TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Marten Mathias Jaani

Inimese asendamine avatariga kasutades poosituvastust – interaktiivne peegel

Bakalaureusetöö (9 EAP)

Juhendaja:
Ardi Tampuu, PhD

TARTU 2025

Inimese asendamine avatariga kasutades poosituvastust – interaktiivne peegel

Lühikokkuvõte:

Tehisnägemine on kiiresti arenenud valdkond, mis tänapäeval võimaldab muuhulgas veebikaamera kaudu täpselt jälgida inimese poosi reaalajas. Käesoleva lõputööga loodi interaktiivne süsteem, mis reaalajas asendab veebikaamera vaateväljas oleva inimese 3D avatariga. Kuigi poosituvastust on laialdaselt kasutatud avatari animeerimiseks, on siin rakendatud segareaalsuslik lähenemine – inimese täielik eemaldamine originaalkaadrist ja tema asendamine avatariga – endiselt katsumusterohke valdkond. Sarnaseid lahendusi on arendatud mobiilseadmetele, kuid nende peamiseks piiranguks on nõrgast riistvarast tulenev ebapiisav jõudlus. Käesolev süsteem on loodud töötama videokaardiga arvuti peal, saavutades tõhusa reaalajalise jõudluse. Tartu Ülikooli Delta õppehoones läbi viidud testimine kinnitas süsteemi hariduslikku ja ekspositsioonilist väärtust ning selle kvaliteeti hinnati positiivselt.

Võtmesõnad: tehisnägemine, poosituvastus, sügavõpe, pildi segmenteerimine, segareaalsus, avatari animeerimine

CERCS: P170 – Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine; P176 – Tehisintellekt; T111 – Pilditehnika

Human Substitution with Avatar Using Pose Estimation - Interactive Mirror

Abstract:

Computer vision has rapidly evolved into a field that nowadays enables precise tracking of human body pose through a webcam in real-time. This thesis presents an interactive system that replaces a person in a webcam's field of view with a 3D avatar in real-time. While pose estimation has been widely used for avatar animation, the mixed reality approach implemented here – completely removing a person from the original frame and replacing them with an avatar – remains a challenging domain. Similar solutions have been developed for mobile devices, but their main limitation is insufficient performance due to weak hardware. The current system is designed to operate on computers with dedicated graphics cards, achieving efficient real-time performance. Testing conducted at the University of Tartu's Delta Centre confirmed the system's educational and expositional value, with users positively rating its quality.

Keywords: Computer Vision, Pose Estimation, Deep Learning, Image Segmentation, Mixed Reality, Avatar Animation

CERCS: P170 – Computer science, numerical, analysis, systems, control; P176 – Artificial intelligence; T111 – Imaging, image processing

Sisukord

Sissejuhatus

Viimastel aastatel on tehisnägemise valdkond muutunud üha võimekamaks. Märkimisväärne on inimese poosituvastuse tehnoloogia areng, mis nüüdseks võimaldab reaalajas tuvastada ja jälgida inimese keha asendit[1]. Seda tehnoloogiat on palju rakendatud liitreaalsuses, tervishoius ja järelevalves ning see on avanud uusi võimalusi inimese ja arvuti vahelisteks interaktsioonideks [1].

Käesolev lõputöö keskendub interaktiivse poosituvastuse süsteemi loomisele Tartu Ülikooli Delta hoone koridori. Süsteem tuvastab veebikaamera vaateväljas olevat inimest, kes reaalajas eemaldatakse kaadrist ning asendatakse tema liigutusi järgiva 3D avatariga.

Poosituvastuse tehnoloogiat on laialdaselt kasutatud 3D avatari animeerimiseks reaalajas[2], kuid inimese täielik asendamine avatariga originaalkaadris on katsumusterohke valdkond. Taolisi lahendusi on uuritud mobiilseadmetele, kuid nende kitsaskohaks on nõrgast riistvarast tulenev jõudlus [2,3]. Käesolev töö aga näitab, et sellist süsteemi on võimalik luua arvutil toimima keskklassi videokaardi jõudlust ning tänapäeva tehnoloogiaid kasutades.

Bakalaureusetöö eesmärk on luua toimiv poosituvastuse süsteem, mis suudab tõhusalt tuvastada ja visualiseerida inimese keha asendit avatarina, tagades seejuures sujuva kasutajakogemuse ja töökindluse.

Projekti olulisus seisneb selle hariduslikul ja ekspositsioonilisel väärtusel. Esiteks toob installatsioon esile poosituvastuse tehnoloogia võimekusi ja tehisintellekti potentsiaali aastal 2025, pakkudes näidet kaasaegsest tehisnägemise rakendusest. Teiseks loob interaktiivne installatsioon haarava kogemuse õppehoone külastajatele, mis võib tekitada suuremat huvi arvutiteaduse instituudi vastu.

Käesoleva töö koostamisel kasutati olemasoleva teksti töötlemiseks tehisintellektil põhinevaid rakendusi (ChatGPT [3]ja Claude[4]).

[1] https://viso.ai/deep-learning/pose-estimation-ultimate-overview/

[2] https://github.com/yeemachine/kalidokit

[3] https://chatgpt.com/

[4] https://claude.ai/new

Mõisted ja terminid

Tehisnägemine (ingl *Computer Vision*) on arvuti võime visuaalseid andmeid tõlgendada ja töödelda.[5]

Masinõpe ehk masinõppimine (ingl *Machine Learning*) on tehisintellekti valdkond, mis uurib tehissüsteemide selliseid algoritme ja mudeleid, millel on võimekus õppida andmete põhjal, leida mustreid ning teha ennustusi ja otsuseid.[6]

Tehisnärvivõrk ehk neurovõrk (ingl *Neural Network*) on masinõppe mudel, mis jäljendab inimaju neuronvõrke.[7]

Konvolutsiooniline närvivõrk (ingl *Convolutional Neural Network*) on närvivõrk, millel on mitu erinevat tüüpi kihte ning mis on laialt kasutusel tehisnägemises.[8]

Sügavõpe (ingl *Deep Learning*) on masinõppe valdkond, kus kasutatakse mitmekihilisi neurovõrke masinõppe ülesannetes.[9]

Transformer (ingl *Transformer*) on närvivõrgu arhitektuur, mis suudab töödelda andmeid paralleelselt, mõistes keerulisi seoseid ja konteksti.[10]

Varjutaja (ingl *Shader*) on kood, mis kasutab arvuti videokaarti, et täita mingi graafikatöötluse ülesanne.[11]

Kvaternioon (ingl *Quaternion*) on andmetüüp, millega on võimalik kolmedimensionaalseid pöördeid esitada.[12]

---

[5] https://akit.cyber.ee/term/10214-tehisnagemine-raalnagemine

[6] https://www.solix.com/et/kb/machine-learning/

[7] https://akit.cyber.ee/term/1638-neurovork

[8] https://jaak.tepinfo.ee/is-5-NN.pdf

[9] https://www.ibm.com/think/topics/deep-learning

[10] https://www.ibm.com/think/topics/transformer-model

[11] https://cgvr.cs.ut.ee/arvutigraafika-terminid/

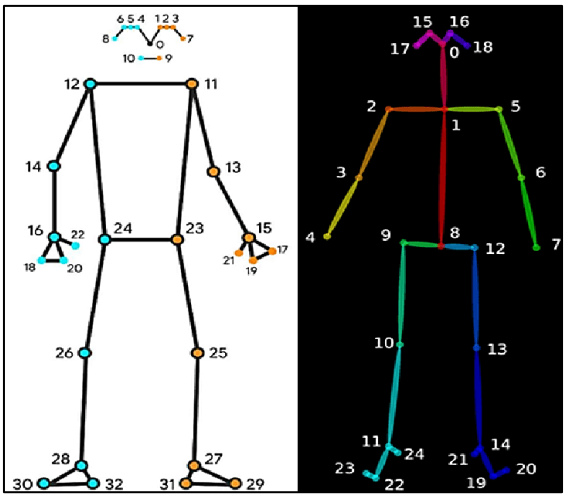
[12] https://en.wikipedia.org/wiki/Quaternions_and_spatial_rotation

1. Taustainfo

Siin peatükis antakse ülevaade kasutatavate tehnoloogiate põhialustest ja tänapäevastest võimekustest. Kirjeldatakse poosituvastuse tehnoloogiat ja arengut, kuidas seda kasutada 3D avatari animeerimiseks, ning inimese segmenteerimise tehnoloogiat.

1.1 Inimese poosituvastuse põhialused

Inimese poosi tuvastus (ingl *human pose estimation*) on tehisnägemise üks pikaaegsemaid ja tuntumaid probleeme [4]. Inimese poosi võib defineerida kui teatud liigeste paigutust mingil ajahetkel, seega tehisnägemise valdkonnas seostub ülesanne inimkeha võtmepunktide asukoha tuvastamisega (vt joonis 1) [5]. Kui 20 aastat tagasi piirdus poosituvastus ühe inimese üldise poosi hindamisele liikumatul pildil [6], siis tänapäeva lahendused demonstreerivad muljet-avaldavaid tulemusi mitme inimese võtmepunktide tuvastamisel videovoos ja reaalajas [7].



Joonis 1. Näited tuvastavatest võtmepunktides[13].

Kümmekond aastat tagasi toimus hüppeline areng poosituvastuse tehnoloogias, kui hakati kasutama neurovõrke. Aastal 2014 ilmunud DeepPose, mis esimesena kasutas poosituvastuses konvolutsioonilist neurovõrku, edastas varasemaid lahendusi märgatavalt [5]. Edasised mudelid hakkasid aina rohkem kasutama sügavõpet[14].

