EKAIA (2024), artikulua prentsan/article in press
https://doi.org/10.1387/ekaia.26293
Behin-behineko bertsioa (euskara-orrazketaren faltan).

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN: 0214-9001
e-ISSN: 2444-3255

Arreta partekatzen Pepper robot sozialarekin

(Shared attention with Pepper)

Miren Samaniego, Igor Rodriguez, Elena Lazkano*
RSAIT ikerketa taldea, Informatika Fakultatea (UPV/EHU), Manuel Lardizabal 1, 20018 Donostia

LABURPENA: Robot sozialek arreta soziala erakutsi behar dute eta, horretarako, ezinbestekoa da giza-ki(ar)ekin elkarrekintzan ari denean intereseko objektua detektatzeko gaitasuna. Lan honetan, pertsonak begiradaren bidez fokatzen duen objektua identifikatu eta lokalizatzeko sistema garatu da eta baita Pepper robotean integratu ere. YOLOv8 erabili da irudi bidez pertsona detektatu eta bere aurpegia lokalizatzeko. Begiradaren fokua lortzeko VGG16 neurona sare konboluzionala entrenatu da, erregresio moduan, emaitza onargarriak erakutsiz. Eredu honen erantzunak bero mapa batean bihurtu eta sakoneraren informazioarekin konbinatuz, ResNet101 sareak emango digu azken erantzuna: begiradaren gunea non den eta zer den. Portaera hau robotera eraman da eta arreta konpartitua erakusteko robotak objektu berari begiratzeko mugimendua gauzatzen du erabiltzaileari ahoz modu egokian erantzuteko. Era honetan, erabiltzaileari arreta partekatuaren sentsazioa transmititzen zaio. Bideo batean erakusten da portaera orokorraren emaitza.
HITZ GAKOAK: Arreta partekatua, Begiradaren estimazioa, Robot sozialak, Ikusmen Artifiziala, Neurona-sare sakonak.

ABSTRACT: *For social robots to show social attention it is essential to be able to detect the object of interest, if any, during human-robot interactions. This work shows an attempt to locate and identify the object in the human's visual focus of attention and integrate the system into the social robot Pepper. YOLOv8 was used to detect the person through the image and extract their face. In order to get the Point of Regard, a VGG16 convolutional network has been adapted and trained for the specific regression task, showing acceptable results. The orientation of the Point of Regard with respect to the head is used to obtain a headmap that allows to extract the object of interest (and its identity) among the ones obtained by the ResNet101, using depth segmentation and the intersection of union strategy. This system has been transferred to the robot that develops a shared attention behavior by performing the proper head motion accompanied with a verbal response that makes the human aware of the situation. A video shows the result of the general behavior*
KEYWORDS: Shared attention, gaze estimation, social robots, computer vision, deep neural networks

***Harremanetan jartzeko/Corresponding author:** Miren Samaniego, Informatika Fakultatea (UPV/EHU), Manuel Lardizabal 1, 20018 Donostia. msamaniego005@ikasle.ehu.eus

Nola aipatu/How to cite: 1. Samaniego, Miren; Rodríguez, Igor; Lazkano, Elena (2024). «Arreta partekatzen Pepper robot sozialarekin », Ekaia, DOI: https://doi.org/10.1387/ekaia.26293

Jasoa: maiatzak 6, 2024; Onartua: uztailak 4, 2024
ISSN 0214-9001-eISSN 2444-3225 / ©2020 UPV/EHU

1. Sarrera

Robotika sozialak gurekin elkarbizitza partekatzeko robotak egitea du helburu. Etorkizun hurbilean, adinekoei eta besteoi lagunartea eta laguntza eskaintzen [1], edota tresna terapeutiko gisa – errehabilitazioan [2] edo autismoaren sindromearen tratamenduetan [3] esaterako – irudikatzen ditugu. Sozialki gure antzekoak diren makinak dira robot sozialak, ez horrenbeste itxura aldetik nahiz eta antzekotasunak laguntzen duen, bai ordea portaera aldetik. Gizakiok badakigu gure artean nola jokatu eta robotekin berdin jokatu ahal izatea nahiko genuke. Hala ere, robotika sozialaren garapen egoera oraindik nahikoa gordina da eta Adimen Artifizialak asko lagun dezake aurrera egiten. Bide horretan, portaera sozialak garatzea ezinbestekoa da.

