ACADEMIA ROMÂNĂ

Şcoala de Studii Avansate a Academiei Române

Institutul de Matematică "Simion Stoilow"

REZUMATUL TEZEI DE DOCTORAT

Percepția Automată a Oamenilor în Video

Conducător de doctorat:
C.S. I Dr. Cristian Sminchișescu

Doctorand:
Mihai Fieraru

București, 2024

Cuprinsul Tezei

1 Introducere **16**

1.1 Problema 16
1.2 Dificultăți 18
1.3 Contribuții 18
1.4 Lista Publicațiilor Relevante 20

2 Reconstrucția Tridimensională a Modelelor de Oameni în Interacțiune **21**

2.1 Introducere 22
2.2 Context 23
2.3 Seturi de Date și Protocoale de Etichetare 25
2.3.1 Protocol de Etichetare 26
2.4 Metodologie 32
2.4.1 Clasificarea Contactului 32
2.4.2 Segmentarea Contactului și a Amprentei 33
2.4.3 Reconstrucție 3D Monoculară 34
2.4.4 Reconstrucție 3D într-o Configurație Controlată - CHI3D . . . 36
2.5 Experimente 39
2.5.1 Estimarea Contactului 39
2.5.2 Rezultate Reconstrucție 3D Monoculară 39
2.5.3 Protocol de Evaluare și Clasament 41
2.6 Concluzie 42

3 Învățarea Autocontactului 3D al Oamenilor **43**

3.1 Introducere 44
3.2 Context 45
3.3 Metodologie 47
3.3.1 Proiecția Autocontactului în Imagine 48
3.3.2 Segmentarea și Amprenta Autocontactului 48
3.3.3 Reconstrucție 3D cu Amprenta Autocontactului 49
3.4 Seturi de Date Propuse 49
3.5 Experimente 51

3.6 Concluzie 57

4 REMIPS: Reconstrucția Fizic-Consistentă 3D a mai Multor Oameni Aflați în Interacțiune sub Supervizare Slabă 58

4.1 Introducere 59
4.2 Context 60
4.3 Metodologie 61
4.3.1 Modele Statistice 3D ale Corpului Uman 61
4.3.2 Modelul Camerei 61
4.3.3 Arhitectură 61
4.3.4 Coliziune Fizică pentru GHUM 64
4.3.5 Funcție de Cost pentru Ordinea în Adâncime 66
4.3.6 Funcție de Cost pentru Contact 66
4.4 Experimente 66
4.4.1 Detalii de Implementare 67
4.4.2 Seturi de Date 67
4.4.3 Rezultate 68
4.5 Concluzie 71

5 AIFit: Modele Automate de Feedback 3D Interpretabil pentru Antrenamentul de Fitness 74

5.1 Introducere 75
5.2 Context 76
5.3 Setul de Date Fit3D 78
5.4 Metodologie 78
5.4.1 Segmentarea Repetițiilor 79
5.4.2 Modelarea Exercițiilor 81
5.4.3 Antrenor Statistic 84
5.4.4 Feedback în Limbaj Natural 84
5.5 Experimente 85
5.6 Concluzie 92

6 Concluzii 94

6.1 Rezumatul Contribuțiilor 94
6.2 Direcții Viitoare 95

Chapter 1

Introducere

Crearea de sisteme inteligente care înțeleg lumea și sunt utile oamenilor necesită dezvoltarea de algoritmi care pot percepe oamenii din date vizuale. În această teză, abordăm reconstrucția tridimensională a oamenilor din imagini monoculare sau videouri. Acesta este un domeniu important, cu potențialul de a permite numeroase aplicații, cum ar fi noi interfețe om-computer, navigație și interacțiuni robotice, securitate și supraveghere, sănătate și divertisment. Totuși, perceperea oamenilor 3D exclusiv din camere monoculare prezintă mai multe provocări: ambiguitatea adâncimii, cadre parțiale, ocluzii cu scena sau auto-ocluzii, variație mare de aspect datorită diversității corpurilor umane, a articulațiilor, a îmbrăcămintii și a diverselor tipuri de contact fizic.

Primele noastre contribuții se concentrează pe înțelegerea interacțiunilor umane in 3D. Introducem modele de predicție a amprentelor de contact 3D și arătăm cum utilizarea acestora într-un cadru de optimizare îmbunătățește reconstrucția oamenilor apropiați unul de celălalt. De asemenea, construim două seturi mari de date pentru scopuri de antrenare și evaluare și propunem metodologie pentru obținerea articulației 3d și a formei corecte a corpului în scene cu oameni ce interacționează într-un mediu controlat.

În continuare, studiem reconstrucția autocontactului uman (contactul fizic ce are loc între părțile corpului aceleiași persoane). Propunem modele pentru prezicerea amprentelor de autocontact din imagini monoculare și arătăm cum pot fi acestea valorificate pentru îmbunătățirea acurateței reconstrucției umane într-un cadru de optimizare. Pentru învățare și evaluare, colectăm două seturi mari de date, unul capturat în laborator și unul constând din imagini din afara laboratorului cu etichete de autocontact 3D.

Apoi, propunem o metodă pentru învățarea integrală a reconstrucției 3D a mai multor oameni ce interacționează unul cu celălalt, sub supervizare slabă. Introducem un model unificat nou pentru funcțiile de cost de auto-coliziune și interpenetrare și folosim în procesul de învățare atât funcții de cost de auto-contact cât și de contact

între persoane. Modelul nostru obține rezultate state-of-the-art chiar și atunci când nu se utilizează supervizare 3D.

