Erkennung von American Sign Language (ASL) mit 3D-Convolutional Neural Networks (C3D & R(2+1)D)

Fabian Piontkowski, Junior Yien

Master Technische Informatik, BHT Berlin

Maschinelles Sehen, Sommersemester 2025
Dozent: Prof. Dr. Hildebrand

25.09.2025

Problem

Menschen mit Hör- und Sprachbehinderung sind auf Gebärdensprache angewiesen.
Eine automatische Erkennung von Handzeichen könnte Barrieren abbauen und Anwendungen in Übersetzungsdiensten, Mensch-Maschine-Interaktion und Assistenzsystemen ermöglichen.

Die von uns verwendeten Modelle (R(2+1)D, C3D) haben relativ wenige Parameter, sind jedoch in der Lage, dynamischeoder statische ASL-Zeichen zu erkennen.

Motivation

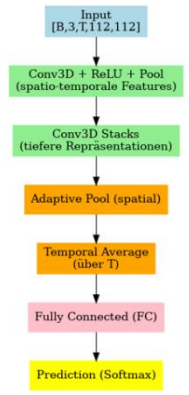
- Klassische Methode wie **SVM** zeigt nur begrenzte Genauigkeit.
- **2D-CNNs** wie ResNet funktionieren gut für Bilder, erfassen jedoch keine zeitlichen Informationen.
- **3D-CNNs** können räumliche und zeitliche Merkmale gemeinsam lernen und sind daher für Gestenerkennung geeignet.

Methodik & Pipeline

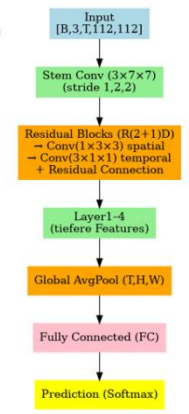
- **Preprocessing:**
  - Skalierung aller Eingaben auf 112×112 Pixel.
  - ROI-Cropping der Hände (MediaPipe),
  - Augmentierung (Color Jitter, Random Resized Crop, Rotation, Gaussian Blur)
- **Modelle:**
  - **C3D:** klassische 3D-Convolutions für Spatio-Temporal Features.
  - **R(2+1)D:** Aufspaltung der 3D-Faltung in 2D (Raum) + 1D (Zeit), dadurch weniger Parameter und bessere Generalisierung.
- **Training:**
  - 20 Epochen mit AdamW (LR=3e-4, Weight Decay=1e-4).
  - Loss: CrossEntropy.
  - Fine Tuning (Kinetics-Pretrain)
- **Evaluation:** Accuracy1, Accuracy5, Macro-F1, Confusion-Matrix

Pipeline

- **C3D**

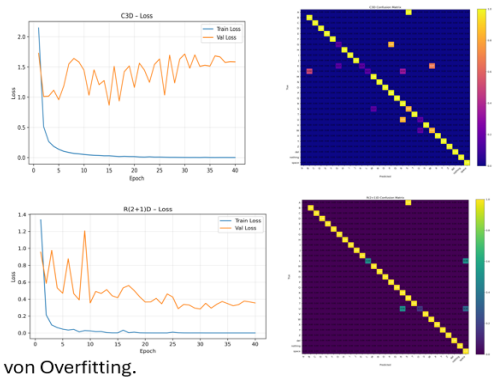


- **R(2+1)D**

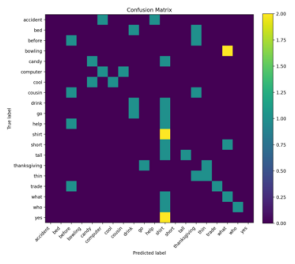


Results-Statisch

- **C3D:**
  - Trainingsloss: 2.14 → 0.0008
  - Beste Validierung:
    - Accuracy1 = **83.9 %**
    - Macro-F1 = **80.6 %**
- **R(2+1)D:**
  - Trainingsloss: 1.33 → 0.0001
  - Beste Validierung:
    - Accuracy1 = **91.9 %**
    - Macro-F1 = **90.1 %**
- **Vergleich:**
  - Beide Modelle lernen schnell, Gefahr von Overfitting.
  - R(2+1)D deutlich robuster und genauer als C3D
  - R(2+1)D hat nur halb so viel Parameter wie C3D



Results-Dynamisch



- **R(2+1)D:**
  - Trainingsloss: 2.0406 → 0.5985
  - Beste Validierung:
    - Accuracy1 = **30 %**
    - Macro-F1 = **27.6 %**

References

- Tran et al., *Learning Spatiotemporal Features with 3D Convolutional Networks*, ICCV 2015
- Tran et al., *A Closer Look at Spatiotemporal Convolutions for Action Recognition*, CVPR 2018
- He et al., *Deep Residual Learning for Image Recognition*, CVPR 2016