Et DeepPose oli loodud vaid üksiku inimese tuvastamiseks liikumatul pildil, hakati sügavõppelist lähenemist rakendama ka keerulisema ülesande jaoks – mitme inimese

[13] https://www.researchgate.net/figure/Dimensional-skeleton-keypoint-topology-for-BlazePose-left-and-OpenPose-right_fig1_358017174

[14] https://nanonets.com/blog/human-pose-estimation-2d-guide/

poosituvastus [8]. Mitme inimese poosituvastuse lahendused jagunevad Zheng *et al.* [1] järgi kaheks:

- Ülalt-alla (ingl *top-down*) – kõigepealt kasutatakse inimesetuvastuse mudelit, mis iga inimese kohta tagastab teda piirava kasti (ingl *bounding box*). Seejärel rakendatakse igale kastile ühe-inimese poosituvastuse mudelit. Selle lähenemise eeliseks on tihti suurem täpsus, kuid see nõuab rohkem arvutuslikku jõudu.
- Alt-üles lähenemine (ingl *bottom-up*) – kõigepealt tuvastatakse pildilt kõik võtmepunktid ja seejärel viiakse need iga inimesega kokku. Eeliseks sel juhul on kiirus, eriti suure arvu inimeste korral.

Varasemad mitme inimese poosituvastuse lahendused, mis keskendusid reaalajalisele toimimisele, kasutasid suures jaos kiiremat, alt-üles lähenemist [9]. Viimastel aastatel on aga ülalt-alla lähenemist aina rohkem rakendatud. Üheks selliseks näiteks on aastal 2023 Jiang *et al.* [7] loodud tipptaseme tulemusi saavutav RTMPose. Nemad toovad välja, et kui varem oli ülalt-alla lähenemise üheks pudelikaelaks inimesetuvastuse mudelite ressursinõudlikkus, siis seda kitsaskohta kiire arengu tõttu enam ei eksisteeri. Lisaks tuuakse välja, et nimetatud lahenduse töövoogu on integreeritud transformeril põhinevat arhitektuuri. RTMPose funktsionaalsust visualiseerib joonis 2.

Joonis 2. RTMPose poolt tuvastatud võtmepunktid[15].

Arvutusvõimsuse kasv ja transformerite kasutuselevõtt on valdkonnas toonud märkimisväärseid edusamme [10]. Kuigi transformer arendati välja aastal 2017 loomuliku keele töötluseks [11], hakati taolise arhitektuuri rakendusi uurima ka tehisnägemise valdkonnas [12]. Aasta 2024 seisuga kasutavad mitmed poosituvastuse lahendused konvolutsiooniliste

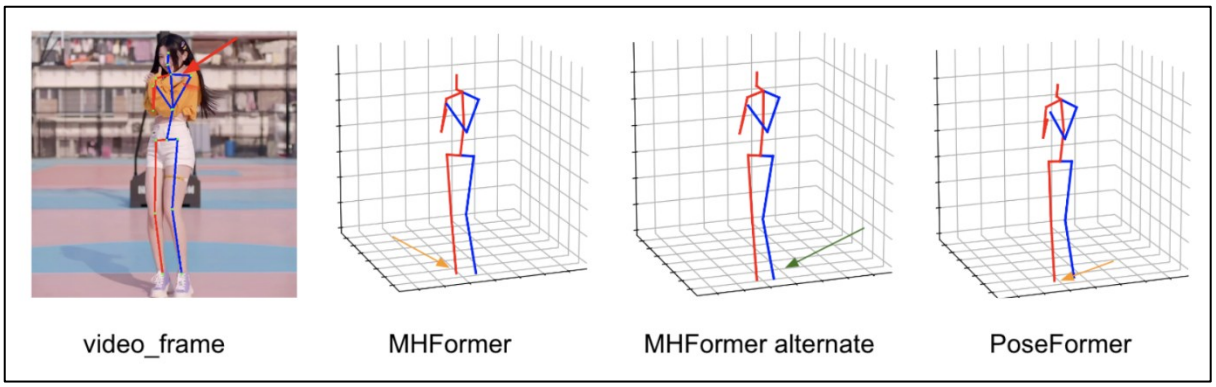
[15] https://github.com/open-mmlab/mmpose/tree/main/projects/rtmpose

närvivõrkude ja transfrormerite hübriid-arhitektuuri [10]. Aktiivselt on uuritud, kuidas transformarhitektuuri saaks kasutada 3D poosituvastuses (vt joonis 3) [13].

Erinavalt 2D poosituvastusele, mis tagastab võtmepunktide x- ja y-koordinaadid pildil, ennustab 3D poosituvastus ka võtmepunktide sügavust ehk z-koordinaati. Tuvastatud võtmepunktide koordinaadid sel juhul lähtuvad inimese juurpunktist või lähtuvad kaamera asukohast (näiteks koordinaat (0,0,0) oleks vastavalt inimese puusa või kaamera asukoht) [13]. Neupane *et al.* [13] järgi teeb 3D poosituvastuse ühest RGB-kaamerast katsumusterohkeks võtmepunktide sügavuse ebaselgus ning inimkeha enesevarjutus (ingl *self-occlusion*). Siiski on nende sõnul tegu väga aktiivse ja kiiresti areneva uurimisvaldkonnaga, ning nad jagavad eksisteerivad lahendused lähenemise poolest kaheks:

- Ühe-astmeline lähenemine (ingl *single-stage*) – mudel ennustab 3D poosi vahesammudeta, olles treenitud andmetel, millel on võtmepunktid märgendatud kolmedimensionaalsetena.
- Kahe-astmeline lähenemine (ingl *two-stage*) – kasutatakse vaheandmetena 2D poosituvastusest saadud võtmepunkte või üldisemaid 2D tunnuseid, et ennustada 3D võtmepunkte.

Nimetatud uurijad veel väidavad, et üheastmelise lähenemise kitsaskohaks on 3D treeningandmete vähesus, ning et sellist lähenemist kasutavad mudelid kipuvad olema arvutuslikult nõudlikud ning ebasobivad reaalajas kasutamiseks. See vastu on võimalik kahe-astmelises lähenemises ära kasutada 2D poosituvastuse tehnoloogia tugevusi – 2D poosituvastust on laialt uuritud valdkond ning 2D treeningandmeid on väga palju [14]. Selle lähenemisega on võimalik 3D poosituvastus taandada probleemile, et kuidas tuvastatud 2D poos tõsta kolme-dimensionaalsesse ruumi (vt Joonis 3) [15].

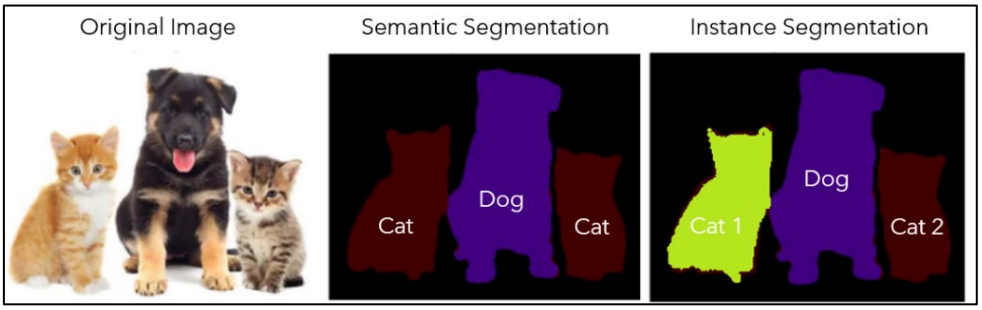


Joonis 3. 2D poosi teisendamine 3D poosiks transformarhitektuuri kasutavate mudelite näitel [15].

Käesoleva töö jaoks on oluline kasutada efektiivset mudelit, mis tuvastab ühe inimese 3D võtmepunkte reaalajas. Lõpuks otsustati kahe-astmelist lähenemist kasutavale BlazePose mudelil [16] põhinevale Google MediaPipe Pose Landmarker lahendusele[16], mille kergekaalulisus võimaldab suurepärast jõudlust ning mida on ka varem kasutatud avatari animeerimiseks (vt peatükk 1.5). Täpsemalt on selle lahenduse kasutamise eelistest juttu 2.2 peatükis.

1.2 Inimese segmenteerimine

Lisaks poosituvastusele on teine pikaaegne tehisnägemise põhikatsumustest olnud pildi segmenteerimine [17]. Pildi segmenteerimine[17] hõlmab pildilt nende pikslite tuvastamist, mis kuuluvad teatud objektile, et selle kaudu tagastada segmentatsioonimaskid, mis kujutavad objektide piiritletud kuju ja asukohta pildil. Varasemad segmenteerimise algoritmid on asendunud sügavõppel põhinevate lahendustega, mis on toonud kaasa kiire arengu selles valdkonnas [18]. On arendatud mudeleid, mis suudavad reaalajas igale kaadri pikslile määrata klassi (semantiline segmentatsioon), ning sama klassi objekte eristada (objekti segmentatsioon) [19]. Semantilist segmentatsiooni ja objekti segmentatsiooni visualiseerib joonis 4.



Joonis 4. Semantiline segmentatsioon ja objekti segmentatsioon[18].

Käesolev lõputöö on aga huvitatud lihtsast inimese segmentatsioonimaskist, mis eristab vaid tuvastatud inimesele kuuluvaid ja mittekuuluvaid piksleid. Kuna inimese poosi tuvastus ja inimese segmenteerimine on üsna seotud ülesanded, siis osad segmentatsiooni mudelid nagu Mask R-CNN suudavad ka inimese poosi hinnata [20], samuti võimaldab näiteks Google MediaPipe Pose Landmarker väljundiks anda ka inimese segmentatsioonimaski[19].

