algoritmi . . . . .	3
1.1.1	Vizualna izkrivljanja slike . . . . .	3
1.1.2	Razlike v zornem kotu . . . . .	6
1.1.3	Primerjava modelov, ki ne temeljijo na nevronskeh mrežah . . . . .	7
1.1.4	Razvrščanje v abstraktne razrede . . . . .	9
1.1.5	Sistematična razlika med modeli in ljudmi . . . . .	10
1.1.6	Preslepitev nevronskeh mrež . . . . .	12
1.2	Izboljšava algoritmov z vnosom človeških karakterističnih značilnosti	14
1.3	Biološke raziskave razpoznavanja objektov . . . . .	19
1.4	Razpoznavanje obrazov . . . . .	23
1.5	Pristranskost podatkovnih baz . . . . .	28

# Chapter 1

## Podobna dela

Globoke nevronske mreže močno vplivajo na raziskave v računalniškem vidu in strojnem učenju, saj po zmoglјivostnih testih premagujejo večino algoritmov, ki jih je razvil človek [1]. Zavedati se moramo, da imajo kljub temu DNN pomankljivosti, zato velikokrat odpovejo v realističnem okolju [1]. Glavni poudarek, zakaj prihaja do razlik med zmoglјivostnimi testi na podatkovnih bazah in realnim okoljem je, da se pri zmoglјivostnih testih raziskovalci osredotočajo na evaluacijske statistike, ki povzemajo zmoglјivost algoritmov [1]. Kot je poudaril Sünderhauf *et al.* [1]: “Če statistika govori, da je bila podatkovna baza rešena, še ne pomeni, da je bil problem rešen.” Trditve, da so nevronske mreže presegle človeško zmoglјivost, temeljijo na zmoglјivostnih testih in so napačne ter zavajujoče. Pri primerjanju zmoglјivosti z ljudmi bi potrebovali drugačen pristop, kot je vizualna psihofizika, ki je predstavljena v nadaljevanju.

Nadaljnje, podatkovne baze ne povzemajo naravnih scenarijov, saj so te zgrajene na slikah in video posnetkih iz interneta [1]. Podatkovne baze vsebujejo inherentno pristranskost, ki vpliva na evaluacijo in interpretacijo rezultatov. Pristranskost je opisana v poglavju .

Nikakor ne razumemo, kako se nevronske mreže obnašajo v scenarijih, kjer se okolje dinamično spreminja in kjer se pojavljajo objekti neznanega razreda [1]. Prav tako se lahko zgodi, da so objekti v realnem scenariju na videz drugačni kot so bili predstavljeni v učni množici. Raziskovalci so pokazali, da lahko že z manjšimi spremembami močno vplivamo na odločitve nevronske mreže. Eden izmed razlogov je ta, da se globoki modeli trenutno uporabljajo v zaprtih sistemih, kjer so razredi popolnoma znani [1]. Nevronske mreže bi morali razviti v bolj fleksibilne strukture, ki bi imele možnost razširiti svoje znanje z majhnim številom podatkov [1].

## 1.1 Primerjava med ljudmi in algoritmi

### 1.1.1 Vizualna izkrivljanja slike

Richard Webster *et al.* [2] se je v svojem delu spraševal, če algoritmi razpoznavanja res delujejo tako dobro, kot mislimo. Glede na trenutno uporabljene statistike bi lahko verjeli, da smo dosegli človeško zmogljivost. Seveda pri takoj veliki količini podatkov težko ugotovimo, kje algoritmi ne delujejo. Zato so avtorji predstavili novo metodo evaluacije, ki temelji na vizualni psihofiziki (angl. Visual Psychophysics). Pri tej metodi obravnavamo odnos med kontrolnim dražljajem in odzivom nanj. S perturbacijami izbranega dražljaja dobimo krivuljo odziva in jo uporabimo za primerjanje med modeli.

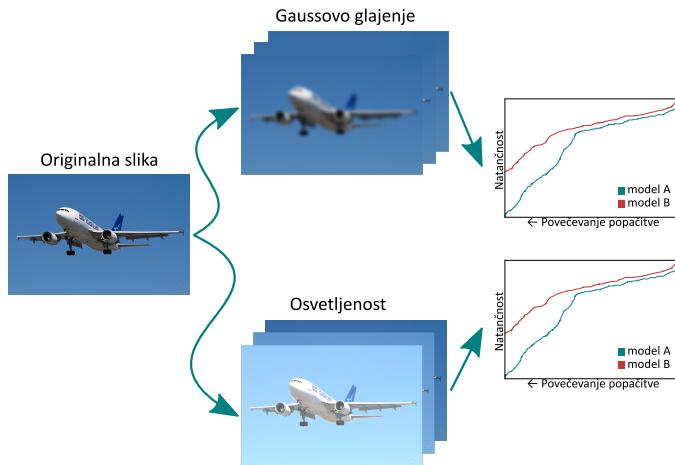


Figure 1.1: Slike iz ImageNet podatkovne baze [3].

V [2] so najprej določili naključen niz slik za izbrano kategorijo. Iz niza so nato izluščili sliko s kanoničnim pogledom. Avtorji so nato perturbirali izbrane kanonične slike in jih uporabili pri testiranju razvrščanja. Za vsako kategorijo so na koncu dobili srednjo natančnost. Točke so interpolirali in zgladili pravokotnim oknom širine 15.

Kanonično sliko so v [2] opredelili kot sliko, pri kateri dobimo največji odziv za privzeto vrednost dražljaja. Pri ljudjeh so to slike, kjer bo večina razpoznala objekt na sliki. Za modele kanonični pogled predstavlja sliko, kjer bo verjetnost razpoznavanja kategorije največja.

Testiranje razvrščanja so razdelili na dva testa. S prvim,  $2AFC$ , so želeli preizkusiti razpoznavanje na temeljni ravni. Z drugim testom,  $MAFC$ , pa so želeli preizkusiti razpoznavanje, ki je bolj primeren več razrednemu razvrščanju izbranih algoritmov.

Pri  $2AFC$  testu so najprej merjencu pokazali vzorčno sliko. Sledil je prikaz dveh alternativ, izmed katerih je ena pozitivna (iz iste kategorije)

in druga negativna slika (iz različne kategorije). Merjenec je nato izmed alternativnih slik moral izbrati najbolj ugodno. Test so nato ponavljali za različne perturbacije kanoničnih slik.

MAFC test so določili kot posplošeno obliko 2AFC testa. V tem testu so za vzorec izbrali vse slike, ki so jih uporabili pri učenju. Za alternativne slike so izbrali kanonične slike iz vseh kategorij. Zaradi velikega števila slik, so pri tem testu subjektom primerno zmanjšali količino.

Perturbacija slike je Richard Webster *et al.* [2] določil s popačitvami slike. Uporabil je Gaussovo glajenje, linearno okluzijo, šum sol in paper, osvetlitev, kontrast in ostrino.

Za podatkovno bazo slik so izbrali ImageNet 2012. Testirali so najbolj pogoste konvolucijske nevronske mreže AlexNet, CaffeNet, GoogleNet, VGG-16 in VGG-19. Študijo so izvedli s pomočjo 24 subjektov.

Ugotovili so, da natančnost modelov v vseh primerih začne hitro upadati pod 80 %. Ljudje so nevronske mreže prekašali v skoraj vseh testih. Največja razlika se je pojavila pri testiranju osvetlitve, kjer so bili ljudje za ~9 % slabši.

Podobno nižjenivojsko primerjavo med modeli in ljudmi, kot v [2], so predstavili v delu [4]. Test s subjekti so razdelili na tri dele. V prvem delu so merjeni videli primerke slik iz vseh izbranih kategorij. S tem so simulirali učenje nevronskih mrež. Sledil je korak validacije. Merjeni so razvrščali čiste slike. V primeru slabega razvrščanja so celoten test označili za osamelec. Tako so uporabili le rezultate merjencev, ki so test skrbno in resno izvedli. Nazadnje je sledil korak testiranja, kjer so merjeni razvrščali popačene slike. Slike so se na zaslonu pojavljale vedno od največje do najmanjše popačitve. S tem so se avtorji znebili efekta spomina, kjer bi lahko merjenec dopolnil manjkajočo informacijo s prejšnjim spoznanjem pri sliki z manjšim popačenjem.

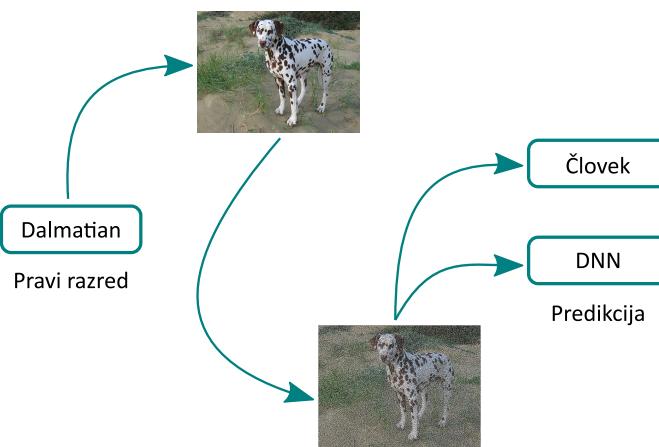


Figure 1.2: Slike iz ImageNet podatkovne baze [3].

Za namen testiranja modelov so avtorji modelom dodali nov sloj za razvrščanje izbranih kategorij. Sloj so naučili z nepopačenimi slikami in ga še bolj natančno nastavili z dodatnim učenjem na popačenih slikah. Testiranje modelov je nato potekalo na enak način, kot pri subjektih.

Pri testiranju so Dodge *et al.* [4] uporabili ImageNet podatkovno bazo. Zaradi velike količine podatkov, so izbrali le 10 kategorij. Vse kategorije so bile vrste psov. Izbiro kategorij avtorji zagovarjajo s tem, da so izbrane kategorije zelo povezane in tako težko ločljive. Za popačitev slike so uporabili Gaussov šum in glajenje. Rezultate na subjektih so dobili s 15. merjenci. Za primerjavo so vzeli modele ResNet, GoogleNet in VGG-16.

Z raziskavo so ugotovili, da je človeška natančnost večja od vseh modelov. Prav tako so poudarili, da ljudje hitreje izgubljajo na natačnosti pri meglenih slikah, modeli pa na pošumljenih slikah.

Skoraj identičen način eksperimentiranja sta avtorja iz [4] uporabila tudi v delu [5], kjer sta testirala nižjenivojski človeški vidni sistem. Za razliko od [4] sta v [5] omejila opazovanje slike na 100 ms. Izbiro sta argumentirala s tem, da se v tako kratkem času oko ne premakne, zato je človekov vizualni sistem omejen na globalno reprezentacijo slike. Tudi v tem primeru sta ugotovila, da je človeška zmogljivost boljša od algoritmov.

Za razliko od ostalih so v [6] teste na subjektih izvajali v kontroliranem laboratorijskem okolju, kjer so na zaslon najprej za 300 ms prikazali prazen okvir. Sledil je prikaz izbrane slike za 200 ms. Takoj zatem so merjencu prikazali masko šuma za 200 ms. Na koncu so imeli merjenci na voljo še 1.5 s, da so izbrali eno izmed ponujenih kategorij.

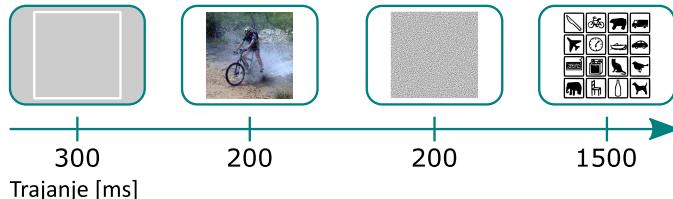


Figure 1.3: Slike iz [7] po licenci <https://creativecommons.org/licenses/by/4.0/legalcode>.

Pred samim testom so merjencem poakazali vse možne kategorije, da so zagotovili jasnost naloge. S prikazom maske šuma so minimizirali povratni vpliv v možganih, ker izbrani modeli nevroskih mrež ne vsebujejo povratnega vpliva.

Pri testiranju so [6] uporabili ImageNet 2012 podatkovno bazo. Testirali so na 16. kategorijah, ki so jih določili s pomočjo MS COCO podatkovne baze. Za popačenje so izbrali naslednje degradacijske metode: barva, kontrast, beli šum in eidolon. Eidolon je parametrično kontrolirana degradacija slike, ki povzroči podobno vizualno zavedanje na človeku [6]. 3 merjenci (moški, starost: 22–28 let, povprečje: 25 let) so sodelovali pri

barvnem testiranju in po 5 merjencev (6 moških in 4 ženske, starost: 19–28 let, povprečje: 23 let) je sodelovalo na ostalih testih. Za modele so izbrali AlexNet, GoogleNet in VGG-16. Geirhos *et al.* [6] so ugotovili, da so CNN-ji manj robustni na popačenja slik kot ljudje.

