Facultatea de **AUTOMATICĂ ȘI CALCULATOARE** din Iași

Detecția și recunoașterea limbajului semnelor din imagini/ secvențe video

Introducere

- Context și problemă: Barierele de comunicare dintre persoanele cu deficiențe de auz și restul populației limitează integrarea socială. Soluțiile existente sunt costisitoare și greu de utilizat.
- Obiectiv: Dezvoltarea unui sistem software non-invaziv, capabil să traducă în timp real Limbajul Semnelor Românești (LSR), adresând atât gesturile statice, cât și cele dinamice.

Dataset

Construirea dataset-ului: Datele au fost colectate prin generarea manuală a secvențelor video specifice fiecărei acțiuni. S-a utilizat librăria **MediaPipe Holistic** pentru maparea coordonatelor.
Optimizarea datelor: Eliminarea trăsăturilor din Face Mesh (1404 puncte) pentru eficientizare și selecția exclusivă a coordonatelor pentru poziție și mâini (159 puncte).

Modelul AI

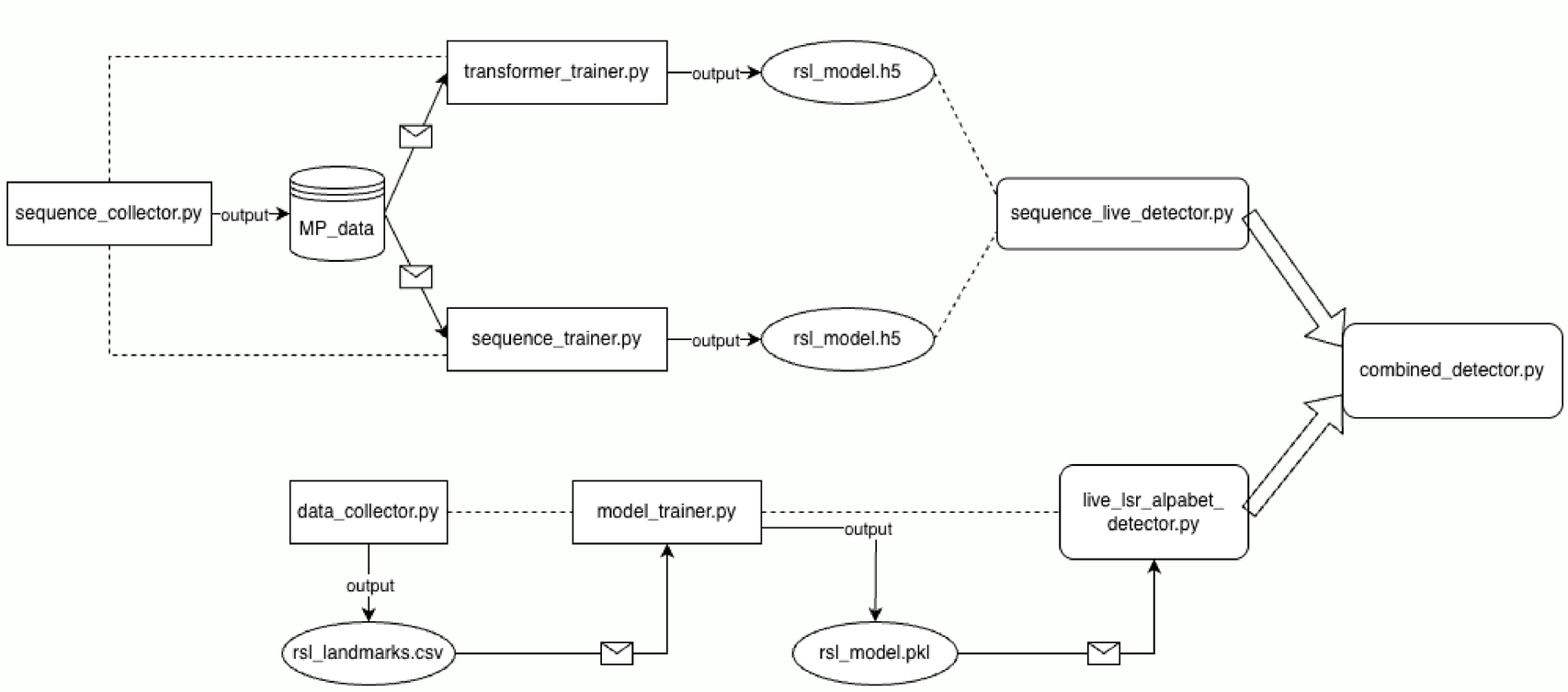
Metodă: Implementarea unui model bazat pe mecanisme de reținere LSTM bidirecțional, specializat pe captarea dependențelor temporale lungi.
Configurație model: un strat de masking pentru a gestiona secvențele de lungimi diferite și trei straturi LSTM, urmate de două straturi dense și un softmax final.
Parametri antrenare: 350 de epoci, optimizator Adam cu rata de învățare 0.001, utilizarea regularizării *"early stopping"* pentru oprirea automată și prevenirea overfitting-ului.

