RAPORT STIINTIFIC

privind implementarea proiectului in perioada iulie 2022 – decembrie 2023

Proiect PN-III-P4-PCE-2021-1959

Acronim: FIGHS

Metode Detaliate de Reconstructie Tridimensionala a Persoanelor

Institutul de Matematica `Simion Stoilow' al Academiei Romane

Grupul de Cercetari in Vedere Artificiala si Invatare Computationala

Director de proiect: Prof-univ. dr. Cristian Sminchisescu

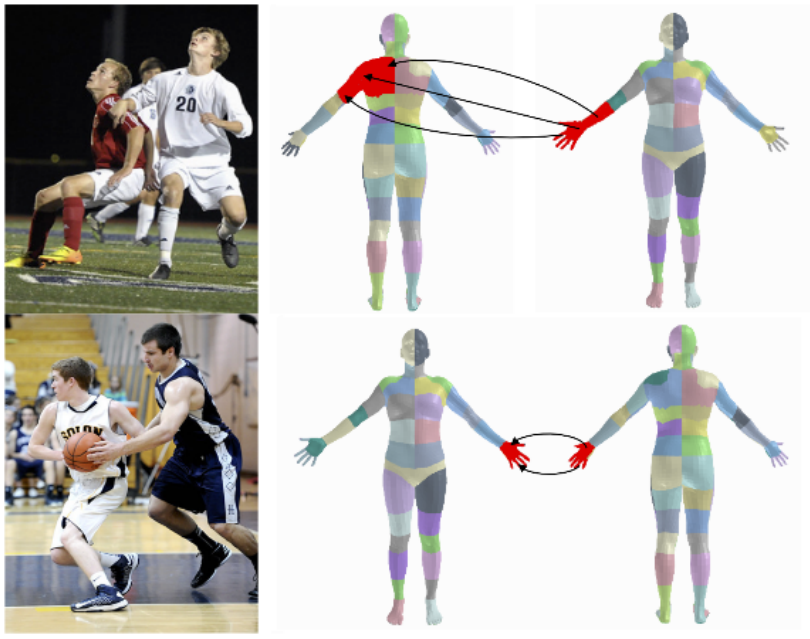
<u>Activitati stiintifice si realizari</u>:

Înțelegerea tridimensionala a interactiunilor umane este fundamentala pentru analiza detaliata a scenei 3d și pentru modelarea comportamentului uman. Cele mai multe dintre modelele existente se concentreaza pe analiza unei singure persoane izolate, iar acelea care proceseaza mai multe persoane pun accentul pe rezolvarea asocierii intre multiple persoane, mai degrabă decât sa înteleaga interactiunile dintre ele. Acest lucru duce la estimari care, chiar și atunci când sunt impresionante în ceea ce privește postura și forma plauzibile de la distanță, ratează esența evenimentului la un scrutin de detaliu, atunci când, de exemplu, două reconstrucții nu reușesc să surprindă contactul în timpul unei strângeri de mână, o atingere pe umăr, sau o îmbrățișare. Astfel de interacțiuni sunt deosebit de dificil de rezolvat, deoarece efectele lor se compun: pe de o parte, incertitudinea privind adâncimea și forma corpului ar putea duce la compensare prin împingerea membrelor în față sau mai departe de poziția lor de adevăr la sol, atunci când se face inferenta 3d din imaginile monoculare; pe de altă parte, ocluzia parțială și detaliul (rezoluția) relativ limitate pentru zonele de contact din imagini, tipice multor interacțiuni umane, pot face dovezile vizuale neconcludente. Astfel, aceste modele conduc la estimari 3d incorecte, nerealiste care pierd din vedere aspectele subtile ale contactului uman și sunt prea puțin utile intelegerii comportamentului uman din imagini.

