Departamentul de Calculatoare
Universitatea POLITEHNICA din București
Facultatea de Automatică și Calculatoare

Recunoașterea Acțiunilor Umane pentru Roboți Sociali

Teză de doctorat – Rezumat

Autor: Mihai NAN

Coordonator Științific: Prof. Adina-Magda FLOREA

Comisia

Președinte	Prof.Dr.Ing. Florin Pop	Universitatea POLITEHNICA din București
Coordonator științific	Prof.Dr.Ing. Adina Magda Florea	Universitatea POLITEHNICA din București
Referent	Prof.Dr.Ing. Sergiu Nedevschi	Universitatea Tehnică din Cluj-Napoca
Referent	Prof.Dr.Ing. Vasile Manta	Universitatea Tehnică din Iași
Referent	Prof.Dr.Ing. Irina Georgiana Mocanu	Universitatea POLITEHNICA din București

---

Universitatea POLITEHNICA din București
București, 2022

Autor:
Mihai Nan

Coordonator Științific:
Prof. Adina-Magda Florea

Instituție:
Universitatea POLITEHNICA din București, România

SUMAR

Capacitatea de a recunoaște acțiunile efectuate de oameni este fundamentală pentru multe aplicații practice din diverse domenii. Multe dintre aceste aplicații necesită ca realizarea recunoașterii să fie făcută în timp real, iar acest aspect impune constrângeri speciale soluțiilor propuse. Am considerat că, în contextul actual, roboții sociali reprezintă unul dintre domeniile în care problema Human Action Recognition (HAR) își demonstrează utilitatea. Un robot nu poate fi considerat social dacă nu este proactiv, iar pentru a îndeplini această proprietate trebuie să fie capabil să înțeleagă acțiunile efectuate de oamenii cu care interacționează. Plecând de la această motivație, cercetătorii au propus deja multe soluții pentru problema HAR, dar pentru că problema este una complexă care depinde de mulți factori, această direcție de cercetare rămâne deschisă și reprezintă un subiect de mare interes în Computer Vision.

Scopul nostru a fost propunerea unei taxonomii care demonstrează ce reprezintă o acțiune umană în concepția noastră și dezvoltarea unor soluții pentru acest tip de problemă. Mai mult, noi am dorit ca soluțiile propuse să atingă un compromis între precizie bună și viteză mare de inferență. Am analizat modalitățile de reprezentare a datelor disponibile și am ajuns la concluzia că este potrivită cea bazată pe scheletul detectat din secvența video, deoarece ne permite atingerea scopului propus și asigură confidențialitatea pentru utilizator. Pentru a ne atinge scopul, am proiectat o serie de abordări bazate pe rețele neurale profunde de dimensiuni mici, care obțin performanțe comparabile cu restul soluțiilor existente pentru unul dintre cele mai complexe benchmark-uri, NTU RGB+D.

Pentru a valida posibilitatea utilizării soluțiilor propuse pentru o platformă robotică, am dezvoltat un pipeline general pe care l-am integrat în cadrul framework-ului AMIRO și l-am testat folosind robotul umanoid Pepper. Cu ajutorul transformărilor introduse în pipeline-ul de integrare, am arătat că modelul nostru neural, antrenat folosind un set de date colectat cu camere fixe, poate recunoaște acțiunile umane într-un scenariu real pornind de la datele furnizate de camerele robotului.

Am propus și implementat soluții multiple pentru problema HAR folosind arhitecturi neurale originale. Am construit aceste modele luând în considerare tehnicile de ultimă generație existente în literatura de specialitate. Principalele cerințe pe care le-am avut în vedere când am dezvoltat aceste abordări sunt viteza mare de inferență și capacitatea bună de generalizare.

Există situații în care corectitudinea predicției oferite de modelul neural este fundamentală. Acest aspect se aplică în special aplicațiilor din domeniul medical. Din acest motiv, în această teză am abordat și problema explicabilității pentru una dintre abordările propuse. Modelul neural propus returnează, pe lângă probabilitățile pentru fiecare clasă, un tensor cu caracteristici pe baza cărora putem determina cele mai importante articulații ale scheletului pentru fiecare cadru.

Am efectuat o analiză amplă a rezultatelor pentru cele mai bune abordări propuse. Această analiză evidențiază clasele pentru care modelele neurale ating cele mai bune performanțe, dar și cele pentru care nu reușesc să identifice corect acțiunile. În urma acestei analize, putem concluziona că modelele neurale au probleme în identificarea corectă a acelor acțiuni pentru care chiar și oamenii eșuează atunci când se bazează doar pe date ce țin de postura umana descrisă prin punctele de articulație care compun scheletul.

CONTENTS

1 INTRODUCERE

Problema HAR necesită analizarea unei secvențe temporale care descrie o acțiune efectuată de un om pentru a o eticheta cu una dintre acțiunile posibile. Această problemă este un subiect de cercetare din domeniul Computer Vision pentru care se acordă o atenție sporită, deoarece are o mare aplicabilitate practică. Există două categorii relevante de probleme practice care necesită soluții care să integreze un modul HAR: probleme de interacțiune om-robot și probleme de monitorizare video. Fiecare categorie prezintă o serie de particularități, iar în această teză ne vom concentra asupra unor abordări care îndeplinesc condițiile necesare pentru aplicarea în problemele de interacțiune om-robot.

1.1 MOTIVAȚIE

Inspirația pentru problema recunoașterii acțiunilor umane vine din percepția umană a informațiilor vizuale pe care ochiul o transmite creierului. Nu putem identifica un algoritm standard pe care creierul uman îl aplică pentru a analiza fluxul video în încercarea de a determina informațiile relevante. Prin urmare, nici aplicațiile care ar putea rezolva problema HAR nu pot depinde de reguli și definiții stricte ale gesturilor, particularitățile oamenilor care acționează, mediul în care este realizată acțiunea etc. În plus, recunoașterea acțiunii este uneori o sarcină dificilă chiar și pentru oameni. Dificultatea vine din faptul că multe aspecte trebuie luate în considerare atunci când se analizează o secvență care descrie o acțiune. Fiecare persoană poate avea un mod distinct de a acționa. Astfel, aceeași acțiune poate să difere de la o persoană la alta sau de la un context la altul.

1.2 OBIECTIVE

Obiectivul principal al acestei teze este de a proiecta un modul capabil să rezolve problema HAR pe care să îl putem integra cu ușurință în platformele folosite pentru rezolvarea problemelor practice. Pentru a atinge acest obiectiv, cercetarea noastră a identificat provocările care există și a fost ghidată de câteva întrebări cheie care apar atunci când ne propunem să evaluăm o astfel de soluție:

- *Care sunt principalele aplicații din lumea reală care necesită utilizarea unui modul capabil să clasifice acțiunile umane?*
- *Ce seturi de date există și cum se leagă ele cu abordările propuse până în acest moment?*

- *Care este reprezentarea adecvată pentru a asigura invarianța față de mediu și persoana care acționează?*
- *Care sunt cele mai potrivite tipuri de straturi neurale pe care le putem folosi în proiectarea rețelei neurale care realizează clasificarea?*
- *Cum putem integra un modul HAR într-o platformă robotică și ce performanță atinge într-un scenariu în timp real?*
- *Cum putem calcula caracteristici suplimentare pornind de la coordonatele articulațiilor din spațiul 3D și folosind formule matematice?*
- *Cum poate un model neural să explice sau să motiveze predicția oferită?*
- *Ce transformate de augmentare putem folosi pentru datele scheletice și cum influențează acestea performanța atinsă de modelele neuronale?*

Am abordat fiecare dintre aceste întrebări în capitolele acestei teze și am oferit răspunsuri pentru ele motivate de rezultatele obținute sau de contribuțiile introduse.

2

RECUNOAȘTEREA ACȚIUNII UMANE SPAȚIO-TEMPORALE CU RNN ȘI TCN

2.1 PROCESAREA DATELOR

Un vector $\mathbf{X}$ este citit din setul de date pentru fiecare acțiunea, unde $\mathbf{X} \in \mathbb{R}^{C \times T \times V \times M}$ ($C = 3$—număr de coordonate, T—număr de cadre, $V = 25$—număr de puncte de articulație, $M \in \{1, 2\}$—număr de persoane) . Pentru ramura ce conține date despre punctele de articulație, pentru fiecare articulație se adaugă 3 valori, determinate pe baza diferenței dintre coordonatele articulației j_i și cele ale punctului de articulație considerat centru de greutate j_c:

$$joint\_features_{j_i} = (x_{j_i}, y_{j_i}, z_{j_i}, x_{j_i} - x_{j_c}, y_{j_i} - y_{j_c}, z_{j_i} - z_{j_c})$$

(centrul de greutate a fost considerat punctul de articulație cu indicele 1 – *baza coloanei vertebrale*). Pentru ramura de viteză au fost determinate diferențele dintre coordonatele punctului de articulație la cadrul $t + 2$ și cele de la cadrul t, precum și diferențele dintre coordonatele punctului de articulație la cadrul $t + 1$ și cele de la cadrul t:

$$velocity\_features^t_{j_i} = (x^{t+2}_{j_i} - x^t_{j_i}, y^{t+2}_{j_i} - y^t_{j_i}, z^{t+2}_{j_i} - z^t_{j_i}, x^{t+1}_{j_i} - x^t_{j_i}, y^{t+1}_{j_i} - y^t_{j_i}, z^{t+1}_{j_i} - z^t_{j_i})$$

Pentru ramura ce descrie oasele, avem tot 6 caracteristici care includ 3 valori pentru lungimi și 3 valori pentru unghiurile pentru axele X, Y, Z:

$$bone\_features_{(j_u, j_v)} = (x_{j_u} - x_{j_v}, y_{j_u} - y_{j_v}, z_{j_u} - z_{j_v}, a_{(j_u,j_v),x}, a_{(j_u,j_v),y}, a_{(j_u,j_v),z})$$

unde punctele de articulație j_u și j_v sunt adiacente, $l_{(j_u,j_v),x} = x_{j_u} - x_{j_v}, l_{(j_u,j_v),y} = y_{j_u} - y_{j_v}, l_{(j_u,j_v),z} = z_{j_u} - z_{j_v}$ și

$$a_{(j_u,j_v),x} = \arccos\left(\frac{l_{(j_u,j_v),x}}{\sqrt{l^2_{(j_u,j_v),x} + l^2_{(j_u,j_v),y} + l^2_{(j_u,j_v),z}}}\right)$$

$$a_{(j_u,j_v),y} = \arccos\left(\frac{l_{(j_u,j_v),y}}{\sqrt{l^2_{(j_u,j_v),x} + l^2_{(j_u,j_v),y} + l^2_{(j_u,j_v),z}}}\right)$$

$$a_{(j_u,j_v),z} = \arccos\left(\frac{l_{(j_u,j_v),z}}{\sqrt{l^2_{(j_u,j_v),x} + l^2_{(j_u,j_v),y} + l^2_{(j_u,j_v),z}}}\right)$$

.