Pertsonen arteko komunikazioan hitzezko zein bestelako keinu eta mezuak erabiltzen ditugu. Arreta soziala (social attention) adimen sozialaren zimendua da, beste pertsonen keinu sozialei (aurpegi espresioak, jestuak) eta ahozko bokalizazioari erantzuten diona. Taldeko portaerak erazagutu eta aztertzeko informazio ezinbestekoa eskaintzen du, beraz. Arlo honek eremu desberdinak ditu, horietako bat da arreta partekatua edo baterakoa (shared eta join). Arreta bateratuaren testuinguruan, bi edo pertsona gehiagok objektu zein gertaera berean zentratzen dute interesa. Amankomuneko interes hori begirada bidez konpartitzen dute; hau da, elkarrekintzan parte hartzen duten banakako horiek bestearen begiradaren noranzkoan jartzen dute arreta, modu sinkronizatuan. Ez hori bakarrik, intereseko objektua detektatzeaz gain informazio sozial anitz eskuratzeko erabiltzen dugu begirada. Begirada ere ikusmira aldatzeko estrategia naturala da, asko egiten dugu informazioa falta dugunean. Baina batez ere gizakiak oso eraginkorrak gara beste gizakien begiradaren noranzkoa estimatzen, oso haurretatik gainera [4]. Aipatzekoa da abilezia honen ezak zailtasun handiak eragiten dituela pedagogiaren hainbat testuingurutan [5].

Robot sozialek beharrezkoa dute arau sozialak eta begiradarekin lotutako portaerak erakutsi eta espektatibak betetzea. Batetik, begiradak gizaki eta roboten arteko komunikazioa errazten du eta, bestetik, robotak begiradaren kontaktua areagotzeak elkarrekikotasuna sortarazten du [6]. Tamalez, prozesu hori sintetikoki erreproduzitzea arras konplikatua da gure ikusmena eta ikusmen artifiziala oso desberdinak direlako. Robotek ingurumena ulertzeko sentsoreetatik jasotako datuak aztertu behar dituzte eta ikusmena, kameren bitartez gauzatzen da batez ere, RGB zein sakonera kamerak gaur egun.

Hari honi helduz, Pepper robotak bere ikusmen sentsoreen bitartez aurrean duen pertsonaren[1] arreta-gunea detektatu eta burua leku berera orientatuko duen portaera inplementatzea da lan honen helburua, aurrerantzean portaera sozial aberatsagoak inplementatzeko mugarri gisa.

Ekarpenak hauek dira:

- RGB zein sakonera irudiak erabiliz, arreta-gunea detektatzeko sistema garatzea.
- Identifikazio-sistema Pepper robotean integratu eta robotaren mugimendua koordinatzen duen kontrola garatzea, sistemak proposatzen duen eskualdera begiratzeko.
- Aurrekoak elkarrizketa-sistemarekin konbinatzea, erabiltzaileari arreta-gunea duen objektua ezagutzen duela jakinarazteko eta elkarrekintza naturalagoa gertatzeko.

2. Begiradaren atentzio gunean dagoen objektuaren identifikazio prozesua

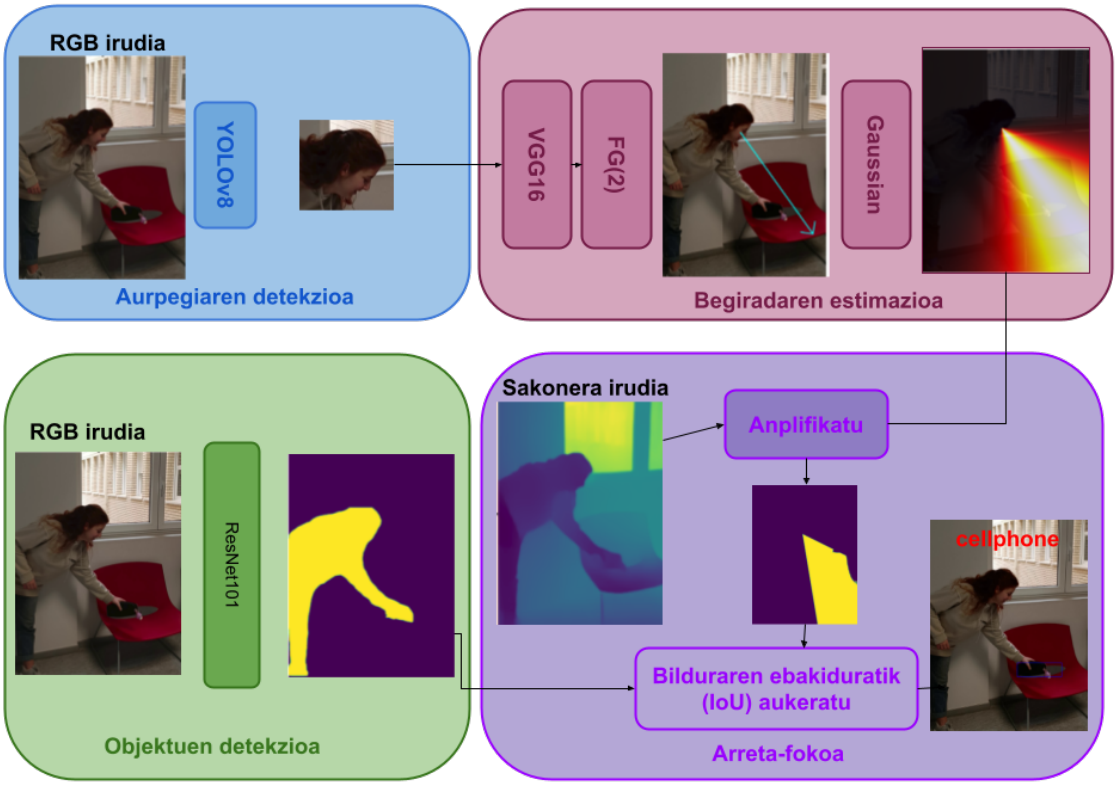
Erabiltzailearen begiradaren fokoan dagoen objektuaren identifikazio prozesuak bi fase nabarmen ditu:

[1] Robot anitzen aukera alde batera utzita, robot-gizaki arteko elkarrekintzan eszenario desberdinak topa daitezke. Pertsona bakarra aritu daiteke robotarekin, banan banakako elkarrekintzan, edo pertsona batek baino gehiago har dezake parte aldiberean. Lan honetan azaldutakoa lehenengo konfiguraziora mugatzen da nahiz eta hainbat kontzeptu eta metodo estrapola daitezkeen bestera.

1. Nora dago begira “irudiaren” ikusmen-eremuan dagoen pertsona? hau da, non dago arreta-gunea?

2. Zer dago arreta-gune horretan? zerk sortu du pertsonaren interesa?

1 irudiak laburbiltzen du atentzio-gunean dagoen objektua detektatzeko sistemaren egitura eta exekuzio sekuentzia, sarrerako eszenatik hasi eta intereseko objektura heldu arte. Goiko zatiak lehenengo galderari erantzuten dio eta, behekoak, aldiz, bigarrenari.



1. irudia: Arreta-gunearen identifikazioaren prozesu-katea

Jarraian, bi urrats nagusi hauen xehetasunak deskribatuko dira.

2.1. Arreta-gunea non da?

Begirada estimatzeko bi aukera agertzen dira literaturan. Batetik, begietan oinarritzen diren metodoak ditugu. Estimazio hori begien ezaugarriak begiaren eredu geometrikoarekin konbinatuz (niniaren zentroa, kornearen isla eta horrelakoak) egiten dute. Begiaren ezaugarri fisikoen datu zehatzak behar dituzte, zehaztasun handia eskatzen dute eta kalibrazioa egin behar da pertsona aldatzean [7]. Errealitate birtual eta areagotuko gailuetan erabiltzen dira hauek, [8] lanean esaterako. Begien ikusmira zehatza behar da hauetan, ez da nahikoa pertsona eta bere ingurumenaren irudi arrunta era horretan begirada estimatzeko eta, beraz, metodo hauek robot sozialen gaitasunetik kanpo daude. Alternatiba gisa, buruaren posizioa erabil daiteke. Gutxi gora beherako metodoak dira hauek, hurbilpenak baino ez, begi eta buruaren arteko orientazioan $\pm 35°$-ko desbiderapena egon daiteke eta [4].

Guk burua erabiliko dugu erreferentzia gisa, bereizmen baxuko irudiak erabili daitezkeelako eta denbora errealean funtzionatzeko ezinbestekoa dugulako. Iruditik abiatuta, burua erabiliz arreta-gunea estimatzeko prozesuak hiru fase ditu (ikus 2 irudia): aurpegiaren kokapena atera, begiradaren noranzkoa (Point of Regard, PoR) estimatu eta bero-mapa lortu.

(a) Aurpegiaren kokapena (b) Begirada-gunea (c) Bero-mapa

2. irudia

Horrela bada, RGB iruditik abiatuz, bertan pertsonarik dagoen aztertu eta, baiezkoan, aurpegia detektatzen du YOLOv8 neurona-sareak [9], nahi izanez gero baita begi, sudur eta ahoaren posizioak, aurpegiaren markak, alegia. Baina guk nahikoa dugu aurpegia barneratzen duen muga-kutxarekin.

Bestalde, PoR bi ikuspuntutik landu daiteke: sailkapena eta erregresioa. Sailkapenak begiradaren eskualdeak bereizten dituen arren, gure helburuarentzat egokiago iruditu zaigu gizakiaren arreta zentratzen duen objektuan dagoen punturen bat lortzea. Gure problema, beraz, erregresioa da. Begiradaren estimazioaz zerbait haratago da helburua, begirada-puntua edo begiradaren fokuaren bila gabiltza (ikus 2 irudia).