În ultimul capitol, prezentăm primul sistem automat care realizează percepția 3D a oamenilor pentru antrenamentul fitness, făcut posibil prin colectarea unui set de date de captură a mișcărilor a mai mult de 37 tipuri de exerciții. Sistemul estimează mișcarea 3D a corpului uman, segmentează repetările exercițiilor și identifică abaterile dintre standardele învățate de la antrenori și execuția unui antrenat, oferind feedback cantitativ în limbaj natural.

Pentru a sprijini cercetarea în acest domeniu, publicăm toate seturile de date colectate în această teză, împreună cu servere de evaluare și clasamente publice.

Chapter 2

Reconstrucția Tridimensională a Modelelor de Oameni în Interacțiune

Acest capitol este bazat pe articolul [1] "Three-Dimensional Reconstruction of Human Interactions." by Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu, publicat în The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. ce a fost augmentat suplimentar în versiunea curentă și este momentat trimis unui jurnal pentru revizuire (disponibil și pe arXiv ca preprint [2]).

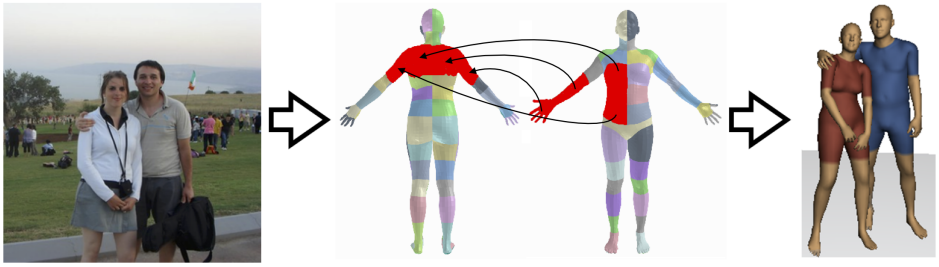

Figure 2.1: Reconstrucția tridimensională monoculară, constrânsă de amprentele de contact, păstrează esența interacțiunii fizice dintre oameni și permite analiza comportamentului.

În acest articol propunem un prim set de elemente metodologice pentru a aborda reconstrucția oamenilor care interacționează unul cu celălalt, bazându-ne pe recunoaștere, segmentare, estimare de corespondențe ale contactului și reconstrucție 3D. Mai precis, împărțim problema producerii de reconstrucții 3D veridice ale oamenilor care interacționează, în (a) detectarea contactului, (b) segmentarea binară a regiunilor de contact de pe suprafețele corpurilor oamenilor care interacționează; (c) predicția am-

prentei de contact (corespondențe între regiunile în contact) și (d) reconstrucție 3D sub funcții de costuri ce penalizează distanța între suprafețele în contact conform unei amprente de contact.

Two people are hugging.

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

Figure 2.2: Text etichetat ce descrie mișcarea și optimizarea modelului GHUM pentru doi subiecți care interacționează în dataset-ul CHI3D. Toate cele 4 cadre suprapuse sunt afișate pentru 3 timpi (de la stânga la dreapta în ordine cronologică): o secvență de îmbrățișare (sus), o secvență în care se stă la poză (jos). Subiectul în verde este subiectul care poartă costum de captură. Subiectul în cărămiziu nu poartă costum de captură.

Pentru a antrena modelele și a evalua tehnicile, introducem două seturi mari de date. Capturăm CHI3D, o bază de date în laborator cu mișcări 3D conținând 631 de secvențe care conțin 2525 evenimente de contact, 728.664 schelete 3D, precum și o descriere textuală a mișcării interacțiunii din fiecare secvență video. Colectăm și FlickrCI3D, un set de date de 11.216 imagini, conținând 14.081 perechi de persoane și 81.233 corespondențe între suprafețele în contact.

În plus, propunem metodologie pentru obținerea articulației și formei corpului oa-

menilor aflați în interacțiuni în setul de date CHI3D. Prin valorificarea informațiilor din senzori de mișcare, camere RGB din mai multe unghiuri și un scaner 3D, dar și din etichete de contact, probabilități ale articulațiilor 3D și constrângeri fizice, obținem reconstrucții 3D de nivel ground-truth (a se vedea fig. 2.2).

Distribuim secvențele de mișcare ground-truth în mai multe formate (parametrii GHUM [3] și SMPLX [4], articulații 3D Human3.6m [5]) la `ci3d.imar.ro` și implementăm un server de evaluare pe un set de date de testare privat, împreună cu un clasament public, cu scopul de a avansa state-of-the-art-ul reconstrucției 3D a oamenilor în contact în interacțiuni.

Evaluăm performanța algoritmului propus pentru detectarea contactului și obținem o acuratețe medie de 0,846, cu 0,844 pentru categoria "contact" și 0,848 pentru categoria "fără contact". Tabelul 2.1 arată evaluarea modelului nostru *ISP* pentru segmentarea contactului și estimarea amprentei folosind metrica intersecție pe uniune ($\text{IoU}_{N_{reg}}$), calculată pentru diferite granulități ale regiunilor. Tabelul 2.2 arată că informațiile de contact etichetate îmbunătățesc acuratețea reconstrucției.