2.2 METODE DE REARANJARE A ARTICULAȚIILOR

Pentru a extrage dependențe spațiale, am propus două variante de reorganizare a articulațiilor: una 2D (prezentată în Figura 2.1) și una 1D (prezentată în Figura 2.2).

Varianta 2D a fost propusă mai devreme în lucrarea noastră [1] și se bazează pe o matrice 5×5. Varianta 2D prezentată în Figura 2.1 permite aplicarea unui strat de tip Temporal Convolutional Network (TCN) bazat pe convoluții 3D. Această variantă de reprezentare consideră cele 5 părți esențiale ale corpului — mâna stângă, trunchi, mâna dreaptă, picior stâng și picior drept.

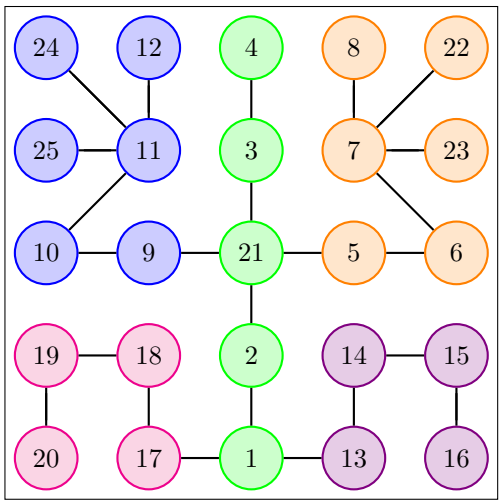

Figura 2.1: O propunere de rearanjare a îmbinărilor într-un format 2D

A doua reorganizare propusă este una liniară și este inspirată de Yang *et al.* [2]. Am ales ca rădăcină pentru arborele propus articulația centrală (cea cu indicele 1), având în vedere că are o importanță deosebită, motiv pentru care a fost folosit și pentru pasul de normalizare. Arborele propus este prezentat în Figura 2.2 și conține toate cele 25 de puncte de articulație. De asemenea, am ținut cont de ordinea în care au fost adăugați sub-arborii pentru rădăcină. Pornind de la acest arbore, am realizat o rearanjare liniară a articulațiilor utilizând parcurgerea DFS (Depth-first search) a arborelui. În acest fel, ne-am asigurat că oricare două noduri care apar unul lângă altul în aranjamentul propus sunt, de asemenea, adiacente în graful scheletului.

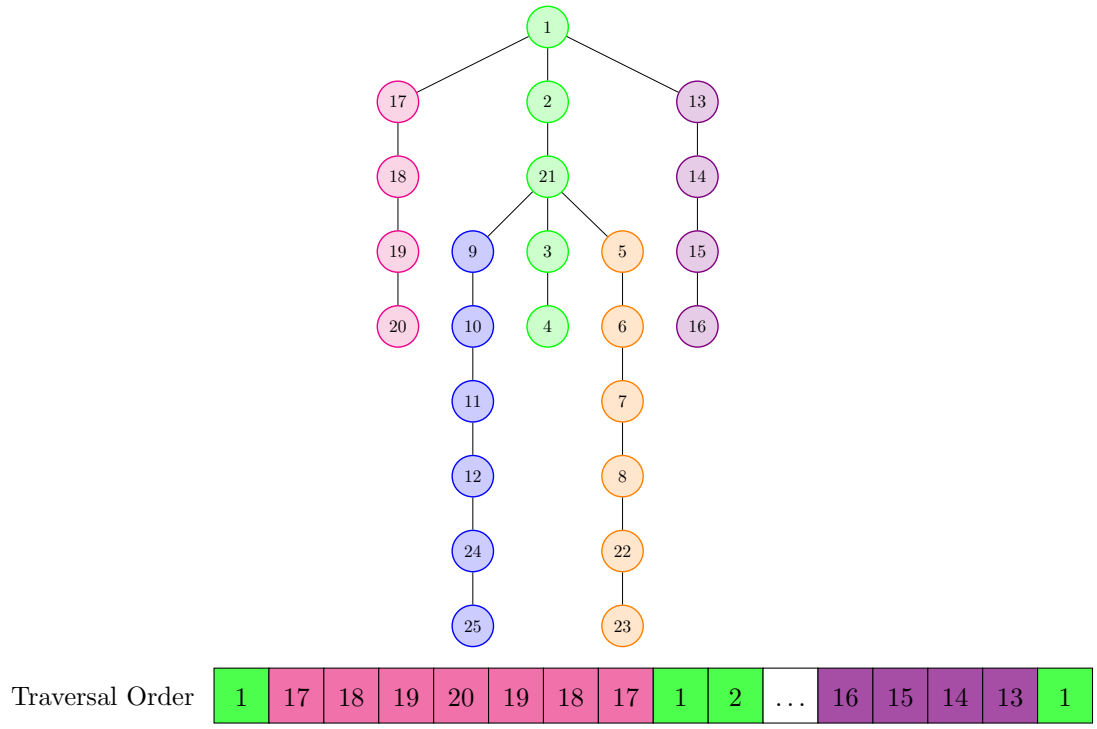

Figura 2.2: Transformarea scheletului într-un arbore având rădăcina articulația 1 (considerat centrul de greutate). Acest arbore este folosit pentru liniarizarea scheletului (realizat pe baza unei parcurgeri în adâncime aplicată arborelui)

2.3 ARHITECTURI BAZATE PE TCN

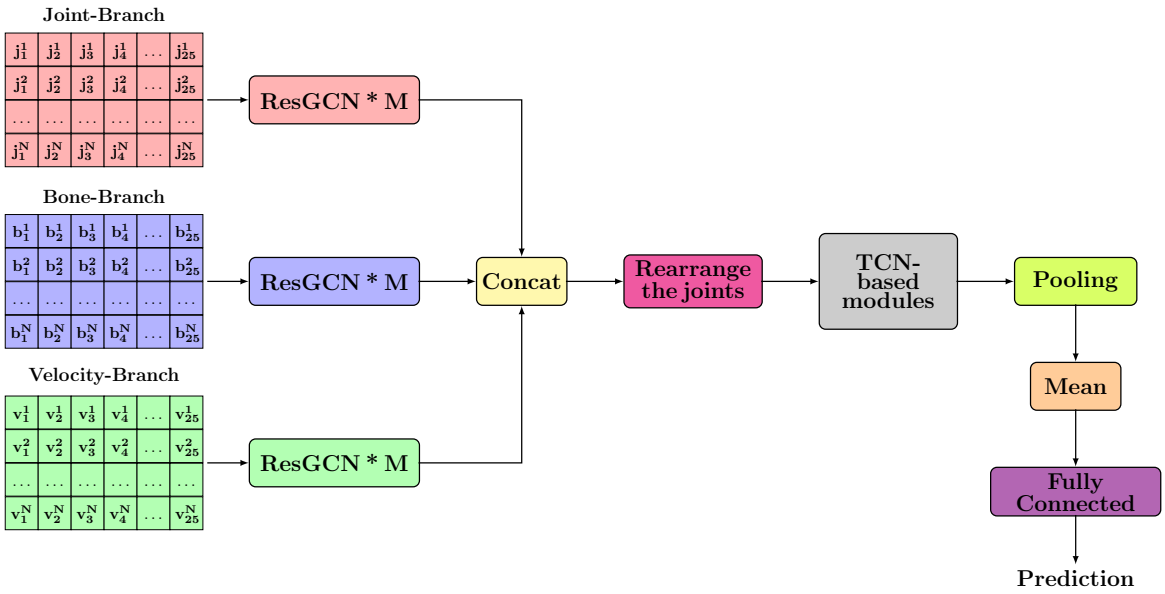

Figura 2.3: Arhitectura generală propusă pentru abordările bazate pe TCN

Schema generală a arhitecturilor propuse bazate pe TCN este prezentată în Figura 2.3. Pentru fiecare ramură, M straturi de tip ResGCN sunt aplicate pentru a extrage caracteristici spațiale. Această parte a extragerii caracteristicilor spațiale este inspirată de arhitecturile

propuse de Song *et al.* [3]. După ce caracteristicile spațiale au fost extrase pentru fiecare ramură, concatenăm toate caracteristicile. Deoarece ne propunem să folosim un modul bazat pe straturi de tip TCN care va putea extrage simultan atât caracteristici temporale cât și spațiale, este necesar să se efectueze o rearanjare a articulațiilor. Rearanjamentele propuse sunt detaliate în Secțiunea 2.2.

Pentru a analiza secvența și a extrage caracteristicile dintr-o perspectivă temporală, am decis să folosim straturi TCN. Astfel, am început prin a testa mai multe tipuri de blocuri bazate pe straturi TCN. Inițial, am folosit un modul bazat pe blocuri TCN inspirat de modelele propuse anterior în [1]. Dezavantajul lor major a fost că nu au păstrat dimensiunea spațială, deoarece unitatea TCN era bazată pe convoluția 1D. Prin urmare, am efectuat concatenarea caracteristicilor extrase pentru fiecare articulație. Apoi, am folosit tensorul 2D rezultat ca intrare pentru unitățile TCN.

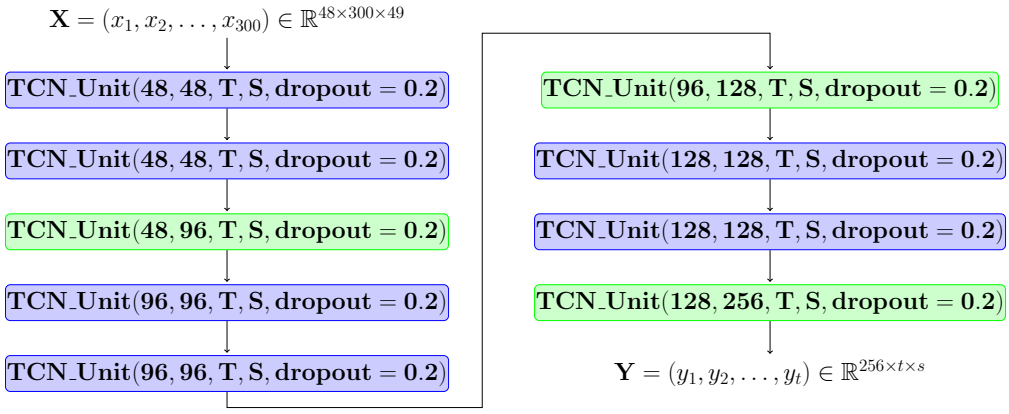

Figura 2.4: T reprezintă dimensiunea ferestrei temporale, iar S reprezintă dimensiunea ferestrei spațiale. Pentru blocurile albastre, pasul are valoarea 1, iar pentru cele verzi, pasul are valoarea 2. 300 reprezintă numărul maxim de cadre, iar 49 reprezintă numărul de articulații analizate (rezultate după rearanjarea liniară descrisă în Secțiunea 2.2). Pentru fiecare unitate de tip TCN, bordarea este determinată pe baza valorilor T și S.