Ataza hau ebazteko, VGG16 [10] neurona-sare sakona aukeratu dugu, 16 geruza dituen konboluzio-sarea. Konboluzio-sareak (CNN) irudietatik ezaugarriak erauzteko eta sailkapenerako oso egokiak dira. Gure kasuan, sare originalaren irteerako geruza, burua deritzona, modifikatu egin da erregresioa egin dezan eta ez sailkapena. VGG16 sarearen burua leundu eta aktibazio linealeko geruza bati lotzen zaio guztiz konektatuta, begirada-gunean dagoen (x, y) koordenatu bat inferitzeko.

Sarea egokitzeaz gain, datu-base berezia sortu behar izan dugu, GazeFollow [11] datu-basetik abiatuta. Datu base honek 122,143 irudi ditu, 130,339 (bakar zein anitz) pertsona agerraldirekin. Irudi horiek prozesatu eta $y_{true} = (px, py)$ irteerak etiketatu ditugu. Dataset-a hiru multzotan banatu da, entrenamendua 80,000 instantziarekin, balidazioa 25,000 instantziarekin eta gainerakoak, eredua probatzeko erabili dira. Sareak erabiliko duen errore funtzioa 1 ekuazioan adierazitakoa da, y_{true} egiazko etiketa eta y_{pred} sareak emandako irteera izanik.

$$f(y_{true}, y_{pred}) = 1 - \frac{y_{true}\, y_{pred}^{T}}{\sqrt{\Sigma y_{true}^{2} \Sigma y_{pred}^{2}}} \tag{1}$$

Entrenamenduan, galera-funtzioa bakarrik erabili beharrean, ereduaren errendimendua neurtzeko batezbesteko errore absolutua (Mean Absolute Error, MAE) eta zehaztasuna erabili dira, bestelako atzeraelikadura emateko entrenamenduari.

Aurpegiaren zentroa eta sarearen (x,y) irteera lotzen dituen zuzenaz abiatuz, $\pm\sigma$ angelu tartean dauden pixelak barneratzen dituen eskualdearekin begiradaren bero-mapa sortzen dugu. Pixel baten berotasun maila kalkulatzeko, pixela eta buruaren zentroideak lotzen dituen zuzena eta erreferentziazko zuzenaren arteko angelua erabiltzen da (θ_i). Zuzenetik aldentzen diren pixelei bero-maila txikiagoa ematen zaie banaketa Gaussiarrai jarraituz, 2 ekuazioan adierazi bezala. Bertan azaltzen den σ baliotzat aurretik aipatutako irekiera angelua hartu dugu. 2c irudian ikus daiteke bero-maparen adierazpena. Kolore argiek bero-maila handiagoa adierazten dute, ilunek, aldiz,

txikiagoa.

$$P(\theta_i) \propto \frac{1}{\sigma} e^{-\frac{\theta_i^2}{2\sigma^2}} \tag{2}$$

Bi pertsona baleude, bi bero-mapen batura egiten da eta hortaz, kasu horretan bi bero-mapen ebakidura, gehiengoak konpartitzen duen begiradaren zona alegia, izango da zona beroena.

2.1.1. Emaitzak

1 taulan ikus daiteke erregresiorako VGG16 ereduaren entrenamenduaren zein balidazioaren galerak ez direla oso desberdinak, baita errore absolutuaren batezbestekoak (MAE) ere. 3a eta 3b irudiek egiaztatzen dute galerak eta MAE-k antzeko eboluzioa erakusten dutela bi faseetan, gainegokitzapen eta azpiegokitzapen eza berretsiz.

Esan, ereduaren entrenamendua 50 epoka egiteko diseinatu dela hasiera batean, baina etete goiztiarrak (*early stopping*) 11. epokan eten duela ikasketa. Eredu hori izan da amaierako sisteman erabili dena.

	Entrenamendua			Balidazioa		
Eredua	Galera	MAE	Zehaztasuna	Galera	MAE	Zehaztasuna
VGG16 + Erregresio burua	0.0193	0.1581	0.6526	0.0245	0.1840	0.6364