	$\text{IoU}_{75}\uparrow$		$\text{IoU}_{37}\uparrow$		$\text{IoU}_{17}\uparrow$		$\text{IoU}_{9}\uparrow$	
Metodă	Segm.	Ampr.	Segm.	Ampr.	Segm.	Ampr.	Segm.	Ampr.
ISP complet	**0.318**	**0.082**	**0.365**	**0.129**	**0.475**	**0.248**	**0.618**	0.408
ISP fără feature-uri semantice 2d la intrare	0.300	0.073	0.350	0.116	0.465	0.240	**0.618**	**0.410**
ISP fără învățarea comună a segm. contactului	-	0.072	-	0.124	-	0.218	-	0.383
ISP fără ascunderea coresp. din afara măștii estimate	-	0.075	-	0.124	-	0.230	-	0.385
Performanța umană	0.456	0.226	0.542	0.370	0.638	0.499	0.745	0.635

Table 2.1: Rezultatele modelului nostru *ISP* pentru segmentarea și estimarea amprentelor de contact, evaluate pe FlickrCI3D pentru diferite niveluri de granularitate a regiunilor pe suprafața 3D a corpului uman (de la 75 la 9 regiuni). Am eliminat diferite componente ale metodei noastre complete pentru a ilustra contribuția lor. Performanța umană reprezintă valorile de consistență dintre adnotatori.

	Grab	Hit	Handshake	Holding hands	Hug	Kick	Posing	Push	Agregat
Funcție	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓	P↓ T↓
de Cost	C↓	C↓	C↓	C↓	C↓	C↓	C↓	C↓	C↓
$L^\star$	**117 390**	**119 367**	**97 388**	101 **380**	**174 400**	**140 419**	**139 364**	**117 381**	**125 368**
	19 (**3.5**)	**8** (**4**)	**12** (**3**)	**20** (**3**)	**62** (**45**)	**32** (**7**)	**41** (**11**)	**14** (**4**)	**26** (**10**)
$L^\star$ **fără** L_G	121 416	128 396	99 406	**100** 389	180 424	155 460	140 377	124 399	131 408
[6]	459 (366)	426 (363)	377 (305)	373 (274)	368 (328)	550 (464)	388 (327)	425 (369)	421 (350)

Table 2.2: Comparație între utilizarea funcției de cost de consistenta a contactului în timpul optimizării ($L^\star$) și neutilizarea ei ($L^\star$ **fără** L_G). Erori de estimare a **posturii** (**P**) si **translației** (**T**) oamenilor 3D, precum și distanța 3D medie (mediana) a **contactului** (**C**), exprimată în mm, pentru setul de date CHI3D. Funcția noastră de optimizare completă, cu termenul de aliniere geometrică pe amprente de contact, reduce erorile de estimare a posturii si translației și distanța 3D dintre suprafețele etichetate ca fiind în contact.

Chapter 3

Învățarea Autocontactului 3D al Oamenilor

Acest capitol este bazat pe articolul [7] "Learning Complex 3D Human Self-Contact." by Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu publicat în Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 1343–1351, 2021.

Majoritatea sistemelor monoculare de reconstrucție 3D a oamenilor nu estimează direct autocontactul, deși rolul său central în recunoașterea corectă a subtilităților multor posturi sau gesturi iconice este larg recunoscut.

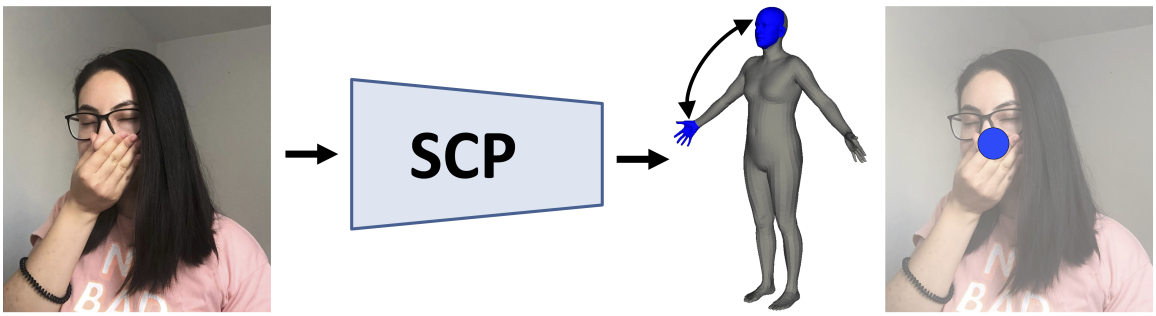

Figure 3.1: Rețeaua noastră de predicție a autocontactului (*SCP*) estimează regiunile corpului aflate în contact, corespondențele lor și poziționarea autocontactului în spațiul imaginii.

Pentru a depăși unele dintre deficiențele metodelor de reconstrucție 3D existente, agnostice la autocontact, propunem să reprezentăm explicit autocontactul și să demonstrăm cum modelele rezultate pot ajuta la înțelegerea comportamentală în aplicațiile de evaluare a atingerii feței. Modelele noastre învață să prezică locația contactului în imagine pentru a ajuta la detectarea zonelor corporale în autocontact, precum și amprenta lor, definită ca setul de corespondențe între regiunile de pe suprafața unui

model de corp uman care se ating. Condiționând reconstrucția de astfel de estimări detaliate, autocontactul poate fi prezis corect în 3D. Pentru a antrena modele și pentru o evaluare cantitativă la scară largă, colectăm și etichetăm două seturi mari de date care conțin imagini cu oameni în autocontact. HumanSC3D este un set de date de captură a mișcării 3D precisă, care conține 1.032 de secvențe cu 5.058 de evenimente de contact și 1.246.487 de poziții 3D ground-truth sincronizate cu imagini capturate din mai multe perspective. De asemenea, colectăm FlickrSC3D, un set de date de 3.969 de imagini, care conține 25.297 de perechi de regiuni ale părților corpului aflate în contact, definite pe un model de suprafață 3D a corpului uman, împreună cu localizarea autocontactului în imagine.