[16] https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker

[17] https://www.ibm.com/think/topics/image-segmentation

[18] https://www.v7labs.com/blog/instance-segmentation-guide

[19] https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker#features

1.3 Avatari animeerimine poosituvastusega

Viimastel aastatel on reaalajas töötava poosituvastuse üks tähelepanuväärsemaid rakendusi olnud avatari animeerimine, mis hõlmab poosituvastusest saadud inimese liikumise informatsiooni ülekandmist 3D avatarile. Populaarseks on muutunud VTube[20], kus inimesed isiksustavad ennast virtuaalse avatariga, keda animeeritakse veebikaamera sisendiga. Lisaks on esile kerkinud ka Metaverse kontseptsioon, mis hõlmab igasuguseid virtuaalseid maailmu, kus inimesed saavad suhelda virtuaalsete teisikutena, ning kus poosituvastuse lahendusi oma teisiku animeerimiseks on palju uuritud [21]. Lõputöö väljund aga kuuluks segareaalsuse (ingl *Mixed Reality*, MR) valdkonda[21], mis hõlmab füüsilise ja digitaalse maailma põimingut, antud lõputöö puhul inimese asendamist avatariga.

1.4 Segareaalsus

Segareaalsus kuulub laiendatud reaalsuse[22] (ingl *Extended Reality*, XR) üldmõiste alla, mis hõlmab lisaks veel virtuaalreaalsust (ingl *Virtual Reality*, VR) ja liitreaalsust (ingl *Augmented Reality*, AR). Kui virtuaalreaalsus hõlmab simuleeritud kogemusi täiesti digitaalses maailmas, ja liitreaalsus hõlmab päris maailmale ainuüksi digitaalse sisu peale kuvamist, siis segareaalsus proovib luua interaktiivse digitaalse- ja pärismaailma segu. Segareaalsuse tehnoloogiaid peetakse laiendatud reaalsuse järgmiseks suureks laineks [22].

Selliseid segareaalsuse tehnoloogiaid, mis asendavad täielikult pärismaailma objekte digitaalsete objektidega, nagu käesoleva lõputöö väljund, on seostatud terminiga asenduslik reaalsus (ingl *Substitutional Reality*, SR) [23] või terminiga objektiasendus (ingl *Object substitution*) [2,3]. Selle saavutamiseks on uuritud liitreaalsuse ja vähendatud reaalsuse[23] (ingl *Diminished Reality*, DR) kombinatsioone [3]. Vähendatud reaalsus hõlmab tehnoloogiaid, mis eemaldavad objekte pildil/videolt, näiteks objekti segmenteerimine ja selle täitmine tasutaga. Kasutades töövoos nii liitreaalsusele omast digitaalse sisu lisamist kui ka vähendatud reaalsusele omast objekti eemaldamist on võimalik saavutada objekti täielik asendamine. Sellist ideed järgib ka käesoleva lõputöö praktilise osa töövoog, kus paralleelselt toimub nii avatari kuvamine kui ka inimese eemaldamine. Täpsemalt on sellest juttu 2.5 peatükis.

[20] https://www.qustodio.com/en/blog/what-is-a-vtuber/

[21] https://learn.microsoft.com/en-us/windows/mixed-reality/discover/mixed-reality

[22] https://blogs.nvidia.com/blog/what-is-extended-reality/

[23] https://atlantis-ar.github.io/infotube/diminished_reality.html

1.5 Seotud tööd

Poosituvastuse kasutamist 3D avatari animeerimiseks on käsitlenud avaliku lähtekoodiga projektid nagu Kalidokit[24], ThreeDPoseUnityBarracuda[25] ja DigiHuman[26]. Kõik need rakendused kasutavad lähenemist, kus tuvastatud võtmepunktide vahel arvutatakse suunavektorid (näiteks suund õlast küünarnukini) ning avatari vastavatele liigestele arvutatakse selle põhjal vajalik pööre, et see vastaks inimese poosile. Järgnev projektide lahtiseletus põhineb vastavate repositooriumite materjalidele, mis on joonealuste viidetena saadaval.

Kalidokit (vt joonis 5) on VTube rakendustele loodud JavaScripti teek, mis kasutab Google MediaPipe'i mudeleid näo, käe ja keha võtmepunktide tuvastamiseks ning tuletab nende kaudu liigeste Euleri nurgad[27], mida rakendatakse avatarile reaalajas. Projekt on aga aegunud ning repositooriumis on kirjas, et nüüdseks on see integreeritud otse Google MediaPipe platvormile. Seisuga 07.05.2025 pole aga avatari animeerimise lahendusi MediaPipe platvormil saadaval[28].

Joonis 5. Kalidokit – tuvastatud poosi ülekandmine avatarile[22].

ThreeDPoseUnityBarracuda on Unity[29] projekt, mis kasutab Barracuda nimelist teeki (nüüdseks asendunud Sentis teegiga), et mängumootori siseselt närvivõrgu mudelit jooksutada. Saadud võtmepunktide abil arvutatakse välja kvaternioonid, mis rakendatakse avatari liigeste pööramiseks. Sarnaselt käesolevale tööle võimaldab projekt kuvada avatari veebikaamera pildi ees, kus avatari asukoht vastab inimese asukohale (vt joonis 6), ning selle projekti üldine lähenemine avatari animeerimiseks ja inimese asukohale asetamiseks Unity keskkonnas inspireeris ka käesoleva töö lähenemist. Projekt on aga samuti nüüdseks mitu aastat aegunud.

[24] https://github.com/yeemachine/kalidokit

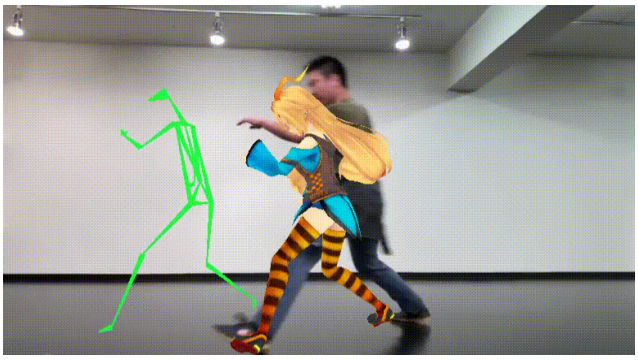
[25] https://github.com/digital-standard/ThreeDPoseUnityBarracuda

[26] https://github.com/Danial-Kord/DigiHuman

[27] https://et.wikipedia.org/wiki/Euleri_nurgad

[28] https://github.com/google-ai-edge/mediapipe/issues/5220

[29] https://unity.com/products/unity-engine



Joonis 6. Kaader ThreeDPoseUnityBarracuda väljundist[23].

DigiHuman (vt joonis 7) on samuti Unity projekt, aga kasutab eraldi *backendi* Google MediaPipe poosituvastuse lahenduse (sh *Holistic* lahendust, mis tuvastab ka näo ja käe võtmepunkte) jooksutamiseks. Üldine loogika avatari animeerimiseks on ThreeDPose-UnityBarracuda projektiga väga sarnane.

Joonis 7. Kaader DigiHuman väljundist[24].

Siiani mainitud projektid on kõik loodud töötama arvuti peal veebikaamera sisendil. Veel on poosituvastust rakendavaid liitreaalsuse süsteeme loodud ka mobiilseadmetele. Liitreaalsuse arendamise platvormid nagu Snapchat Lens Studio[30] ja Unity AR Foundation[31] (kasutades Apple ARKit[32] või Google ARCore[33] pluginaid) pakuvad võimalusi luua mobiilseadmetele süsteeme inimese keha tuvastamiseks ja jälgimiseks ning liitreaalsuse elementide (sh avataride) kuvamist inimesele peale (vt joonised 8 ja 9).

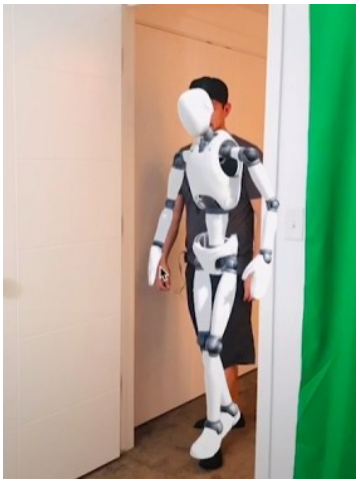
[30] https://developers.snap.com/lens-studio/overview/getting-started/lens-studio-overview

[31] https://docs.unity3d.com/Packages/com.unity.xr.arfoundation@5.1/manual/index.html

[32] https://developer.apple.com/augmented-reality/arkit/

[33] https://developers.google.com/ar

Joonis 8. Snapchat 3D Body Tracking[34].



Joonis 9. Unity AR Foundation rakendus[35].

Reaalajas inimese täielik asendamine avatariga mobiilseadmetel on veel katsumusterohke valdkond. Ke ja Wang [3] ning Kari *et al.* [2] (vt joonised 10 ja 11) uuringute sõnul osutub pudelikaelaks mobiilseadmete nõrk riistvara, sest kui poosituvastusele ja avatari animeerimisele veel juurde lisada inimese segmenteerimine ja taustaga täitmine, on arvutuslik nõudlus liiga suur. Liikuva kaamera tõttu muutub kaadri sisu pidevalt ning tuleb kasutada keerukaid tasuta täitmise mudeleid. Need kaks projekti kasutasid pudelikaela lahendamiseks vastavalt servtöötlust[36] (ingl *Edge Computing*) ja pilvtöötlust[37] (ingl *Cloud Computing*), kuigi sellise lähenemisega kaasnevat viivitust tuuakse samuti probleemiks.

Joonis 10. TransforMR - inimese asendamine avatariga pilvtöötlust kasutades [2].

[34] https://developers.snap.com/lens-studio/features/ar-tracking/body/body-templates/3d-body-tracking

[35] https://www.youtube.com/watch?v=cfKzUYH4i7A

[36] https://www.ibm.com/think/topics/edge-computing

[37] https://www.ibm.com/think/topics/cloud-computing

Joonis 11. Mobiilseadmel inimese asendamine avatariga servtöötlust kasutades [3].