Într-un articol aflat în evaluare la IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), aliniat cu obiectivele FIGHS, sunt investigate și propuse urmatoarele contributii: (1) modele pentru estimarea semnaturii interactiunilor care cuprind dectectia contactului, segmentarea lui și predictia semnaturii 3d a contactului; (2) demonstrarea utilitatii unor asemenea modele pentru a produce funcții de loss extinse pentru a asigura consistenta reprezentarii contactului uman în procesul de reconstructie 3d a persoanelor; (3) propunerea unei metodologii pentru recuperarea posturii și formei ground-truth a persoanelor într-un mediu controlat; (4) prezentarea unor modele și algoritmi baseline pentru a ilustra felul în care estimarea contactului uman sprijina reconstructia 3d superioara a persoanelor atunci când sunt capturare interacțiuni esentiale intre ele.

Articolul propune un prim set de elemente metodologice pentru a aborda reconstrucția persoanelor care interacționează, într-o manieră mai principială, bazându-se pe recunoaștere, segmentare, mapare și reconstrucție 3d. Mai precis, problema furnizarii de reconstrucții 3d veridice ale persoanelor care interacționează este descompusa în (a) detectia contactului, (b) segmentarea binară a regiunilor de contact pe suprafețele corespunzătoare asociate oamenilor care interacționează; (c) predicția semnăturii contactului pentru a produce estimări ale hărții potențiale de corespondență many-to-many dintre regiunile aflate în contact; și (d) reconstrucție 3d folosind funcții de loss extinse construite folosind constrângeri suplimentare de contact de suprafață, atunci când exista o semnătură de contact. Cu ajutorul unui set extins de experimente, s-au evaluat toate componentele sistemului și s-au oferit comparații cantitative și calitative care arată modul în care abordarea propusă poate surprinde interacțiunile umane reprezentate 3d în mod realist.

Pentru primele trei task-uri de mai sus (a-c) s-au folosit metode de invatare bazate pe retele neuronale profunde care primesc ca date de intrare bounding box-urile a doua persoane care interactioneaza împreună cu posturile 2d detectate. Clasificarea contactului presupune estimarea existenței unui contact fizic intre cele doua persoane din imagine, care se face cu o rețea de clasificare binara care prezice existența contactului sau a absentei lui. Segmentarea contactului și predictia semnaturii de contact se fac la nivel de regiune și sunt antrenate împreuna o rețea mulți-task care folosește adnotari ground-truth despre contact și segmentarea sa, reprezentabile ca în figura de mai jos în care se poate observa cum adnotatorii marcheaza regiuni ale corpurilor celor doua persoane aflate în contact. Corespondentele intre regiuni pe care ei le marcheaza pe mesh-urile 3d ale persoanelor din imagine constitutie o segmentare automata a regiunilor aflate în contact.



Odată estimata semnatura contactului dintre doua persoane aflate într-o imagine, aceasta informație poate fi folosită pentru a reconstrui postura 3d și parametrii mesh-urilor celor doua persoane prin extinderea unui framework de reconstructie 3d a mai multor persoane dintr-o imagine cu funcții de loss care folosesc semnatura de contact. În primul rând, având în vedere ca mâinile sunt cel mai adesea implicate în interacțiuni de contact, s-a folosit modelarea lor pentru a augmenta functiile de loss care cuantifica eroarea de reproiectie 2d fata de evidenta vizuala extrasa din imagine sub forma posturii 2d și a etichetarii semantice a partilor corpului, respectiv functia de regularizare a posturii și formei mesh-ului. Suplimentar a fost introdus un nou termen de loss specific informatiei de contact care măsoară alinierea geometrica a regiunilor puse în corespondenta. Mai exact, termenul încearcă sa minimizeze suma distantelor dintre toate perechile de regiuni aflate în contact și, respectiv, sa forteze alinierea orientarilor suprafetelor acestor regiuni.