Contribuțiile principale ale articolului sunt următoarele:

- Introducem un prim model pentru detectarea regiunilor corporale în autocontact și a amprentei autocontactului. Noua noastră rețea neurală profundă *SCP* este ajutată de un predictor intermediar de localizare a imaginii de auto-contact, folosit atât la antrenament, pentru selectarea featere-urilor locale, cât și la testare, prin impunerea consistentei cu amprenta de contact 3D estimată. Rezultatele evaluării sunt prezentate în Tabelul 3.1.

- Seturi de date noi, de scară largă și valoroase comunității, care capturează persoane în autocontact, împreună cu adnotări dense pe un model 3D al corpului uman cu regiunile în contact, precum și adnotări în spațiul imaginii asociate punctelor de contact observate. Bazele de date sunt puse la dispoziția comunității de cercetare.

- Rezultate cantitative (a se vedea Tabelul 3.2) și calitative (a se vedea Figura 3.2) ale reconstrucțiilor 3D mai precise metric și perceptiv veridice pe baza amprentelor de autocontact.

- Un model de bază pentru o mare clasă de aplicații care ar beneficia de reprezentări 3D precise de autocontact, cum ar fi monitorizarea stării răspândirii posibilelor infecții atunci când mâinile ating părți ale feței (gură, nas, ochi) în spitale sau în timpul unei pandemii, sau înțelegerea subtilă a gesturilor pentru terapia copiilor cu autism asistată de roboți.

Method	$\text{IoU}_{75}\uparrow$		$\text{IoU}_{37}\uparrow$		$\text{IoU}_{17}\uparrow$		$\text{IoU}_{9}\uparrow$	
	Segm.	Ampr.	Segm.	Ampr.	Segm.	Ampr.	Segm.	Ampr.
SCP	**0.469**	**0.301**	0.507	**0.339**	0.591	**0.442**	0.693	**0.550**
ISP [1] (adaptat pt. autocontact)	0.462	0.133	0.503	0.186	0.583	0.305	0.688	0.460
Performanța umană	0.528	0.422	0.564	0.475	0.664	0.579	0.768	0.692

Table 3.1: Rezultatele metodei noastre *SCP* pentru segmentearea de autocontact și a amprentei acestuia, pe FlickrSC3D, evaluate pentru diferite granularități ale regiunilor de pe suprafața 3D a corpului uman (de la 75 la 9 regiuni). Performanța umană reprezintă valorile de consistență între adnotatori.

Funcție de cost	W/o chair - standing				W/o chair - sitting				W/ chair				Overall			
	P↓	T↓	V↓	C↓	P↓	T↓	V↓	C↓	P↓	T↓	V↓	C↓	P↓	T↓	V↓	C↓
L	**94**	**408**	**77**	**13**	**116**	**424**	**93**	**27**	**107**	**426**	**85**	**24**	**98**	**414**	**80**	**16**
L fără L_G	106	419	121	210	145	436	147	183	132	432	123	189	114	423	124	203

Table 3.2: Erori 3D de estimare a posturii (P), a translației (T) și a vârfurilor mesh-ului (V), precum și distanța medie 3D de contact (C) exprimată în mm pentru setul de date HumanSC3D. Utilizarea funcției de cost complete, cu termenul de aliniere geometrică pe semnaturi de auto-contact etichetate, reduce erorile de estimare a posturii, a translației și a vârfurilor mesh-ului, precum și distanța 3D între suprafețe etichetate ca fiind în contact.