Kirjeldatud projektidest on näha, et poosituvastuse tehnoloogiat on laialt varem kasutatud 3D avatari animeerimiseks reaalajas. Lisaks on uuritud segareaalsuse süsteemide arendamist mobiilseadmetele, mis suudavad inimest täielikult asendada avatariga reaalajas. Selliste lahenduste nõrkusteks on aga mobiilseadmete üldiselt nõrk riistvara. Antud lõputöö lähenemine, mis kasutab videokaardiga arvuti jõudlust, paigal seisvat kaamerat ning keskendub ühe inimese tuvastamisele, võimaldab luua usaldusväärse ja efektiivselt toimiva interaktiivse süsteemi inimese täielikuks asendamiseks.

2. Metoodika

Töö hõlmab erinevaid valdkondi ning parima lähenemise leidmine nõudis palju katsetamist. Selles peatükis kirjeldatakse käesoleva töö metoodikat ja tehnilisi valikuid, mis võimaldasid inimese poosituvastuse ja segmentatsiooni tehnoloogiaid rakendada Unity keskkonnas, et luua reaalajas toimiv inimest avatariga asendav süsteem.

2.1 Platvorm

Projekt on üles ehitatud Unity mängumootoril, mis on laialt kasutusel olev ja väga mitmeotstarbeline platvorm[38]. Lisaks autori varasemale kogemusele langes valik Unity kasuks mitmete tehniliste eeliste tõttu.

Esiteks võimaldab Unity tehisnärvivõrgu mudelite käivitamist otse mängumootorist, kasutades Unity loodud Sentis[39] teeki. See eemaldab vajaduse luua eraldi *backend* süsteem mudeli jooksutamiseks, mis tunduvalt lihtsustab süsteemi arhitektuuri ning aitab vähendada viivitust, mis tuleneks võtmepunktide edastamisest ühest protsessist teise. Saadaval on ka laialdane tugi 3D avataride importimiseks ja nende kontrollimiseks. Unity Quaternion nimeline klass võimaldab poosituvastusest saadud informatsiooni teisendada kvaterniooniks, mida saab rakendada avatari liigete pööramiseks. Lisaks võimaldab Unity varjutajaid luua, mis kasutavad GPU jõudlust, et efektiivselt töödelda inimese segmentatsioonimaski ja saavutada inimese eemaldamine.

2.2 Poosituvastuse mudel

Käesolev töö kasutab ühte inimest tuvastavat Google Mediapipe Pose Landmarker (Heavy)[40] lahendust, mis kasutab kahest osast koosnevat BlazePose mudelit:

- ***Pose detection model*** tuvastab kaadrilt inimese olemasolu ja asukoha, tagastades inimest piiritleva kasti (*bounding box*) ja mõned võtmepunktid. Juhul, kui on mitu inimest kaadris, tagastab neist kõige kindlamalt tuvastatuma.
- ***Pose landmarker model*** saab eelmiselt mudelilt sisendiks kärbitud kaadri inimesest, ning väljundiks annab inimekeha 33 kolmemõõtmelist võtmepunkti ja inimese segmentatsiooniandmed.

[38] https://www.pubnub.com/guides/unity/

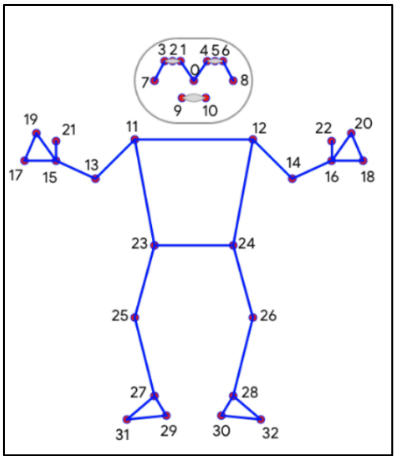
[39] https://unity.com/products/sentis

[40] https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker#models

Kuigi Google MediaPipe'i lahendused jäävad uuemate ja keerukamate mudelite täpsustele pigem alla [24], siis projekti kasutuskonteksti eeldades (staatiline kaamera, inimene mitte liiga kaugel, hea valgustus) on täiesti rahuldavad tulemused saavutatavad [25].

Põhjuseid mudeli kasutamise jaoks oli mitmeid. Esiteks võimaldab see MediaPipe'i lahendus lisaks võtmepunktidele tagastada ka inimese segmentatsiooni, mis eemaldab vajaduse eraldi segmentatsiooni väljastava mudeli järele, säästes nii arvutusressursse kui ka lihtsustades süsteemi arhitektuuri. Lisaks on mudel kergekaaluline, mis võimaldab efektiivset reaalajalist toimimist. Veel oli kasuks asjaolu, et Unity on avalikustanud HuggingFace platvormil näidisprojekti Sentis-blaze-pose[41], mis aitas kaasa mudeli integreerimisel käesolevasse projekti.

Tuvastatud 33 võtmepunkti (vt joonis 12) on võimalik viia avatari liigestega vastavusse. Siiski esineb ka piiranguid – sõrmede detailne asend jääb tuvastamata, mudel annab vaid käelaba peamiste punktide asukohad. Autor katsetas ka Mediapipe Hand Landmarks Detection[42,43] mudeli integreerimist, et avatar suudaks jäljendada ka inimese näppude liigutusi, kuid see ei osutunud edukaks. Probleem seisnes selles, et mudel ei suutnud piisavalt kindlalt tuvastada käsi kaadritel, kus inimene oli kaamerast mõne või enama meetri kaugusel. Lisaks proovis autor ka tipptaseme RTMW [24] mudelit rakendada, mis suudab mitmel inimesel tuvastada 133 võtmepunkti (kaasaarvatud näpud), kuid selle efektiivne integreerimine oli oodatust raskem, ning sellel lähenemisel puudusid MediaPipe mudeli kasutamise eelised.



Joonis 12. BlazePose tuvastatavad võtmepunktid [16].

2.3 Avatar

Avatarina kasutati Unity poolt loodud ja Unity Asset Store'ist kättesaadavat Unity-Chan![44] 3D mudelit (vt joonis 13). Valik langes Unity-Chan! kasuks selle tõttu, et mudel on tasuta ja selle litsentsitingimused lubavad seda vabalt kasutada. Lisaks on tegu kvaliteetse, luustikuga varustatud (ingl *rigged*) mudeliga, mis sobib hästi animeerimiseks. Samuti reageerivad avatari juuksed ja riietus dünaamiliselt liikumisele, lisades tulemusele elavust.

[41] https://huggingface.co/unity/sentis-blaze-pose

[42] https://ai.google.dev/edge/mediapipe/solutions/vision/hand_landmarker

[43] https://huggingface.co/unity/sentis-blaze-hand

[44] https://assetstore.unity.com/packages/3d/characters/unity-chan-model-18705

Autor katsetas alguses ka Mixamo[45] tasuta pakutavaid 3D avatare, mis kasutavad omavahel ühilduvat luustiku ülesehitust (ingl *rig*), ning mis oleks teinud ka mitme erineva avatari vahel vahetamise lihtsasti teostatavaks. Nendel avataridel aga puudus Unity-Chan! mudeli dünaamilisus, ning kvaliteedi poolest olid lihtsakoelisemad.

Joonis 13. Unity-Chan! Unity stseenis.

2.4 Tehniline ülesehitus

Süsteem koosneb mitmest omavahel suhtlevast komponendist (C# skriptist). Süsteemi lähtekoodi saab lähemalt uurida lisade I sektsioonis oleva repositooriumi lingi kaudu. Iga komponent vastutab kindlate ülesannete täitmise eest:

PoseDetection tegeleb järgmiste poosituvastusega seotud ülesannetega:

- Initsialiseerib ja käivitab Unity Sentis teegi abil ONNX-formaadis[46] närvivõrgumudelid
- Töötleb veebikaamerast saadud kaadreid, rakendades Google Mediapipe Pose Landmarker (Heavy) lahendust: esmalt tuvastab inimese asukoha (*poseDetector*) ja seejärel leiab täpsemad 33 võtmepunkti ning segmentatsiooniandmed (*poseLandmarker*).
- Rakendab tuvastatud võtmepunktidele ajalisi filtreid (Kalman[47] ja One Euro filter [26]), et vähendada müra ja muuta liikumine sujuvamaks.

SegmentationRenderer vastutab segmentatsiooni töötlemise eest, kasutades varjutajaid:

- Töötleb *PoseDetection*-ist saadud segmentatsiooniandmeid *SegmentationMask.compute* nimelise varjutajaga, et luua segmentatsioonimask.
- Haldab dünaamilist taustamudelit, mida uuendatakse pidevalt varjutaja *BackgroundModel.compute* abil. Tausta uuendatakse järk-järgult nendes piirkondades, kus segmentatsioonimask näitab inimese puudumist. Siinkohal on programmi jooksutamisel oluline, inimene ei püsiks kogu protsessi vältel ühe koha peal kaadris. Vastasel juhul jääb kogu aeg inimese poolt kaetud ala jäädvustamata ning meil pole infot kuidas selles asukohas taust välja näeb.