Seturile de date folosite au fost extinse prin adnotarea întregii perioade temporale a contactului fizic în fiecare secvență video. De asemenea, metodologia propusa pentru a obține postura și forma ground-truth a persoanelor care interacționează depășește simpla extrapolare a pozițiilor markerilor 3d la parametrii unui model de forma a corpului, deoarece: (a) doar unui singur subiect din fiecare videoclip ii este urmărita mișcarea cu ajutorul sistemului de captura a miscarii și (b) subiecții inregistrati nu sunt doar apropiați unul de celalalt, dar sunt, de cele mai multe ori, chiar în contact fizic. În plus, sistemul de captura a miscarii nu permite urmarirea articulatiei mainii în cea mai mare parte a ei. Prin valorificarea informațiilor de la senzorii de detectie a mișcarii, a celor de la mai multe camere RGB și respectiv a celor de la un scaner 3d, dar și din adnotări de contact, informații a priori despre postura corpului si constrângeri fizice, s-au realizat reconstrucții 3d similare dpdv al corectitudinii cu cele rezultate din reprezentarile ground-truth (vezi fig. de mai jos).

Pentru atingerea acestui tel a fost nevoie de un model de optimizare care sa ia în considerare suma unor termeni de loss care sa cuantifice diversele constrangeri mentionate mai sus. În primul rând, scanuri 3d ale subiectilor au fost folosite pentru a potrivi un model parametrizat (GHUM) la ele. Diferența dintre reconstructia 3d folosind markeri ground-truth și cea rezultata din GHUM a fost folosită doar pentru subiectul urmărit de sistemul mocap. În 2d, functia de loss aferenta a cuantificat pentru fiecare dintre cele 4 viewpoint-uri RGB ale scenei eroarea dintre proiectia 2d a articulatiilor 3d estimate ale modelului GHUM și reprezentarile 2d ale articulatiilor estimate direct din imagine (pentru subiectul cu markeri, estimarea se face într-o zona decupata strâns în jurul subiectului, zona obtinuta prin proiectarea markerilor 3d ground-truth în imagine folosind parametrii cunoscuți ai camerei; pentru celelalt subiect s-a folosit un detector de keypoints pe întreaga imagine). Pentru prevenirea auto-intersectiilor și respectiv a interpenetrarii a doua mesh-uri s-au folosit teste de tip generalized winding numbers pentru a stabili relația vertex-urilor interioare unui mesh fata de cele mai apropiate vertex-uri vecine (nearest neighbors) fie din propriul mesh (self-collision) fie din celălalt mesh (interpenetration). De asemenea, s-a exercitat modelarea corectului contact cu solul ca și constrangerea impusa de o funcție de loss care cuantifica semnatura de contact bazata pe interpretarea fatetelor în contact ale celor doua mesh-uri.

Pentru a avea o reprezentare calitativa a importantei modelarii interactiunilor intre persoane atasam mai jos reconstructii 3d ale unor persoane aflate în interacțiune conform unor varii scenarii. Prima coloana prezinta imaginile RGB, urmata de coloana a doua care prezinta reconstructia 3d fără informație de contact (interacțiune). Coloanele 3 și 4 prezinta rezultatele reconstructiei 3d folosind informatie de contact la nivel de regiuni (abordare coarse-grain) la diferite granularitati ((35 și respectiv 75 de regiuni utilizate). Ultima coloana prezinta rezultatele reconstructiei 3d atunci când se folosește informație de contact la nivel fine-grain, mai exact la nivel de fatete de mesh 3d. După cum se observa, chiar dacă folosirea constrangerilor la nivel de fateta produce cele mai bune rezultate, rezultate rezonabile se obțin și la un nivel de utilizare coarse-grain al informatiei de contact.

De asemenea, s-au adnotat mișcările de interacțiune cu descrieri text deschizand astfel calea pentru antrenarea și evaluarea modelor care genereaza miscari de interacțiune 3d din text. Fiecarei perechi de mesh-uri în interacțiune i-au fost asociate adnotari text cu o lungime medie de 12 cuvinte în engleza care ofera și posibilitatea unor augmentari suplimentare (de ex, schimbarea ordinii miscarilor celor doua persoane fără a schimba adnotarea textuala). Un exemplu de adnotare poate fi văzut mai jos:

A man holds his right arm around somebody's shoulder and raises his left hand for a picture.

