Ekaia, ale berezia 2025, 153-166
https://doi.org/10.1387/ekaia.26293

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN 0214-9001 – eISSN 2444-3255

Arreta partekatzen Pepper robot sozialarekin

(Shared attention with Pepper)

Miren Samaniego, Igor Rodriguez, Elena Lazkano*

RSAIT ikerketa taldea, Informatika Fakultatea (UPV/EHU)

LABURPENA: Robot sozialek arreta soziala erakutsi behar dute eta, horretarako, ezinbestekoa da gizaki(ar)ekin elkarrekintzan ari direnean intereseko objektua detektatzeko gaitasuna izatea. Lan honetan, pertsonak begiradaren bidez enfokatzen duen objektua identifikatu eta lokalizatzeko sistema bat garatu dugu, baita Pepper robotean integratu ere. YOLOv8 erabili dugu irudi bidez pertsona detektatu eta haren aurpegia lokalizatzeko. Begiradaren fokua lortzeko VGG16 neurona-sare konboluzionala entrenatu dugu, erregresio moduan, eta emaitza onargarriak lortu ditugu. Eredu horren erantzunak bero-mapa batera bihurtu eta, sakoneraren informazioarekin konbinatuz, ResNet101 sareak emango digu azken erantzuna: begiradaren gunea non den eta zer den. Portaera hori robotera eraman dugu, eta robotak, arreta partekatua erakusteko, objektuari berari begiratzeko mugimendu bat egiten du erabiltzaileari ahoz modu egokian erantzuteko. Era horretan, erabiltzaileari arreta partekatuaren sentsazioa transmititzen zaio. Bideo batean erakutsi dugu 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 – https://orcid.org/

Nola aipatu / ***How to cite:*** Samaniego, Miren; Rodriguez, Igor; Lazkano, Elena (2025). «Arreta partekatzen Pepper robot sozialarekin». *Ekaia*, ale berezia 2025, 153-166. (https://doi.org/10.1387/ekaia.26293).

Jasotze-data: 2024, maiatzak 06; Onartze-data: 2024, uztailak 04.

ISSN 0214-9001 - eISSN 2444-3255 / © UPV/EHU Press 2025

Lan hau *Creative Commons Aitortu-EzKomertziala-PartekatuBerdin 4.0 Nazioartekoa* lizentzia baten mende dago

1. **SARRERA**

Robotika sozialak gurekin elkarbizitza partekatzeko robotak egitea du helburu. Etorkizun hurbilean, adinekoei eta besteoi lagunartea eta laguntza eskaintzen [1] irudikatzen ditugu, edota tresna terapeutiko gisa —errehabilitazioan [2] edo autismoaren sindromearen tratamenduetan [3], esaterako—. 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 garapenaren egoera nahikoa gordina da oraindik eta adimen artifizialak asko lagundu 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, eta beste pertsonen keinu sozialei (aurpegi-espresioak, jestuak) eta ahozko bokalizazioari erantzuten die. Taldeko portaerak erazagutu eta aztertzeko informazio ezinbestekoa eskaintzen du, beraz. Arlo horrek zenbait eremu ditu, eta haietako bat da arreta partekatua edo bateratua. Arreta bateratuaren testuinguruan, bi pertsonak edo gehiagok objektu edo gertaera berean zentratzen dute interesa. Amankomuneko interes hori begirada bidez partekatzen dute; hau da, elkarrekintzan parte hartzen duten banako horiek bestearen begiradaren noranzkoan jartzen dute arreta, modu sinkronizatuan. Ez hori bakarrik, intereseko objektua detektatzeaz gain, informazio sozial anitz eskuratzeko erabiltzen dugu begirada. Begirada ikusmira aldatzeko estrategia naturala ere bada; asko egiten dugu informazioa falta dugunean. Baina, batez ere, gizakiak oso eraginkorrak gara beste gizakien begiradaren noranzkoa estimatzeko orduan; oso haurretatik, gainera [4]. Aipatzekoa da abilezia ez horrek zailtasun handiak eragiten dituela pedagogiaren hainbat testuingurutan [5].