2.4 ARHITECTURI BAZATE PE RNN

Arhitectura utilizată pentru abordarea bazată pe Recurrent Neural Network (RNN) este prezentată în Figura 2.5. Această arhitectură este similară cu cea prezentată anterior în Secțiunea 2.3.

În această arhitectură, straturile inițiale au fost aplicate independent pentru fiecare schelet. În cele din urmă, a fost calculată o medie a caracteristicilor extrase. Apoi sunt păstrate doar caracteristicile corespunzătoare stării ascunse finale pentru fiecare eșantion și sunt trecute printr-un strat Complet Conec pentru a realiza clasificarea.

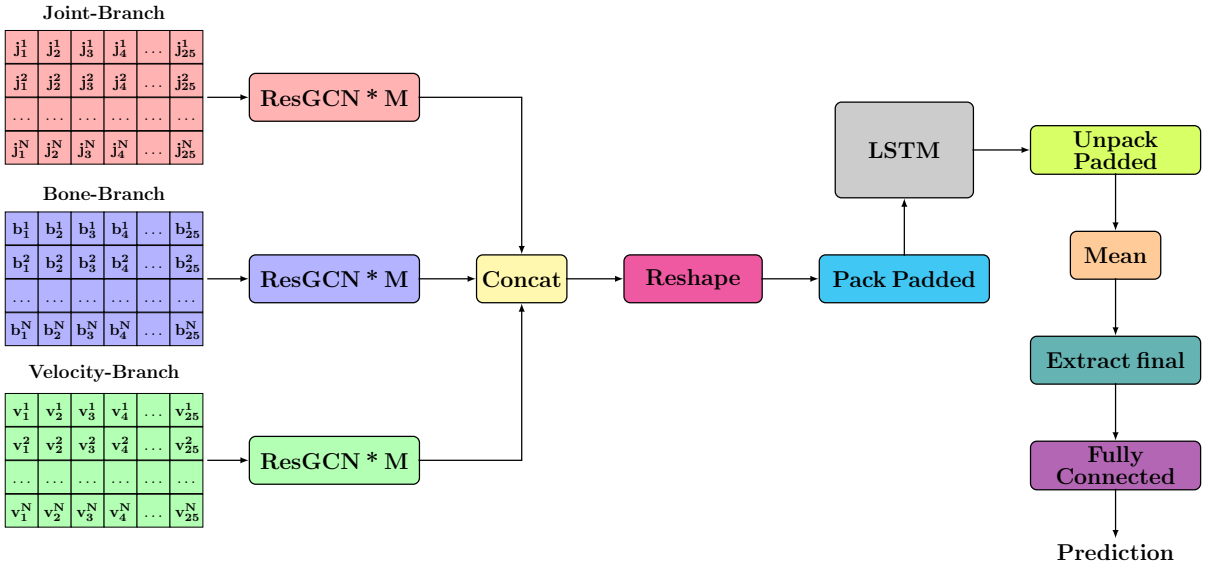

Figura 2.5: Arhitectura propusă pentru abordarea bazată pe RNN

2.5 DISCUȚIE

Am testat un model arhitectural similar bazat pe Long short-term memory (LSTM) pentru a evidenția importanța metodelor noastre bazate pe o unitate de tip TCN extinsă. Chiar dacă rata de inferență este mai mică pentru abordarea bazată pe TCN, acest aspect ar putea fi îmbunătățit dacă s-ar folosi proprietatea de paralelizare a acestui tip de rețea neuronală. Spre deosebire de RNN, unde calculele pentru marcajele de timp ulterioare trebuie să aștepte ca predecesorii lor să se finalizeze, convoluțiile pot fi calculate în paralel chiar și pe arhitecturi SIMD simple și puternice, precum cele găsite în plăcile grafice, deoarece același nucleu sau nuclee foarte asemănătoare sunt independente. folosit în fiecare strat în mod repetat. În schimb, în ceea ce privește performanța, cele mai bune rezultate au fost obținute pentru abordările bazate pe TCN.

În cazul metodelor bazate pe TCN, pentru sample-urile care au conținut mai puțin de 300 de cadre, se aplică operația de bordare cu valoarea 0. În schimb, pentru abordările bazate pe LSTM, acest aspect este evitat prin utilizarea unor operații de optimizare specifice. Acesta poate fi unul dintre motivele pentru care viteza de inferență obținută pentru abordările bazate pe TCN este mai mică decât cea obținută atunci când se utilizează LSTM.

Un avantaj important al arhitecturilor bazate pe TCN este capacitatea de a-și schimba dimensiunea câmpului receptiv în multe feluri. De exemplu, stivuirea unor straturi convoluționale dilatate (cauzale), utilizarea unor factori de dilatare mai mari sau creșterea dimensiunii nucleului sunt toate opțiuni posibile, fiecare cu avantajele și dezavantajele sale în funcție de detaliile mai fine ale fiecărei implementări. Acest lucru ne permite să folosim diferite valori pentru câmpul receptiv în funcție de domeniu. Cele mai bune performanțe au fost obținute atunci când am folosit dimensiunea kernel-ului egală cu 5 pentru domeniul spațial și egală cu 9 pentru domeniul temporal.

3

RECUNOAȘTEREA ACȚIUNII UMANE ÎN CADRUL PLATFORMEI ROBOTICE SOCIALE AMIRO

Integrarea unui modul specializat în recunoașterea acțiunii într-un framework reprezintă o problemă provocatoare. În acest capitol, prezentăm platforma robotică AMIRO, subliniind modul în care am dezvoltat componenta pentru HAR în cadrul acestei platforme. Astfel, descriem pipeline-ul general de integrare propus împreună cu două modele neurale specializate în recunoașterea a 8 acțiuni umane. Aceste rezultate nu ar fi fost posibile fără ajutorul Alexandrei Ștefania Ghiță.

3.1 PLATFORMA AMIRO

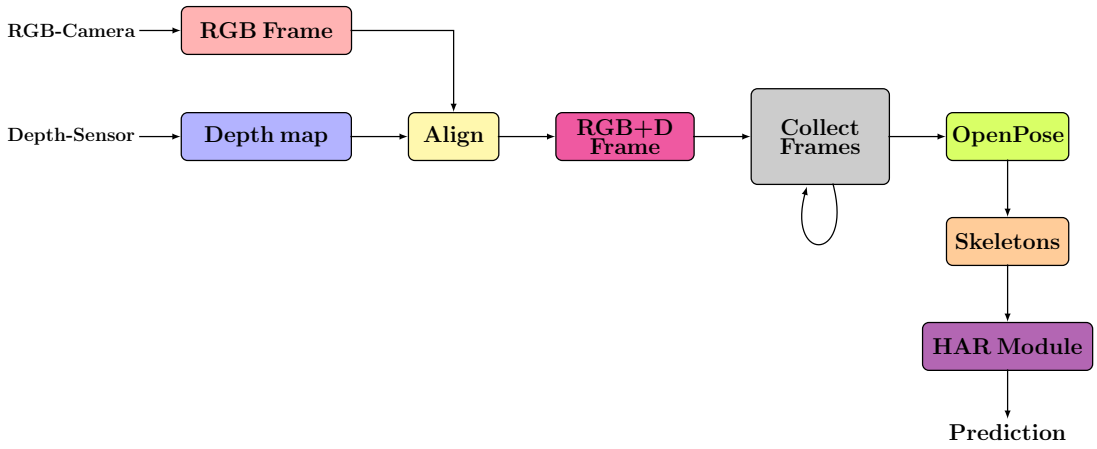

Figura 3.1: Pipeline-ul componentei care recunoaște acțiunile umane (această componentă este integrată în modulul Vision)

Arhitectura componentei utilizate pentru recunoașterea acțiunilor umane este prezentată în Figura 3.1. După cum se poate observa, această componentă cuprinde 3 submodule importante:

1. Modul de achiziție de date;
2. Modul de extragere a caracteristicilor;
3. Modul pentru HAR.

Primul modul este cel care colectează și prelucrează datele furnizate de cei doi senzori importanți ai robotului: RGB-Camera și Depth-Sensor. Deoarece robotul nu este echipat cu

un singur senzor care să permită colectarea ambelor tipuri de date, a fost necesară alinierea datelor din perspectivă temporală și spațială. După ce au fost prelucrate, aceste date sunt colectate și, când devin suficiente, ele încep să fie analizate de către modulul următor. Deoarece modulul HAR funcționează cu date scheletice, a fost necesar să se introducă un modul intermediar care are ca scop transformarea cadrelor RGB+D în date scheletice.

3.2 PIPELINE GENERAL PENTRU RECUNOAȘTEREA ACȚIUNII

Pentru a avea un robot care poate interacționa cu oamenii, robotul trebuie să identifice acțiunile efectuate de persoana pe care o monitorizează sau căreia trebuie să îi trimită o notificare. Astfel, modulul de recunoaștere a activităților umane a fost inclus în framework-ul propus și testat utilizând robotul Pepper, pentru a obține un robot de asistență socială.