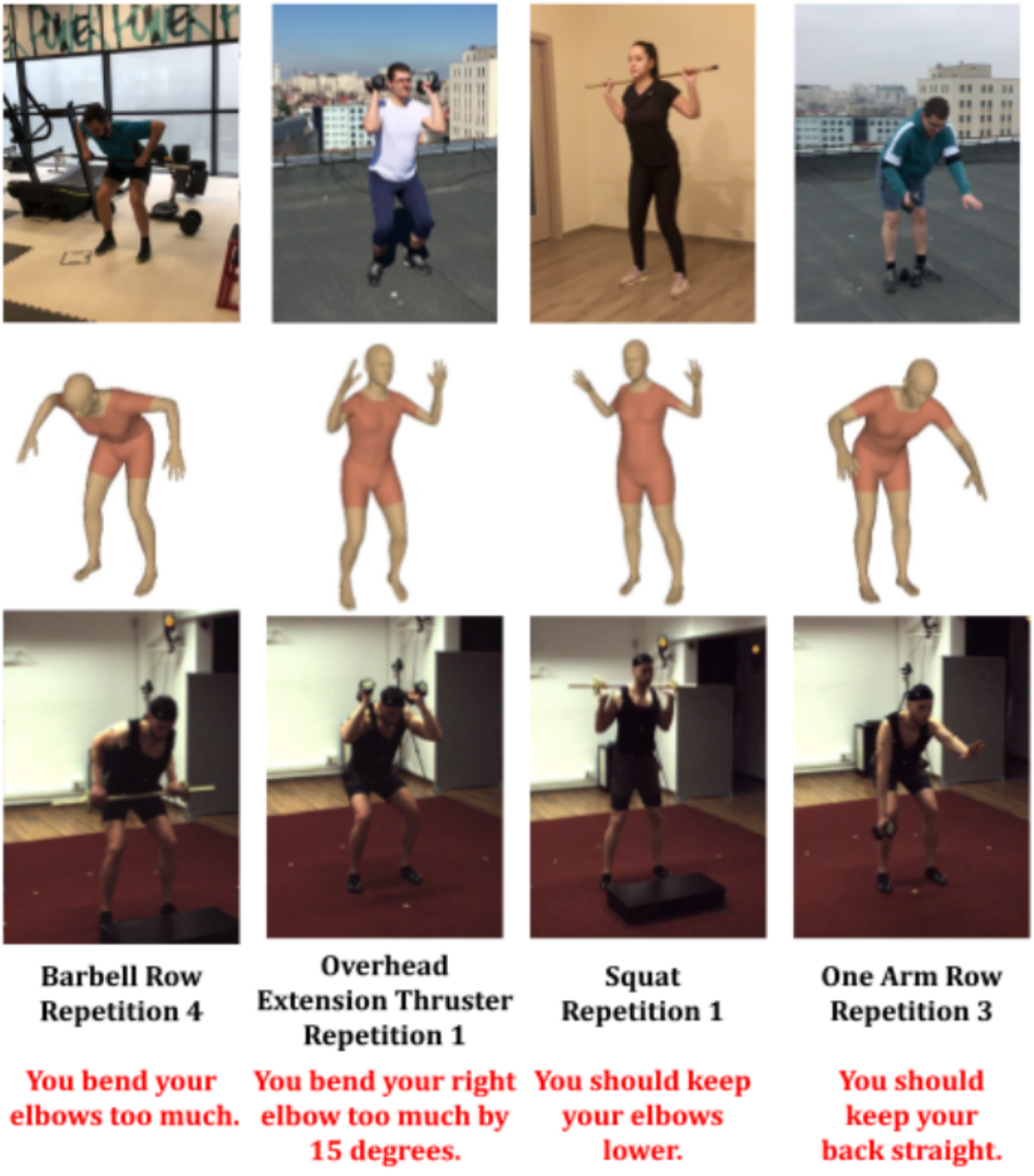
Pe lângă secvențele comune 3d, articolul face publice secvențele de mișcare ground-truth atat în format GHUM cât și SMPLX și pune la dispoziția comunității de cercetare un server de evaluare cu set de test ascuns, împreună cu un benchmark public, cu scopul de a avansa stadiul cunoasterii în reconstrucția 3d a persoanelor aflate în interacțiune.