Zelo podoben test, kot [6], so izvajali tudi v delu [8]. Tu so testirali DNN modelle, če njihova zmogljivost degradira podobno kot človeška zmogljivost. V ta namen so primerjali DNN in ljudi pri spremnjanju kontrasta slike. Kontrast so izbrali zato, ker je relativno dobro razumljen vidik človeškega vida. Ker ljudje po navadi kategorizirajo slike po osnovnih kategorijah, so za testiranje uporabili MS COCO podatkovno bazo ?. Merjencem so vsako sliko pokazali za 200 ms. Sledil je prikaz roza šuma. Nazadnje so morali merjenci v roku 1500 ms izbrati eno izmed 16 kategorij.

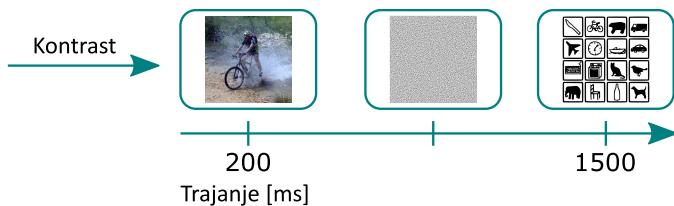


Figure 1.4: Slike iz [7] po licenci <https://creativecommons.org/licenses/by/4.0/legalcode>.

Avtorji so s svojo raziskavo ugotovili, da AlexNet, GoogleNet in VGG-16 razpoznavajo polno kontrastne slike podobno kot ljudje. Podobnosti z zmanjševanjem kontrasta hitro izginejo, kar se opazi tako na krivulji natančnosti glede na kontrast kot tudi na konfuzijski matriki. Na podlagi konfuzijskih matrik so ugotovili, da DNN-ji vseeno niso dobri modeli človeškega ventralnega procesiranja, saj bi v nasprotnem primeru pričakovali, da oboji delajo podobne napake.

### 1.1.2 Razlike v zornem kotu

Kheradpisheh *et al.* [9] so preizkušali delovanje glede na pozicijo, velikost in orientacijo objektov. S tem so želeli ugotoviti, ali se nevronске mreže naučijo invariantnosti na izbrane parametre. Za testiranje so renderirali slike iz 3D računalniških modelov, kjer so za parametre 3D transformacij uporabljali naključno število. Modelom so dodajali tudi različna ozadja iz realnih slik.

Pri testiranju subjektov so uporabljali podobno metodologijo, kot [6]. Prvih 500 ms so prikazovali črn križec ali naravno ozadje. Sledil je prikaz izbrane slike za 25 ms in nato prazen zaslon za enak časovni interval. Nato so prikazali masko šuma za 100 ms in na koncu še slike kategorij, izmed katerih so merjenci eno izbrali.

Pri evaluaciji modelov so prednaučenim modelom podali naključno izbran niz učnih in testnih slik. Za vsako sliko so za vsak sloj modela dobili

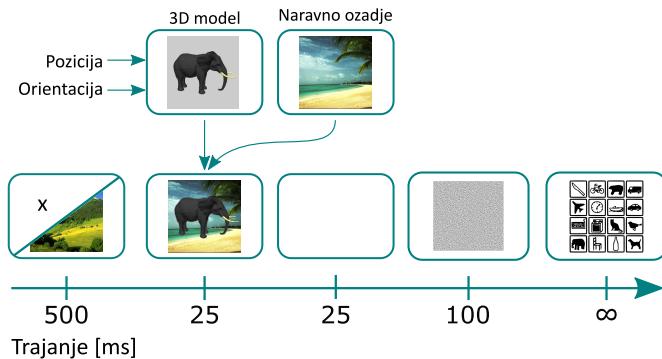


Figure 1.5: Slike naravnega ozadja so iz [10]. 3D modeli so iz [11]. Slika kategorij je iz [7] po licenci <https://creativecommons.org/licenses/by/4.0/legalcode>.

vektor značilk. Vektorje so nato uporabili za učenje in testiranje SVM razvrščevalnika. Tak način testiranja so  $15 \times$  ponovili za vse modele, sloje in vrednosti parametrov 3D transformacij. S tem so dobili povprečno vrednost in standardni odklon natančnosti delovanja modela in posameznih slojev modela za izbran parameter transformacije.

Za testiranje so uporabili že prej opisane renderirane slike z objekti iz 5. različnih kategorij (600 slik na kategorijo). Vsak objekt so transformirali na 7 različnih načinov, pri tem pa so uporabili po 2 različna ozadja. Tako so dobili 14 podatkovnih baz s 3000 slikami. Testirali so 26 subjektov (17 moških in 9 žensk, starost 21–32 let, povprečje starosti 26 let) in 9 različnih modelov nevronskeh mrež.

V raziskavi so ugotovili, da delujejo modeli skoraj tako dobro kot ljudje, če imajo objekti enakomerno sivo ozadje. Pokazali so tudi, da dodajanje slojev v nevronske mreže poveča natančnost glede na parametre 3D transformacij.

Z manjšimi spremembami so avtorji dela [9] ponovili poskus v [12]. Tu so razdelili podatkovno bazo na 3 dele. V vsakem delu so spremenjali izbrano število parametrov pozicije in rotacije. Prav tako so deloma spremenili psihofizioliške eksperimente, ki so bolj podrobno predstavljeni na sliki ??.

V [12] so raziskovalci ugotovili, da ljudje razpoznavajo objekte tudi pri velikih variacijah, so zelo natačni in imajo kratek odzivni čas. Natančnost in odzivni čas sta močno odvisna od tipa variacije. Raziskave so pokazale da je rotacija v globino najtežja. Za DNN-je so ugotovili, da so močno korelirani z ljudmi. Tudi nevroske mreže so imele težave pri rotacijski v globino.

### 1.1.3 Primerjava modelov, ki ne temeljijo na nevronskeh mrežah

Borji *et al.* [13] so v svojem delu izvedli obsežno raziskavo primerjave med človeškim in računalniškim razpoznavanjem objektov in prizorov. Pri tem

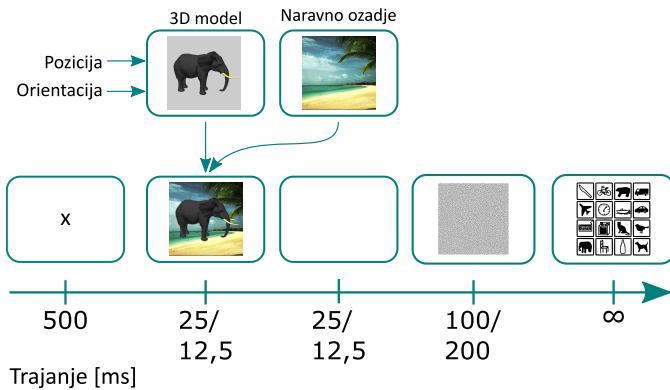


Figure 1.6: Slike naravnega ozadja so iz [10]. 3D modeli so iz [11]. Slika kategorij je iz [7] po licenci <https://creativecommons.org/licenses/by/4.0/legalcode>.

so uporabili 14 modelov, 7 podatkovnih baz in 5 različnih testov.

Prvi test so uporabili za preverjanje moči vizualnega razlikovanja in reprezentacije med ljudmi in algoritmi. V tem testu so preverjali kako dobro algoritmi in ljudje razlikujejo med različnimi prizori in če zaznajo razlike kaj je žival in kaj ni. Drugi test so namenili razpoznavanju črtnih slik (nizko nivojsko razpoznavanje). V tretjem testu so se osredotočili na analizo invariantnosti pri razpoznavanju tipa žival ne-žival. Pri tem so uporabili skaliranje in vrtenje v ravnini. Razpoznavanje lokalne in globalne informacije iz slik so testirali v četrtem testu. Tu so uporabili slike, kjer so njeni deli med seboj premešani. Zadnji test so uporabili za večrazredno razpoznavanje objektov na velikih podatkovnih bazah.

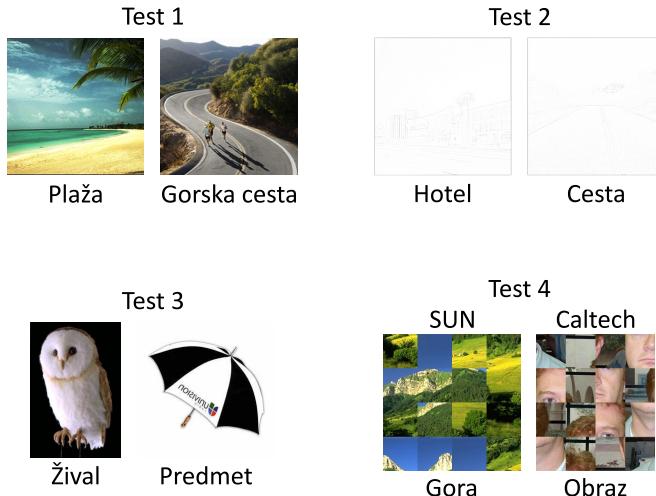


Figure 1.7: Slike iz [10] in [14].

Za modele so izbrali enorazredne SVM razvrščevalnike z 10 kratno križno validacijo. Izbrani deskriptorji so temeljili na HOG in SIFT deskriptorjih in Gaborjevih filtrihi za razpoznavanje prizorov. Natančnost modelov so izračunali kot povprečje konfuzijskih matrik za posamezno križno validacijo.

Pri raziskovanju so ugotovili da v prvem testu najmanj polovica algoritmov deluje z natančnostjo nad 70 % medtem ko je natančnost ljudi okoli 80 %. V manjši meri so algoritmi celo dosegali natančnost ljudi. V drugem testu so ljudje bili sposobni razpoznati prizor iz črtnih slik z natančnostjo 66 %, algoritmi pa z natančnostjo preko 70 %. Pri testiranju invariantnosti so ugotovili, da rotacija na ljudi zelo malo vpliva glede na algoritme. Na zmešanih slikah so se algoritmi večinoma bolje odrezali kot ljudje, kjer slike predstavljajo zunanjji prizor in slabše na slikah z notranjim prizorom. Njihovo slabo delovanje za zmešane slike z objekti je nakazovalo, da modeli večinoma temeljijo na zbiranju globalne informacije iz slik. V zadnjem testu so raziskovalci ugotovili, da najboljši modeli ne dosegajo sposobnosti človeškega razpoznavanja. Razlika med njimi je bila okoli 17 %.

#### 1.1.4 Razvrščanje v abstraktne razrede

Objekti različnih kategorij, okluzija in nepričakovane postavitve so tipični pojni, pri katerih današnji algoritmi, glede na človeka, delujejo slabo [15]. Za dobro delovanje pri tako veliki kompleksnosti avtorji iz [15] zato zagovarjajo, da algoritmi potrebujejo neko vrsto globalnega presojanja, s katerim lahko zgradijo višenivojske koncepte - abstrakcije.

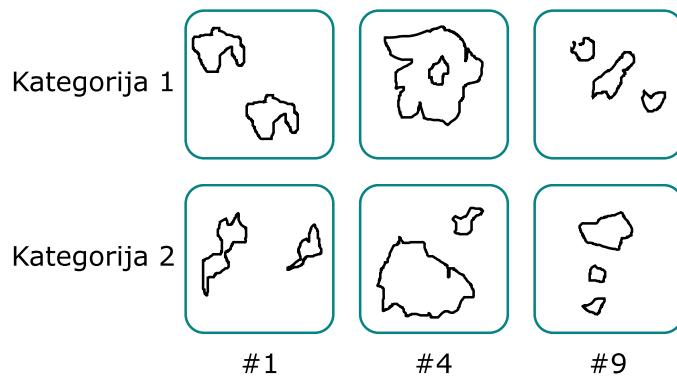


Figure 1.8

V ta namen so razvili kontroliran eksperiment, SVRT, s katerim bi lahko izmerili razliko v abstraktnem presojanju med ljudmi in algoritmi. Pri tem gre za serijo 23 različnih problemov, ki vsebujejo slike z naključno generiranimi oblikami brez kakršnega koli pomena. S takim testom se znebimo problemov razvrščanja naravnih slik pri algoritmih, kot so osvetlitev, tekstura in šum. Prav tako zmanjšamo prednost človeka pred algoritmi, ker imajo veliko izkušenj z naravnimi objekti v 3D svetu [15].

Vsek problem v SVRT je sestavljen iz dveh izključujočih razredov slik. Posamezen razred vsebuje abstraktno lastnost, ki ni vključena v drugem. Abstraktne lastnosti temeljijo samo na odnosih med oblikami.