[45] https://www.mixamo.com/#/?page=1&type=Character

[46] https://onnx.ai/

[47] https://thekalmanfilter.com/kalman-filter-explained-simply/

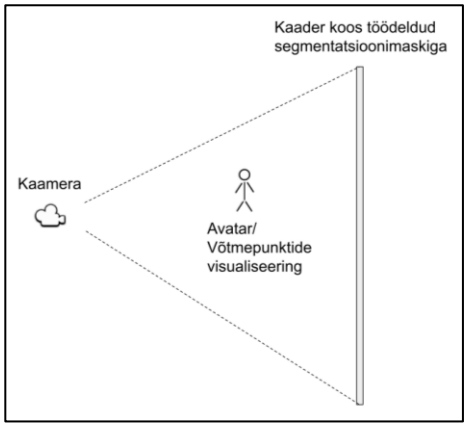
- Genereerib lõpliku väljundkaadri, kasutades varjutajat *SegPaint.shader*, mis kombineerib originaalkaadri, segmentatsioonimaski ja dünaamilise tausta, võimaldades inimest kuvada segmenteerituna (kaetud rohelise maskiga) või teda kaadrist eemaldada (täites segmentatsioonimaski taustaga). Selline meetod inimese eemaldamiseks töötab tänu asjaolule, et kaamera on fikseeritud ning taust seega enam-vähem staatiline.

AvatarController teisendab tuvastatud inimese võtmepunktid avatari liigutusteks:

- Seostab tuvastatud võtmepunktid avatari liigestega ja rakendab edasisuunalist kinemaatikat[48] (ingl *forward kinematics*), liikudes hierarhiliselt juurpunktist (puusast) väljapoole. Iga liigesepaari vahel arvutatakse vastavate võtmepunktide vahelised suunavektorid ning selle põhjal leitakse vajalikud liigeste pöörded, kasutades Unity *Quaternion.LookRotation()* funktsiooni.
- Viib avatari asukoha (avatari puusa) ja mõõtmed vastavusse tuvastatud inimesega.

ViewController - haldab erinevaid visualiseerimisrežiime ja võimaldab nende vahel lülituda.

- Vaade 1: Inimene on segmenteeritud ning talle on peale kuvatud võtmepunktid.
- Vaade 2: Inimese segmentatsioon täidetakse taustaga, mille tulemusel eemaldatakse inimene kaadrist, võtmepunktid püsivad nähtavana.
- Vaade 3: Inimene on asendatud avatariga.



Joonis 14. Unity stseeni diagramm.

Joonis 15. Unity stseen nurga alt vaadatuna.

Unity stseen on ülesehitatud nii, et võtmepunktide visualiseering ja avatar asetsevad virtuaalses ruumis kaamera ja töödeldud kaadri vahel. Nii on võimalik inimese peale kuvada võtmepunktid ja teda asendada avatariga (vt joonised 14 ja 15).

[48] https://www.whizzystudios.com/post/forward-kinematics-vs-inverse-kinematics-in-3d-character-rigging

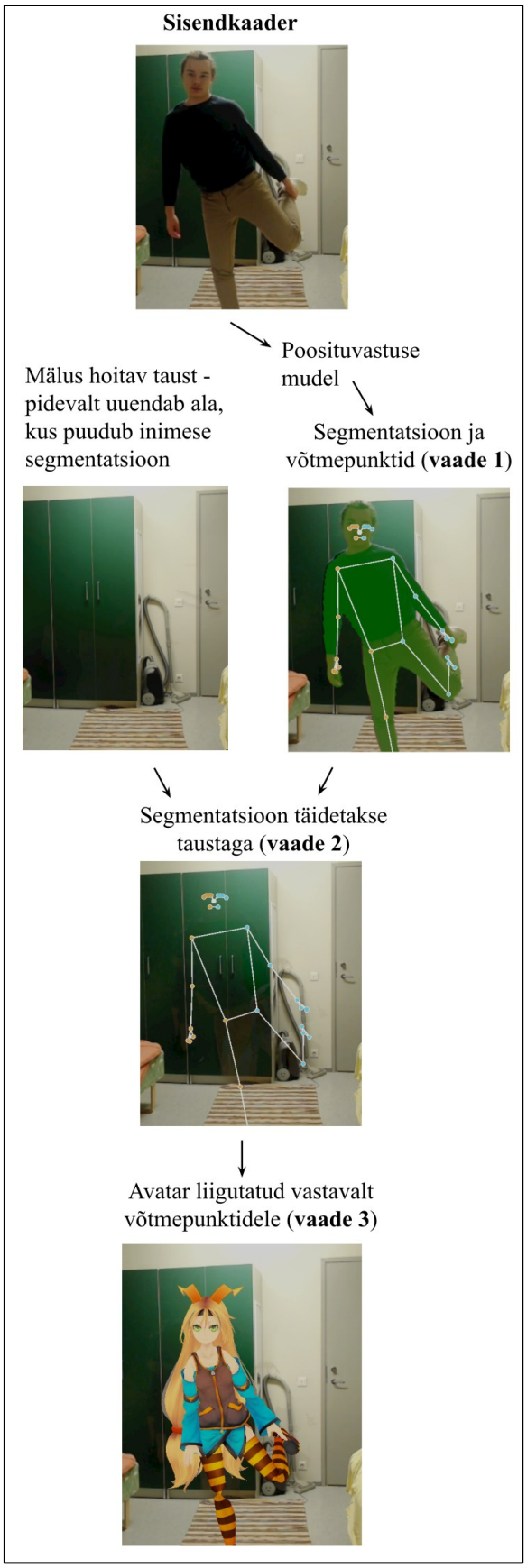
2.5 Töövoog

Süsteem järgib igas kaadris järgmist töövoogu (vt joonis 16):

1. Hangib kaadri veebikaamerast.
2. *PoseDetection* töötleb kaadrit, rakendades närvivõrke. See genereerib segmentatsiooniandmed ja filtreeritud 3D võtmepunktid.
3. *AvatarController* kasutab võtmepunkte, et arvutada ja rakendada avatarile vastavad liigutused. Paralleelselt kasutab *SegmentationRenderer* segmentatsiooniandmeid ja kaamera kaadrit, et genereerida mask, uuendada taustamudelit ning koostada töödeldud väljundkaader.
4. *ViewController* määrab, milline visualiseerimisrežiim on aktiivne (vaated 1, 2 ja 3). Vastavalt vaatele kuvatakse eraanile kas avatari, võtmepunktide või segmentatsiooni visualiseering.

2.6 Jõudluse ja kvaliteedi mõõdikud

Lõplik lahendus peab nii jõudluse kui ka kvaliteedi poolest olema rahuldav. Jõudlust saame hinnata sellega, mitu kaadrit suudab süsteem sekundis töödelda (FPS) ning kui suur on viivitus. Sujuva töötamise jaoks on oluline, et FPS oleks vähemalt 20-30 ringis ning viivitus alla 50 millisekundi. Kvaliteeti saab hinnata testimisel saadava tagasiside järgi.



Joonis 16. Töövoog visualiseerituna.

3. Tulemused ja testimine

Selles peatükis esitatakse loodud süsteemi tulemused, analüüsitakse selle jõudlust ning käsitletakse kasutajatelt saadud tagasisidet.

3.1 Süsteemi jõudluse mõõtmised

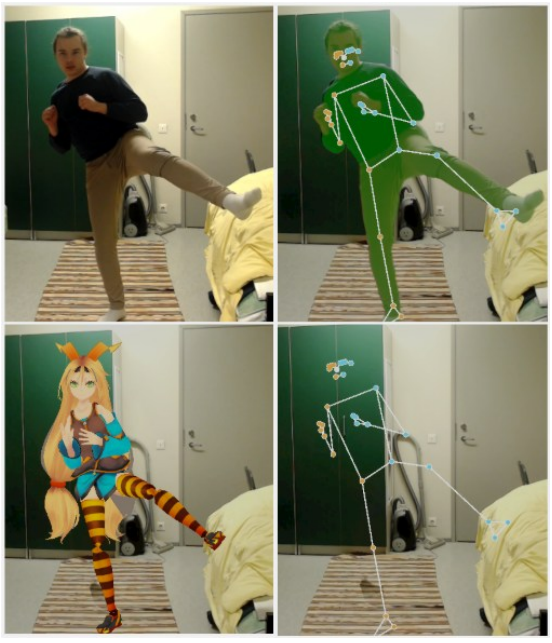
Süsteemi reaalajas toimimiseks on vajalik videokaardi olemasolu, kuna nii närvivõrkude käitamine kui ka segmentatsiooni ja tausta eemaldamise töötlus kasutavad videokaardi ressursse. Lisaks on oluline kvaliteetse veebikaamera olemasolu.

Süsteemi arendamisel kasutati NVIDIA GeForce RTX 3060 Ti videokaarti ning saavutati jõudlus 60-80 kaadrit sekundis (FPS). Närvivõrgu protsessi põhjustatud viivitus jäi vahemikku 15-25 millisekundit, teistest protsessidest lisandub alla 10 ms. Need tulemused näitavad, et süsteem on optimeeritud piisavalt hästi, et pakkuda sujuvat reaalajas kasutuskogemust ka kesktaseme riistvaral.

3.2 Tulemused

Süsteem genereerib kolm erinevat reaalajalist visuaalset väljundit (vt joonis 17), mis regulaarselt vahetuvad:

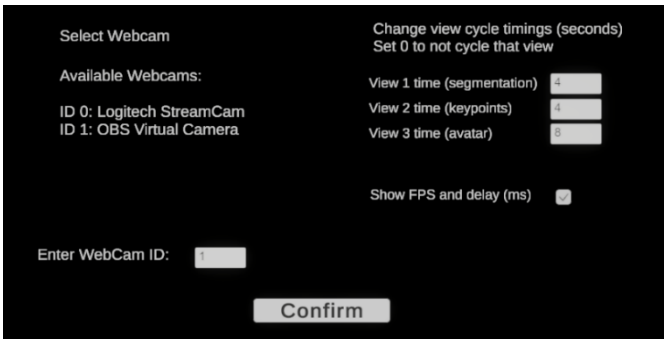
- Inimene eemaldatud ja asendatud tema liikumist jäljendava 3D avatariga.
- Inimesele on kuvatud segmentatsiooni-mask koos võtmepunktidega.
- Inimene eemaldatud ja asendatud teda kujutavate võtmepunktidega.