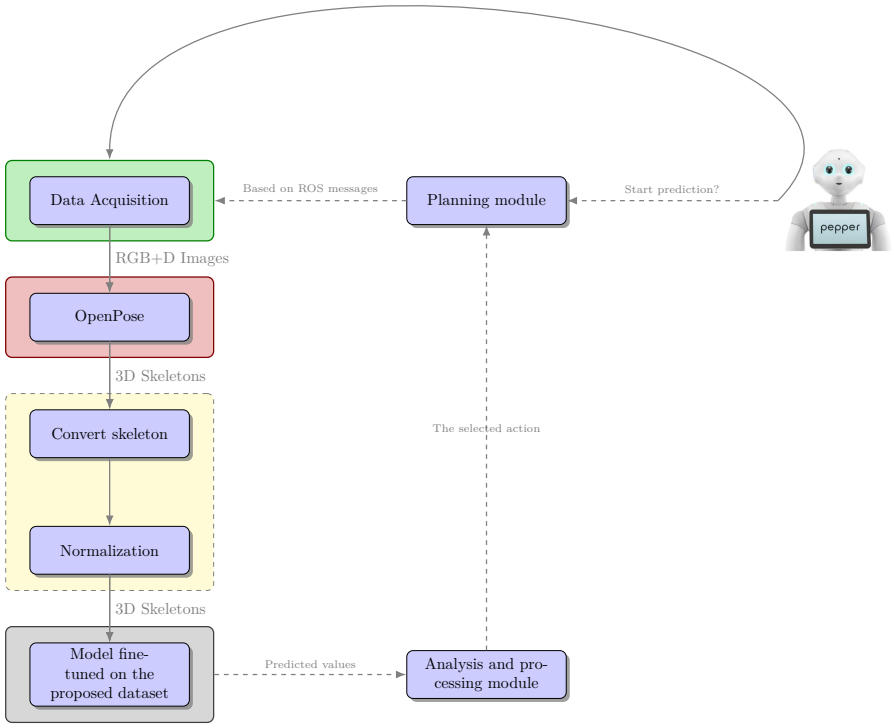

Figura 3.2: Arhitectura procesului complet de integrare a modulului de recunoaștere a acțiunii umane

3.3 ARHITECTURA MULTI-STAGIU

Pipeline-ul întregului proces de integrare este prezentat în Figura 3.2, iar arhitectura de rețea folosită drept clasificator al acțiunii umane este prezentată în Figura 3.3. În lucrarea [4], am prezentat o arhitectură care nu a reușit să diferențieze corect acțiuni similare (cum ar fi *bea apă* și *tuși* sau *a face cu mână* și *indicând spre ceva*). Astfel, am decis să propunem o

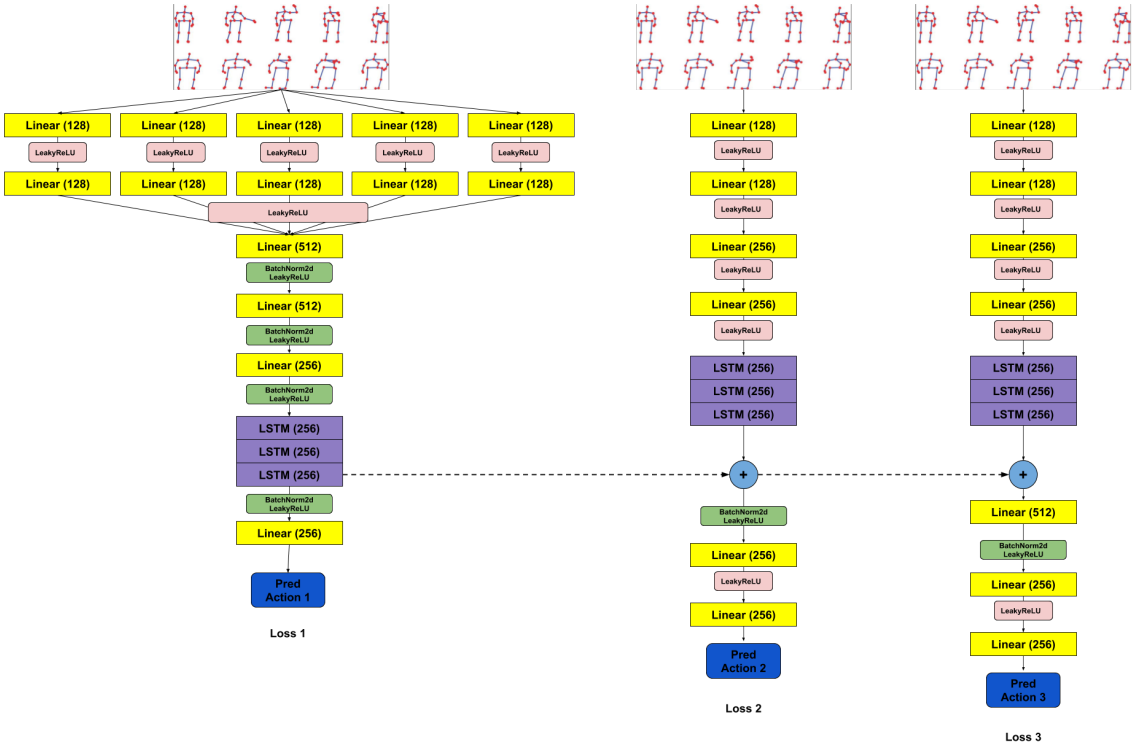

Figura 3.3: Arhitectura rețelei neurale folosită pentru recunoașterea acțiunii umane

nouă arhitectură extinsă care să conțină două etape suplimentare care primesc ca intrare secvența procesată cu coordonatele scheleților, dar ia în considerare și informații din etapa anterioară. Fiecare etapă folosește o serie de straturi liniare pentru a extrage caracteristici din secvența de coordonate scheletice și o rețea LSTM pentru analiza acestor secvențe temporale. A fost aplicată o funcție de eroare pentru fiecare etapă. Clasificatorul de acțiuni utilizat a fost antrenat pe setul de date NTU RGB+D [5] și apoi s-a specializat pe un set de date colectat folosind robotul Pepper. Setul de date colectat cu robotul Pepper conține un subset de 8 acțiuni considerate a fi relevante și provocatoare pentru un robot folosit ca asistent personal. Deoarece există acțiuni foarte asemănătoare (de exemplu, *jucat cu telefonul/tableta* și *tastarea pe o tastatură*), în cadrul acestui subset de acțiuni selectate, era necesar un model complex pentru a putea diferenția corect acțiunile. Astfel, am propus o arhitectură compusă din trei etape. După ce acest clasificator a fost antrenat, rezultatul furnizat de etapa a treia a fost folosit ca predicție finală. Arhitectura clasificatorului este o versiune îmbunătățită a unui model testat și analizat în lucrarea noastră anterioară [1].

3.4 ARHITECTURĂ BAZATĂ PE TCN

O altă abordare utilizată pentru clasificatorul acțiunii umane a fost bazată pe straturi TCN. Arhitectura acestui tip de clasificator este prezentată în Figura 3.4. Inițial se aplică transformări care asigură preprocesarea datelor. Aceste transformări au două scopuri fundamentale: normalizarea datelor și adăugarea de descriptori legați de mișcare (viteză și accelerație). După etapele de preprocesare, datele sunt trecute prin niște strat-

uri convoluționale responsabile cu extragerea informațiilor spațiale relevante. Aceste informații sunt analizate, din perspectivă spațială și temporală, prin straturi de tip TCN. Straturile TCN colorate în albastru din Figura 3.4 sunt cele care vor păstra dimensiunile, iar cele colorate în verde sunt straturile care vor modifica dimensiunile (cel spațial – numărul de articulații și cel temporal - numărul de cadre). Caracteristicile determinate după aplicarea straturilor de tip TCN sunt trecute printr-un strat de pooling și apoi clasificate folosind un strat complet conectat.

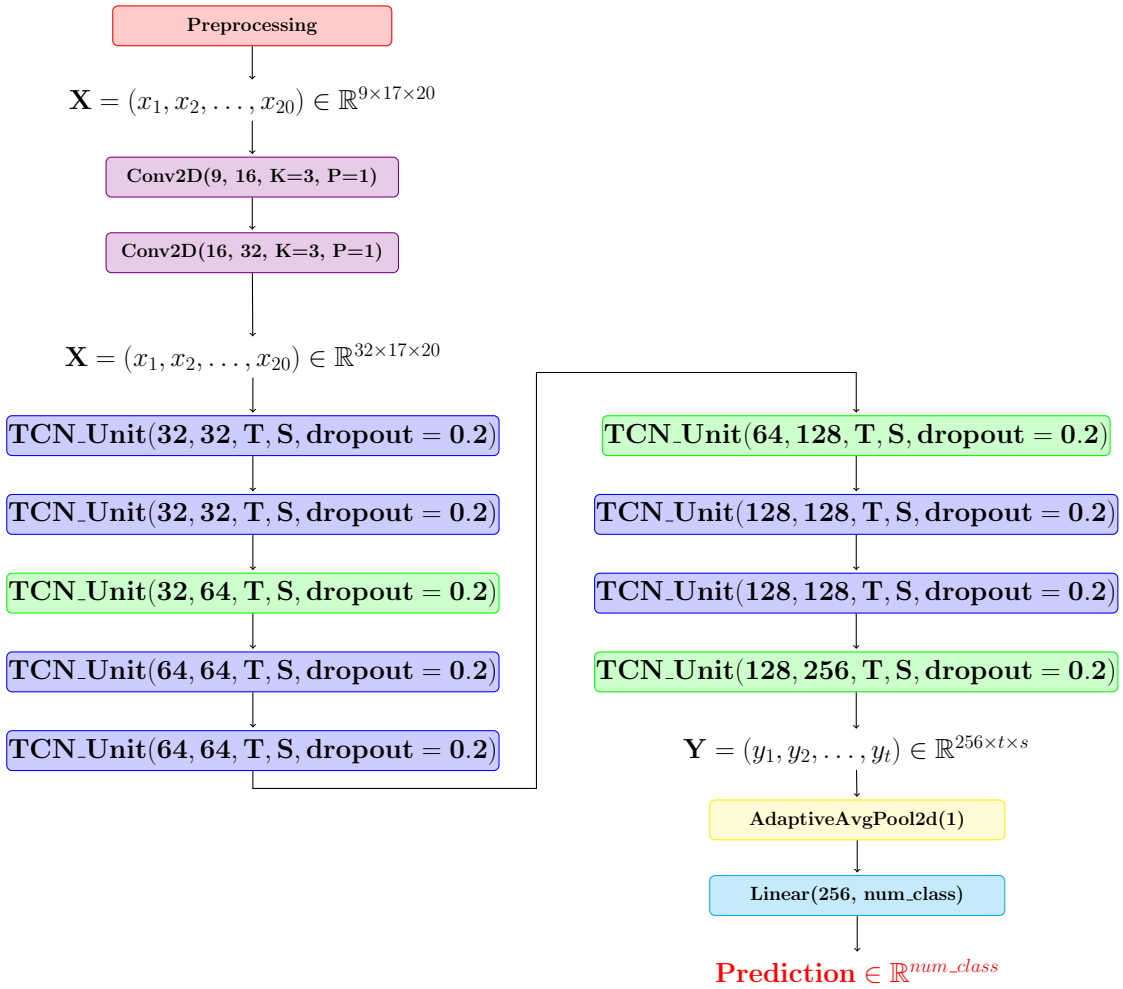

Figura 3.4: Arhitectura clasificatorului bazat pe TCN. În cazul straturilor convoluționale simple, K reprezintă dimensiunea nucleului, iar P reprezintă valoarea pentru bordare. T este valoarea utilizată pentru *kernel_size* corespunzătoare dimensiunii temporale, iar S este valoarea *kernel_size* corespunzătoare dimensiunii spațiale.