1. taula: VGG16 ereduaren entrenamenduko emaitzak

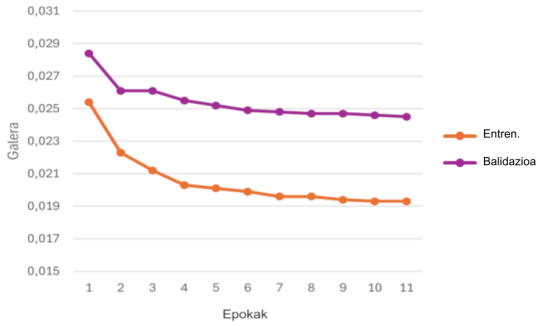
Aitzitik, entrenamenduan eta balidazioan ereduak eboluzio gorakorra eta konbergentzia erakutsi arren (3c irudia), zehaztasun aldetik lortutako emaitzak onargarriak baino ez dira izan. 1 taulan erakusten da entrenamendu eta balidazioan %65 inguruko zehaztasuna eman duela sareak. Emaitza hauek oso esanguratsuak ez diren arren, eskuartean dugun atazak ez du eskatzen sekulako doitasuna puntuaren zehaztasunean. Hurrengo fasean kalkulatzen den bero-mapak norabidearen kalkuluan gertatzen den ziurgabetasuna erlaxatuko du.

2.2. Zer dago gune bero horretan?

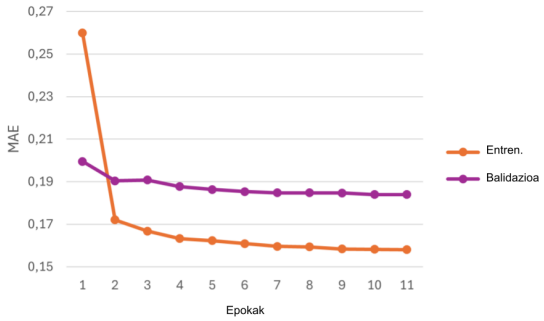
Bigarren urratsa, eskualde beroenean zer dagoen jakitea da, 3D espazioan intereseko objektua atzematea alegia. Hemen, sakonera irudiak hartzen du garrantzia. Bero mapa horretan sakonera berean dauden puntuek objektu bera adierazten dute. Horregatik, bero-mapa horri dagokion zonaldearen sakoneraren segmentazioa behar dugu. Bi urrats behar ditugu horretarako. Lehenengo, iada entrenatuta eta denbora errealean funtzionatzeko optimizatuta dagoen ResNet101 [12] sare batek, irudi originaletik abiatu eta bertan dauden objektuen maskarak eta identifikazioak itzultzen dizkigu. Bigarrenez, sakonera irudia bero maparekin gainjartzen da eskualde gertuenak kontuan hartzeko bakarrik; eskualde horiek eta objektuen maskarak edukita, bilduraren ebakidura (Intersect of Union, IoU) metrika erabilitzen da pertsonak ikusten duen objektua inferitzeko. Beste modu batean esanda, IoU balio handiena duena eratortzeko (ikus 4 irudia).

2.2.1. Emaitzak

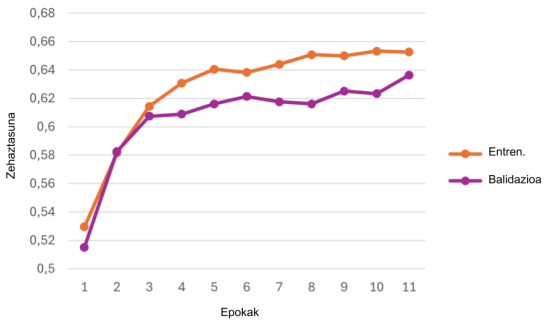
Intereseko objektua identifikatzeko sekuentzia osoa ebaluatu dugu *Video-CoAtt* izeneko bideoen datubasean [13]. Datu multzo hau arreta partekatua aztertzeko aproposa da, eszena sozial desberdinak barneratzen ditu, eta herrialde, genero, arraza, fiseonomia, janzkera eta kultura desberdinak estaltzen ditu. Telebista-saio zein filmetatik ateratako 360 bideo-sekuentziaz osatuta dago.



(a) Galera



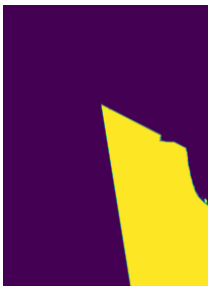
(b) MAE



(c) Zehaztasuna

3. irudia: Egokitutako VGG16 ereduaren ikasketa joerak denboran zehar

(a) Sakonera irudia



(b) Bilduraren ebakidura (IoU)

(c) ResNet101-en detekzioak

(d) Arreta-gunean dagoena: mugikorra

4. irudia: Azken fasea: intereseko objektuaren aukeraketa eta identifikazioa

Bideo sekuntziek 20 segundu zein minutu luzeko iraupena dute eta 25 fotograma/s maiztasuna. Guztira, 320×480 tamainako 492,100 irudi ditu.