Joonis 17. Visuaalsed väljundid.

Programmi käivitamisel on kasutajaliidese (vt joonis 18) kaudu võimalik valida kaamera allikas, et süsteemi seadistajal oleks võimalik valida õige sisend mitme kaamera olemasolul. See on kasulik ka olukorras, kus kaamera sisend on suunatud läbi OBS VirtualCamera[49], kust kaudu on võimalik kaamera sätteid muuta. Lisaks on võimalik valida, kui kaua mõni vaade ajaliselt kestab. Pärast nende valikute tegemist töötab programm kasutajaliideseta.

[49] https://obsproject.com/kb/virtual-camera-guide



Joonis 18. Käivitamise kasutajaliides.

Autori ning juhendaja soov oli, et kasutajad ei omaks hiire või klaviatuuriga ligipääsu arvutile, sest süsteem on loodud töötama avalikus keskkonnas. Seega vahelduvad need kolm vaadet regulaarselt omavahel, kasutaja ei saa vaadet valida. Lahenduse tööd demonstreerivad videoklipid on kättesaadavad lisade I sektsioonis asuva repositooriumi lingi kaudu. Samuti leiab sealt installeerimis- ja kasutusjuhendi.

3.3 Testimine

Süsteemi töökindluse testimiseks ja kasutajapoolseks hindamiseks seati see üles avaliku demona Tartu Ülikooli Delta hoone 3. korrusel ajavahemikul 21.–24. märts 2025. Tegemist oli suure avatud alaga, kus tudengitel ja töötajatel oli võimalik demo testida. Nimetatud ajavahemikul toimis süsteem pidevalt ega vajanud peale paigaldamist kõrvalist sekkumist, millest saab järeldada süsteemi töökindlust. Oluline on mainida, et demo oli lõplikule versiooniga väga sarnane, kuid erinevate vaadete kuvamine polnud veel implementeeritud ning kasutajad nägid vaid enda asendamist avatariga.

Joonis 19. Kaader demost.

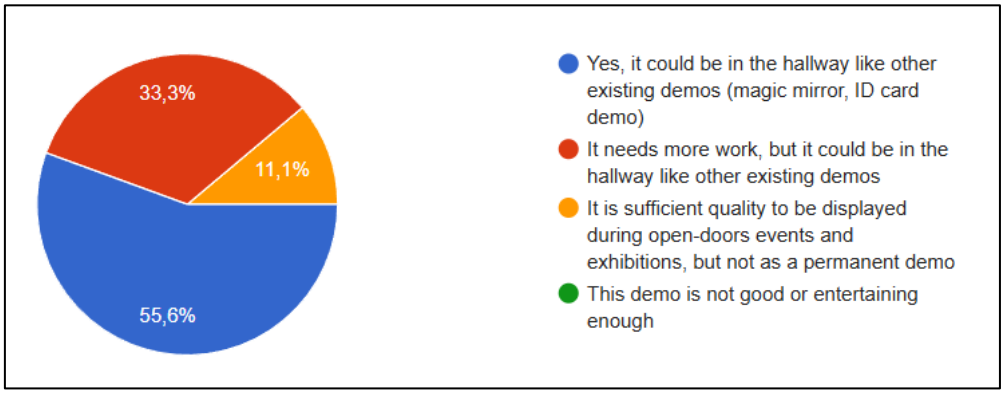
Süsteem seati üles Nvidia GeForce RTX 2070 videokaardiga arvuti peal, mis jõudluse poolest andis võrreldava tulemuse autori poolt testitud Nvidia RTX GeForce 3060 Ti arvutiga. Monitorile fikseeriti Logitech StreamCam veebikaamera, millel lülitati välja automaatne valge tasakaal, säritus ja fookus. Fikseeritud kaamerasätted osutusid oluliseks kvaliteetse tausta eemaldamise saavutamiseks, kuna need tagasid taustapikslite stabiilsuse dünaamilise taustamudeli jaoks ja aitasid saavutada kvaliteetsema inimese eemaldamise. Kasutajad nägid monitorist reaalajalist väljundit, mille näitlikustamiseks on toodud välja ka kaader (joonis 19).

Testimisel tuli arvestada mitme piiranguga. Esiteks on süsteem loodud vaid ühe inimese tuvastamiseks. Juhul, kui mitu inimest on korraga kaadris, võib poosituvastuse mudel hakata vahetama fookust erinevate inimeste vahel. Teiseks on oluline, et inimene ei oleks kaamerale liiga lähedal ega liiga kaugel. Sobivaks distantsiks osutus umbes 2 kuni 5 meetrit. Sellisel distantsil on inimene piisavalt kaugel, et täielikult kaadrisse mahtuda, ning piisavalt lähedal, et poosituvastuse mudel suudab kõrge kindlusega kõik võtmepunktid õigesti tuvastada ja neid jälgida. Selleks, et kasutajad saaksid süsteemi võimalikult õiglaselt hinnata, asetati demo kõrvale ka juhend (vt Lisad II – Demo kasutusjuhend), mis kirjeldas nimetatud piiranguid.

3.4 Kasutajate tagasiside

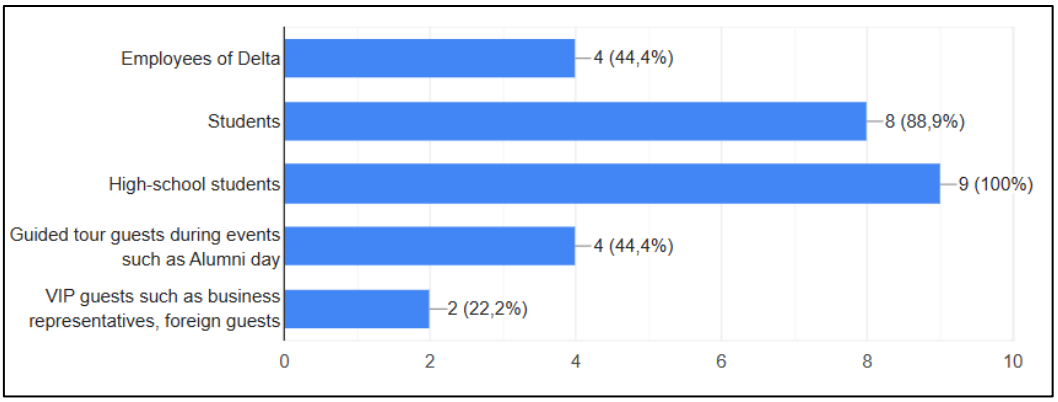
Delta hoones toimunud demo testimisel oli kasutajatel võimalik anda tagasisidet anonüümse Google Forms veebiküsitluse kaudu. Küsimused on täies mahus välja toodud lisade III sektsioonis. Et paljud instituudi töötajad ei räägi eesti keelt, oli küsitlus inglise keelne. Küsitlusele vastas kokku 9 inimest ning üldine tagasiside oli positiivne

Kõigepealt küsiti arvamust selle kohta, kas demo oleks huvitav eksponaat Delta hoone koridori jaoks (vt joonis 20). Vastanutes hulgast 56% leidis, et see sobiks Delta koridoridesse püsieksponaadiks sarnaselt olemasolevate eksponaatidega. Kolmandik arvasid, et see vajaks enne püsivat paigaldamist veel arendustööd. Üks vastaja veel arvas, et see pole püsiva eksponeerimise jaoks piisavalt huvitav või kvaliteetne, kuid sobiks eriüritusteks nagu lahtiste uste päevad.



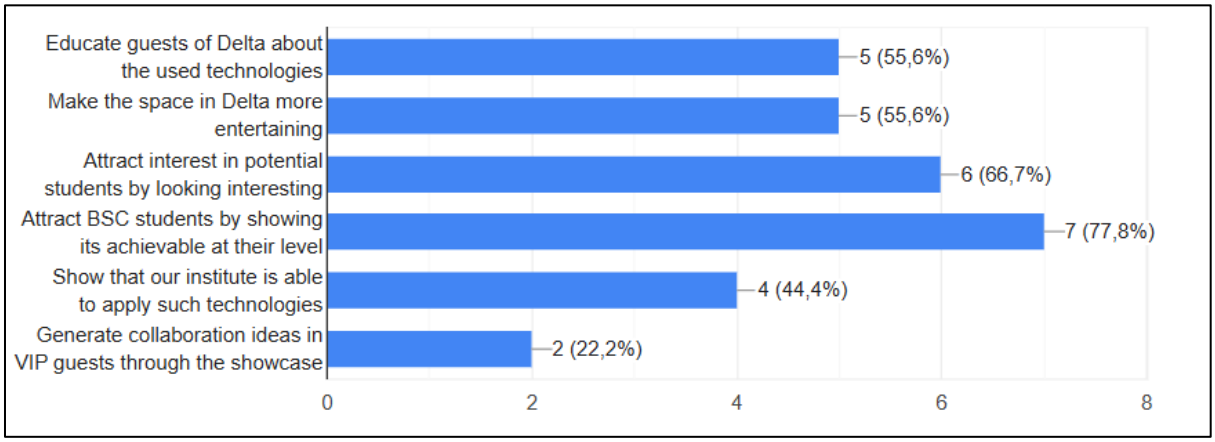
Joonis 20. Kas demo sobiks Delta koridori?

Edasi küsiti, mis sihtrühmadele eksponaat kõige enam huvi või meelelahutust pakuks (vt joonis 21). Enamik tõid sihtrühmadena välja tudengid ja gümnaasiumiõpilased. Pea pooled tõid välja ka Delta töötajaid, giidiga ekskursioonide külalisi (nt vilistlaspäevadel) ning 2 vastajat pakkusid veel välja VIP-külalisi nagu äriesindajad või väliskülalised.