Dezvoltarea recenta spectaculoasă a modelelor lingvistice pe scara larga (Large Language Models, LLM) de tipul GPT (Generative Pre-trained Transformer) și a modelelor derivate text-to-image de tip DALL-E care se bazează pe ele au motivat așa cum spuneam mai sus efortul de augmentare a inregistrarilor miscarilor de interacțiune 3d ale persoanelor cu text descriptiv, cu intenția de a putea genera miscari de interacțiune 3d din descrieri text. În acest context, am trimis spre publicare la IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) un articol care prezinta modelarea reconstructiei 3d a persoanelor care fac fitness. Articolul prezinta un sistem automat care poate fi folosit acasă, la sala sau în aer liber pentru a reconstrui postura, forma și mișcarea persoanei, segmenteaza repetarea exercitiilor cu acuratete și identifica în timp real deviatiile miscarilor persoanei fata de mișcările standard recomandate de antrenori calificati (instructori de fitness). Ca rezultat, sistemul ofera feed-back cantitativ pentru executia corecta a exercitiilor ceea ce reduce riscul de accidentare și permite imbunatatirea continua a performatei. Pentru a sprijini cercetarea și evaluarea, articolul introduce un set de date pe scara larga cu peste 3 milioane de imagini și configuratiile aferente de captura a miscarii pentru 37 de exerciții de fitness executate deopotriva de amatori/elevi și instructori. Sistemul ofera asistenta unui antrenor virtual statistic care evalueaza critic performanța elevilor pe baza unui parametru care permite atât adaptarea recomandarilor critice nivelului de pregătire al elevului (incepator, avansat, expert) cât și ajustarea acuratetei estimate a metodei de reconstructie 3d a posturii umane. Antrenorul virtual ofera feed-back în limbaj natural sprijinit de evidenta spatio-temporala în imaginile video ale exercitiilor elevului.

Metodologia propusa în articol include evaluarea monoculara, multi-view pe scara larga a reconstructiei 3d a posturii umane, modele pentru identificarea automata a repetitiilor unui exercițiu în video, ca și metode de comparație intre performantele elevilor și cele ale instructorilor potrivit unor politici de evaluare statistice. În funcție de parametrul care guverneaza gradul de exigenta critica al antrenorului virtual cu privire la performanța elevului, parametru care variaza intre 0 și 1 (0 insemnand atitudine foarte critica a instructorului în vreme ce 1 reprezintă o evaluare relaxata a performatei elevului), sistemul de feed-back bazat pe estimarea posturilor 3d atinge o acuratete înaltă, comparabila cu aceea a unui asemenea sistem bazat pe o reconstructie 3d bazate pe captura miscarii cu markeri ground-truth.