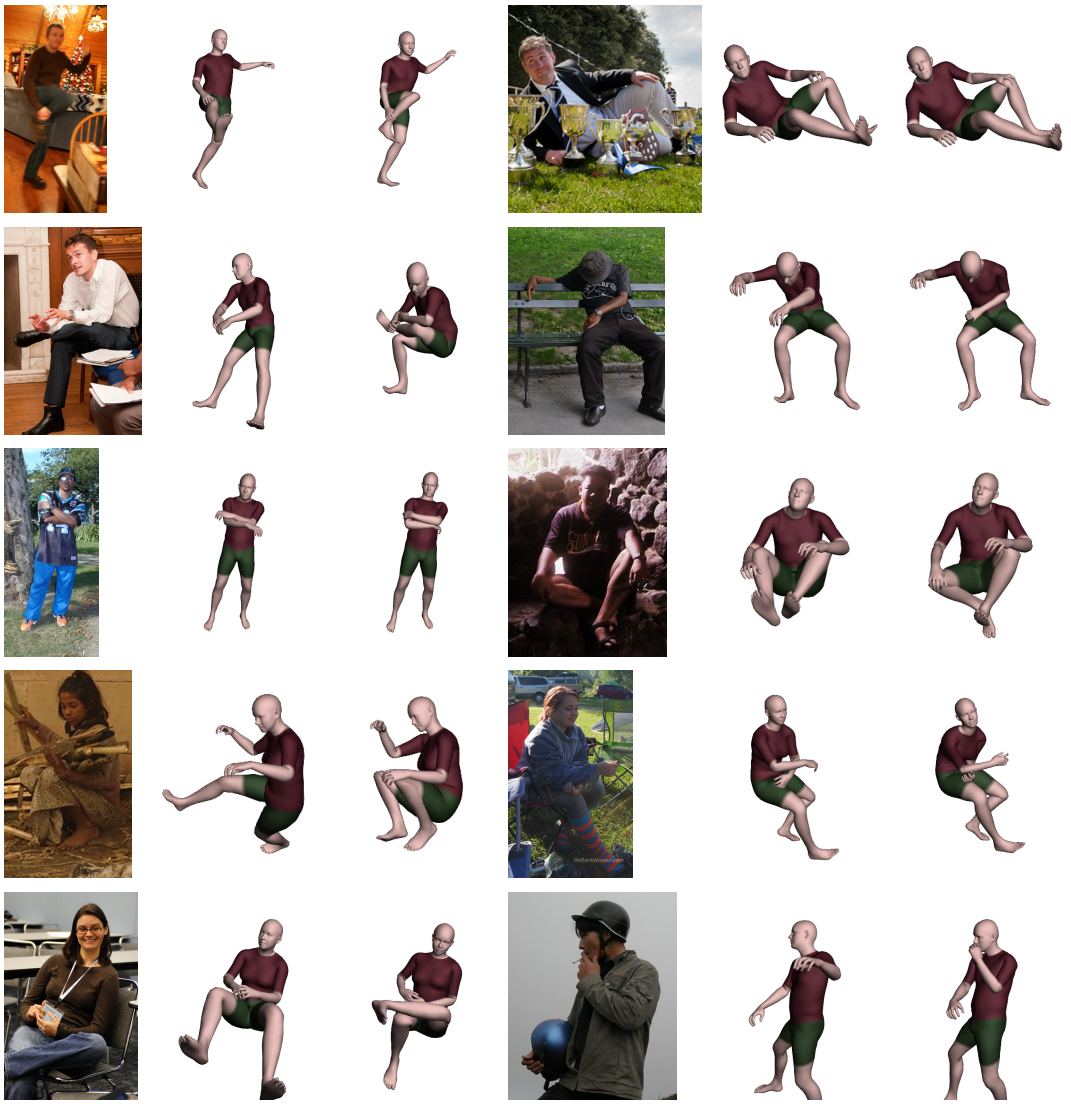

Figure 3.2: Estimarea 3D a poziției si formei corpului, folosind datele noastre de autocontact etichetate. Imagine originală (stânga). Reconstrucția fără a lua în considerare autocontactul în funcția de cost (mijloc). Reconstrucția care utilizează etichetele de autocontact în funcția de cost (dreapta).

Chapter 4

REMIPS: Reconstrucția Fizic-Consistentă 3D a mai Multor Oameni Aflați în Interacțiune sub Supervizare Slabă

Acest capitol este bazat pe articolul [8] "REMIPS: Physically Consistent 3D Reconstruction of Multiple Interacting People under Weak Supervision." by Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu, publicat în Advances in Neural Information Processing Systems, volume 34, pages 19385–19397. Curran Associates, Inc., 2021.

Reconstrucția tridimensională a mai multor persoane care interacționează pe baza unui singur cadru monocular este crucială pentru problema generală de înțelegere a scenei, deoarece capturarea subtilităților interacțiunii este adesea motivul pentru care se face o fotografie. Modelele actuale de reconstrucție 3D a corpului uman fie tratează fiecare persoană independent, ignorând majoritatea contextului, fie reconstruiesc oamenii împreună, dar nu pot estimează interacțiunile corect atunci când oamenii sunt în apropiere unul de celălalt. În acest studiu introducem **REMIPS**, un model pentru reconstrucția 3D a mai multor persoane care interacționează, antreat în condiții de supervizare slabă. **REMIPS** poate reconstrui un număr variabil de persoane direct din imagini monoculare. În centrul metodologiei noastre se află o rețea transformer nouă care combină tokenuri de persoane neordonate (câte o persoană pentru fiecare om detectat) cu tokenuri codificate cu poziție din patch-uri de featere-uri de imagine (a se vedea Figura 4.1). Introducem un nou model unificat pentru auto-coliziuni și interpenetrari bazat pe o aproximare a mesh-ului calculată prin aplicarea operatorir de reducere a rezoluției. Ne bazăm pe funcții de cost ce folosesc supervizarea ușoară pentru flexibilitate și generalizare și integrăm funcții de cost de autocontact și contact

din interacțiuni în procesul de învățare. Cu **REMIPS**, raportăm rezultate cantitative state-of-the-art (a se vedea Tabelul 4.1) în clasamente populare, chiar și în cazurile în care nu se utilizează nicio supervizare 3D. În plus, rezultatele vizuale calitative (a se vedea Figura 4.2) arată că reconstrucțiile noastre sunt plauzibile în ceea ce privește poziția și forma corpului uman și coerente pentru imagini complexe, colectate în afara laboratorului, unde oamenii se află in proximitate unul de celălalt.

Metodă	MPJPE ↓	MPVPE ↓	Eroare Translație ↓	#2D	#3D
Fieraru *et al.* [1]	125.4	–	368.0	N/A	N/A
Jiang *et al.* [9]	136.0	N/A	N/A	100K	300K
REMIPS	**120.8**	**134.7**	**284.1**	115K	0

Table 4.1: Performanță pe setul de date **CHI3D** [1] pentru metodele de reconstrucție a posturii și formei corpului multiplelor persoane. În coloane 2, 3 și 4 sunt afișate erorile medii pe poziția fiecărui punct de articulație (**MPJPE**), erorile medii pe poziția fiecărui vârf al mesh-ului (**MPVPE**) și eroarea de translație. Toate erorile sunt raportate în milimetri și sunt relative în raport cu articulația pelvisului. Metoda noastră are erori mai mici în comparație cu celelalte metode bazate pe optimizare și inferență. Comparăm și numărul de adnotări **#2D** și **#3D** utilizate pentru supervizare în timpul antrenamentului. Modelele noastre nu utilizează date **#3D** și obțin performanțe mai bune pe setul dificil de date **CHI3D** [1].