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

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

Ekarpenak hauek dira:

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

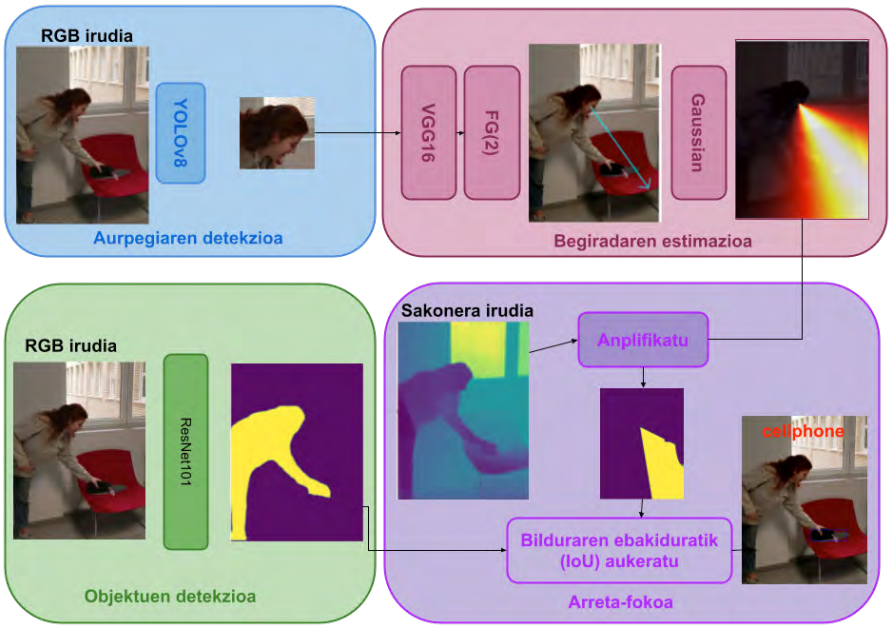
2. BEGIRADAREN ARRETA-GUNEAN DAGOEN OBJEKTUA IDENTIFIKATZEKO PROZESUA

Erabiltzailearen begiradaren fokuan dagoen objektua identifikatzeko prozesuak bi fase nabarmen ditu:

1. Nora begira dago irudiaren «ikusmen-eremuan» dagoen pertsona? Hau da, non dago arreta-gunea?
2. Zer dago arreta-gune horretan? Zerk sortu du pertsonaren interesa?

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

[1] Robot anizkunen aukera alde batera utzita, roboten eta gizakien arteko elkarrekintzan hainbat eszenatoki topa daitezke. Pertsona bakarra aritu daiteke robotarekin, bana-banako elkarrekintzan, edo pertsona batek baino gehiagok har dezakete parte aldi berean. Lan honetan azaldutakoa lehenengo konfiguraziora mugatzen da, nahiz eta hainbat kontzeptu eta metodo estrapola daitezkeen bestera.



1. irudia. Arreta-gunea identifikatzeko prozesu-katea.

Jarraian, bi urrats nagusi horien xehetasunak deskribatuko ditugu.

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 da. Begiaren ezaugarri fisikoen datu zehatzak behar dira; zehaztasun handia eskatzen dute, eta kalibrazioa egin behar da pertsona aldatzean [7]. Errealitate birtual eta areagotuko gailuetan erabiltzen dira; [8] lanean esaterako. Begien ikusmira zehatza behar dute, hau da, ez da nahikoa pertsona baten eta haren ingurunearen irudi arrunta izatea era horretan begirada estimatzeko, eta, beraz, metodo hauek robot sozialen gaitasunetik kanpo daude. Alternatiba gisa, buruaren posizioa erabil daiteke. Gutxi gorabeherako metodoak dira halakoak, hurbilpenak baino ez, begien eta buruaren arteko orientazioan ±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 proze-

suak 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 ba ote dagoen dagoen aztertu, eta, baiezkoan, aurpegia detektatzen du YOLOv8 neurona-sareak [9], eta, nahi izanez gero begi, sudur eta ahoen posizioak —aurpegiaren markak, alegia— ere. Baina guk nahikoa dugu aurpegia barneratzen duen muga-kutxarekin.