3.5 SET DE DATE CU PERCEPȚIE ROBOTICĂ

Setul nostru de date conține înregistrări de 720, aproximativ 15% din dimensiunea setului de date NTU RGB+D, cu 90 de înregistrări pentru fiecare acțiune selectată. Acțiunile au fost realizate de 10 participanți. Fiecare acțiune a fost înregistrată de 3 ori, din 3 unghiuri diferite, pentru a simula diferitele unghiuri din care robotul poate vedea o persoană.

Fiecare set de înregistrări pentru acțiune a fost filmat în 3 scene diferite, scenele fiind în interiorul unei clădiri, cu lumină artificială. Am introdus acest set de date în lucrarea [4].

3.6 DISCUȚIE

Scenariul de evaluare propus evidențiază o serie de dezavantaje prezentate de versiunea actuală a modulului de recunoaștere a acțiunii umane. Dacă robotul este prea aproape de subiect, atunci scheletul prezis de OpenPose [6] este incomplet. În plus, dacă există ocluzii cu alte obiecte, atunci scheletul prezis este incomplet sau coordonatele unor articulații sunt incorect prezise. Având în vedere că modulul de recunoaștere a activităților umane a fost antrenat folosind mostre în care au apărut coordonatele tuturor articulațiilor, în astfel de situații cu schelet parțial se obțin rezultate slabe. De asemenea, atunci când robotul se apropie de utilizator, aspectul poate fi modificat, ceea ce duce la o recunoaștere mai dificilă a activității. Acest lucru poate fi atenuat în două moduri: prin antrenament cu un set de date mai divers, devenind mai robust împotriva ocluziilor articulare sau a distanței de observare, precum și prin îmbunătățirea sarcinii de recunoaștere a activității.

4

REȚEA NEURALĂ SPAȚIO-TEMPORALĂ CU CARACTERISTICI CALCULATE MANUAL

În acest capitol, prezentăm abordarea rețelei neurale spațio-temporale cu caracteristici geometrice determinate manual, care constă într-o fază de preprocesare a datelor urmată de aplicarea unui model neural spațio-temporal pentru a recunoaște acțiunea. Modelul neural introdus în această secțiune este unul cu mai multe ramuri de intrare bazate pe straturi TCN și GCN. Spre deosebire de arhitecturile neurale existente, modelul nostru prezintă un timp de inferență redus, obține rezultate comparabile cu metodele de ultimă generație și oferă posibilitatea de a determina o hartă de activare care poate fi utilă în procesul de explicabilitate.

4.1 MODELUL SPAȚIO–TEMPORAL

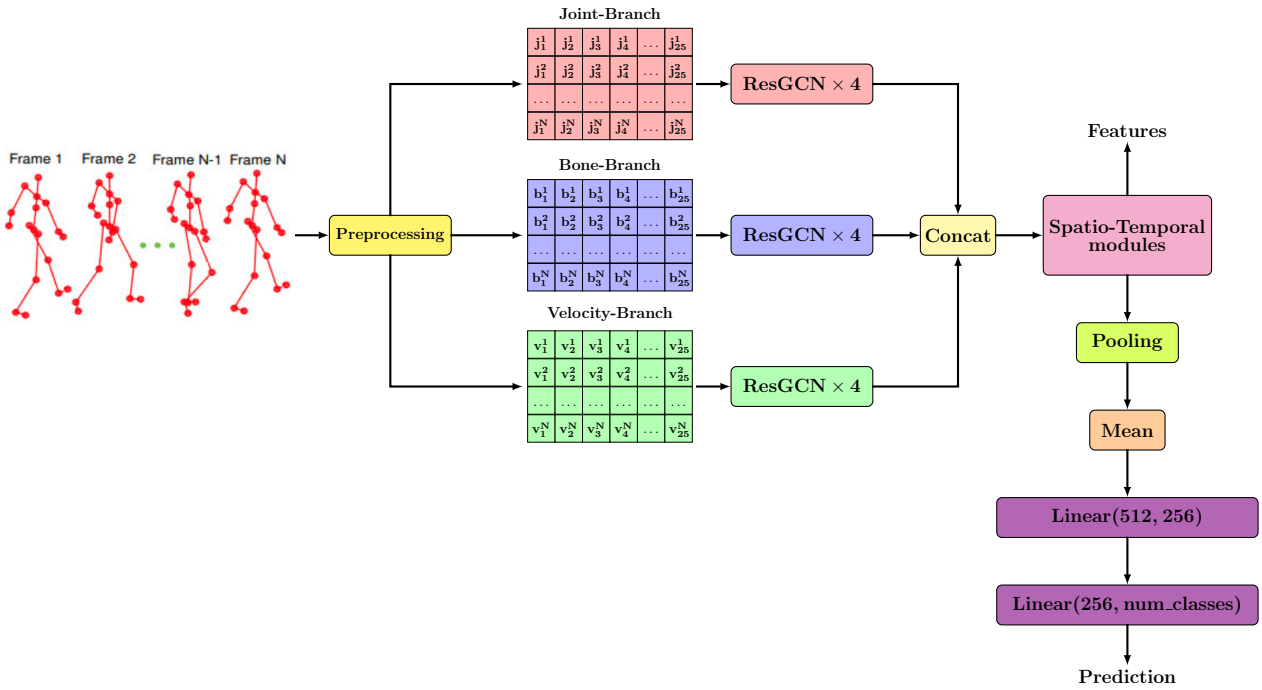

Figura 4.1: Structura propusă pentru modelul neural utilizată pentru a rezolva problema recunoașterii acțiunii umane

Arhitectura generală pentru abordarea propusă este prezentată în Figura 4.1. Pentru a normaliza și extinde caracteristicile rezultate din procesul de preprocesare, am folosit 4

straturi ResGCN pentru fiecare ramură. Concatenăm datele rezultate din aplicarea acestor straturi și folosim tensorul rezultat ca intrare pentru un modul Spațio-Temporal. Arhitectura propusă returnează două tipuri de rezultate: caracteristicile obținute în urma aplicării modulului Spațio-Temporal și predicția finală.

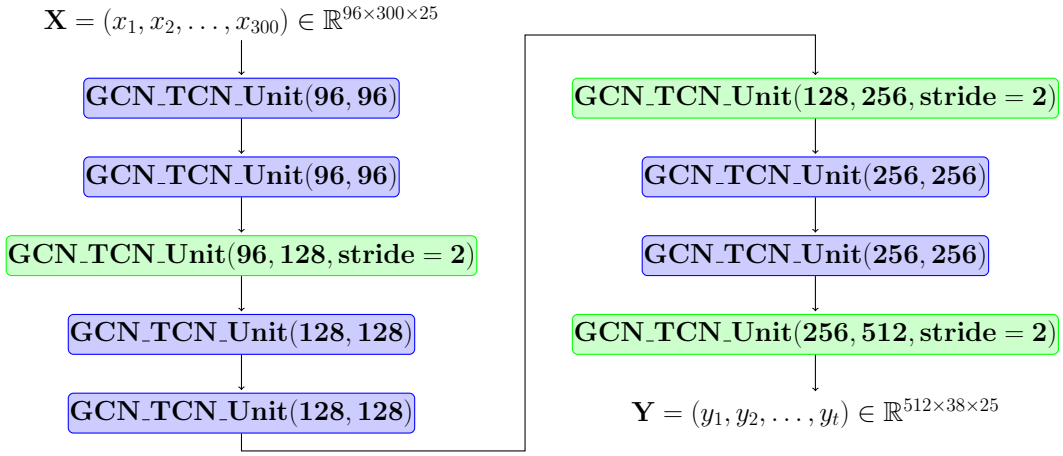

Figura 4.2: Structura modulului Spațio-Temporal inclusă în arhitectura generală prezentată în Figure 4.1. Straturile evidențiate în albastru folosesc pași cu valoarea egală cu unu și păstrează ambele dimensiuni (spațială și temporală)

Pentru proiectarea modulului Spatio-Temporal am folosit unități de tip GCN–TCN. Arhitectura unui astfel de bloc este prezentată în Figura 4.3. Aceste blocuri au fost propuse de Chen *et al.* în [7].

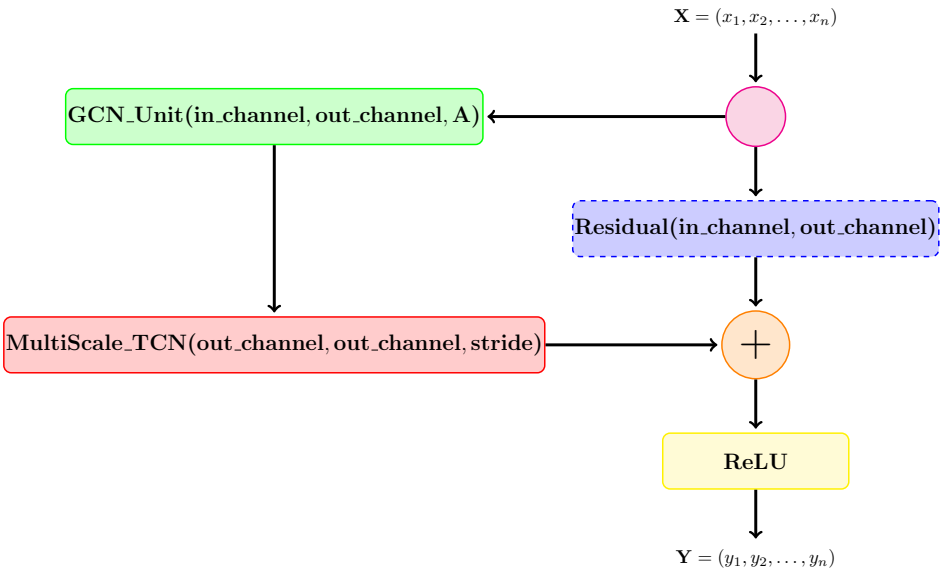

Figura 4.3: Arhitectură utilizată pentru unitățile de tip GCN–TCN. **A** reprezintă matricea care descrie graful. Stratul rezidual este aplicat numai dacă *in_channel* $\neq$ *out_channel*. Această arhitectură a fost propusă de Chen *et al.* în [7]

4.2 EXPLICAREA PREDICȚIEI REȚELEI

O explicație a motivului pentru care rețeaua neurală prezice un o clasă poate aduce informații valoroase atât pentru cei care dezvoltă aplicații bazate pe HAR, cât și pentru progresele viitoare în rezolvarea acestei probleme. Spre deosebire de metodele care explică predicția de rețea pentru imagini, există destul de puține propuneri care încearcă să explice HAR, iar majoritatea sunt limitate la modele 3D Convolutional Neural Network (CNN). Doar o lucrare recentă [3] propune o vizualizare a scheletelor, care încearcă să descopere cele mai esențiale părți ale corpului pe o întreagă secvență de acțiuni, în încercarea de a obține reprezentări mai explicabile pentru diferite secvențe de acțiuni. Considerăm că explicabilitatea HAR bazată pe un model schelet este o direcție foarte promițătoare și solidă pentru înțelegerea predicției rețelei și modelul nostru propus a fost dezvoltat pornind tot de la această premisă.