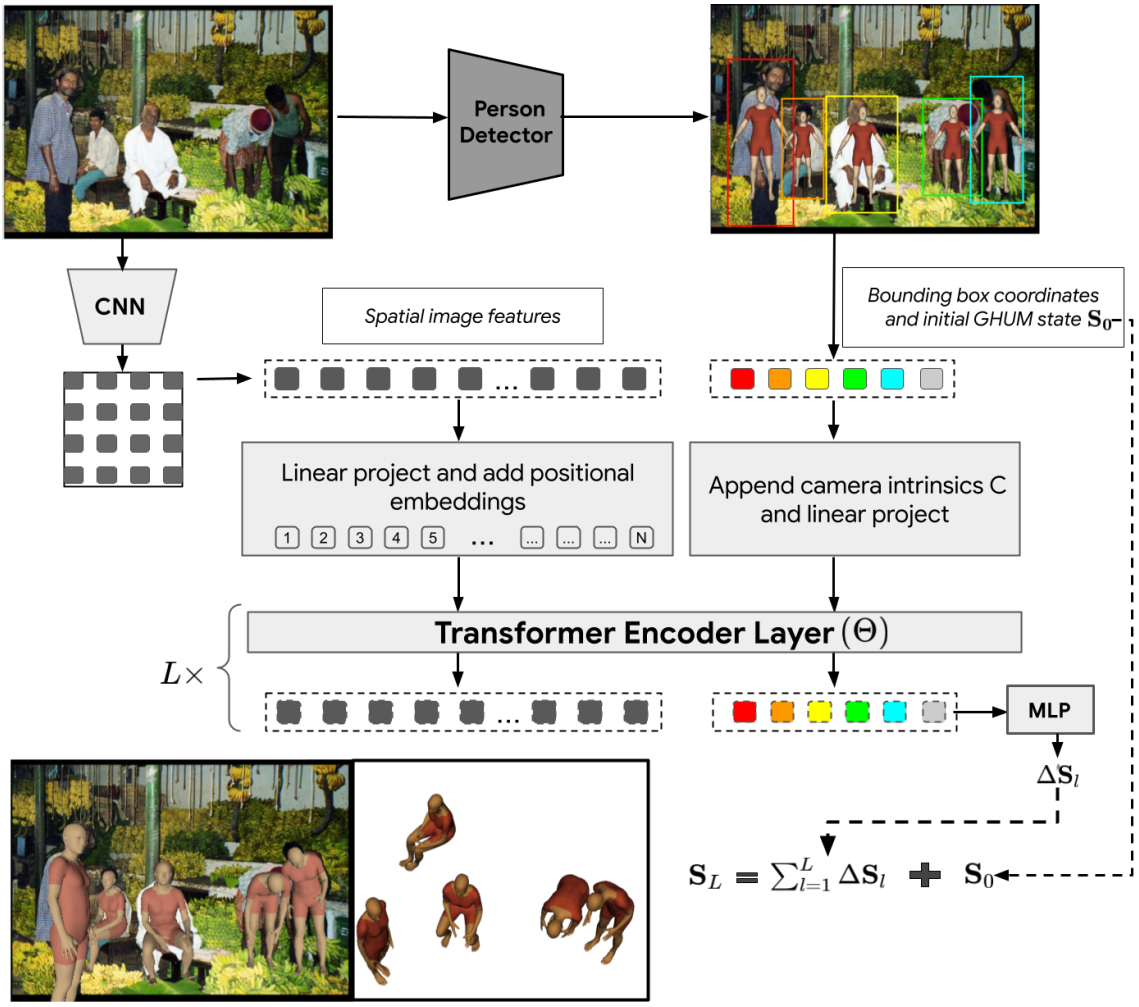

Figure 4.1: Prezentare generală a **REMIPS**, arhitectura noastră propusă pentru reconstrucția articulației și formei 3D a mai multor persoane care interacționează. Pornind de la o singură imagine de intrare, folosim un detector preantrenat pentru a extrage detecții de oameni. Creăm o secvență de token-uri de persoană din aceste detecții cărora le atașăm o estimare inițială a stării GHUM $\mathbf{S}_0$. Pe o ramură separată, pornind de la imagine, rulăm o arhitectură de rețea neurală convoluională și creăm o secvență suplimentară de token-uri de feater-uri spațiale ale imaginii, $\mathbf{F}_s$. Concatenăm cele două secvențe de reprezentări și rafinăm iterativ această reprezentare în L pași de transformer encoder. La sfârșitul fiecărui pas l, colectăm reprezentarea transformată pentru secvențele de tokenuri asociate cu persoanele și aplicăm un strat neuronal liniar (MLP) pentru a regresa estimările reziduale de stare GHUM $\Delta\mathbf{S}_l$. Estimarea noastră finală este dată de suma ponderată a tuturor rezidurilor de stare și a stării inițiale. Rețeaua este antrenată cu supervizie slabă pe diverse seturi de date cu annotări 2D. Utilizăm funcții de cost de contact și de coliziune pentru a asigura plauzibilitate fizică.