Za testiranje subjektov so uporabili 20 ljudi (6 moških in 14 žensk, starost: 18–21). Merjencu so po vrsti prikazovali eno naključno sliko iz izbranega problema, ta pa jih je zlagal v eno ali drugo kategorijo. Po vsaki zložitvi je merjenc dobil odgovor o pravilnosti razvrščanja. Prav tako so bile na zaslonu prikazane vse pravilno razvrščene slike. S tem so merjencem pomagali pri dojemanju abstraktnih lastnosti. Za testiranje algoritmov so uporabili Adabost in SVM z Gaussovim jedrom.

S tovrstnim testiranjem so ugotovili, da ljudje za popolno razvrščanje rabijo manj kot 20 slik za učenje medtem, ko se delovanje algoritmov izboljšuje z večanjem števila učnih slik nad 10 000. Več kot 90 % ljudi je rešilo 14 od 23 problemov medtem ko so algoritmi rešili le 11 problemov z napako razvrščanja pod 10 %.

Enak test so uporabili tudi Stabinger *et al.* [16], s katerim so želeli ugotoviti kakšen napredek smo dosegli s CNN pri razvrščevanju slik v abstraktne kategorije. Primerjali so delovanje dveh konvolucijskih mrež GoogleNet in LeNet.

S preprostim primerjanjem natančnosti so ugotovili, da v 25 letih ni bilo praktično nobene razlike. Povprečje obeh mrež je bila okoli 77 % (človek 93 %). Dodatno testiranje je pokazalo, da CNN niso sposobni reševati problemov primerjave oblik. Brez te primerjave je natančnost GoogleNet prekašala tako LeNet kot človeka. Seveda so pri tem poudarili, da CNN potrebuje vsaj 4000 slik za učenje.

Problem takega testiranja je, kot so že poudarili v [15], da ni prilagojen za otroke, ki še nimajo dovolj razvitega logičnega razmišljanja. Prav tako ni razvidno, kako poteka abstraktno dojemanje konceptov na naravnih slikah. Seveda je prav tako težko določiti ali se razvrščevalniki sploh naučijo abstraktnih lastnosti, ki jih podamo v testu.

### 1.1.5 Sistematična razlika med modeli in ljudmi

V [17] se avtorji sprašujejo, če razlika med ljudmi in stroji obstaja zaradi sistematičnih razlik. V ta namen so uporabili primerjavo razdalj med objekti v prostoru značilk.

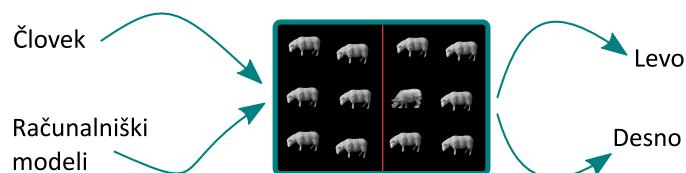


Figure 1.9: Slike iz [17].

Podatkovno bazo so zgradili iz naravnih objektov in silhuet. Vse slike so vsebovale izolirane objekte, brez ozadja. Objekti so bili iz različnih kategorij kot so živali, orodja in vozila.

V raziskavi je sodelovalo 269 ljudi (starost: 20–30 let). Vsak test se je začel s pritrditvenim križcem za 500 ms. Sledila je slika z eno ciljno enoto in več identičnimi motilnimi enotami. Ciljna enota se je razlikovala od motilnih enot. Merjenci so morali, čim hitreje in čim bolj natančno določiti ali se ciljna enota nahaja levo ali desno. Avtorji so nato povprečno frekvenco iskalnega časa uporabili kot približek zaznavanja različnosti med ciljno enoto in motilnimi enotami.

Za primerjavo so v [17] testirali 19 računalniških algoritmov, ki temeljijo na slikovnih elementih, robovih, značilkah, statistikah ali nevronskih mrežah. Za zaznavane različnosti pri računalniških algoritmih so uporabili razdaljo med vektorji značilk za vsako sliko.

Z raziskavo so ugotovili, da imajo vsi modeli podobno obliko odklona od človeškega zaznavanja. Odklon se pojavlja za specifičen tip slik, in sicer, za simetrične objekte in objekte, ki imajo podobne značilnosti. Tako so modeli zaznali manj različnosti med dvema simetričnima objektoma, kot je bilo to značilno za ljudi. Pri objektih s podobnimi značilnostmi pa so modeli zaznali preveč različnosti.

V delu [18] so preverjali, če ljudje in CNN uporabljajo podobno vizualno predstavitev. V ta namen so ustvarili spletno igro za identifikacijo značilk, ki jih ljudje ali CNN uporabljajo med razpoznavanjem. Igra sta igrala po dva igralca, učitelj in učenec. Učitelj je videl celotno sliko, učenec slike ni videl. S klikanjem je učitelj postopoma razkrival dele slike učencu, ta pa je moral ugotoviti v katero kategorijo spada. Na tak način so določili zemljevid pomembnosti.

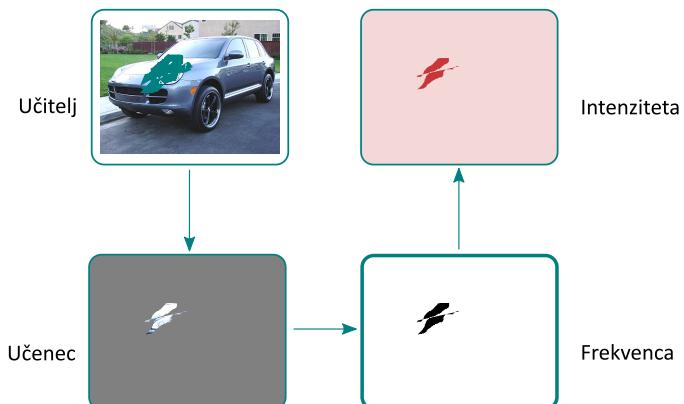


Figure 1.10: Slike iz [3].

V eksperimentu je sodelovalo 60 ljudi za CNN pa so uporabili VGG-16. 20 kategorij slik so izbrali iz ImageNet 2012 podatkovne baze.

Avtorji so ugotovili, da je korelacija med človeškim in CNN zemljevidom zelo majhna. To pomeni, da strategija razpoznavanja objekta ni enaka med človekom in CNN.

V delu [19] so izvedli študijo človeške in strojne pozornosti, z namenom, da bi razumeli, kam človek in stroj gledata, ko odgovarjata na vprašanja. S tem namenom so 800 merjencem pokazali meglene slike in jim naročili naj izostrijo dele slike tako, da bodo lahko odgovorili na postavljena vprašanja. Kot rezultat so dobili zemljevid pozornosti. Za določanje strojne pozornosti so izbrali nevronske mreže, ki predvidevajo, kam človek gleda. Ugotovili so, da imajo najbolj natačni modeli s človeško pozornostjo korelacijski koeficient 0.26, kar pomeni, da trenutni modeli ne gledajo na ista področja kot ljudje.

### 1.1.6 Preslepitev nevronskeih mrež

Da nevronske mreže še zdaleč ne dosegajo človeške sposobnosti so prikazali v delu [20]. Če imamo sliko  $\mathbf{x}'$ , ki leži v  $\epsilon$  okolini učne slike  $\mathbf{x}$  tako da velja,  $\mathbf{x}' = \mathbf{x} + \mathbf{r}$  pri pogoju  $\|\mathbf{r}\| < \epsilon$ , bo po teoriji algoritom določil visoko verjetnost pravilnega razreda tej slike. Z drugimi besedami, majhne, za človeka neopazne razlike, ne bi smele spremeniti njenega razreda.

Z rezultati pa so [20] pokazali ravno nasprotno. Z optimizacijskim problemom (1.1), kjer je  $f : \mathbb{R}^m \rightarrow \{1 \dots k\}$  razvrščevalnik in  $l \in \{1 \dots k\}$  učne labele, so razvili metodo, s katero lahko poiščemo slepe točke v okolini učne slike  $\mathbf{x}$ . Slike iz slepih točk predstavljajo nizko verjentost za izbran razred, zato so jih poimenovali kontradikotrne slike.

$$\begin{array}{ll} \min & \|\mathbf{r}\|_2 \\ \hline \text{p. p.} & f(\mathbf{x} + \mathbf{r}) = l \\ & \mathbf{x} + \mathbf{r} \in [0, 1]^m \end{array} \quad (1.1)$$

Seveda bi lahko rekli, da je problematika pretirana, saj robustnost modelov lahko izboljšamo z deformacijami učnih slik. Raziskovalci so v [20] argumentirali, da teh slepih točk ne moremo najti s preprostim naključnim vzorčenjem okoli učne slike, saj so te preveč medsebojno korelirane in tako statistično nezanesljive.

Nedelovanje nevronskeih mrež, pri rahlo spremenjenih slikah so prikazali tudi v delu [21]. Tu so razložili, da problem kontradiktornih slik nastane zato, ker so modeli preveč linearни. Prav tako so argumentirali, da pravilno razpoznavanje objektov še ne pomeni, da model popolnoma razume nalogo, ki smo mu jo zadali. Seveda lahko delno popravimo napačno delovanje modelov, vendar pa bi bilo bolje, če bi razvili nove metode optimizacije nelinearnih modelov, ki bi bili bolj stabilni [21].

Avtorji dela [22] so demonstrirali, da DNN modele lahko z lahkoto preslepimo tudi s slikami, ki so za človeka nerazpoznavne. Avtorji so izbrali DNN modele (AlexNet in LeNet), ki dobro delujejo na ImageNet ali MNIST

podatkovni bazi. Nato so z evolucijskimi algoritmi generirali slike, ki so jih DNN modeli razvrstili v kategorijo z visokim zaupanjem  $\geq 99.6\%$ .

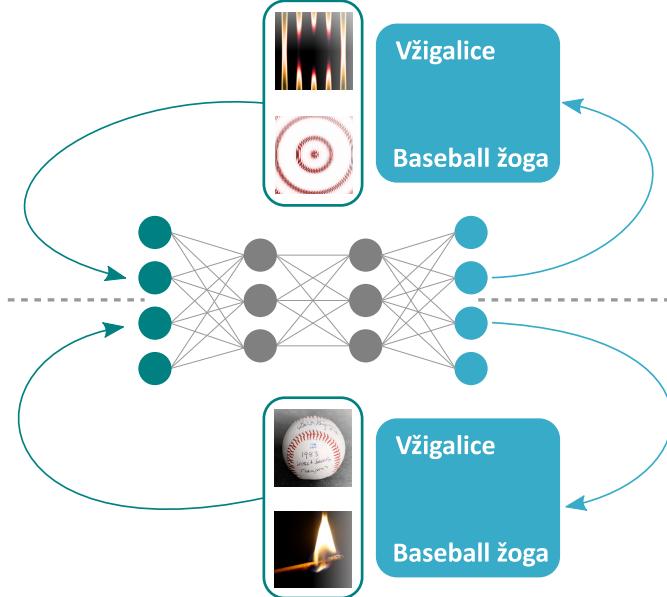


Figure 1.11: Slike iz [23].

Tako so dobili slike, kjer so DNN razpoznali številke, ampak te niso vsebovale številk in niso bile v ničemer povezane z ročno napisanimi številkami iz MNIST podatkovne baze. Podobne rezultate so dobili z naravnimi slikami iz ImageNet podatkovne baze.

Z generiranjem slik so ugotovili, da te pogosto vsebujejo značilnosti, ki se pojavljajo v ciljni kategoriji. To pomeni, da so evolucijski algoritmi za dobro delovanje generirali značilnosti, ki so diskriminativne, se uporabljajo za razlikovanje med razredi. To vodi v sklepanje, da se DNN-ji učijo nizko in srednje nivojskih značilk, kot pa globalne strukture objektov. V nasprotnem bi dobili nižje rezultate zaupanja, saj so bili v generiranih slikah ponavljajoči vzorci, ki v naravi redko obstajajo.

Pri nadaljnem testiranju so avtorji preverili, če lahko tako generirane slike generaliziramo na več vrst DNN. Rezultati so pokazali, da se različni DNN-ji naučijo podobnih značilnosti, obstajajo pa tudi slike ki so natančno nastavljene na določeno vrsto DNN.

Problem preslepitve DNN-jev bi lahko rešili z dodatnim učenjem, kjer bi generirane slike določili pod negativno kategorijo. Seveda so v [22] to tudi preverili. Kljub učenju na generiranih slikah, so z evolucijskimi algoritmi še vedno lahko generirali nove slike, ki so jih DNN modeli razvrstili v pozitivno kategorijo z visokim zaupanjem.