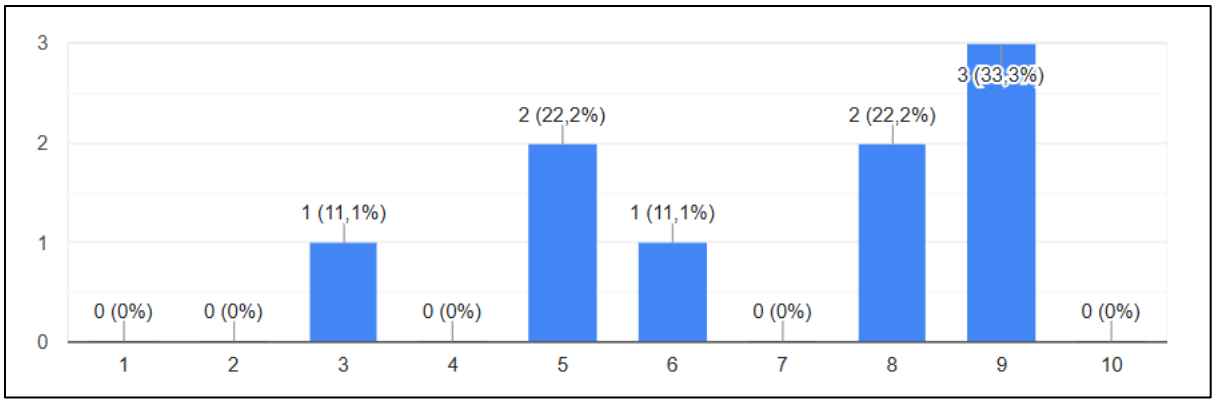
Joonis 21. Millistele Delta külalistele eksponaat kõige enam huvi või meelelahutust pakuks?

Järgmisena küsiti, mis eesmärke see Deltas püsieksponaadina täidaks (vt joonis 22). Vastanutest suurim osa ehk 7 olid nõus väitega, et see oleks huvitav informaatika bakalaureuse õppeastme tudengitele, näitamaks, et selline projekt on teostatav nende haridustasemel. Enamik vastanutest veel valisid, et sellel oleks hariduslik mõju kasutatava tehnoloogia kohta (poosituvastus, pilditöötlus), ning et see tõmbaks tähelepanu interaktiivsusega ja teeks Delta avaliku ruumi huvitavamaks. Enamik aga ei olnud valinud väidet, et eksponaat demonstreeriks Arvutiteaduse Instituudi ja valdkonna võimekusi, ning et eksponaat võiks luua külalisetes koostööideid.

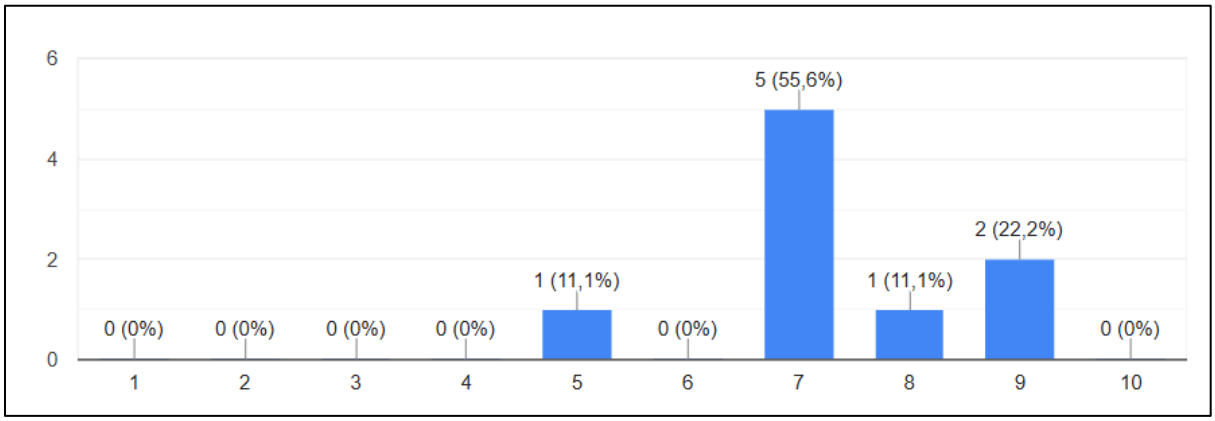


Joonis 22. Mis eesmärke see püsieksponaadina täidaks?

Kasutajatel oli võimalik ka hinnata demo kvaliteeti 10-palli skaalal (vt joonised 23 ja 24). Eraldi sai hinnata inimese eemaldamise ja avatari animeerimise/poosituvastuse kvaliteeti, kus 1 tähendas väga halba tulemust, 10 väga head. Jooniste järgi selgub, et mõlema kvaliteeti suures jaos heaks või rahuldavaks. Inimese eemaldamise kvaliteedi keskmine hinnang oli 6.9 ning avatari animeerimise keskmine hinnang oli 7.3.



Joonis 23. Kuidas hindad inimese eemaldamise kvaliteeti?



Joonis 24. Kuidas hindad poosituvastuse ja avatari animeerimise kvaliteeti?

Edasi oli vastajatel vabas vormis võimalik kirjeldada, kuidas süsteemi võiks edasi arendada ja mida saaks juurde lisada. Paljud vastasid, et süsteem võiks töötada mitme inimesega ning mitu erinevat avatari võiks olla saadaval. Toodi ka näiteid, milliseid tegelasi võiks avatarina kuvada. Veel lisati, et võtmepunktide tuvastus võiks olla täpsem ning et näpud võiksid samuti animeeritud olla.

3.5 Järeldused

Tagasiside kinnitas ideed, et süsteem sobib eksponaadiks Delta hoonesse ning see oleks huvitav nii noortele õpilastele kui ka töötajatele. Kasutajad tõid välja, et süsteem täidaks mitmeid erinevaid eesmärke, omades selget ekspositsioonilist, meelelahutuslikku ja hariduslikku väärtust. Süsteem oli töökindel ning kasutajate hinnangul hea või rahuldava kvaliteediga. Nimetatud asjaolude põhjal saab eeldada, et projekt täitis kõik püstitatud eesmärgid, pakkudes haaravat interaktiivset kogemust kaasaegse tehisnägemise rakenduse näol.

Kasutajate soovid nagu mitme inimese tuvastus, näppude animeerimine ja erinevate avataride kuvamine jäävad paraku projekti mahust välja. Mitme inimese ja sõrmede animeerimine on

tulevikus teostatav - seda võimaldavaid närvivõrgumudeleid katsetati projekti jooksul, kuid nendega kaasnenud kitsaskohti jõudluse ja töövoogu integreerimise osas ei jõudnud autor käesoleva töö raames lahendada. Samuti on mitme erineva avatari kuvamine tulevikus teostatav. Avataride vahel vahetamiseks oleks kõige tarvilikum eksponaadile lisada mingi sisendseade, mille nupulevajutused avatare ja/või vaateid vahetaks. Seevastu tehti täiustus erinevate vaadete kuvamise näol (vt joonis 17), mis iseloomustab süsteemi töövoogu ning seeläbi tõstab hariduslikku väärtust.

Kokkuvõte

Käesoleva lõputööga loodi interaktiivne süsteem, mis kasutab tehisnärvivõrgu mudelid inimese poosi tuvastamiseks ja reaalajas 3D avatariga asendamiseks. Süsteem on üles ehitatud Unity mängumootoriga ning kasutab Google MediaPipe'i Pose Landmarker poosituvastuse mudelit, mis tuvastab inimese võtmepunktid ja segmentatsiooni. Avatar animeeritakse vastavalt tuvastatud võtmepunktidele ning inimene eemaldatakse segmentatsioonimaski töötlemise kaudu.

Lahendus on üles ehitatud nii, et see toimiks sujuvalt reaalajas keskklassi videokaardiga varustatud arvutil, tagades piisava kaadrisageduse (60 FPS) ja väikese viivituse (25–35 ms).

Süsteemi testiti avaliku demona Tartu Ülikooli Delta hoones, kus kasutajad said tagasisidet anda süsteemi toimimise kohta. Tagasiside oli positiivne, ning toodi välja süsteemi potentsiaal püsieksponaadina. Samuti tõsteti esile võimalikku atraktiivsust sihtrühmade nagu tudengite ja gümnaasiumiõpilaste jaoks. Süsteemi arendamise lõppfaasis lisati ka erinevate vaadete kuvamise võimalus, mis demonstreerivad süsteemi töövoogu interaktiivselt, lisades hariduslikku väärtust. Töö juhendaja hinnangul on tõenäone, et loodud lahendusest saab eksponaat, mida Delta hoones ka tulevikus näha saab.

Töö tulemused kinnitavad, et selline interaktiivne segareaalsuse lahendus on teostatav ja pakub eksponaadina haaravat meelelahutuslikku ja hariduslikku väärtust, demonstreerides tehisnägemise valdkonna võimekusi. Tulevikus oleks huvitav teostada projekt, mis töötaks mitme inimese korral ning suudaks teisendada ka näppude ja näo liigutused avatarile. Üks alternatiivne suund on luua ka ainult näo avatar. Praegusele programmile saaks lisada ka erinevate avatarid kuvamise, või võimaldada kasutajatel luua ise avatar. Kaugemas tulevikus oleks huvitav näha taolist projekti mobiilseadmetes kasutades liitreaalsuse platvorme.

Viidatud kirjandus

[1] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, ja M. Shah, „Deep Learning-Based Human Pose Estimation: A Survey“, 3. juuli 2023, *arXiv*: arXiv:2012.13392. doi: 10.48550/arXiv.2012.13392.