Pentru a realiza o explicație a acțiunii recunoscute, am folosit caracteristicile rezultate din aplicarea modulului Spațio-Temporal și ponderile ultimelor două straturi liniare. Pornind de la acestea, am determinat pentru fiecare cadru care sunt cele mai importante articulații luate în considerare de rețea în ceea ce privește activările și le-am ilustrat într-o Hartă de Activare. Am luat în considerare și importanța pe care rețeaua o acordă fiecărui cadru. La fel, în cazul în care apar două schelete, am verificat importanța fiecăruia.

Figura 4.4: Eșantion din subsetul de testare pentru acțiunea *spălarea dinților*. Rețeaua prezice corect această acțiune cu o probabilitate de 100%

Pentru a evidenția unele rezultate calitative obținute la probele din setul de testare, folosind protocolul pentru care rețeaua a obținut cel mai mic punctaj, Cross-Subject v2, am efectuat testarea din perspectiva a 3 acțiuni: *spălarea dinților* (Figura 4.4), *bea apă* (Figura 4.5) și *eat meal or snack* (Figura 4.6).

Am prezentat în Figura 4.4 câteva cadre dintr-un eșantion pentru acțiunea *spălarea dinților*. Am selectat aceste cadre folosind un pas de 10 cadre. Pentru aceste cadre, punem în evidență scheletul uman și alegem culorile pentru fiecare articulație în funcție de importanța generată de modelul neural. Articulațiile considerate de model neimportante le-am colorat în albastru, iar, în rest, am folosit culoarea roșie având în vedere intensitatea generată de model. Rețeaua recunoaște corect momentul cheie al acțiunii. Intensitatea articulațiilor este evidențiată în imaginile 4, 5, 6, 7, 8.

Figura 4.5: Eșantion din subsetul de testare pentru acțiunea *bea apă*. Rețeaua prezice corect această acțiune cu o probabilitate de 100%. Am eșantionat cadrele cu un pas de 5 cadre

Subliniem în Figura 4.5 un exemplu care prezintă o eroare generată de senzorul Kinect. Pentru fiecare cadru apar două schelete, chiar dacă o singură persoană realizează acțiunea. Senzorul a confundat scaunul din imagine cu un schelet, pentru care prezice unele date distorsionate. În abordarea noastră, oferim datele din cele două schelete ca intrare pentru rețea. Rețeaua recunoaște corect scheletul de interes, ignorându-l complet pe cel fals. Acest exemplu evidențiază robustețea modelului nostru și motivează corectitudinea predicției. În toate exemplele, distingem o diferență între intensitatea culorii pentru articulațiile mâinii care efectuează acțiunea și celelalte. Acest lucru este vizibil mai ales în imaginile 6 și 7. De asemenea, observăm că, în ultimele cadre, importanța asociată îmbinărilor scade deoarece coordonatele nu se schimbă considerabil. Merită subliniat că, pentru acest eșantion, modelul face o predicție corectă cu o încredere de 100%.

În ultimul exemplu calitativ selectat, am prezentat o acțiune repetitivă. După cum se vede din cadrele incluse în Figura 4.6, rețeaua surprinde acest aspect. În aceste imagini este prezentată acțiunea *mănâncă*. Modelul prezice corect această acțiune cu o încredere foarte mare (99.7%) și recunoaște cadrele în care persoana începe să mănânce. De data aceasta, fiind o acțiune repetitivă, dimensiunea temporală este cea căreia modelul îi acordă mai multă atenție. Prin urmare, doar în ultimele imagini s-a putut distinge o discrepanță între importanța asociată fiecărei articulații la același moment de timp.

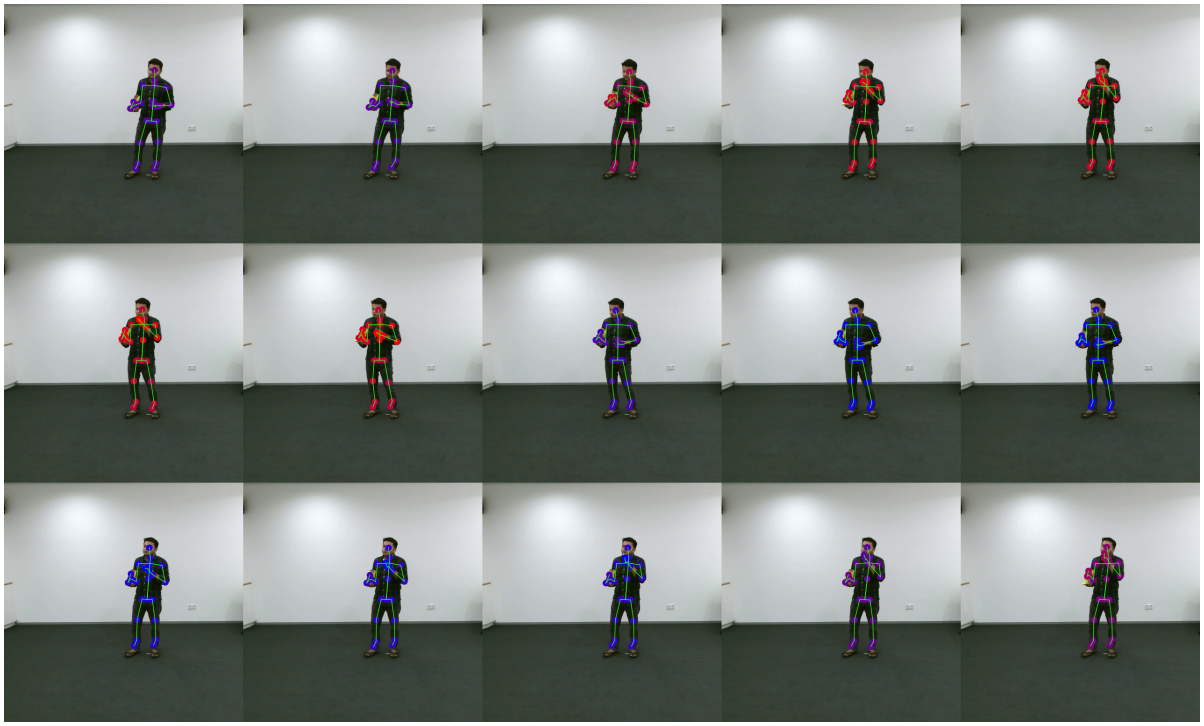

Figura 4.6: Eșantion din subsetul de testare pentru acțiunea *mănâncă*. Rețeaua prezice corect această acțiune cu o probabilitate de 99,7%. Am eșantionat cadrele cu un pas de 5 cadre

4.3 DISCUȚIE

În acest capitol am propus o metodologie de recunoaștere a acțiunii umane care constă într-o etapă de preprocesare, în care sunt determinate caracteristici geometrice și sunt aplicate tehnici de normalizare a datelor pentru a obține o performanță mai bună, urmată de o arhitectură de rețea neurală spațio-temporală care combină straturile TCN și GCN pentru a captura atât dimensiunea spaţială cât şi cea temporală a acțiunii. Am arătat că modelul propus de noi este capabil să obțină rezultate de acurateţe similare celor de ultimă generație și are un timp de procesare mai mic, un comportament robust în cazul unor schelete identificate incorect de către senzori și capacitatea de a explica datele recunoscute prin evidențierea celor mai importante articulații considerate de rețea în ceea ce privește activările și importanța pe care rețeaua o acordă fiecărui cadru.

Am efectuat o analiză amănunțită a comportamentului rețelei pe cele 2 versiuni ale setului de date NTU RGB+D (60 și 120 de acțiuni) pentru toate cele 4 protocoale propuse pentru testare, atât în cazul acțiunilor corect recunoscute, cât și în cazul celor incorect recunoscute. De asemenea, am evidențiat faptul că erorile de clasificare sunt generate de acțiuni foarte asemănătoare (unele chiar greu de identificat și pentru om).

Pe baza caracteristicilor oferite de model și a ponderilor din ultimele două straturi de tip strat Liniar, putem genera statistici care prezintă cele mai importante articulații luate în considerare de model și cele mai relevante cadre în acțiunea efectuată. Astfel, am efectuat analiza și vizualizarea motivelor din spatele predicțiilor pentru acțiunile efectuate de una sau două persoane, pentru acțiuni individuale sau repetitive, arătând cum rețeaua acordă o importanță dimensiunii spațiale și/sau dimensiunii temporale.

5

MODEL RAPID CU CONVOLUȚIE TEMPORALĂ PE GRAFURI PENTRU RECUNOAȘTEREA ACȚIUNILOR

În acest capitol, prezentăm abordarea propusă pentru problema recunoașterii acțiunii umane, care constă dintr-o etapă de preprocesare și un model neural bazat pe diferite tipuri de straturi de convoluții. Această abordare reprezintă o îmbunătățire față de contribuția anterioară descrisă în Capitolul 2.

5.1 MODELUL TEMPORAL GRAPH CONVOLUTIONAL

Pipeline-ul complet proiectat pentru soluția noastră este prezentat în Figura 5.1. Conține două etape: o etapă pentru calcularea caracteristicilor și o etapă pentru aplicarea modelului neural propus.