Podobno kot [20] in [21] so se s kontradiktornimi slikami ukvarjali tudi [24]. Avtorji so se osredotočili na rahlo spremenjene slike, s katerimi lahko

preslepimo tudi ljudi. Pri psihofizioloških testih, so omejili opazovanje slike na 71 ms. Z majhnim časom opazovanja so želeli procese v možganih bolj približati delovanju umetnih nevronskih mrež in meriti majhne razlike v natačnosti ljudi. Ugotovili so, da primeri slik, ki ne delujejo med različnimi algoritmi prav tako vplivajo na človeško percepциjo. S tem so želeli pokazati, da imajo tudi pri preslepitvah ljudje in algoritmi podobne lastnosti.

## 1.2 Izboljšava algoritmov z vnosom človeških karakterističnih značilnosti

Ljudje se lahko hitro naučimo novih konceptov iz nekaj pozitivnih slik, kar pa je za računalniški model zelo težavna naloga [25]. Naučene koncepte lahko fleksibilno uporabimo na različnih primerih, v drugačnem okolju, kar pa za algoritme ne velja [26]. V ta namen je Jia *et al.* [25] predlagal nov izziv za strojno učenje—učenje vizualnih konceptov. V tem izzivu mora sistem določiti, če izbrana slika spada v določen koncept.

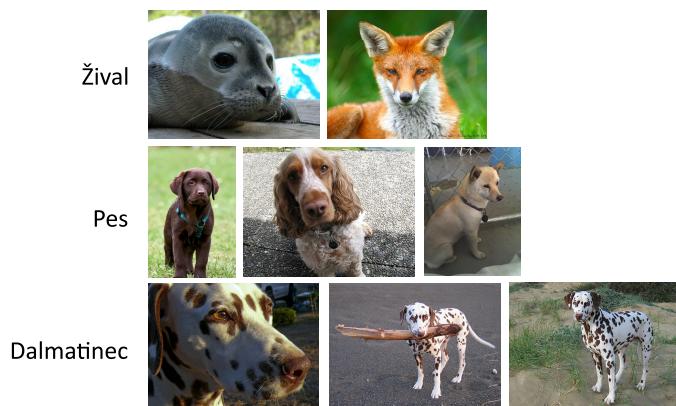


Figure 1.12: Slike iz [3].

Problematiko so osvetlili s razvojem algoritma, ki bolje oponaša učenje otrok in se lahko nauči novih vizualnih konceptov iz majhnega števila slik. Algoritem temelji na linearinem logističnem regresorju z minibatch Adagrad algoritmom in posplošenim Bayesovim algoritmom. Regresor so učili na ImageNet 2010 podatkovni bazi. Z njim so dosegli 41.28 % top-1 natančnost in 61.69 % top-5 natančnost na testnih podatkih. Bayesov algoritem temelji na računanju verjetnosti, da slika iz neznane kategorije spada v kategorijo vzorčnih slik. Tu so uporabili konfuzijsko matriko, s katero so dodali vizualno nejasnost, ki se lahko pojavi zaradi nepopolnosti razvrščevalnikov.

Za testiranje svoje hipoteze so avtorji uporabili slike iz ImageNet 2010 podatkovne baze. Pri tem so izbrali 4 tipe kategorij iz pomenskega drevesa. Prva kategorija je bila najbolj specifična (borovnica), nadaljnje pa so bile

vedno bolj posplošene (sadje, rastlina). Pri izbiranju kategorij so pazili, da je pridobljena informacija med njimi največja.

Podatke za referenco, so v [25] pridobili s pomočjo ljudi. Vsak subjekt so testirali tako, da so mu podali 5 vzorčnih slik, ki naj bi predstavljale nek koncept. Subjekt je nato za novih 20 slik (12 pravilnih in 8 napačnih) moral določiti, ali so povezane z vzorčnimi slikami.

Z eksperimenti so pokazali, da dobijo za okoli 10 % boljše rezultate glede na ostale metode in ravno za toliko slabše rezultate glede na človeka.

Problematiko učenja konceptov so opredelili tudi v delu [26]. Razvili so novo metodo Bayesovega programskega učenja (BPL), ki se lahko nauči velika števila konceptov samo iz enega primera. Osnovni koncepti so predstavljeni kot verjetnostni programi, nove koncepte pa se BPL uči z združevanjem starih na podlagi vzročnosti in sestave objektov iz realnega sveta [26].

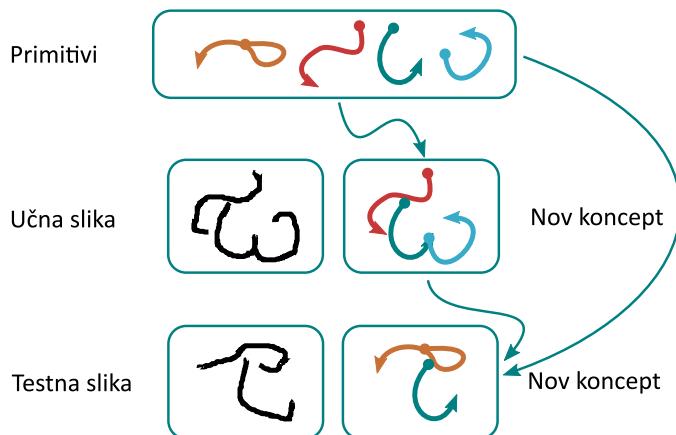


Figure 1.13

BPL so preizkusili na ročno pisanih črkah, kjer so osnovne koncepte (primitive) predstavljale zaključene krivulje, ki jih naredimo s peresom, sestavo objektov pa so določili na podlagi prostorskih relacij med njimi. Algoritmom so raziskovalci naučili tako, da so mu podali učno sliko ročno napisane črke, ta pa je s svojimi koncepti rekonstruirala črko, ki je bila najbolj verjeten približek. Naučeno strukturo je BPL nato uporabil kot osnovo za rekonstrukcijo testnih slik.

BPL algoritmom so primerjali z rezultati 40 ljudmi, ki so med testiranjem morali za vsak primer črke najti najbolj podobno izmed 20 ponujenih. Raziskovalci so za BPL dobili podobno stopnjo napake, kot pri človeku, medtem ko so bile globoke nevronske mreže veliko slabše. Povzetek rezultatov iz [26] je prikazan v tabeli 1.1.

Poleg primerjanja stopnje napake so v [26] izvedli še poenostavljen Turingov test, kjer je posameznik poskušal razpozнатi, katero črko je narisal človek in katero je generiral algoritmom. Ljudje so razlike zelo težko razpoznali, kar

Metoda	Stopnja napake [%]
Ljudje	4.50
BPL	3.30
ConvNet	13.50
SiameseNet	8.00

Table 1.1

nakazuje na približevanje BPL algoritma človeku. Seveda so Lake *et al.* [26] poudarili, da BPL algoritem še vedno vsebuje slabše koncepte kot človek, saj strukture niso tako abstrakne in kompleksne, da bi jih lahko uporabili za razpoznavanje vsakodnevnih objektov, kot so vozila in drevesa.

V [27] so se izboljšanja modelov lotili z neposrednim koriščenjem človeške sposobnosti. S spletnim testiranjem ljudi so pridobili vzorce napak, ki so jih prevedli v človeške utežne funkcije. Te so nato uporabili v SVM, kjer funkcija dodaja kazni mejam, ki niso skladne s podatki o ljudeh.

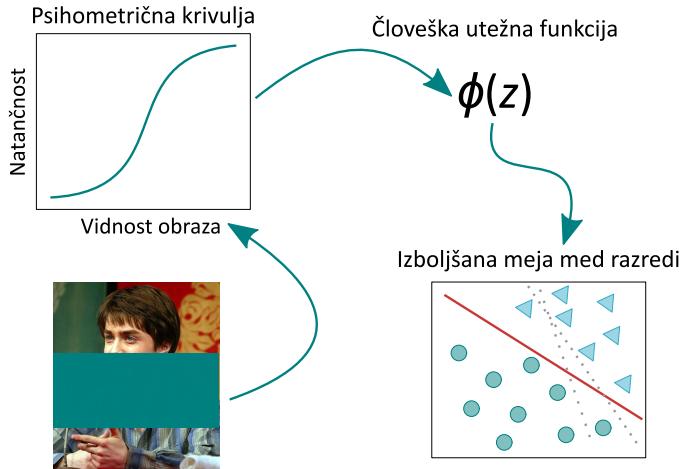


Figure 1.14: Slike iz [28].

Utežno funkcijo so pridobili s pomočjo dveh testov. V prvem testu so avtorji vsakemu merjencu  $102 \times$  pokazali 3 slike, izmed katerih je merjenec moral izbrati sliko z obrazom. Pri tem so spremenjali območje vidnosti obraza. Tako so lahko dobili krivuljo človekove natačnosti glede na vidno območje obraza.

V drugem testu so uporabili sivinske slike, ki so se prikazale za 50 ms. Merjenec je moral nato določiti, ali je na sliki videl obraz ali ne.

Scheirer *et al.* [27] je z eksperimenti pokazal, da z uporabo take utežne funkcije močno izboljšamo delovanje detektorjev obraza. Izboljšavo lahko še povečamo če uporabimo značilke, ki temeljijo na biologiji (Cox in Pinto).

Avtorji dela [29] so se koriščenja človeške sposobnosti lotili še bolj elementarno. Ti so v učni proces implementirali meritve človeških možganskih aktivnosti fMRI. Meritve so izvajali na enem merjencu, ko je ta opazoval vse slike iz izbrane podatkovne baze. Z merjenjem so dobili nevronski odziv na zazanavanje slike in jih uporabili kot značilke za učenje SVM razvrščevalnika z RBF jedrom. Rezultate razvrščanja so nato transformirali v verjetnosti preko logistične funkcije. Na ta način so dobili utežno funkcijo aktivnosti, s katero so kaznovali napačno razvrščanje pri učenju novih modelov.

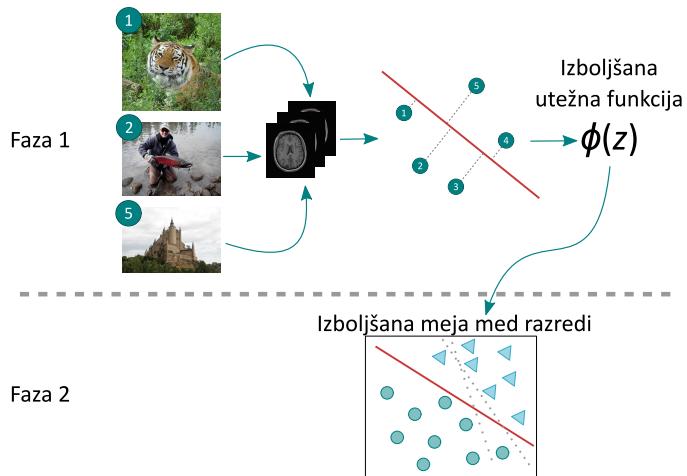


Figure 1.15: Avtor fMRI slike je [30] po licenci <https://creativecommons.org/licenses/by/2.0/legalcode>. Ostale slike smo dobili iz ImageNet podatkovne baze [3].

Pri eksperimentiranju so uporabili podatkovno bazo 1386 naravnih prizorov ljudi, živali, zgradb, hrane in vozil. Za testiranje so uporabili SVM razvrščevalnik s RBF jedrom, ki so ga učili na HOG in CNN značilkah. CNN značilke so dobili iz prednaučenega AlexNet modela.

Njihovi eksperimenti so pokazali, da lahko s takim načinom izboljšamo delovanje razvrščevalnikov. Izboljšave so najbolj opazne na ročno izdelanih značilkah, kot je HOG.

Branson *et al.* [31] je prav tako koristil človeške sposobnosti, vendar na popolnoma drugačen način. Sestavil je hibrid med modelom in človekom, ki ga imenujemo tudi človek-v-zanki (angl. Human-in-the-loop) [31]. Pri tem načinu, model uporabimo za razvrščanje enostavnih problemov, človeka pa za ostalo. Delovanje so prikazali na podatkovni bazi Birds-200, kjer s pomočjo razvrščevalnika in vprašanj, ki jih postavimo subjektu določimo kategorijo ptice, ki je prikazana na sliki.

V primeru, da so za razvrščanje ptic v kategorije uporabili samo vprašalnik, so ljudje, pred pravilnim razvrščanjem, v povprečju odgovorili na 12 vprašanj. Ob pomoči računalniškega vida pa se je število vprašanj zmanjšalo na 7.

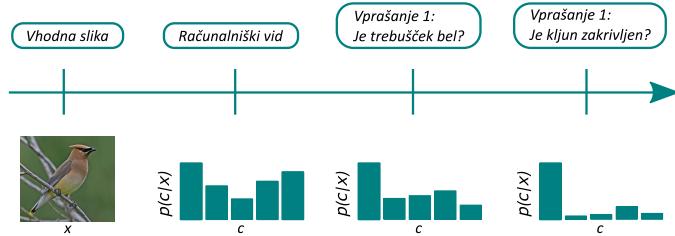


Figure 1.16: Slike iz [32].