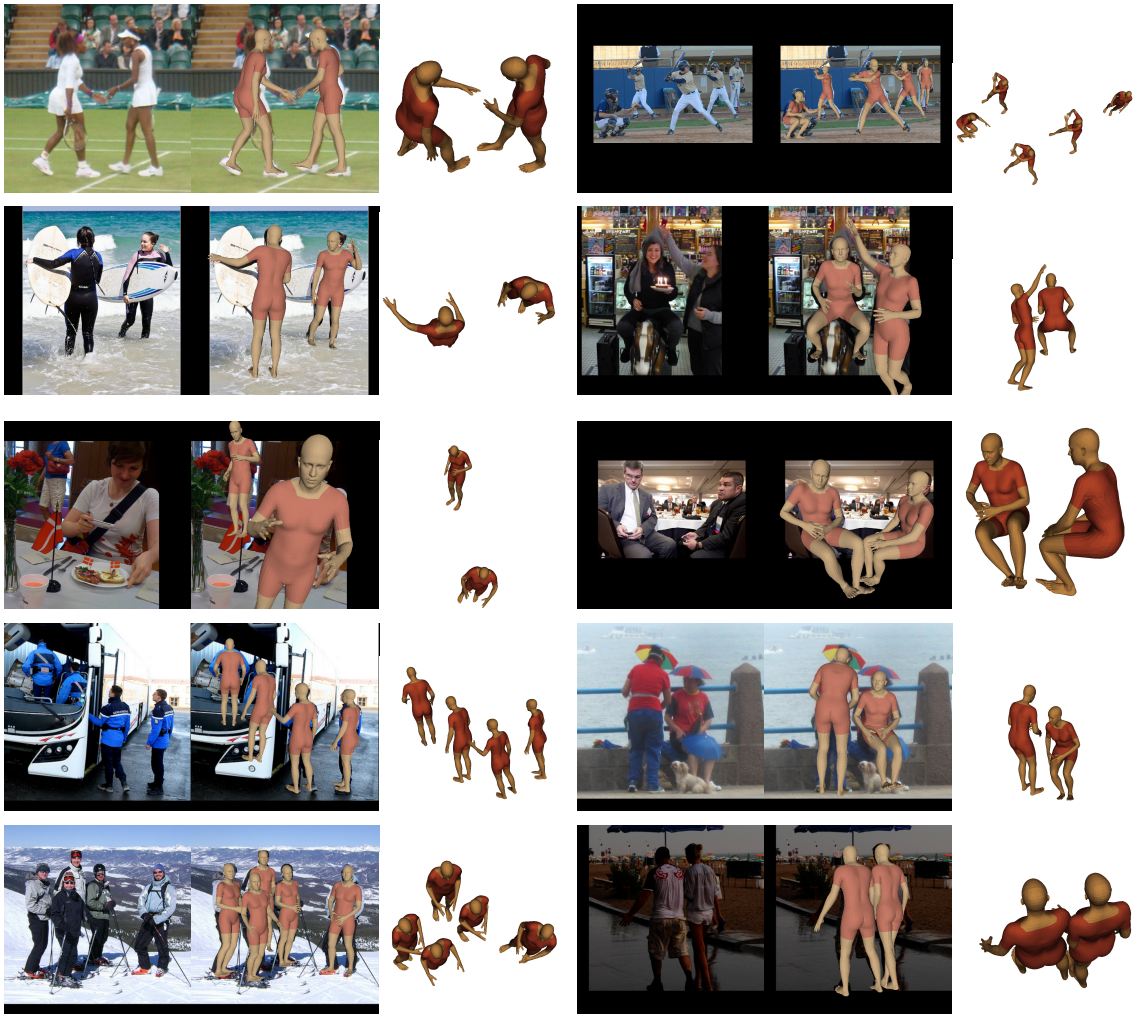

Figure 4.2: Reconstrucții 3D ale corpului uman pe setul de validare **COCO** (liniile 1-5) pentru imagini capturate în afara laboratorului. Afișăm imaginea inițială împreună cu o reconstrucție suprapusă a mesh-urilor, precum și o redare dintr-un unghi de vizualizare diferit, care ilustrează mai bine consecvența fizică a reconstrucțiilor **REMIPS**.

Chapter 5

AIFit: Modele Automate de Feedback 3D Interpretabil pentru Antrenamentul de Fitness

Acest capitol este bazat pe articolul [10] "AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training." by Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu, publicat în The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021.

În acest articol propunem *AIFit*, primul sistem de antrenament îmbunătățit cu AI pentru fitness. Sistemul este capabil să reconstruiască poziția 3D dinamică a corpului uman, să numere repetările și să ofere feedback localizat automat, ancorat în spațiul imaginii, cât și exprimat în limbaj natural (a se vedea Fig 5.1).

Pentru a susține cercetarea și evaluarea, introducem *Fit3D*, un set de date de mare amploare cu peste 3 milioane de imagini și articulații 3d ground-truth, colectate de la 13 subiecți (inclusiv un instructor de fitness licențiat și un subiect de fitness avansat), observați de 4 camere RGB diferite, împreună cu scanări 3D ale fiecărui subiect. Setul de date cuprinde 37 de exerciții constând din mișcări simple și compuse, acoperind toate grupele principale de mușchi și tipurile de articulații, inclusiv, printre altele, încălzirea corpului, halterele, gantere, flotări sau yoga.

Metodologia propusă de noi include o evaluare de mare amploare monoculară și din mai multe camere a reconstrucției poziției 3D a corpului uman pentru antrenamentul de fitness folosind *Fit3D*, modele pentru identificarea automată a repetărilor exercițiilor, precum și metode pentru compararea performanțelor instructorilor și practicanților în funcție de statistici definite pe caracteristici extrase automat (pasive și active) care definesc exercițiul și explică cea mai mare parte a mișcării. Antrenorul nostru statistic este guvernat de un parametru global care modelează cât de critic este în ceea ce privește performanța unui elev. În practică, parametrul ajută antrenorul să se adapteze

la nivelul de fitness al unui elev (adică începător vs avansat vs expert) sau la acuratețea metodei de reconstrucție 3D a poziției corpului. În cele din urmă, antrenorul nostru statistic oferă feedback spațio-temporal ușor de înțeles, vizual ancorat, dar și în limbaj natural. O imagine de ansamblu a sistemului este prezentată în figura 5.2.