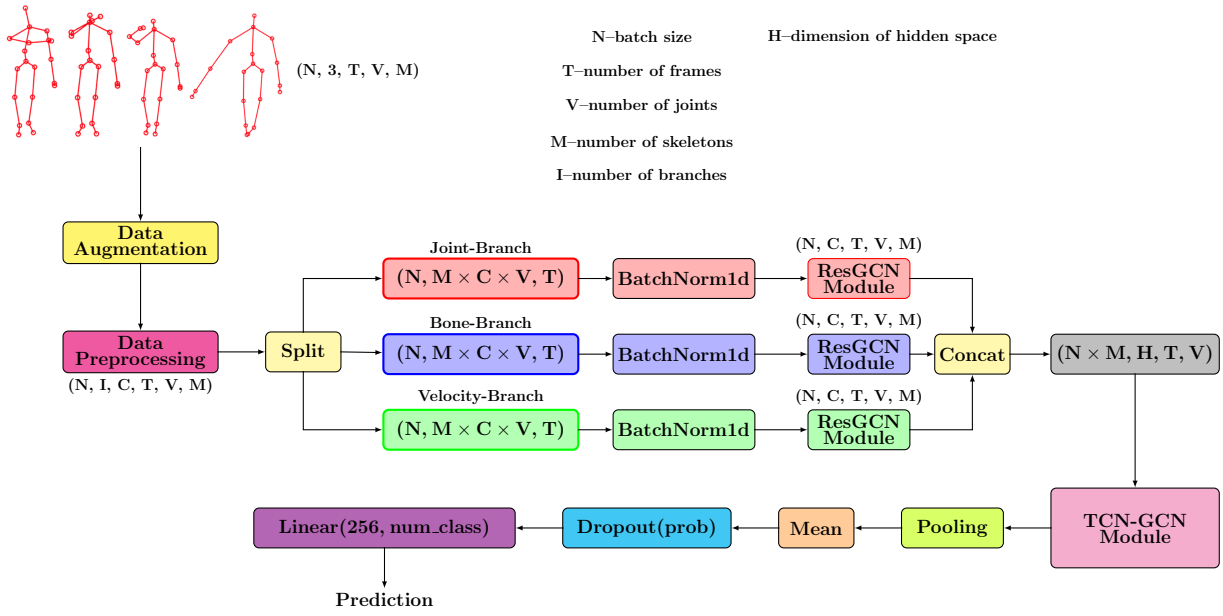

Figura 5.1: Pipeline-ul propus constă din două componente fundamentale: etapa de prelucrare a datelor și modelul convoluțional

5.1.1 *Modulul ResGCN*

În pipeline-ul propus în cadrul metodei noastre, folosim independent un modul bazat pe blocuri ResGCN pentru a extrage dependențe spațiale sau temporale din datele corespunzătoare fiecărei ramuri. În figura 5.2, evidențiem structura propusă pentru acest modul. Tensorul furnizat ca intrare are numărul de canale $C = 6$, iar modulul îl remodelează pentru a fi aplicat independent pentru fiecare schelet. Astfel, dimensiunea batch-ului devine egală cu $N \cdot M$, unde N reprezintă dimensiunea inițială a lotului și M este numărul de schelete.

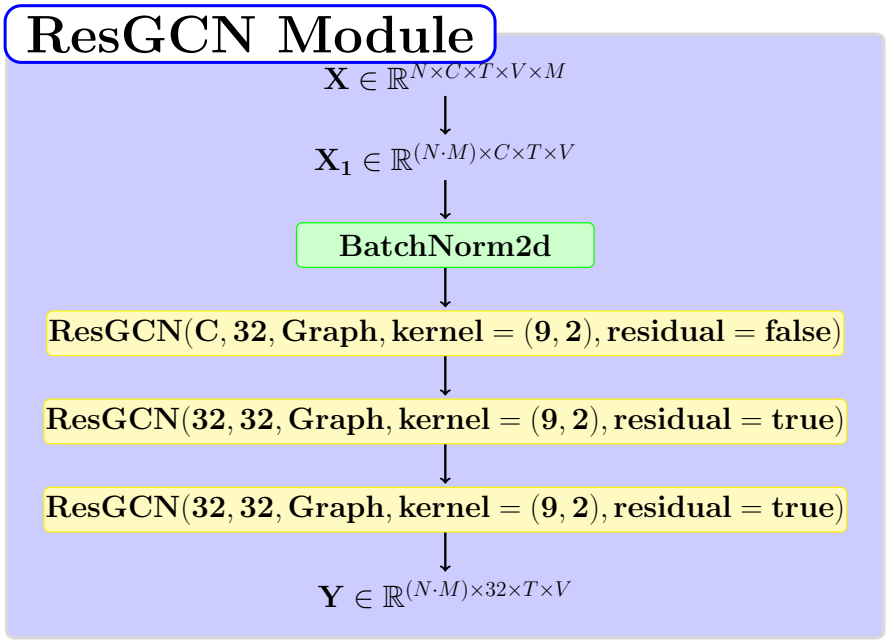

Figura 5.2: Structura modulului bazat pe straturi ResGCN

Graful folosit pentru aceste unități de tip ResGCN este unul în care distanța maximă este setată la 2. Structura de tip bottleneck utilizată pentru blocurile spațiale și temporale asigură un model cu viteză mare de inferență și care necesită un număr redus de epoci pentru antrenare. Mai mult, modulul prezintă o optimizare suplimentară, obținută prin modificarea dimensiunii batch-ului și aplicarea de operații în paralel pentru cele două schelete.

5.1.2 *Modulul TCN-GCN*

Modulul spațio-temporal reprezintă partea fundamentală a modelului neural propus, iar structura acestuia este descrisă în Figura 5.3. Unitatea de bază a acestui modul este blocul GCN-TCN propus de Chen et al. [7]. Acest bloc este compus din două dintre cele mai relevante tipuri de straturi neurale pentru problema recunoașterii acțiunii umane: Graph Convolutional Network (GCN) și TCN. În abordarea noastră, am folosit pentru aceste blocuri un graf în care vecinătatea fiecărui nod conține întregul graf ce descrie scheletul. Acest modul este compus din 9 straturi aplicate secvențial și fiecare este o unitate GCN-TCN. Trei

dintre aceste straturi înjumătățesc dimensiunea temporală utilizând o valoare a pasului de 2. Straturile de culoare albastră din Figura 5.3 păstrează dimensiunea temporală.

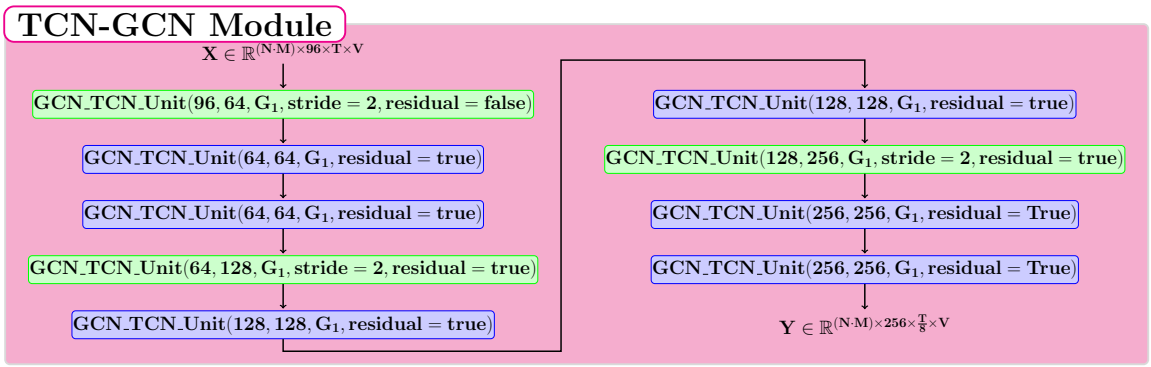

Figura 5.3: Structura propusă pentru modulul TCN-GCN

5.2 REZULTATE EXPERIMENTALE

Table 5.1: Performanțele modelului propus pentru protocolul Cross-Subject (v1—60) în funcție de numărul de epoci utilizat la antrenare

Metoda	Numărul total de epoci	Top 1	Top 5	Epoca Scorului
Fast Convolutional (200 frames, dropout = 0.3, theta = 0.3)	100	89.37	98.25	97
Fast Convolutional (200 frames, dropout = 0.3, theta = 0.3)	70	89.37	98.17	66
Fast Convolutional (200 frames, dropout = 0.3, theta = 0.3)	50	89.05	98.20	49
Fast Convolutional (200 frames, dropout = 0.3, theta = 0.3)	30	87.65	98.09	29
Fast Convolutional (200 frames, dropout = 0.3, theta = 0.3)	10	82.73	97.13	10

În Tabelul 5.1, am inclus o analiză a rezultatelor experimentale obținute prin modelul propus pentru protocolul de testare Cross-Subject. Am obținut aceste rezultate prin modificarea numărului de epoci utilizate în procesul de instruire. Pe baza acestora, am evidențiat faptul că modelul propus nu necesită un număr mare de epoci pentru antrenare. După cum se poate observa, nu există nicio diferență între acuratețea Top 1 obținută pentru 100 de epoci și cea obținută pentru doar 70 de epoci. Mai mult, diferența dintre precizia Top 5 pentru aceste două configurații de antrenare este nesemnificativă. Datorită numărului

mic de parametri și transformărilor propuse în această abordare, se poate obține o precizie de 82,73% folosind doar 10 epoci. Cu alte cuvinte, utilizarea unui număr mai mare de epoci de antrenare ajută modelul să facă distincția între clase similare pentru care prezice probabilități apropiate. Mai mult, putem observa că prin introducerea transformărilor de augmentare și folosind un strat de tip Dropout cu o probabilitate de 0,3, am reușit să obținem un model cu o putere de generalizare bună, care nu mai este supra-specializat. Această observație este certificată de rezultatele obținute pentru 100 epoci care nu sunt mai slabe decât cele pentru 70.

5.3 DISCUȚIE

Această contribuție introduce o abordare a problemei recunoașterii acțiunii umane folosind date scheletice. Metoda propusă se bazează pe o rețea neurală proiectată folosind unele dintre cele mai utilizate tipuri de straturi convoluționale: GCN și TCN. Inovația abordării noastre actuale constă în proiectarea unui pipeline performant și rapid care augmentează datele, determină caracteristicile geometrice și utilizează o rețea neurală pentru a identifica acțiunea. Modelul neural inclus în acest pipeline este, de asemenea, inovator și combină avantajele rețelelor propuse anterior.

Versiunea actuală reprezintă o îmbunătățire față de metoda noastră anterioară propusă în lucrarea [8]. Metoda actuală pleacă de la aceleași categorii de caracteristici dar, prin introducerea unor optimizări la nivelul structurii modelului neural și prin aplicarea tehnicilor de augmentare, îmbunătățește viteza de inferență, crește acuratețea atinsă de model și reduce numărul de parametri. Pipeline-ul propus în versiunea actuală conține suplimentar etapa de augmentare care nu exista în metoda anterioară. Modelul neural propus este, de asemenea, diferit de cele prezentate și analizate în capitolele anterioare. Arhitectura bazată pe TCN introdusă anterior conține un modul în care integrăm doar straturi de tip TCN. Pentru metoda descrisă în capitolul curent, am folosit un modul bazat pe straturi GCN-TCN. În modelul actual, straturile de tip GCN au extras dependențele spațiale, spre deosebire de soluția noastră anterioară, în care straturile de tip TCN analizau secvența atât temporal, cât și spațial. Mai precis, de data aceasta dimensiunea ferestrei spațiale utilizate de straturile TCN este 1 și, în acest fel, rețeaua păstrează dimensiunea spațială până la aplicarea blocului Polling.

6

CONCLUZII

Scopul principal al acestei teze a fost construirea unui modul robust pentru rezolvarea problemei HAR. În acest sens, am identificat provocările care caracterizează această problemă și neajunsurile abordărilor existente. Pornind de la aceste observații, am dezvoltat o serie de abordări axate pe principalele proprietăți pe care am considerat că o soluție pentru o platformă robotică trebuie să le aibă: capacitate de generalizare, dimensiune redusă a modelului, timp mic de preprocesare și viteză mare de inferență. Modelele neurale propuse în aceste abordări se bazează pe cele mai utilizate tipuri de rețele neurale profunde pentru abordarea problemei HAR: GCN, TCN și LSTM.