Gledano z druge perspektive je človek v povprečju povečal natančnost algoritma iz  $\sim 19\%$  na  $\sim 66\%$ .

Dodajanje človeka za izboljšanje sistema so raziskovali tudi v [33]. Tu so se osredotočili na CRF modelle s katerimi izvajamo segmentacijo, detekcijo, analizo oblike, razpoznavanje prizora in razumevanje konteksta. V raziskavi, v kateri je sodelovalo več kot 500 oseb, so zamenjali različne segmente CRF modela s človekom in preizkusili njegovo delovanje na MSRC-21 podatkovni bazi. Rezultati so pokazali, da obstaja potencial izboljšanja modelov pri lokalni segmentaciji.

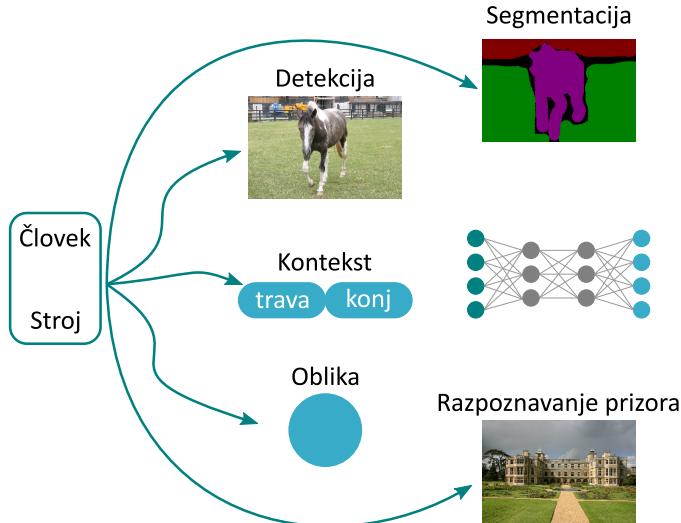


Figure 1.17: Slike iz MSRC podatkovne baze [34]

Razširitev dela [33] so izvedli v [35], kjer so se osredotočili na izboljšanje sistema detekcije objektov in razpoznavanja prizora. Sistem segmentacije so ponovno preizkusili na bolj zahtevni PASCAL podatkovni bazi in prišli do podobnih zaključkov.

V delu [36] so se raziskovalci osredotočili na združevanje človeka in algoritmov za razpoznavanje obrazov. Rezultate algoritma in ljudi so združili z regresijo delnih najmanjših kvadratov. Merjenci so v eksperimentu določevali

verjetnost, da para slik predstavlja isto osebo. Slike so lahko opazovali 2 s. V eksperimentu je sodelovalo 49 oseb (25 moških in 24 žensk). Tudi njihovi rezultati so pokazali, da združevanje pripomore k izboljšanju rezultatov.

Zanimivo raziskavo so izvedli Vondrick *et al.* [37], kjer so raziskovali pristranskoost človeka v razpoznavanju objektov. Argumentirali so, da razlika v delovanju algoritmov in človeka nastaja tudi zaradi človeške pristranskoosti. Pokazali so, da pristranskoost lahko v določeni meri ovrednotimo in jo tudi uporabimo za izboljšanje algoritmov.

Pristranskoost so določili tako, da so ljudem pokazali barvne slike belega šuma in jim naročili naj jih kategorizirajo v izbrane kategorije. Z velikim številom ljudi, ki so sodelovali na Amazon Mechanical Turk, so dobili povprečne slike, ki so nakazovale oblike, podobne izbranim kategorijam. Pridobljene slike so nato klasificirali in primerjali s testno množico realnih slik iz podatkovne baze PASCAL VOC 2011. Ugotovili so, da jih algoritmi lahko dobro klasificirajo, saj so bile povprečne natančnosti večje od naključnosti. Povprečne slike so Vondrick *et al.* [37] zato določili kot predloge človeške pristranskoosti za izbrano kategorijo.

### 1.3 Biološke raziskave razpoznavanja objektov

Intuicija nam govori, da bomo z razpoznavanjem 3D objektov iz 2D slik zelo težko dosegli človeške zmogljivosti, saj s slikami izgubimo določeno informacijo o naravnih objektih. Kljub temu vemo, da ljudje in opice s pogledom na sliko hitro razpoznaobjekte [38]. Zato so v delu [39] raziskovali biološko verodostojnost razpoznavanja 3D objektov iz 2D slik. Primerjali so dva načina modeliranja za razpoznavanje objektov. Slikovni način naj bi bil razmeroma boljši od strukturnega opisa, saj pri slednjem potrebujemo postopek rekonstrukcije [39]. Seveda je strukturni opis, s katerim rekonstruiramo 3D objekte bolj robusten od slikovnega načina, saj ta predstavlja izbran objekt samo iz specifičnega pogleda [39]. Avtorji [39], vseeno zagovarjajo slikovni način, saj so z raziskavami na opicah pokazali, da v možganih poleg od pogleda neodvisnih nevronov obstajajo tudi taki, ki so aktivni za specifične poglede izbranega objekta. Za razpoznavanje 3D objektov iz 2D slik bi zato za najboljše rezultate morali poleg slikovnih značilk uporabiti struktурno informacijo objekta [39].

Dodatno potrditev, da so algoritmi razpoznavanja objektov iz 2D slik biološko pomenljivi, so raziskovalci pokazali v delu [40]. Raziskovali so, kako dobro različni modeli razpoznavanja posnemajo vidno pot v možganih. Pri tem so se osredotočali na starejše modele, ki temeljijo na SIFT značilkah in Gaborjevih filtrih.

V eksperimentih so uporabili 60 slik objektov na sivem ozadju. Sodelovalo je 5 oseb (moški spol: 4, ženski spol: 1, starost: 20–24 let). Vsako sliko so prikazali za 2 s. Sledila je sivinska slika z naključnim trajanjem

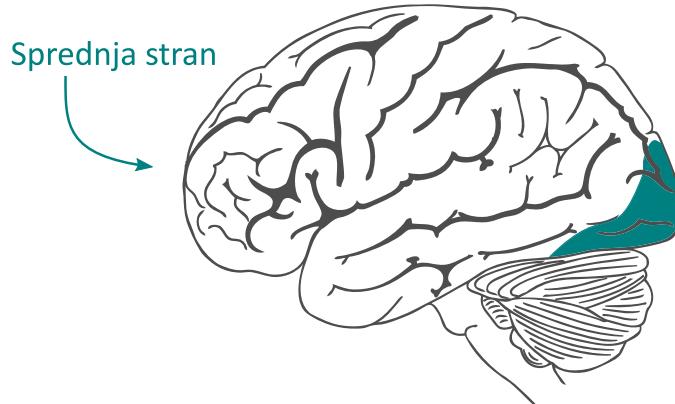


Figure 1.18: Avtor možganov je [41].

med 500 ms in 3000 ms. Na koncu so prikazali še sredinski križec. Pasivno opazovanje slik so merili z MRI napravo. fMRI slike merjencev so povezali z računalniškimi modeli preko reprezentativne disimilarne matrike, ki vsebuje razdalje med objekti. Ker z razdaljo merimo podobnost dveh objektov, disimilarna matrika predstavlja razvrščanje objektov v razrede.

Leeds *et al.* [40] je ugotovil, da SIFT daje najboljše rezultate, čeprav so korelacije med modelom in korteksom dokaj nizke (pod 0.3).

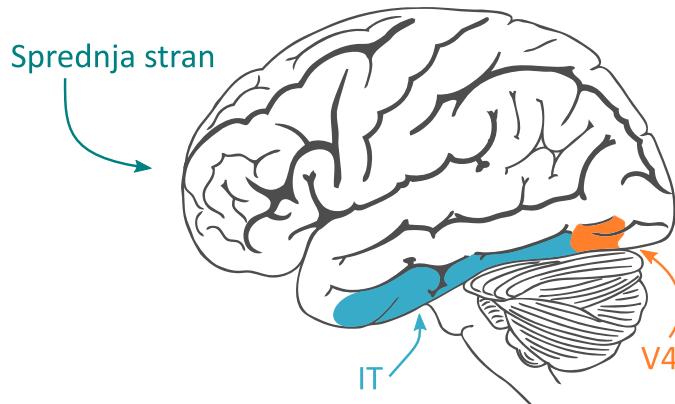


Figure 1.19: Avtor možganov je [41].

V novejših nevroznastvenih raziskavah so raziskovalci ugotovili, da se kategorična informacija za razpoznavanje pojavlja v inferiorni senčni skorji (tudi inferiorni temporalni korteks, inferotemporalni korteks ali IT korteks) [42]. Z razvojem modelov, ki posnemajo ta korteks, bi tako lahko lažje razumeli kako ljudje razpoznavajo objekte [42]. V ta namen so v [42] zgradili model ventralnega toka informacij. Kvaliteto modela so določili z reprezentativno matriko razlik, kjer za vizualni stimulant izmerimo razdaljo v možganih (fMRI) in v računalniških modelih z znacilkami.

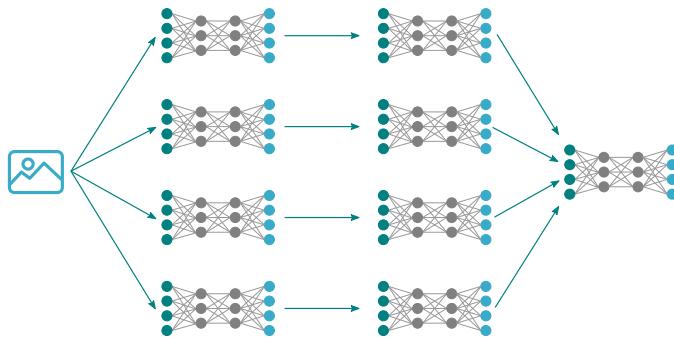


Figure 1.20: Avtor ikone slike je [43]. Ikono smo pridobili po licenci <https://creativecommons.org/licenses/by/3.0/legalcode>.

Model so zgradili iz hierarhičnega sklada enostavnih nevronskeih mrež in ga optimizirali z modularno optimizacijo. Pri tej optimizaciji enostavne nevronske mreže najprej specializiramo na del celotnega problema (razpoznavanje določene kategorije). Mreže nato združimo in popravimo uteži tako, da ima celotna mreža najmanjš napako.

Model so testirali na NRB podatkovni bazi, ki je bila originalno namenjena primerjanju nevronskeih značilnosti opic in ljudi [42]. Matrice razlik so pokazale, da HMO algoritmom dobro opisuje ventralni tok ljudi in opic.

Podobno so se z modeliranjem IT korteksa ukvarjali v [44]. Tudi tu so modele zgradili iz hierarhičnega sklada enostavnih nevronskeih mrež. S primerjavo med modelom in odzivom opičjih možganov so ugotovili, da se natančnost predikcije povečuje z vsakim novim slojem nevronov. Prav tako so z rezultati pokazali, da so, poleg IT korteksa, modeli podobni tudi odzivom vidnega prodročja V4 v vizualnem korteksu.

Seveda pa je sama primerjava med umetnimi nevroskimi mrežami in možganskim vidnim zaznavanjem zelo težka, saj po eni strani nimamo enotne metrike pri eksperimentiranju, po drugi strani pa obstajajo računalniške omejitve [45]. Zato so v [45] uvedli nov način merjenja zmogljivosti pri razpoznavanju kategorij objektov. Natančnost razpoznavanja so določili kot funkcijo kompleksnosti razvrščevalnika in primerjali delovanje DNN-jev s primati.

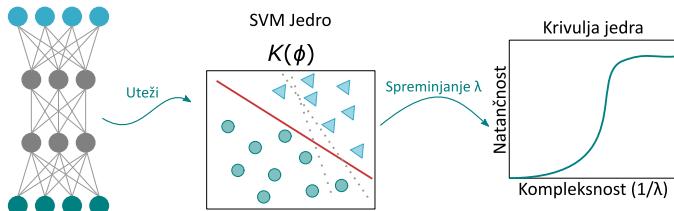


Figure 1.21

V svoji študiji so izoblikovali 1960 sintetičnih slik, s katerimi so kontrolirali variacije pozicije in orientacije objektov na slikah. Ozadje je bilo določeno naključno. Eksperimente so izvajali s pomočjo dveh primatov. Vsako sliko so jima prikazali za 100 ms s 100 ms periodo brez slike. Pri tem so merili odziv v IT korteksu in opazovali pozicijo očesa živali.