Imaginea de mai jos ofera o ilustrare a feed-back-ului oferit de antrenorul virtual atunci când sistemul opereaza asupra unor video-uri capturate în medii reale cu camera video a unui telefon mobil. În rândul

de sus este prezentata imaginea elevului care executa o mișcare de fitness în mod greșit, reconstructia sa 3d pe rândul următor urmata de imaginile executiei corecte a exercitiului de către un instructor calificat și feed-back-ul textual care puncteaza eroarea de execuție. Aceasta descriere text este punctul de plecare pentru un sistem generativ care pe baza indicatiilor text ale antrenorului cu privire la imbunatatirea performantei elevului sa genereze interactiuni 3d intre elev si antrenor care sa reflecte vizual sfaturile antrenorului. Ca un exemplu concret, inspirat din imaginea de mai jos, se poate genera mesh-ul antrenorului intr-o interactiune virtuala cu mesh-ul elevului incercand sa-i indrepte coloana vertebrala atunci cand nu executa corect o mișcare (v. exercițiul din ultima coloana) pornind de la feed-back-ul text („you should keep your back straight”). Un astfel de sistem generativ va fi antrenat pe baza setului de date de captura a miscarilor de interacțiune 3d a persoanelor augmentat cu descrieri text ale interactiunilor, așa cum s-a menționat anterior.



Gradul de realizare al obiectivelor

Obiectivele etapelor de pana acum au fost atinse în întregime după cum urmează. S-a realizat setup-ul colectiei de date atât în contextul augmentarii unor seturi de date existente cât și pentru folosirea noului sistem de captura a miscarii achizitionat. Metodologia folosită pentru organizarea și setup-ul datelor în ambele cazuri este principial aceeași, diferențele survenind exclusiv în privința metodei tehnice de achiziție a datelor care este, în mod evident, particulară echipamentului de captura a miscarii umane folosit.

S-au folosit aceste baze de date augmentate pentru evaluarea metodologiei state-of-the-art de estimare 3d a interactiunii dintre persoane și, așa cum s-a menționat anterior, s-a extins aceasta metodologie prin propunerea unui prim set de elemente metodologice care presupun o abordare principiala a reconstructiei 3d a persoanelor care interacționează, bazata pe recunoaștere, segmentare, mapare și reconstrucție 3d. În plus, este în curs de dezvoltare metodologia de estimare a interactiunilor umane în scene 3d din natura pe baza datelor inregistrate cu noul sistem de captura a miscarii achizitionat. Toate aceste dezvoltari sunt aliniate cu obiectivele proiectului și contribuie la elaborarea metodologiei de estimare a interactiunilor umane în scene 3d din natura și realizarea bazei de date, activitati care care se extind pana la finalul proiectului.

Primele rezultate ale proiectului sunt disponibile într-o varianta sumara (dezvoltarile metodologice descrise mai sus apar in articole aflate în evaluare sau trimise spre publicare și nu pot fi publicate încă în detaliu pe site) la adresa de web https://vision.imar.ro/fighs

În aceste condiții, consideram obiectivele curente ale proiectului ca fiind indeplinite.

Sumarul progresului înregistrat în proiect

1. setup-ul colectiei de date disponibil
2. evaluarea și extinderea metodologiei de estimare 3d a interactiunilor umane (incluzand și extinderea unor seturi de date existente)
3. demararea activitatilor de culegere a datelor cu noul sistem de captura a miscarii umane și de dezvoltare a metodologiei necesare pentru estimarea interactiunilor umane în scene 3d din natura
4. transmiterea unui articol care descrie progresele inregistrate spre publicare la IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), articol aflat momentan în plin proces de evaluare

5. transmiterea spre publicare a unui alt articol care va permite extinderea modelarii interactiunilor umane către sisteme generative la IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

Concluzii:

Activitățile proiectului s-au desfasurat conform calendarului stabilit si s-au adaptat conditiilor contractuale întâi prin reorientarea către folosirea unor baze de date existente pentru continuarea cercetarii în domeniul analizei scenelor 3d cu mai multe persoane care interactioneaza intre ele, urmata de demararea achizitionarii de date cu noul sistem de captura a miscarii umane disponibil în proiect și inceperea elaborarii metodologiei specifice de estimare a interactiunii persoanelor în scene 3d din natura. Ca rezultate concrete, s-au trimis spre publicare doua articole la IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), revista cu unul dintre cei mai mari factori de impact din domeniul stiintei calculatoarelor (IF actual: 23.6).