Am folosit unul dintre cele mai provocatoare seturi de date, NTU RGB+D, pentru a ne evalua modelele. Reprezentarea aleasă pentru abordările propuse este sub formă de date scheletice, iar caracteristicile utilizate sunt determinate folosind coordonatele 3D prezise de senzorul Kinect. Soluțiile propuse ating performanțe comparabile cu cele mai avansate pentru toate protocoalele de testare propuse de setul de date NTU RGB+D. Mai mult, am adaptat două abordări pentru a fi evaluate folosind un set de date colectat din perspectiva robotică. În acest fel, am evidențiat capacitatea de generalizare a modelelor propuse și am arătat că putem aplica conceptul de învățare prin transfer la această problemă.

Am realizat o analiză profundă a problemei HAR constând într-o prezentare a principalelor tipuri de modalități utilizate și o prezentare generală a aplicațiilor practice care necesită un astfel de modul. Pentru fiecare tip de modalitate am evidențiat câteva caracteristici și limitările prezentate de soluțiile bazate pe reprezentarea respectivă. Pentru a evidenția importanța fiecărui domeniu de aplicare, am inclus câteva abordări reprezentative. Pornind de la toate acestea, am subliniat provocările care există în acest domeniu de cercetare și am identificat cinci categorii de sub-sarcini pentru problema HAR.

6.1 CONTRIBUȚII

Principalele contribuții originale ale acestei teze sunt următoarele:

- În lucrarea originală în care sunt introduse rețelele de tip TCN, autorii implementează aceste blocuri folosind straturi convoluționale 1D. În cazul problemei HAR, transformarea secvenței într-una compatibilă cu convoluții 1D nu permite păstrarea dimensiunii spațiale. Pentru a evita acest inconvenient, am propus o versiune modificată pentru straturile TCN care utilizează convoluții 2D. În plus, am sugerat extinderea conceptului de convoluție dilatată la nivel spațial. Rezultatele experimentale

obținute au demonstrat că această modificare este benefică pentru performanța modelului.

- Dimensiunea spațială este fundamentală pentru problema HAR. De aceea ne-am concentrat asupra ei propunând o modalitate de a integra două metode de rearanjare a articulațiilor. Una dintre metodele de rearanjare se bazează pe o matrice 2D și este propusă de noi în [1]. A doua metodă este de tip 1D și pleacă de la scheletul uman perceput ca un arbore a cărui rădăcină este articulația considerată cea mai apropiată de centrul de greutate.

- Integrarea unui modul de recunoaștere a acțiunii umane într-o platformă robotică este o sarcină complexă, deoarece trebuie să luăm în considerare multe aspecte. Am proiectat o metodă generală de integrare în cadrul platformei AMIRO și am testat-o folosind robotul umanoid Pepper. În această metodă am folosit modelul OpenPose [6] pre-antrenat pentru a extrage coordonatele 2D din imaginile RGB, am extras a treia coordonată din harta de adâncime corespunzătoare și am aplicat o arhitectură neurală originală pentru a clasifica acțiunile. Am obținut rezultate bune în ceea ce privește acuratețea și am arătat că această metodă funcționează în scenarii în timp real. Integrarea modulului de recunoaștere a acțiunii umane în cadrul platformei AMIRO a fost realizată în colaborare cu Alexandra Ștefania Ghiță.

- Am proiectat două arhitecturi neurale pe care le-am testat pe un set de date colectat din perspectivă robotică. Prima arhitectură este una cu mai multe etape care utilizează straturi liniare pentru extragerea caracteristicilor și celule LSTM pentru analiza secvenței temporale. Această rețea neurală returnează 3 predicții și eroarea este calculată separat pentru fiecare etapă. Scopul principal pe care l-am urmărit în proiectarea sa este de a îmbunătăți predicția de la o etapă la alta. Astfel, etapa 2 folosește rezultatul obținut din aplicarea celulelor LSTM din etapa 1, iar etapa 3 procedează în mod similar cu rezultatul din etapa 2. A doua arhitectură este una în care folosim straturi convoluționale 2D pentru a extrage caracteristici și apoi le procesăm cu ajutorul unui modul temporal format din straturi de tip TCN.

- Generalizarea și supraadaptarea nu sunt suficient analizate pentru abordările propuse pentru rezolvarea problemei HAR. De aceea am decis să investigăm această direcție de cercetare. Pentru aceasta, am folosit două arhitecturi neurale pe care le-am antrenat folosind setul de date NTU RGB+D și apoi le-am testat pe un set de date colectat folosind robotul Pepper. Astfel, am analizat conceptul de învățare prin transfer din perspectiva diferitelor seturi de date. A fost necesar să se introducem o operație de conversie a scheletului din formatul prezis de OpenPose [9] la formatul folosit de senzorul Kinect.

- Caracteristicile de la care pornește analiza unui model neural influențează performanțele obținute. De aceea ne-am concentrat pe acest subiect și propunem o metodă de determinare a caracteristicilor geometrice care pot ajuta modelul neural

să obțină performanțe mai bune. Această metodă conține și o variantă de normalizare a datelor.

- Prezentăm o arhitectură neurală spațio-temporală, care combină straturile TCN și GCN. Am raportat rezultatele modelului propus pe benchmark NTU RGB+D [5, 10], pentru toate protocoalele de testare. Această abordare este o soluție non-black-box în care modelul scoate un tensor de caracteristici împreună cu rezultatele, putând astfel explica predicțiile.

- Recent, domeniul Explainable Artificial Intelligence (XAI) a primit multă atenție și au fost propuse diferite metode pentru a determina o explicație pentru predicția rețelei. Cu alte cuvinte, cercetătorii nu mai doresc să perceapă modelele neurale precum o cutie neagră. Urmând această tendință, am arătat cum putem proiecta modelul neural astfel încât să fie capabil să explice predicția din perspectiva importanței fiecărei articulații. Mai mult, am prezentat câteva exemple elocvente pentru a evidenția importanța unor astfel de explicații.

- Augmentarea datelor este o tehnică prin care putem crește artificial dimensiunea setului de date pentru a obține un model care funcționează cu o precizie mai bună. În ceea ce privește problema HAR, este complicat să colectezi suficiente mostre pentru acțiuni precum *cădere*. De asemenea, dorim ca modelul antrenat să devină invariant față de poziția sau caracteristicile fizice ale persoanei. Pentru a reduce aceste limitări, am proiectat un pipeline în care am integrat o etapă de augmentare a datelor. Am efectuat o validare experimentală extinsă a abordării, demonstrând potențialele acesteia în comparație cu restul metodelor existente.

6.2 PERSPECTIVE ȘI DEZVOLTĂRI VIITOARE

Putem concluziona că problema HAR este o sarcină complexă care nu poate fi rezolvată folosind un algoritm determinist clasic. Abordările bazate pe tehnici de Deep Learning reușesc să rezolve parțial această problemă. Considerăm că această problemă rămâne o temă de cercetare deschisă deoarece soluțiile existente sunt specializate în identificarea subgrupului de acțiuni posibile și prezintă diverse limitări. Mai mult, procesul de integrare a acestor soluții în platforme utilizate pentru rezolvarea scenariilor din viața reală aduce provocări suplimentare.

O primă îmbunătățire care poate fi realizată la nivel arhitectural constă în încercarea de a îmbunătăți partea finală a modelului neural care se ocupă de clasificare prin propunerea unui model arhitectural mai complex, eventual bazat pe o arhitectură în mai multe etape. Pentru aceasta, putem construi o arhitectură neurală complexă formată din mai multe submodele, fiecare specializat în identificarea unui subset de acțiuni.

A doua direcție de cercetare este o încercare de a îmbunătăți înțelegerea modelului neural prin introducerea de date suplimentare pentru a oferi mai mult context despre mediul în care subiectul efectuează acțiunea. Considerăm că adăugarea de informații de

context poate îmbunătăți atât performanța arhitecturii, cât și capacitățile de explicație ale modelelor pentru problema HAR.

ACRONIME

CNN	Convolutional Neural Network
GCN	Graph Convolutional Network
HAR	Human Action Recognition
LSTM	Long Short-term Memory
RNN	Recurrent Neural Network
TCN	Temporal Convolutional Network
XAI	Explainable Artificial Intelligence

REFERINȚE

[1] M. Trăscău, M. Nan and A. M. Florea, 'Spatio-temporal features in action recognition using 3d skeletal joints', *Sensors*, vol. 19, no. 2, p. 423, 2019.

[2] Z. Yang, Y. Li, J. Yang and J. Luo, 'Action recognition with spatio-temporal visual attention on skeleton image sequences', 2018.

[3] Y.-F. Song, Z. Zhang, C. Shan and L. Wang, 'Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition', in *Proceedings of the 28th ACM International Conference on Multimedia (ACMMM)*, New York, NY, USA: Association for Computing Machinery, 2020, pp. 1625–1633, ISBN: 9781450379885. DOI: 10.1145/3394171.3413802. [Online]. Available: https://doi.org/10.1145/3394171.3413802.

[4] M. Nan, A. S. Ghiță, A.-F. Gavril, M. Trascau, A. Sorici, B. Cramariuc and A. M. Florea, 'Human action recognition for social robots', in *2019 22nd International Conference on Control Systems and Computer Science (CSCS)*, IEEE, 2019, pp. 675–681.

[5] A. Shahroudy, J. Liu, T. Ng and G. Wang, 'NTU RGB+D: A large scale dataset for 3d human activity analysis', *CoRR*, vol. abs/1604.02808, 2016.

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, 'Openpose: Realtime multi-person 2d pose estimation using part affinity fields', *arXiv preprint arXiv:1812.08008*, 2018.

[7] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng and W. Hu, 'Channel-wise topology refinement graph convolution for skeleton-based action recognition', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 359–13 368.

[8] M. Nan, M. Trăscău, A. M. Florea and C. C. Iacob, 'Comparison between recurrent networks and temporal convolutional networks approaches for skeleton-based action recognition', *Sensors*, vol. 21, no. 6, p. 2051, 2021.

[9] Z. Cao, G. Hidalgo, T. Simon, S. Wei and Y. Sheikh, 'Openpose: Realtime multi-person 2d pose estimation using part affinity fields', *CoRR*, vol. abs/1812.08008, 2018. arXiv: 1812.08008. [Online]. Available: http://arxiv.org/abs/1812.08008.

[10] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan and A. K. Chichung, 'Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding', *IEEE transactions on pattern analysis and machine intelligence*, 2019.