Cadieu *et al.* [45] je po svoji metodi ugotovil, da imajo DNN modeli podobno, če ne celo boljšo natančnost.

Študija primerjave med nevronskimi mrežami in primati je potekala tudi v delih [46] in [38]. V prvem delu so raziskovalci večinoma primerjali opice in ljudi v drugem delu pa so se osredotočili na primerjavo med DNN-ji in primati. Tu so generirali sintetične slike na enak način kot v delu [9]—3D modeli so bili z naključno pozicijo in orientacijo postavljeni na naravne slike.

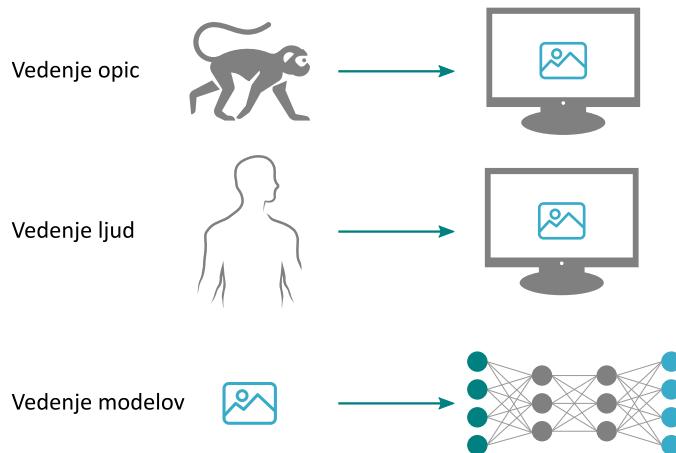


Figure 1.22: Avtor opice je [47], avtor človeka je [48] avtor ikone slike je [43] in avtor ekrana je [49]. Ikone smo pridobili po licenci <https://creativecommons.org/licenses/by/3.0/legalcode>.

Teste so v [38] izvajali tako, da so najprej prikazali črno piko za 500 ms. Sledila je testna slika (100 ms) z naključno pozicijo in orientacijo. Na koncu sta se prikazali še dve sliki objektov s kanoničnim pogledom izmed katerih je ena slika vsebovala testni objekt. Merjenec je nato moral izbrati eno izmed slik. Testirali so 1476 ljudi na AMT, 5 odraslih opic (*Macaca mulatta*) in DNN modele.

Za testiranje so uporabili naslednje vedenjske metrike:

1. *B.O1* Razlikovanje objekta proti vsem ostalim.
2. *B.O2* Razlikovanje objekta proti distraktorju.
3. *B.I1* Razlikovanje slike proti vsem ostalim.
4. *B.I2* Razlikovanje slike proti distraktorju.

Rezultate so ovrednotili na podlagi normalizirane Pearsonove korelacije, kjer je izbrani vizualni sistem identičen človeškemu pri vrednosti 1.0, četudi obstaja šum v podatkih. Rezultati za posamezen test so zbrani v tabeli 1.2. Avtorji so odkrili, da so DNN-ji natančno določili vzorce primatov o zmedenosti med objekti (kako pogosto zmedeno razpoznamo kamelo za psa). Vendar pa, ko so raziskovali za posamezne slike, so ugotovili, da noben DNN model ne deluje dobro.

Metoda	B.O1	B.O2	B.I1	B.I2
Ljudje	0.97	0.94	0.96	0.77
Opice	0.90	0.80	0.77	0.75
Inception-V3	~0.90	>0.80	0.62	0.53

Table 1.2

## 1.4 Razpoznavanje obrazov

Razvoj algoritmov z namenom preseganja človeške zmogljivosti poteka tudi na področju razpoznavanja ljudi. V letih 2007–2015 so se vrstile številne študije primerjave med zmogljivostjo ljudi in takratnih najboljših algoritmov. Sprva so se raziskovalci osredotočali na enostavnejše probleme, kot so različni pogoji osvetlitve na slikah frontalnih obrazov. V [50] so tako primerjali natačnost algoritmov in ljudi na slikah FRGC izziva (angl Face Recognition Grand Challenge). Za problem so izbrali ujemanje obrazov na parih slik, ki so jih z algoritmom po PCA (Principal components analysis) metodi razdelili na lahke in težke primere. Z različnimi eksperimentalnimi protokoli, ki so podrobneje opisani v [50], so nato testirali 91 ljudi (45 žensk in 46 moških). Vsak merjenec je moral oceniti ujemanje para slik na lestvici 1–5. Tako so dobili ROC krivulje, ki so jih nato primerjali s krivuljami algoritmov.

Njihovi rezultati so pokazali, da algoritmi prekašajo ljudi na enostavnih primerih, na težkih primerih pa so primerljivi z ljudmi. Z dodatno analizo, so nato pokazali, da razlike niso nastale, ker bi bili ljudje utrujeni. Eden izmed razlogov zakaj obstajajo razlike, bi lahko bil ta, da merjenici niso poznali obrazov na slikah [50]. Drugi razlog bi lahko bil, da so značilke na enostavnih slikah dobro primerljive [50]. Ta dodatna informacija, bi lahko bila bistvena za prekašanje ljudi. S tem so v [50] prišli do sklepa, da so po natančnosti algoritmi primerljivi s človeško zmogljivostjo.

Zelo podobno primerjavo med algoritmi in ljudmi so naredili v [51], kjer so poleg FRGC uporabili še podatkovno bazo obrazov FRVT 2006. Tudi v tem delu so se osredotočili na primerjavo pod različnimi osvetlitvenimi pogoji in prišli do podobnih zaključkov, da so algoritmi uspešnejši od ljudi. Z

združevanjem človeka in algoritmov z regresijo delnih najmanjših kvadratov so zmanjšali stopnjo napake iz 0.12 na 0.008 in tako izboljšali rezultate. Pri tem so prišli do sklepa, da obstajajo razlike v napakah, ki jih delajo ljudje in algoritmi [51].

Nadaljnja testiranja na bolj zahtevnih podatkovnih bazah (GBU Phillips et al 11) so pokazala, da je sklepanje o prekašanju ljudi napačno. Zato so se v [52] osredotočili na bolj sofisticirano primerjavo, saj so poleg nekontrolirane osvetlitve uporabili tudi dnevno spreminjanje videza ljudi, kot so frizura, obrazna mimika, uporaba očal in pokrival. Enako kot v prejšnjih delih, so tudi tu preverjali zmogljivost pri primerjavi parov slik. Slike so z najboljšim algoritmom za razpoznavanje obrazov razdelili na tri težavnostne stopnje. Ljudi so testirali na podoben način kot [50] in primerjali ROC krivulje ljudi in algoritmov.

Rezultati ROC krivulj iz [52] so prav tako pokazali, da algoritmi prekašajo ljudi na enostavnejših slikah in da so primerljivi na bolj težavnih. V nadaljnem testu so preverili še korelacijski koeficient med ljudmi in izbranim algoritmom in ugotovili, da so korelacije negativne in statistično pomembne. Prav tako so ugotovili, da med korelacijami obstaja raztros, kar nakazuje, da obstajajo razlike med ljudmi in algoritmi.

S tovrstno primerjavo so se ukvarjali tudi v delu [53]. Tu so avtorji predstavili izsledke raziskovanja človeške natančnosti razpoznavanja obrazov na slikah in video posnetkih. Za eksperimente so uporabili LFW in YTF podatkovno bazo. Podatke so zbrali s pomočjo Amazon Mechanical Turk spletnne strani, kjer so merjenci morali določiti, če se na parih slik ali posnetkov pojavi ista oseba. V raziskavi je sodelovalo 307 oseb (27.4 % iz ZDA in 55.1 % iz Indije). Rezultate so primerjali s [54] in z algoritmom DeepFace [55].



Figure 1.23: Avtor človeka je [48] po licenci <https://creativecommons.org/licenses/by/3.0/legalcode>. Slike iz [56].

Za slike so [53] dobili 98.3 % človeško natančnost, medtem ko so v [54] za človeško natančnost določili 99.2 %. Algoritem DeepFace [55] je dosegel 97.5 % natančnost. Pri testiranju na video posnetkih je bila natančnost ljudi veliko slabša od algoritma (89.7 % ZDA in 88.6 % Indija proti 91.4 %), vendar so ljudje prekašali algoritem s stališča TAR pri nizkih vrednostih FAR. Rezultati so povzeti v tabeli 1.3.

<b>Metoda \ FAR</b>	0.4 %	1.0 %
Ljudje (ZDA)	71.20	80.60
Ljudje (Indija)	44.90	63.70
DeepFace [55]	25.90	54.80

Table 1.3

Raziskovalci so v [53] ugotovili, da v podatkovnih bazah obstaja demografska pristranskost, saj večino ljudi izhaja iz zahodne Evrope. Prav tako so opazili vpliv znanih obrazov. Ljudje so se pri znanih obrazih odrezali bolje.

Z rezultati so v [53] prikazali, da obstajajo razlike v interpretaciji uspešnosti algoritmov ob uporabi različnih metrik. To lahko nakazuje na neprimeren način testiranja, ki vodi v napačno sklepanje. S tovrstno problematiko so se ukvarjali v [57], kjer so predstavili novo ogrodje za primerjalno analizo. Relativno zmogljivost ljudi in algoritmov so karakterizirali z AUC statistiko in jo grafično predstavili na točkovnem grafu, kjer  $x$  os predstavlja AUC za zmogljivost ljudi in  $y$  os AUC za zmogljivost algoritmov. Točke pod diagonalo pomenijo, da so ljudje uspešnejši, točke nad diagonalo pa, da so uspešnejši algoritmi. Na tak način so ugotovili, da algoritmi prekašajo ljudi, ko obdelujemo frontalne obraze na slikah. V primeru video posnetkov pa so ljudje še vedno najboljši. To so razložili tako, da ljudje pri razpoznavanju dobro uporabljamo tudi druge značilnosti na sliki, kot so deli telesa, medtem ko algoritmi tega ne zmorejo.

V [58] so ravno tako poskušali izboljšati protokole primerjanja. Predstavili so nov način pridobivanja človeške zmogljivosti (zlivanje), s katerim dobimo višje ocene. Po standardnem načinu (združevanje) vsak merjenec oceni ujemanje parov slik ali posnetkov na skali 1–5. Iz zbranih podatkov nato izračunamo ROC krivuljo. Po novem načinu pa pred izračunom ROC krivulje zbrane ocene še povprečimo. Primerjavo med načinoma za težje primere slik lahko vidimo v tabeli 1.4. Podatki so zbrani iz [58]. Z novim načinom dobimo višje ocene človeške zmogljivosti tako za slike kot video posnetke.

Z novo metodo računanja človeške zmogljivosti so v [58] pokazali, da v težkih primerih človeško zmogljivost doseže le en algoritem. V ekstremno težkih primerih pa rezultati kažejo na to, da je zmogljivost ljudi boljša od

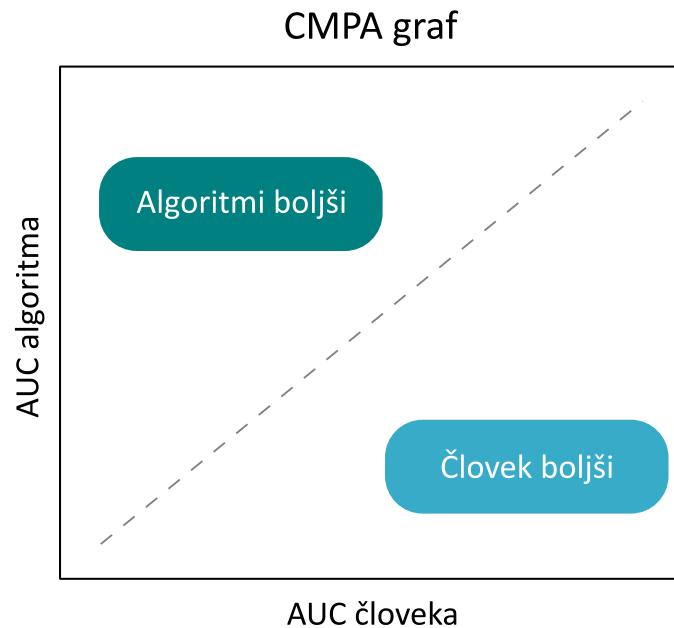
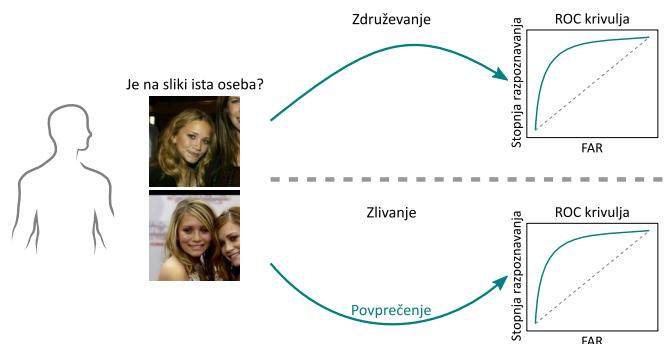


Figure 1.24

Figure 1.25: Avtor človeka je [48] po licenci <https://creativecommons.org/licenses/by/3.0/legalcode>. Slike iz [56].