Figure 5.1: Feedback textual și vizual generat de *AIFit* pe videoclipuri reale, capturate cu o cameră foto obișnuită a smartphone-ului. Folosim MubyNet-FT pentru a estima poziția în 3D a persoanei antrenate. Pentru fiecare exemplu, prezentăm următoarele: o imagine cu eroarea identificată a persoanei antrenate (***primul rând***), reconstrucția 3D (***al doilea rând***), imaginea corespunzătoare cu execuția corectă a instructorului (***al treilea rând***) și feedback-ul textual (***al patrulea rând***). Cele două exemple din (***stânga***) arată feedback-ul pentru caracterisicile active, în timp ce cele două din (***dreapta***) arată feedback-ul pentru caracterisicile pasive. Observați generalizarea la diferite persoane umane în diferite medii și unghiuri diverse ale camerei.

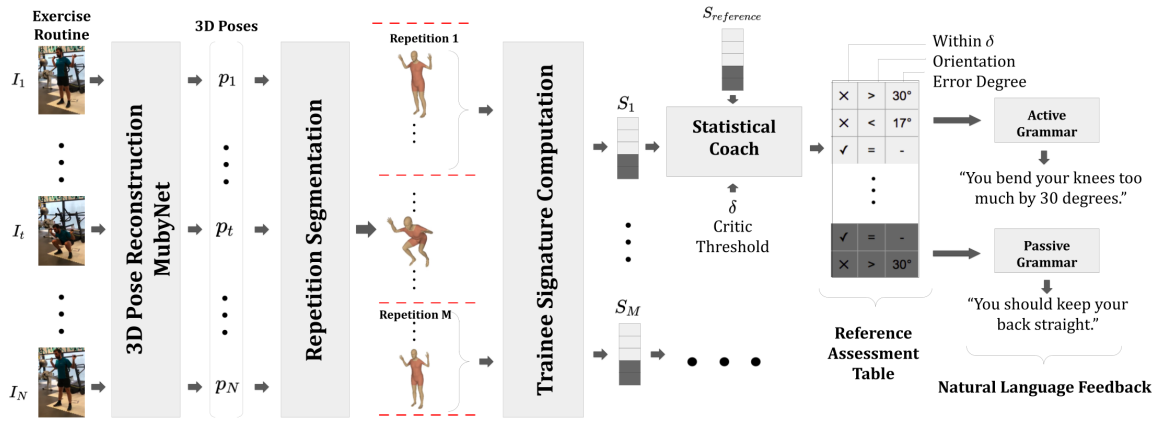

Figure 5.2: Prezentare generală a *AIFit*. Fiind dat un videoclip al unui practicant de fitness care efectuează un exercițiu, (a) sistemul realizează o **reconstrucție a posturii 3D** în fiecare cadru și apoi (b) aplică o **segmentare a repetărilor** pentru a număra automat numărul de repetări ale posturii 3D și pentru a determina fiecare interval de repetare. În continuare, (c) **modelarea exercițiului** calculează o *semnătură a exercițiului* utilizând caracteristicile angulare ale fiecărei repetări a practicantului (a se vedea fig. 5.3 pentru o prezentare detaliată). (d) **antrenorul statistic** compară fiecare semnătură de repetare cu semnătura de referință a instructorului sub un parametru care permite diferite grade de eroare. Rezultatele comparației sunt introduse într-o **tabelă de evaluare față de referință** care specifică care abateri sunt mari, semnul abaterii și gradul de eroare. În cele din urmă, pe baza tabelului, (e) *AIFit* produce **feedback în limbaj natural** pentru practicant, folosind fie o gramatică activă, fie o gramatică pasivă.

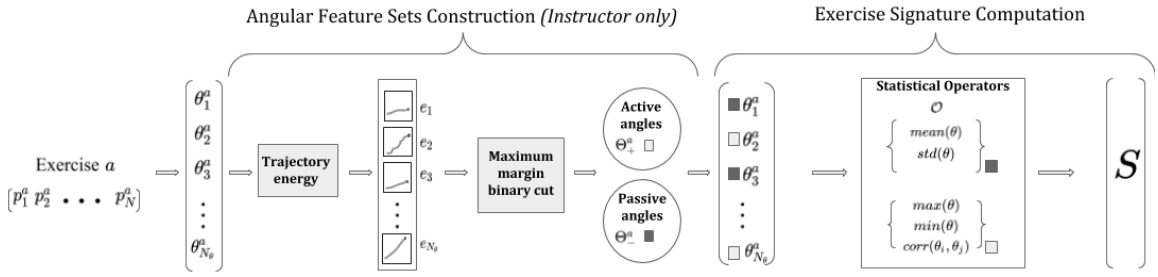

Figure 5.3: *Modelarea exercițiilor:* **(Stânga) Construirea seturilor de caracteristici angulare active și pasive** (numai pentru instructor). Pentru un *exercițiu a* și pentru fiecare funcție angulară, integrăm traiectoria sa de mișcare peste secvența de posturi 3D a instructorului și obținem energia mișcării. Grupăm energiile în două seturi, active Θ_+^a (asociate cu energie ridicată) și pasive Θ_-^a (asociate cu energie scăzută) utilizând algoritmul *maximum margin binary cut*. **(Dreapta) Calcularea semnăturii exercițiului.** Atât pentru exercițiile persoanelor instruite, cât și pentru cele ale instructorului, se generează o semnătură formată din: caracteristicile angulare calculate, tipul de energie (activă sau pasivă) și operatorii statistici prestabiliți.

Bibliography

[1] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[2] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Reconstructing three-dimensional models of interacting humans. *arXiv preprint arXiv:2308.01854*, 2023.

[3] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *CVPR*, 2020.

[4] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

[5] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[6] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *CVPR*, 2018.

[7] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1343–1351, 2021.

[8] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. In M. Ranzato, A. Beygelzimer,

Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19385–19397. Curran Associates, Inc., 2021.

[9] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2020.

[10] Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.