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

Ataza hori ebazteko, VGG16 [10] neurona-sare sakona aukeratu dugu, 16 geruza dituen konboluzio-sarea. Konboluzio-sareak (CNN) oso egokiak dira irudietatik ezaugarriak erauzteko eta sailkatzeko. 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 berez bata sortu behar izan dugu, GazeFollow [11] datu- basetik abiatuta. Datu-base horrek 122.143 irudi ditu, eta 130.339 pertsona-agerraldi (bakar zein anizkun). Irudi horiek prozesatu eta $y_{\text{true}} = (px, py)$ irteerak etiketatu ditugu. Datu-basea hiru multzotan banatu dugu: entrenamendua, 80.000 instantziarekin; baliozkotzea,

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.

$$f\left(y_{true}, y_{pred}\right) = 1 \frac{y_{true} y_{pred}^{T}}{\sqrt{\sum y_{true}^{2} \sum y_{pred}^{2}}} \tag{1}$$

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

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

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

Bi pertsona baleude, bi bero-mapen batura eta, hortaz, kasu horretan bi bero-mapen ebakidura —gehiengoak partekatzen duen begiradaren aldea, alegia— izango da alde beroena.

2.1.1. *Emaitzak*

1. taulan ikus daitekeenez, erregresiorako VGG16 ereduaren entrenamenduaren zein baliozkotzearen galerak ez dira oso desberdinak, ezta errore absolutuaren batezbestekoak (MAE) ere. 3a eta 3b irudiek egiaztatzen dute galerak eta MAEk antzeko eboluzioa dutela bi faseetan, hau da, gainegokitzapen eta azpiegokitzapen eza berresten da.

Esan dezagun ereduaren entrenamendua 50 aro egiteko diseinatu genuela hasiera batean, etete goiztiarrak (*early stopping*), baina, 11. aroan eten du ikasketa. Eredu hori erabili dugu amaierako sisteman.

Dena den, entrenamenduan eta baliozkotzean ereduak eboluzio gorakorra eta konbergentzia erakutsi arren (3c irudia), zehaztasun aldetik lortutako emaitzak onargarriak baino ez dira izan.

1. taula. VGG16 ereduaren entrenamenduaren emaitzak

Eredua	Entrenamendua			Baliozkotzea		
	Galera	MAE	Zehaztasuna	Galera	MAE	Zehaztasuna
VGG16 + Erregresio burua	0,0193	0,1581	0,6526	0,0245	0,1840	0,6364

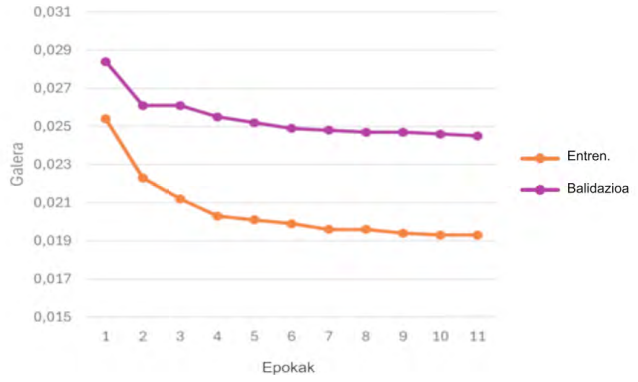
1. taulan erakutsi dugu entrenamenduan eta baliozkotzean % 65 inguruko zehaztasuna eman duela sareak. Emaitza horiek oso esanguratsuak ez diren arren, eskuartean dugun atazak ez du eskatzen sekulako doitasuna puntuaren zehaztasunean. Hurrengo fasean kalkulatuko dugun den bero-mapak norabidearen kalkuluan gertatzen den ziurgabetasuna erlaxatuko du.

2.2. **Zer dago gune bero horretan?**

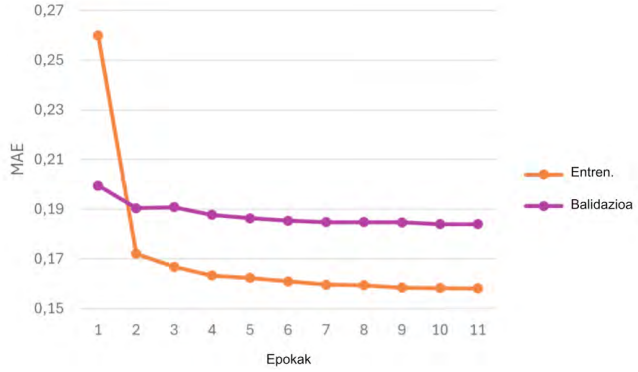
Bigarren urratsa, alde beroenean zer dagoen jakitea da; 3D espazioan intereseko objektua atzematea, alegia. Hemen, sakonera-irudiak hartzen du garrantzia. Bero-mapa horretan sakonera berean dauden puntuek objektu bera adierazten dute. Horregatik, bero-mapa horri dagokion aldearen sakoneraren segmentazioa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