Metoda	AUC (slike)	AUC (posnetki)
Združevanje	0.86	0.94
Zlivanje	0.99	1.00

Table 1.4

naključja, zmoglјivost algoritmov pa ne.

Pomembna odkritja na področju nestandardne primerjalne analize so dobili tudi [59]. V razpoznavanje obrazov so uvedli vizualno psihofiziko, kot je bila predstavljena že v poglavju 1.1.1. Delovanje algoritmov so predstavili s krivuljo odziva v odvisnosti od izbrane popačitve slike. Po nedavnih rezultatih na podatkovnih bazah bi verjeli, da se najbolje obnesejo algoritmi, ki temeljijo na nevronske mrežah. Presnetljivo pa so izbrani stresni testi pokazali, da eden izmed slabših algoritmov OpenBR **a**, ki temelji na ročno izdelanih značilkah LBP **a** in SIFT **a** prekaša DNN algoritma FaceNet **a** in OpenFace **a**. Kot so poudarili v [59], rezultati nakazujejo na to, da ni vedno mogoče določiti robustne značilke z učenjem na veliki množici podatkov.

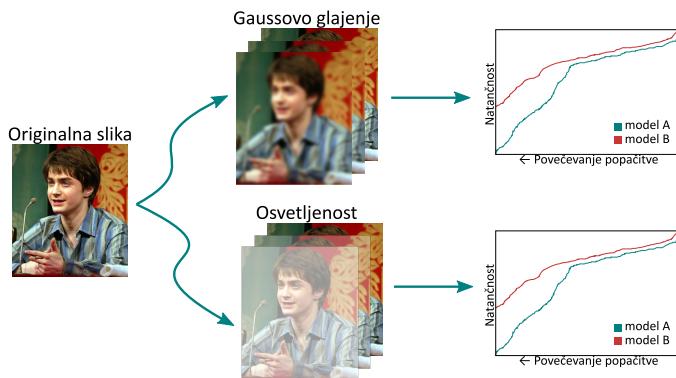


Figure 1.26: Slike iz [28].

V [59] so predstavili tudi izsledke pri primerjavi z ljudmi. Merjencem so najprej prikazali sliko za 50 ms in nato barvni šum za 500 ms. Zatem so dobili 3 slike izmed katerih so morali izbrati najbolj podobno sliko prvi. Ugotovili so, da je delovanje algoritmov in ljudi konsistentno le pri Gaussovem glajenju in zmanjševanju kontrasta.

Vsa omenjena dela na področju razpoznavanja obrazov lepo nakazujejo na dve temeljni problematiki, ki se pojavljata pri primerjanju zmoglјivosti ljudi in algoritmov. Prva izhaja iz same podatkovne baze. S podatkovnimi bazami zelo težko zajamemo kompleksnost... Kot sta pokazala že **a** in **b** so baze pristranske...

Druga problematika temelji na načinu primerjanja človeške in algoritmične zmoglјivosti. Kot sta pokazala [53] in [57], metrike, ki so splošno uporabljene na področju razpoznavanja obrazov, lahko vodijo v napačno interpretacijo rezultatov...

## 1.5 Pristranskost podatkovnih baz

Na vseh stopnjah raziskovanja in razvoja algoritmov razpoznavanja in detekcije objektov potrebujemo primerne podatkovne baze [60]. So glavni razlog za razvoj računalniškega vida, saj vsebujejo veliko podatkov za učenje, prav tako pa z njimi lahko primerjamo algoritme med seboj [61]. Ravno zaradi podatkovnih baz lahko rečemo, da je računalniški vid eksperimentalna znanost [61]. Čeprav si brez njih razvoja računalniškega vida težko predstavljam, pa v njih obstajata dva temeljna problema, vedenjske napake in pristranskost.

S podatkovnimi bazami se raziskovalci osredotočajo na premagovanje pogostokrat ene številke, ki predstavlja zmogljivost algoritma [61]. Kot so pokazali v ?, bolj zmogljiv algoritem ni nujno statistično signifikanten. Podatkovne baze, ki bi morale biti vzorec realnega sveta tako postanejo same sebi namen [61]. Zadnji očiten vedenjski problem pa je t.i. plazeča prenasičenost, kot jo imenuje [61]. Če je podatkovna baza dolgo na razpolago, algoritme že tako dobro nastavimo na bazo, da izgubi možnost generalizacije [61].

Še večji problem podatkovnih baz je njihova kvaliteta. V delih, ki jih bomo podrobnejše predstavili v nadaljevanju, opisujejo ravno ta aspekt podatkovnih baz. Kvaliteta se najpogosteje odraža v pristranskosti podatkovnih baz in močno vpliva na algoritme. Lahko bi rekli, da je zmogljivost algoritma odvisna od kvalitete podatkovne baze, ki jo uporabljamo za razvoj. Seveda obstaja veliko pristranskosti na račun različnega namena podatkovnih baz [61]. Nekatere vsebujejo samo prizore, druge profesionalne fotografije in tretje amaterske fotografije iz interneta. Četudi bi izločili namensko pristranskost, so v [61] ugotovili, da v določeni meri pristranskost še vedno obstaja. To so pokazali na primeru avtomobilov. Caltech tako vsebuje avtomobile s stranskim pogledom, ImageNet pa večinoma športne avtomobile [61]. Skozi zgodovino razvoja podatkovnih baz lahko vidimo, da je razvoj potekal večinoma tako, da so se z novimi podatkovnimi bazami hoteli znebiti pojavljajoče pristranskosti [61]. Kljub trudu pristranskost ostaja.

V [60] so večinoma raziskovali Caltech 101 [62] in PASCAL VOC podatkovno bazo ?. Za Caltech 101 so poudarili, da ne obstaja medrazredna variabilnost. Večino objektov izbranega razreda ima enako velikost, orientacijo in zorni kot pogleda. To lahko enostavno prikažemo s povprečenjem slik izbranega razreda. Če bi imeli veliko variacijo, bi bile slike ?? homogene.

PASCAL VOC podatkovna baza, naj bi rešila probleme Caltech 101. Ima večjo medrazredno variabilnost, prav tako pa se pojavlja šum v ozadju in okluzija objektov [60]. Za zelo dobre metode so se pri tej podatkovni bazi izkazale tiste, ki iščejo globalno informacijo v sliki. To vodi v problem slabe generalizacije [60]. Modeli se lahko naučijo, da so avtomobili povezani s cesto in potem ne delujejo na primerih kjer je avto na polju. Na tak način ne vemo, ali algoritmi razpoznavajo objekt ali ozadje [60].

V delu [63] so se prav tako osredotočali na podatkovno bazo Caltecch 101

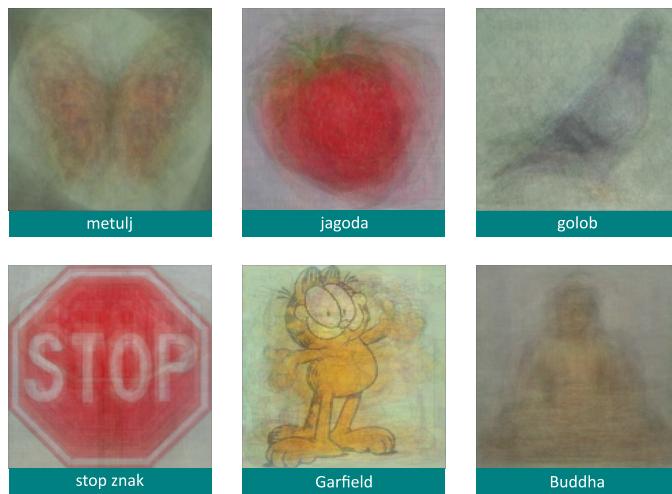


Figure 1.27

[62]. Preverili so delovanje sistema V1. Sistem V1, temelji na Gaborjevih filterih in predstavlja prvo procesno enoto v vidnem sistemu primatov [63]. Z zmogljivostjo 67 % je dokaj primitiven algoritem presegel večino do tedaj najboljših algoritmov. S tem so dokazali, da testiranje na tovrstnih naravnih slikah ni dovolj dobro. Predlagali so, da objektivno evaluacijo težavnosti nalog razpoznavanja preverjamo z t.i. ničtimi modeli. Podatkovne baze bi morale biti zelo velike in nepristranske, da bi zaobjeli čimvečjo populacijo. Zaradi težavnosti pridobivanja takih podatkovnih baz so predlagali uporabo sintetičnih slik.

S težavami modernejših podatkovnih baz so se ukvarjali v [61]. S preprostim poskusom so najprej pokazali, da razvrščevalnik, ki razpoznavata podatkovne baze deluje razmeroma dobro (39 %). Z opazovanjem konfuzijske matrike, so ugotovili, da obstaja združevanje podatkovnih baz na tiste, ki se osredotočajo na prizore in tiste, ki se osredotočajo na objekte. Kljub majhnemu vzorcu je bila močno izražena diagonalna, kar pomeni, da imajo podatkovne unikatne značilnosti [61].

Nadalje so uvedli nov način testiranja zmogljivosti algoritmov z generalizacijo med bazami. S križnim testom so učili na eni bazi in testirali na drugi. Ker podatkovne baze predstavljajo vzorec naše realnosti, bi moral biti test za algoritme enostaven [61]. Kot so pokazali, je pri testiranju z drugo podatkovno bazo zmogljivost algoritmov močno padla. Tako je na primeru razvrščanja avtomobilov padla iz 53.4 % na 27.5 % [61]. Argumentirali so, da to nastane zaradi naslednjih razlogov [61]:

1. *Pristranskost izbire:* Podatkovne baze preferirajo slike s specifičnimi lastnostmi.
2. *Pristranskost vzorčenja:* Ljudje fotografiramo objekte na podoben

način.

3. *Pristranskost kategorij*: Semantične kategorije so slabo definirane in isti tipi objektov imajo lahko različne labele.
4. *Pristranskost negativnega vzorca*: Representativnost negativnih primerov slik je slaba.

Še posebej so poudarili, da ima največji vpliv *pristranskost negativnega vzorca*. Če recimo želimo najti vse slike z avtomobili, kako vemo, da razvrščevalnik najde avtomobile in ne ceste, ki je s prevoznimi sredstvi močno korelirana [61]? Tu pride prav negativni vzorec, v katerega damo primere cestič in tako prisilimo algoritme v pravilno delovanje. S preizkusom na negativni množici primerov, združeni iz vseh baz, so v [61] pokazali, da so negativni primeri iz različnih baz povezanimi s pozitivnimi primeri testne podatkovne baze (slaba reprezentativnost negativnih primerov).

Pomislili bi lahko, da bi se z združevanjem podatkovnih baz rešili problematike pristranskosti. Vendar pa, kot so poudarili v [61], če želimo podatkovno bazo izboljšati, bi morali močno povečati učno množico, da bi bila ta signifikantna. Tako bi recimo za izboljšanje detektorja avtomobila na 1250 primerov PASCAL VOC 2007 morali dodati 50000 LabelMe vzorcev [61]. Sama vrednost vrednost podatkovnih baz za delovanje algoritmov v resničnem svetu je tako po mnenju Torralba *et al.* [61]: “boljša kot nič, vendar ne velika”.

Problematika kako lahko uporabimo znane podatke za generalizacijo na še nepoznanih podatkih je v strojnem učenju poimenovana kot premostitev domen (angl. Domain Shift) [64]. Domena je množica podatkov z lastno distribucijo glede na izbrane labele [64]. Domene so tako lahko podatkovne baze ali resnični svet. Logično je, da algoritem lahko deluje v drugi domeni (nepoznana množica) samo v primeru, če dobro deluje v svoji bazični domeni (poznana množica podatkov) [64]. Ker aspekta pri križnem testiranju med podatkovnimi bazami v [61] niso upoštevali, so zato v [64] predstavili novo metriko križnega testiranja generalizacije modelov. V [61] so uporabili le relativno metriko (procentualni upadec zmogljivosti), v [64] pa so obravnavali delovanje tudi na učni množici z enačbo (1.2), kjer je  $s$  zmogljivost znotraj podatkovne baze in  $o$  zmogljivost med podatkovnimi bazami. Vrednosti metrike se nahajajo na intervalu  $[0, 1]$ . Če je  $CD$  večja od 0.5 nakazuje, da obstaja pristranskost. Vrednosti pod 0.5 pa govorijo o tem, da je  $o \geq s$  ali pa so rezultati  $s$  zelo nizki.

$$CD = \frac{1}{1 + \exp\{-\frac{s-o}{100}\}} \quad (1.2)$$

Pri analizi pristranskosti ob uporabi DeCAF značilk so v [64] dobili rezultate zmogljivosti večrazrednega razvrščanja, katerih povprečja so povzeta v tabeli 1.5. Vidimo lahko, da metrika  $CD$  nakazuje na to, da imajo vse

podatkovne baze približno enako pristranskost ne glede na izbiro značilke BOWSift ali DeCAF7. Po drugi strani pa z metriko upada vidimo večje razlike med podatkovnimi bazami.