Irudi bakoitza etiketatuta dago. Etiketak eszenan arreta partekaturik gertatzen den edo ez adierazten du; baiezkoan, komunikazioan parte hartzen duten gizakien buruak non dauden eta arreta horren elementuak barneratzen dituzten muga-kutxak. Ezkutuan edo estalita daudenak ez dira kontuan hartzen. Irudien kalitatea nahikoa kaskarra denez, internetetik HD formatuan hainbat bideo eskuratu dira eta datu-base berri bat osatu eta etiketatu da modu berean. Guztira, biak bilduz, 500 bat mila irudi erabili ditugu sarearen orokortzeko gaitasuna ebaluatzeko.

Ereduaren entrenamendu desberdinak probatu dira eraginkortasun handiagoa bilatzeko. Batetik, begiradaren bero-maparen irekiera angelua σ-ren balio desberdinak testatu dira bero-mapa zabaltzeak duen eragina neurtzeko. Bestetik, sakoneraren irudiaren beharra ebaluatu nahi izan dugu, konputazio zama alperrik ez handitzeko. 2 taulan lortutako hiru konbinazio hoberenak azaltzen dira, lehenengo biek ez dute sakonera irudirik erabiltzen, Resnet sarearen maskara eta bero-mapa konbinatzen dituzte zuzenean. Hirugarrenak, aldiz, sakonera gainjartzen du bero-mapan eta, ondoren, balio handieneko IoU elementua lortzen du. Ikus daiteken moduan, σ handitzeak helburu-objektua detektatzeko gaitasuna areagotzen du. Sakonera erabiltzeak ordea, zehatasuna zertxobait pobretzen du baina errore tasa dezente txikitzen du. Hau oso baliozkoa zaigu, egiazko objektua detektatzeko gaitasuna areagotzen delako, okerreko identifikazioak ekidinez.

Eredua	Zehaztasuna	Espezifikotasuna	Doitasuna	F1	Errorea
Segmentazio maskarak, $\sigma = 0.2$	0.937	0.81	0.915	0.9	0.72
Segmentazio maskarak, $\sigma = 0.4$	**0.975**	**0.92**	**0.964**	1	0.68
Sakonenaren segmentazio maskarak, $\sigma = 0.4$	0.962	0.88	0.947	1	**0.4**

2. taula: Gune beroenean dagoen objektuaren identifikaziorako eredu desbedinen emaitzak

Emaitza hauek guztiak aztertuta, azken eredua izango da Pepper robotera eramango duguna; emaitza okerragoak lortzen baditu ere zehaztasun, zehaztapen eta prezisio neurrietan, zarata gutxiagoko emaitzak itzultzen ditu, eta horrek du garrantzi handiagoa gure helburuan.

3. Detekzio sistema Pepper robotean integratzea

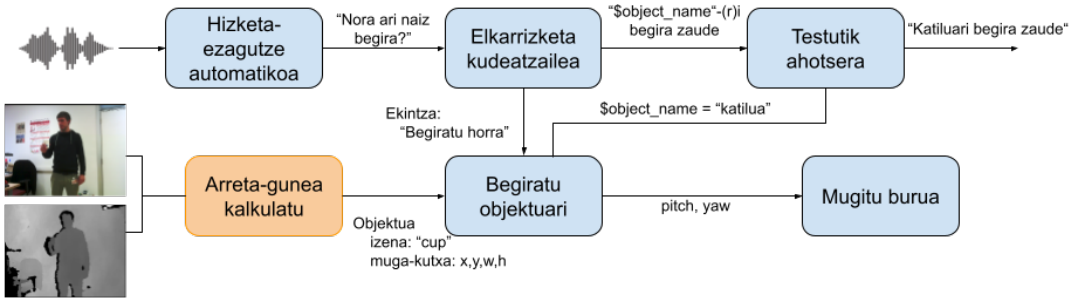
Pepper[2] aski ezaguna den eta giza itxura duen robot soziala da, 1.20m altuera duena. Kopetean kokatuta duen RGB-D kamerak, irudiaz gain, sakonera ere ematen du. Bi irudi horiek dira robotarekin elkarrekintzan dabilen pertsonaren begirada non kokatzen den eta bertan zer dagoen identifikatzeko muina. Kamerak hainbat bereizmenetan konfigura badaiteke ere, bi gailuetan 320x240 pixeleko bereizmena ezarri da; baxua beraz, denbora errealean aritzeko murrizpenari aurre egin ahal izateko.

Azpimarratzekoa da robotaren beraren kamera(k) erabiltzeak zailtasuna areagotzen duela. Batetik, eszena osoa harrapatzerik ez dago. Kanpo-kamara bat erabili beharko litzateke horretarako, Duffner eta Garcia-k egiten duten antzera [14]. Bestetik, robotaren mugimenduak perspektiba aldaketak ekartzen ditu, elkarrekintzan parte hartzen duen pertsona ikusmiratik kanpo ateratzeko arriskuarekin.