2.2.1. *Emaitzak*

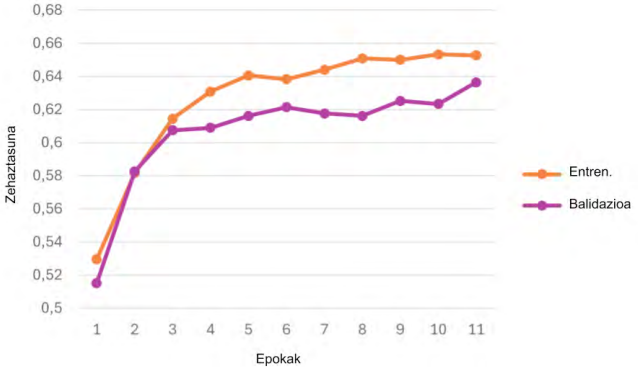
Intereseko objektua identifikatzeko sekuentzia osoa ebaluatu dugu, *Video-CoAtt* izeneko bideoen datubasean [13]. Datu multzo hori aproposa da arreta partekatua aztertzeko; hainbat es-zena sozial barneratzen ditu, eta hainbat herrialde, genero, arraza, fisonomia, janzkera eta kultur estaltzen. Telebista-saio zein filmetatik ateratako 360 bideo-sekuentziaz osatuta dago. Bideosekuentziek 20 segundo edo minutu luze bateko iraupena dute, eta 25 fotograma/s-ko maiztasuna. Guztira, 320 × 480 tamainako 492.100 irudi ditu.



(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 objektua aukeratzea eta identifikatzea.

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

2. taula. Gune beroenean dagoen objektua identifikatzeko ereduen emaitzak.

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

Ereduaren zenbait entrenamendu probatu ditugu, ahalik eta eraginkortasun handiena bilatzeko. Batetik, begiradaren bero-maparen σ irekiera-angeluaren zenbait balio testatu ditugu, bero-mapa zabaltzeak duen eragina neurtzeko. Bestetik, sakonera-irudiaren beharra ebaluatu nahi izan dugu, konputazio-zama alferrik ez handitzeko. 2. taulan lortutako hiru konbinazio hoberenak azaltzen dira: lehenengo biek ez dute sakonera-irudirik erabili, eta Resnet sarearen maskara eta bero-mapa konbinatzen dituzte zuze-

nean. Hirugarrenak, aldiz, sakonera gainjarri du bero-mapan, eta, ondoren, balio handieneko IoU elementua lortu du. Ikus daitekeen moduan, σ handitzeak helburu-objektua detektatzeko gaitasuna areagotzen du. Sakonera erabiltzeak, ordea, zehaztasuna zertxobait pobretzen du, baina errore-tasa dezente txikitzen du. Hori oso baliozkoa zaigu, egiazko objektua detektatzeko gaitasuna areagotzen delako eta okerreko identifikazioak ekiditen direlako.

Emaitza horiek guztiak aztertuta, azken eredua eramango dugu Pepper robotera; emaitza okerragoak lortzen baditu ere zehaztasun-, zehaztapen- eta prezisio-neurrietan, zarata gutxiagoko emaitzak itzultzen ditu, eta horrek du garrantzi handiena gure helburuan.

3. DETEKZIO-SISTEMA PEPPER ROBOTEAN INTEGRATZEA

Pepper[2] aski ezaguna den eta giza itxura duen robot soziala da, 1,20 m-ko altuera duena. Kopetan kokatua 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 badaitezke ere, 320 × 240 pixeleko bereizmena ezarri dugu bi gailuetan; baxua, beraz, denbora errealean aritzeko murrizketari aurre egin ahal izateko.

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

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.