Etapele implementării

1. Etapa statică (Random Forest): Pentru detecția literelor a fost utilizat un clasificator Random Forest (RFC) ales pentru eficiența algoritmului pe seturi de date de dimensiuni reduse și viteza de răspuns excelentă în mediul live. Modelul a reușit o clasificare aproape perfectă a literelor statice (ex: 'g', 'h' cu F1-score 1.00), însă a prezentat limitări în recunoașterea semnelor care implică mișcare, precum 'J' sau 'Z'.
2. Tranziția către secvențe (LSTM): Pentru a integra componenta temporală, s-a trecut la utilizarea rețelelor LSTM (Long Short-Term Memory). Inițial s-a testat o variantă simplă, evoluând rapid către un LSTM bidirecțional. Această schimbare a fost esențială pentru a permite modelului să analizeze gestul în ansamblul său, fără a fi constrâns de o fereastră fixă de cadre, învățând tiparele mișcării din ambele direcții temporale.
3. Modelul Transformer: În final s-a implementat un model bazat pe arhitectura Transformer, pentru a explora capacitatea sistemului de a identifica corelații complexe între cadrele video. Deși neadecvat dataset-ului mic, pe viitor ar putea fi combinat cu un LLM pentru a elimina detecțiile false și „halucinațiile".

Arhitectura soluției

Abordare hibridă: Dezvoltarea unui sistem care integrează recunoașterea statică (litere) și cea dinamică (cuvinte) într-un flux video unificat.
Flux de procesare: Transformarea input-ului video brut în date semantice prin vectorizare și clasificare secvențială, permițând interpretarea limbajului semnelor în timp real.



Figură 1. Flowchart arhitectură

	Timp mediu de inferență	Avantaje	Limitări
Random Forest Classifier	1.32 ms	Ideal pentru un dataset mic	Nu recunoaște semnele dinamice
LSTM	31.38 ms	Capturează contextul temporal	Constrâns de o fereastră fixă de cadre
LSTM Bidirecțional	42.60 ms	Atenție bidirecțională temporală	Predispus la detecții false
Transformer	16.94 ms	Relații globale între cadre	Necesită volume mari de date

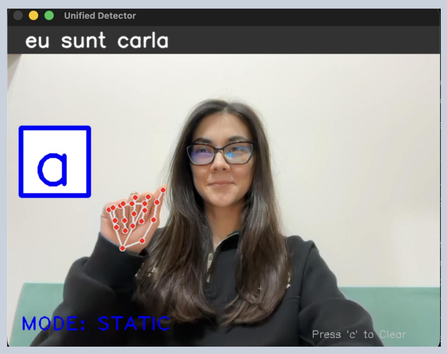
Tabel 1. Comparație modele

Concluzii

- S-a demonstrat fezabilitatea utilizării rețelelor neuronale recurente (LSTM) în conjuncție cu tehnici de procesare video în timp real pentru o soluție completă de traducere a limbajului semnelor.
- Optimizarea vectorului de trăsături prin eliminarea punctelor faciale a redus dimensiunea datelor de intrare, permițând rularea modelului în timp real fără necesitatea unui hardware dedicat costisitor.
- Limitări: Dependența de condițiile de iluminare și detecția parțială a mâinilor în gesturi complexe.
- Dezvoltări ulterioare: Extinderea setului de date pentru un vocabular amplu și integrarea unui LLM care să asiste generarea predicțiilor în propoziții.

Autori

Velnic Vlad-Andrei
Epure Carla-Maria



Referințe

1. Alphabet Recognition of Sign Language Using Machine Learning (2023) - AVINASH KUMAR SHARMA, ABHYUDAYA MITTAL, AASHNA KAPOOR, ADITI TIWARI
2. Hand Gesture Recognition using Image Processing and Feature Extraction Techniques (2020) - Ashish Sharma, Anmol Mittal, Savitoj Singh, Vasudev Awatramani
3. A Comprehensive Review of Sign Language Recognition: Different Types, Modalities, and Datasets (2022) - M. MADHIARASAN, PARTHA PRATIM ROY
4. Application of transfer learning to sign language recognition using an inflated 3D deep convolutional neural network (2021) - ROMAN TOENGI
5. Score-level Multi Cue Fusion for Sign Language Recognition (2020) - Cagrı Gokce, Ogulcan Ozdemir, Ahmet Alp Kındıroglu, Lale Akarun