[2] M. Kari, T. Grosse-Puppendahl, L. F. Coelho, A. R. Fender, D. Bethge, R. Schutte, ja C. Holz, „TransforMR: Pose-Aware Object Substitution for Composing Alternate Mixed Realities“, *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Bari, Italy: IEEE, okt 2021, lk 69–79. doi: 10.1109/ISMAR52148.2021.00021.

[3] H. Ke ja H. Wang, „Poster: Real-Time Object Substitution for Mobile Diminished Reality with Edge Computing“, *Proceedings of the Eighth ACM/IEEE Symposium on Edge Computing*, dets 2023, lk 279–281. doi: 10.1145/3583740.3628422.

[4] H. Chen, X. Jiang, ja Y. Dai, „Shift Pose: A Lightweight Transformer-like Neural Network for Human Pose Estimation“, *Sensors*, kd 22, nr 19, Art. nr 19, sept 2022, doi: 10.3390/s22197264.

[5] A. Toshev ja C. Szegedy, „DeepPose: Human Pose Estimation via Deep Neural Networks“, *2014 IEEE Conference on Computer Vision and Pattern Recognition*, juuni 2014, lk 1653–1660. doi: 10.1109/CVPR.2014.214.

[6] D. Ramanan, „Learning to parse images of articulated bodies“, *Advances in Neural Information Processing Systems*, MIT Press, 2006. Vaadatud: 6. mai 2025. [Online]. Available at: https://papers.nips.cc/paper_files/paper/2006/hash/a209ca7b50dcaab2db7c2d4d1223d4d5-Abstract.html

[7] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, ja K. Chen, „RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose“, 3. juuli 2023, *arXiv*: arXiv:2303.07399. doi: 10.48550/arXiv.2303.07399.

[8] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, ja B. Schiele, „DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model“, 30. november 2016, *arXiv*: arXiv:1605.03170. doi: 10.48550/arXiv.1605.03170.

[9] M. Ben Gamra ja M. A. Akhloufi, „A review of deep learning techniques for 2D and 3D human pose estimation“, *Image and Vision Computing*, kd 114, lk 104282, okt 2021, doi: 10.1016/j.imavis.2021.104282.

[10] H. Yunusa, S. Qin, A. H. A. Chukkol, A. A. Yusuf, I. Bello, ja A. Lawan, „Exploring the Synergies of Hybrid CNNs and ViTs Architectures for Computer Vision: A survey“, 5. veebruar 2024, *arXiv*: arXiv:2402.02941. doi: 10.48550/arXiv.2402.02941.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, ja I. Polosukhin, „Attention Is All You Need“, 2. august 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, ja N. Houlsby, „An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale“, 3. juuni 2021, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[13] R. B. Neupane, K. Li, ja T. F. Boka, „A survey on deep 3D human pose estimation“, *Artificial Intelligence Review*, kd 58, nr 1, Art. nr 1, nov 2024, doi: 10.1007/s10462-024-11019-3.

[14] Q. Nie, Z. Liu, ja Y. Liu, „Lifting 2D Human Pose to 3D with Domain Adapted 3D Body Concept“, *International Journal of Computer Vision*, kd 131, nr 5, Art. nr 5, veebr 2023, doi: 10.1007/s11263-023-01749-2.

[15] V. Patel, J. Chen, ja A. B. Koranteng, „Lifting 2D Keypoints to 3D Human Pose Estimation“, 2022, Vaadatud: 5. mai 2025. [Online]. Available at: https://api.semanticscholar.org/CorpusID:250406572

[16] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, ja M. Grundmann, „BlazePose: On-device Real-time Body Pose tracking“, 17. juuni 2020, *arXiv*: arXiv:2006.10204. doi: 10.48550/arXiv.2006.10204.

[17] G. Wang, Z. Li, G. Weng, ja Y. Chen, „An overview of industrial image segmentation using deep learning models“, *OAE Publishing Inc.*, kd 5, veebr 2025, doi: 10.20517/ir.2025.09.

[18] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, ja D. Terzopoulos, „Image Segmentation Using Deep Learning: A Survey“, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, kd 44, nr 7, lk 3523–3542, juuli 2022, doi: 10.1109/TPAMI.2021.3059968.

[19] J. Hu, L. Huang, T. Ren, S. Zhang, R. Ji, ja L. Cao, „You Only Segment Once: Towards Real-Time Panoptic Segmentation“, 26. märts 2023, *arXiv*: arXiv:2303.14651. doi: 10.48550/arXiv.2303.14651.

[20] K. He, G. Gkioxari, P. Dollár, ja R. Girshick, „Mask R-CNN“, 24. jaanuar 2018, *arXiv*: arXiv:1703.06870. doi: 10.48550/arXiv.1703.06870.

[21] S. Tanberk, D. B. Tukel, ja K. Acar, „The Design of a 3D Character Animation System for Digital Twins in the Metaverse“, 10. juuli 2024, *arXiv*: arXiv:2407.18934. doi: 10.48550/arXiv.2407.18934.

[22] S. Snyder, „Mixed reality.“, *Salem Press Encyclopedia of Science*. Salem Press, 25. aprill 2025. Vaadatud: 5. mai 2025. [Online]. Available at: https://research.ebsco.com/linkprocessor/plink?id=5d7ae968-f1d2-3ee5-a57a-a91135fcf133

[23] A. L. Simeone, „Substitutional Reality“, 2018. doi: 10.1007/978-3-319-08234-9_254-1.

[24] T. Jiang, X. Xie, ja Y. Li, „RTMW: Real-Time Multi-Person 2D and 3D Whole-body Pose Estimation“, 11. juuli 2024, *arXiv*: arXiv:2407.08634. doi: 10.48550/arXiv.2407.08634.

[25] S. Dill, A. Rösch, M. Rohr, G. Güney, L. D. Witte, E. Schwartz, ja C. H. Antink, „Accuracy Evaluation of 3D Pose Estimation with MediaPipe Pose for Physical Exercises“, sept 2023, doi: 10.1515/cdbme-2023-1141.

[26] G. Casiez, N. Roussel, ja D. Vogel, „1 € filter: a simple speed-based low-pass filter for noisy input in interactive systems“, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’12. New York, NY, USA: Association for Computing Machinery, mai 2012, lk 2527–2530. doi: 10.1145/2207676.2208639.

Lisad

I. Lähtekood

https://github.com/martenjaani/poseanim

Antud lingi kaudu pääseb ligi projekti avalikule lähtekoodile, mille *README* sisaldab demonstratiivseid videoklippe, installeerimis- ja kasutamisjuhendit ning lähtekoodi kirjeldust.

II. Demo kasutusjuhend

Siin on välja toodud kasutusjuhend, mis oli Delta õppehoones toimunud avaliku testimise ajal eksponaadi kõrval.

Try the “PoseAnimator Demo"

This is a work-in-progress result of a BSc thesis, please give feedback via the QR code (Google Forms questionnaire, 8 questions).

HOW TO USE THE DEMO?

1. Stand at **least 2 meters** from the camera, so that your legs are in the frame at least from below the knees. The round metal socket hole on the ground is a good marker behind which to stand
2. For best results, there should be **only one person in the frame**
3. If you see the 3D avatar matching your movements, it is working.
4. Do not go too far (**not beyond 5 meters**, far edge of the ping-pong table).
5. If the segmentation mask doesn't fill (i.e you see your ghost), then moving side to side helps. This happens because the camera changes exposure due to your distance from the camera among other factors.

FYI technologies used: Unity, 3D human pose detection, human segmentation, 3D avatar animation.

III. Küsimustik

Feedback on BSc thesis project "PoseAnimator Demo"

Please answer these 10 short questions to evaluate the quality and give feedback to the project for the next iteration.

1. Do you think this demo would be an interesting exhibit in the hallways of Delta?
    a. Yes, it could be in the hallway like other existing demos (magic mirror, ID card demo)
    b. It needs more work, but it could be in the hallway like other existing demos
    c. It is sufficient quality to be displayed during open-doors events and exhibitions, but not as a permanent demo
    d. This demo is not good or entertaining enough
2. Which guests of Delta do you think this demo would interest or be entertaining for?
    a. Employees of Delta
    b. Students
    c. High-school students
    d. Guided tour guests during events such as Alumni day
    e. VIP guests such as business representatives, foreign guests
3. If this demo was placed permanently in Delta, what goals would it serve in your opinion?
    a. Educate guests of Delta about image processing technologies (would it teach anything new?)
    b. Make the space in Delta more entertaining
    c. Attract interest in this technology in potential students by looking interesting
    d. Attract interest in students by showing it is achievable at their competence level (at Bsc level)
    e. Demonstrate that our institute is capable of applying such technologies (is this demonstration is needed?)
    f. Generate collaboration ideas in VIP guests by demonstrating the technology
4. From a scale of 1 to 10, rate the quality of the removal of you from the image?
    a. The ghost that sometimes appears dirturbed me a lot (1)
    b. I did not see the "ghost" or it did not disturb me at all (10)

5. From a scale of 1 to 10, rate the quality of the pose detection and avatar animation?
    a. my movements were missed and/or displayed with a lag (1)
    b. my movements were perfectly detected and tracked in real time (10)
6. Give feedback on the current UI and printed "how to use" guidelines. What could make it better?
7. What are some ways you would like to see it developed further (i.e multiple people, multiple avatars, just the pose stick figure etc)?
8. If you could switch avatars by going out of the frame, which types of avatars would you like to see the most?

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Marten Mathias Jaani ,

(autori nimi)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

„Inimese asendamine avatariga kasutades poosituvastust – interaktiivne peegel“ ,

(lõputöö pealkiri)

mille juhendaja(d) on Ardi Tampuu ,

(juhendaja nimi)

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
3. olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
4. kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Marten Mathias Jaani

10.05.2025