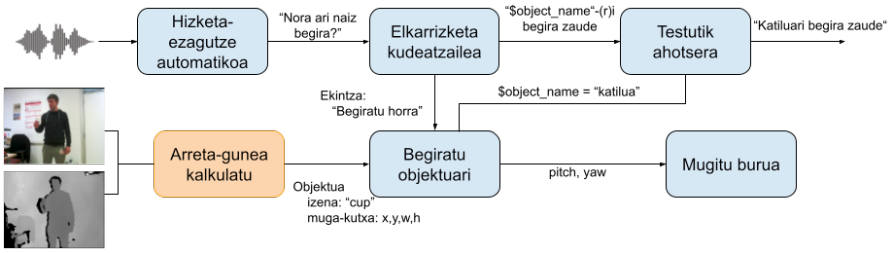
Arreta-gunea identifikatzeko prozesua nahiko astuna denez konputazio aldetik (hainbat sare eta funtzio aplikatu behar dira, arestian azaldu dugun moduan), kanpoko prozesu-unitate bat erabili dugu (NVIDIA GeForce RTX 4080, 32GB RAM eta Intel Core i9-13900K). Robota eta kanpoko ordenagailuak ROS (Robot Operating System)[3] bidez komunikatu ditugu, irudiak (RGB + sakonera) bidaliz lehenengoa, eta buruaren mugimendua

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

[3] www.ros.org

eta bota behar duen audioa aginduz bigarrena. ROS arkitektura hori zenbait moduluz osatuta dago (ikus 5. irudia). Hiru funtzionalitate ditu muinean:

1. Aho bidez agintzen zaiona ulertu eta erantzun egokia sortzeko gaitasuna.
2. Arreta-gunea detektatu eta bertan dagoen objektua identifikatzea (posizioa eta izena), hau da, 2. atalean deskribatutako prozesua.
3. Buruaren kontrola gauzatzea: objektuari begiratzeko orientazioa kalkulatzea eta mugimendu hori exekutatzea.
4. Erantzuna osatzea: ahoskatu beharreko testuan objektuaren izena txertatzea eta audio moduan botatzea.



5. irudia. Sistema osoaren arkitektura.

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 hori 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 kalkulatzen du. Horretarako, kameraren berezko parametroak kontuan izanda, objektua zenbat dagoen desplazatuta irudiaren zentroarekiko (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.

[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 zenbait sare mota erabili ditugu robot sozial batekin arreta partekatzeko. YOLOv8 eta ResNet101 bere horretan erabili ditugu, egokitzapenerako entrenamendu gehigarririk egin gabe. VGG16 sarea, aldiz, erregresiorako entrenatu behar izan dugu gure atazara egokitzeko. Bideo honetan[6] ikus daitezke portaera globalaren nondik norakoak. Modulu guztiak integraturik, sistema gai da denbora errealean erantzuna emateko. Esan dezakegu funtzionamendu orokorra onargarria dela. Nabarmendu behar dugu ezen, 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 kopetan duen kamera erabiltzea eragozpena da. Burua mugitzerakoan pertsona ikusmen-eremutik kanpo gelditzen da; horregatik, erantzuna eman ahala, posizio neutrora eramaten dugu berriro. Ez dugu pertsonaren jarraipena egiterik. Horretarako soluzioa bularrean Realsense D435[7] moduko RGBD kamera bat lotzea izango litzateke eta, hala, ikusmira mantentzeaz gain, mugimenduak berak sortzen dituen irudi lausoak ekidin.

Objektuak identifikatzeko erabili dugun ResNet101-ak objektu asko nahasten dituela egiaztatu dugu; irudi berriekin aritzeko zailtasunak dituela, alegia. Esaterako, mailua, urruneko kontrola eta beste hainbat objektu telefono mugikor gisa etiketatu ditu, baita katilua ardo-kopa moduan ere. Aurpegiak identifikatzeko erabili dugun YOLOv8 sareak objektuak etiketatzeko aukera ere eskaintzen du eta etorkizun hurbilean ordeztu beharko genuke ResNet sarea.

Elkarrekintza 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, Pepper-ek euskara ulertzea eta euskaraz hitz egitea da gure erronka premiazkoena, eta, horretarako, hizpide ditugu Elhuyar-en ADITU hizketa-ezagutzailea eta *ORAI NLP Teknologiak* enpresak eskaintzen duen TTSa. Badugu lehen mailako hezkuntzarako zein solasaldirako aplikazio baterako erroa; izan ere, baditugu ikusi-makusi joko didaktiko bat inplementatzeko bloke guztiak. Garapen maila horretara iristean, askotariko populazioekin (ikasle eta irakasleekin, adibidez) tes-

[6] https://youtu.be/Wnh2qNDnVRs

[7] https://www.intelrealsense.com/depth-camera-d435/

tatu beharko genuke sistema eta haren 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.

[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.