[2]https://www.aldebaran.com/es/pepper

Arreta partekatuan bi fase daude: hasieratzea eta erantzutea. Elkarrekintza aurrera eraman nahi duen pertsonak emango dio hasiera prozesuari, robotari zeri begira dagoen galdetuz. Erantzuna, berriz, robotak berak emango du, begirada identifikatu duen objektu horretara orientatuz eta ahoz identifikazioaren emaitza komunikatuz erabiltzaileari.

Arreta-gunea identifikatzeko prozesua dezente astuna denez konputazio aldetik (hainbat sare eta funtzio aplikatu behar dira, arestian azaldu den moduan), kanpoko prozesu-unitate bat erabili dugu (NVIDIA GeForce RTX 4080, 32GB RAM eta Intel Core i9-13900K). Robota eta kanpoko ordenagailuak ROS (Robot Operating System) [3] bidez komunikatuko dira, irudiak (RGB+sakonera) bidaliz lehenengoa, eta buruaren mugimendua eta bota behar duen audioa aginduz bigarrena. ROS arkitektura hau modulu desberdinez osatuta dago (ikus 5 irudia). Hiru funtzionalitate dira muin:



5. irudia: Sistema osoaren arkitektura

1. Aho bidez agintzen zaiona ulertu eta erantzun egokia sortzeko gaitasuna
2. Arreta-gunea detektatu eta bertan dagoen objektua identifikatu (posizioa eta izena), hau da, 2. atalean deskribatutako prozesua
3. Buruaren kontrola gauzatu: kalkulatu objektuari begiratzeko orientazioa eta exekutatu mugimendu hori
4. Erantzuna osatu: ahozkatu beharreko testuan txertatu objektuaren izena, eta audio moduan bota

Lehenengo fasean, Vosk[4] hizketa-ezagutze automatikorako tresnaren bidez audioa jaso eta testu bihurtzen da. Ondoren, elkarrizketa-kudeatzaile baten bidez eskaera jaso eta erantzunerako testu-txantiloia sortzen da. Txantiloi hau azken urratsean osatuko da, objektuaren etiketa ezagutzen denean.

Bigarren pausoan, 2. atalean deskribatutakoa gauzatzen da; RGB eta sakonera irudietatik abiatuz, arreta-gunearen objektuari dagokion posizioa, tamaina eta identitatea eskuratzen dira. Jarraian, "Objektuari begiratu"moduluak buruaren mugimendua kalkultatzen du. Horretarako, kameraren berezko parametroak kontuan izanda, irudiaren zentroarekiko objektuaren desplazamendua (angelu horizontala eta bertikalak) kalkulatzen da. Burua angelu horietara mugitzeko agindua emango zaio robotari.

Begirada objektura bideratu ondoren, berriro jatorrizko egoera neutrora itzuli eta erantzuna ahoz ematen zaio erabiltzaileari Nuance[5] sintetizadoreak duen TTS ("testutik ahotsera") moduluaren bitartez.

[3] www.ros.org
[4] https://alphacephei.com/vosk/
[5] https://www.nuance.com/es-es/omni-channel-customer-engagement/voice-and-ivr/text-to-speech/vocalizer.html

4. Eztabaida eta etorkizuneko lanak

Lan honetan literaturan punta-puntakoak diren sare mota desberdinak erabili dira arreta partekatzeko robot sozial batekin. YOLOv8 eta ResNet101 bere horretan erabili dira, egokitzaenerako entrenamendu gehigarririk egin gabe. VGG16 sarea, aldiz, erregresiorako entrenatu behar izan dugu gure atazara egokitzeko. Bideoan[6] ikus daitezke portaera globalaren nondik norakoak. Modulu guztiak integraturik sistema denbora errealean erantzuna emateko gai da. Funtzionamendu orokorra onargarria dela esan dezakegu. Nabarmendu behar dugu, erakutsitako esperimentuan pertsona bakar baten begirada aztertu arren, sistema gai dela pertsona baten baino gehiagoren arreta foku konpartitua estimatzeko.Badira, noski, zenbait koska.

Batetik, eta arestian aipatu bezala, robotak kopetean duen kamera erabiltzea eragozpena da. Burua mugitzerakoan pertsona ikusmen eremutik kanpo gelditzen da, horregatik erantzuna eman ahala berriro posizio neutrora eramaten dugu. Ez dugu pertsonaren jarraipena egiterik. Honetarako soluzioa bularrean Realsense D435[7] moduko RGBD kamera lotzea izango litzateke. Ikusmira mantentzeaz gain, mugimenduak berak sortzen dituen irudi lausoak ekidin daitezke.