Podatkovna baza	BOWSift		DeCAF7	
	Upad [%]	CD	Upad [%]	CD
Caltech 256	51.50	0.53	47.90	0.58
ImageNet	34.00	0.52	33.20	0.55
SUN	42.10	0.51	25.90	0.52

Table 1.5

# Bibliography

- [1] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, and P. Corke, “The limits and potentials of deep learning for robotics,” *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 405–420, 2018.
- [2] B. RichardWebster, S. E. Anthony, and W. J. Scheirer, “PsyPhy: A Psychophysics Driven Evaluation Framework for Visual Recognition,” 2017, pp. 1–9. arXiv: 1611.06448.
- [3] J. Deng, W. Dong, R. Socher, K. Li, Li-Jia and Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255, 2009.
- [4] S. Dodge and L. Karam, “A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions,” in *2017 26th International Conference on Computer Communication and Networks (ICCCN)*,, 2017, pp. 1–7.
- [5] ——, “Can the Early Human Visual System Compete with Deep Neural Networks?” In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2798–2804, ISBN: VO -.
- [6] R. Geirhos, D. H. J. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, “Comparing deep neural networks against humans: object recognition when the signal gets weaker,” 2017. arXiv: 1706.06969.
- [7] ——, *Data and materials from "Comparing deep neural networks against humans: object recognition when the signal gets weaker"*, 2017. [Online]. Available: <https://github.com/rgeirhos/object-recognition> (visited on 05/08/2018).
- [8] F. A. Wichmann, D. H. J. Janssen, R. Geirhos, G. Aguilar, H. H. Schütt, M. Maertens, and M. Bethge, “Methods and measurements to compare men against machines,” *Electronic Imaging*, vol. 2017, no. 14, pp. 36–45, 2017.

- 
- [9] S. R. Kheradpisheh, M. Ghodrati, M. Ganjtabesh, and T. Masquelier, “Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition,” *Scientific Reports*, vol. 6, pp. 1–24, 2016.
  - [10] J. Xiao, J. Hays, K. A. Ehinger, and A. Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” in *2010 IEEE conference on Computer vision and pattern recognition (CVPR)*, IEEE, 2010, pp. 3485–3492.
  - [11] R. C. O'Reilly, D. Wyatte, S. Herd, B. Mingus, and D. J. Jilk, “Recurrent processing during object recognition,” *Frontiers in psychology*, vol. 4, p. 124, 2013.
  - [12] S. R. Kheradpisheh, M. Ghodrati, M. Ganjtabesh, and T. Masquelier, “Humans and Deep Networks Largely Agree on Which Kinds of Variation Make Object Recognition Harder,” *Frontiers in Computational Neuroscience*, vol. 10, no. August, p. 92, 2016.
  - [13] A. Borji and L. Itti, “Human vs. computer in scene and object recognition,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2014, pp. 113–120.
  - [14] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” p. 20, 2007.
  - [15] F. Fleuret, T. Li, C. Dubout, E. K. Wampler, S. Yantis, and D. Geman, “Comparing machines and humans on a visual categorization test,” in *Proceedings of the National Academy of Sciences*, vol. 108, National Acad Sciences, 2011, pp. 17 621–17 625.
  - [16] S. Stabinger, A. Rodriguez-Sanchez, and J. Piater, “25 years of CNNs: Can we compare to human abstraction capabilities?” In *Artificial Neural Networks and Machine Learning – ICANN 2016*, Springer, 2016, pp. 380–387.
  - [17] R. T. Pramod and S. P. Arun, “Do Computational Models Differ Systematically from Human Object Perception?” In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1601–1609.
  - [18] D. Linsley, S. Eberhardt, T. Sharma, P. Gupta, and T. Serre, “What are the visual features underlying human versus machine vision?” In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2706–2714.
  - [19] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, “Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?” *Computer Vision and Image Understanding*, vol. 163, no. September 2016, pp. 90–100, 2017.

- [20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *2nd International Conference on Learning Representations (ICLR)*, 2014, pp. 1–10.
- [21] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” in *3th International Conference on Learning Representations (ICLR)*, 2015, pp. 1–11.
- [22] A. Nguyen, J. Yosinski, and J. Clune, “Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 427–436.
- [23] ———, *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Imagesle*, 2015. [Online]. Available: <http://www.evolvingai.org/fooling> (visited on 05/07/2018).
- [24] G. F. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein, “Adversarial Examples that Fool both Human and Computer Vision,” 2018. arXiv: 1802.08195.
- [25] Y. Jia, J. Abbott, J. Austerweil, T. Griffiths, and T. Darrell, “Visual Concept Learning: Combining Machine Vision and Bayesian Generalization on Concept Hierarchies,” in *Advances in Neural Information Processing Systems*, 2013, pp. 1842–1850.
- [26] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, “Human-level concept learning through probabilistic program induction,” *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.
- [27] W. J. Scheirer, S. E. Anthony, K. Nakayama, and D. D. Cox, “Perceptual Annotation: Measuring Human Vision to Improve Computer Vision,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1679–1686, 2014.
- [28] V. Jain and E. Learned-Miller, “Fddb: A benchmark for face detection in unconstrained settings,” University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.
- [29] R. C. Fong, W. J. Scheirer, and D. D. Cox, “Using Human Brain Activity to Guide Machine Learning,” *Scientific Reports*, pp. 1–10, 2017.
- [30] Kelley, *fMRI one*, 2009. [Online]. Available: <https://www.flickr.com/photos/twitchcraft/3223196785/in/photolist-9hiPC-5GhxMn-6hYaXa-5UPJ7n-9yLTZw-c27Be9-c27vR7-5Vi6z4-5Vnsx3-dYocn5-5VnsDf-7kHkjn-9yHNe2-62TpLD-mxeWUb-4uNsCX-yBNSqL-Fuz2cu> (visited on 05/09/2018).

- [31] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie, “Visual Recognition with Humans in the Loop,” in *Computer Vision — ECCV 2010*, Springer, 2010, pp. 438–451.
- [32] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, “Caltech-UCSD Birds 200,” California Institute of Technology, Tech. Rep. CNS-TR-2010-001, 2010.
- [33] R. Mottaghi, S. Fidler, J. Yao, R. Urtasun, and D. Parikh, “Analyzing semantic segmentation using hybrid human-machine CRFS,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2013, pp. 3143–3150.
- [34] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “TextonBoost: Joint Appearance, Shape and Context Modeling for Multit-Class Object Recognition and Segmentation,” in *European conference on computer vision*, Springer, 2006, pp. 1–15.
- [35] R. Mottaghi, S. Fidler, A. Yuille, R. Urtasun, and D. Parikh, “Human-Machine CRFs for Identifying Bottlenecks in Scene Understanding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 74–87, 2016.
- [36] A. J. O’Toole, H. Abdi, F. Jiang, and P. J. Phillips, “Fusing Face-Verification Algorithms and Humans,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 5, pp. 1149–1155, 2007.
- [37] C. Vondrick, H. Pirsiavash, A. Oliva, and A. Torralba, “Learning visual biases from human imagination,” *Advances in neural information processing systems*, pp. 289–297, 2015.
- [38] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, and J. J. DiCarlo, “Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks,” *bioRxiv*, 2018.
- [39] M. J. Tarr and H. H. Bülthoff, “Image-based object recognition in man, monkey and machine,” *Cognition*, vol. 67, pp. 1–20, 1998.
- [40] D. D. Leeds, D. A. Seibert, J. A. Pyles, and M. J. Tarr, “Comparing visual representations across human fMRI and computational vision,” *Journal of Vision*, vol. 13, no. 13, pp. 25–25, 2013.
- [41] T. W. Mills, *Brain Drawing*, 1893. [Online]. Available: [https://commons.wikimedia.org/wiki/File:Brain\\_Drawing.svg](https://commons.wikimedia.org/wiki/File:Brain_Drawing.svg) (visited on 05/15/2018).
- [42] D. L. Yamins, H. Hong, C. Cadieu, and J. J. DiCarlo, “Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream,” *Advances in Neural Information Processing Systems*, pp. 3093–3101, 2013.

- 
- [43] Ananth, *Image*. [Online]. Available: <https://thenounproject.com/icon/922827> (visited on 05/15/2018).
  - [44] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, “Performance-optimized hierarchical models predict neural responses in higher visual cortex,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014, ISSN: 0027-8424.
  - [45] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, “Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition,” *PLoS Computational Biology*, vol. 10, no. 12, 2014.
  - [46] R. Rajalingham, K. Schmidt, and J. J. DiCarlo, “Comparison of Object Recognition Behavior in Human and Monkey,” *Journal of Neuroscience*, vol. 35, no. 35, pp. 12 127–12 136, 2015.
  - [47] L. Riffault, *Monkey*. [Online]. Available: <https://thenounproject.com/icon/352929> (visited on 05/15/2018).
  - [48] A. Skowalsky, *Human*. [Online]. Available: <https://thenounproject.com/icon/755682> (visited on 05/15/2018).
  - [49] I. Papa, *Computer*. [Online]. Available: <https://thenounproject.com/icon/1720982> (visited on 05/15/2018).
  - [50] A. J. O’Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Penard, H. H. Abdi, N. Pénard, and H. H. Abdi, “Face recognition algorithms surpass humans matching faces over changes in illumination,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 9, pp. 1642–1646, 2007.
  - [51] A. J. O’Toole, P. J. Phillips, and A. Narvekar, “Humans versus algorithms: Comparisons from the face recognition vendor test 2006,” in *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008*, IEEE, 2008, pp. 1–6.
  - [52] A. J. O’Toole, X. An, J. Dunlop, V. Natu, P. J. Phillips, and Unknown, “Comparing face recognition algorithms to humans on challenging tasks,” *ACM Transactions on Applied Perception (TAP)*, vol. 9, no. 4, p. 16, 2012.
  - [53] L. Best-Rowden, S. Bisht, J. C. Klontz, and A. K. Jain, “Unconstrained Face Recognition: Establishing Baseline Human Performance via Crowdsourcing,” in *2014 IEEE International Joint Conference on Biometrics (IJCB)*, 2014, pp. 1–8.
  - [54] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, “Describable Visual Attributes for Face Verification and Image Search,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, 2011.

- 
- [55] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the gap to human-level performance in face verification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708. [Online]. Available: [https://www.cs.toronto.edu/%7B~%7Dranzato/publications/taigman%7B%5C\\_%7Dcvpr14.pdf](https://www.cs.toronto.edu/%7B~%7Dranzato/publications/taigman%7B%5C_%7Dcvpr14.pdf).
  - [56] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.
  - [57] P. J. Phillips and A. J. O’Toole, “Comparison of human and computer performance across face recognition experiments,” *Image and Vision Computing*, vol. 32, no. 1, pp. 74–85, 2014.
  - [58] P. J. Phillips, M. Q. Hill, J. A. Swindle, and A. J. O’Toole, “Human and algorithm performance on the PaSC face Recognition Challenge,” in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems, BTAS 2015*, IEEE, 2015, pp. 1–8.
  - [59] B. RichardWebster, S. Y. Kwon, C. Clarizio, S. E. Anthony, and W. J. Scheirer, “Visual Psychophysics for Making Face Recognition Algorithms More Explainable,” pp. 1–20, 2018.
  - [60] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman, “Dataset Issues in Object Recognition,” *Toward Category-Level Object Recognition*, pp. 29–48, 2006.
  - [61] A. Torralba and A. A. Efros, “Unbiased Look at Dataset Bias,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE, 2011, pp. 1521–1528, ISBN: 1457703955.
  - [62] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” *Computer vision and Image understanding*, vol. 106, no. 1, pp. 59–70, 2007.
  - [63] N. Pinto, D. D. Cox, J. J. DiCarlo, N. Pinto, and D. D. Cox, “Why is real-world visual object recognition hard?” *PLoS Computational Biology*, vol. 4, no. 1, p. 27, 2008.
  - [64] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, “A Deeper Look at Dataset Bias,” in *Domain Adaptation in Computer Vision Applications*, Springer, 2017, pp. 37–55.