Objektuen identifikaziorako erabili dugun ResNet101-ak objektu asko nahasten dituela egiaztatu dugu, irudi berriekin aritzeko zailtasunak dituela alegia. Esaterako, mailua, urruneko kontrola eta beste hainbat, telefono mugikor gisa etiketatzen ditu, baita katilua ardo kopa moduan ere. Aurpegien identifikaziorako erabili dugun Yolov8 sareak objektuak etiketatzeko aukera ere eskaintzen du eta etorkizun hurbilean ordeztu beharko genuke ResNet sarea.

Elkarrekintza orain amaitutzat ematen dugu objektua identifikatu denean. Berez, erabiltzaileak fokua aldatu bitartean, robotak ere interes bera erakutsi beharko luke, modu jarraituan. Horregatik, objektuaren jarraipena egin beharko luke Pepper-ek, arreta-gunean aldaketarik gertatzen ez den bitartean.

Azkenik, Pepperrek euskaraz ulertu eta hitzegitea da gure erronka premiazkoena, Elhuyar-en ADITU hizketa-ezagutzailea eta *ORAI NLP Teknologiak* enpresak eskaintzen duen TTS-a hizpide ditugu horretarako. Lehen mailako hezkuntza zein solasaldirako aplikazio baterako erroa badugu, ikusi-makusi bezalako joko didaktikoa inplementatzeko bloke guztiak ditugu eta. Garapen maila horretara iristean, populazio desberdinekin (ikasle eta irakasle adibidez) testatu beharko genuke sistema eta eraginkortasuna eta egokitasuna neurtu.

Erreferentziak

[1] Y. JUNG eta S. HAHN, 2023, «Social robots as companions for lonely hearts: The role of anthropomorphism and robot appearance», *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE.
URL `http://dx.doi.org/10.1109/RO-MAN57019.2023.10309617`

[2] L. R. DA COSTA, J. CASTRO, C. LINS, J. KELNER, M. LENCASTRE eta Ó. PASTOR, 2023, «On the use of social robots for rehabilitation: The case of nao physio», Á. ROCHA, C. FERRÁS eta W. IBARRA,*Information Technology and Systems*, 507–517, Springer International Publishing, Cham.

[3] R. VAGNETTI, A. D. NUOVO, M. MAZZA eta M. VALENTI, 2024, «Social robots: A promising tool to support people with autism. a systematic review of recent research and critical analysis from the clinical perspective», *Review Journal of Autism and Developmental Disorders*.

[4] B. MASSÉ, 2019, *Gaze direction in the context of social human-robot interaction*, PhD Thesis, Université Grenoble Alpes.

[6] `https://youtu.be/Wnh2qNDnVRs`
[7] https://www.intelrealsense.com/depth-camera-d435/

[5] P. MUNDY eta L. NEWELL, 2007, «Attention, joint attention, and social cognition», *Curr Dir Psychol Sci*, **16**(5), 269–274.

[6] T. L. XU, H. ZHANG eta C. YU, 2016, «See you see me», *ACM Transactions on Interactive Intelligent Systems (TiiS)*, **6**, 1 – 22.
URL https://api.semanticscholar.org/CorpusID:13358452

[7] A. A. AKINYELU eta P. BLIGNAUT, 2020, «Convolutional neural network-based methods for eye gaze estimation: A survey», *IEEE Access*, **8**, 142581–142605.

[8] A. PLOPSKI, T. HIRZLE, N. NOROUZI, L. QIAN, G. BRUDER eta T. LANGLOTZ, 2022, «The eye in extended reality: A survey on gaze interaction and eye tracking in head-worn extended reality», *ACM Computing Surveys (CSUR)*, **55**(3), 1–39.

[9] G. JOCHER, A. CHAURASIA eta J. QIU, «Ultralytics yolov8».
URL https://github.com/ultralytics/ultralytics

[10] K. SIMONYAN eta A. ZISSERMAN, 2015, «Very deep convolutional networks for large-scale image recognition», *3rd International Conference on Learning Representations*, 1–14, Computational and Biological Learning Society.

[11] A. RECASENS, A. KHOSLA, C. VONDRICK eta A. TORRALBA, 2015, «Where are they looking?», *Advances in Neural Information Processing Systems*, Bolumena 28, Curran Associates, Inc.

[12] K. HE, X. ZHANG, S. REN eta J. SUN, 2015, «Deep residual learning for image recognition», *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Orrialdeak 770–778.
URL https://api.semanticscholar.org/CorpusID:206594692

[13] L. FAN, Y. CHEN, P. WEI, W. WANG eta S.-C. ZHU, 2018, «Inferring shared attention in social scene videos», *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] S. DUFFNER eta C. GARCIA, 2016, «Visual focus of attention estimation with unsupervised incremental learning», *IEEE Transactions on Circuits and Systems for Video Technology*, **26**(12), 2264–2272.