Testergebniss von Envisionhgdetector: Tool für binary classification

Betreuer: Prof. Dr. Alexander Mehler, Dr. Andy Lücking, Dr. Alexander Henlein
Student: Jaehyun Shin
Datum: 20.08.2025
Ort: TTLab (RMS-10) Goethe Universität Frankfurt
Github: `https://github.com/autoshin0322/Jayproject/blob/main/README.md`

Aktueller Stand

- Modell für binary classfication : envisionhgdetector [1]
- Datensatz: Multiperspektive Videos (TTLab Va.Si.Li-Lab) / Evolving artificial sign languages in the lab: from improvised gesture to systematic sign (dataset) [2]

Einführung

Um vor der eigentlichen Programmierung mit einem vLLM ein besseres Verständnis für Klassifikationsaufgaben zu gewinnen, wurde zunächst ein relativ einfaches Tool für eine binäre Klassifikation sowie eine BIO-Schema-Klassifikation verwendet.

Als Tool kam Envisionhgdetector zum Einsatz, das auf Empfehlung von Dr. Henlein ausgewählt wurde. Die verwendeten Datensätze stammen einerseits aus dem Multiperspektive Video-Korpus des Va.Si.Li-Lab, andererseits aus Videos des Centre for Language Evolution an der Universität Edinburgh.

Der zweite Datensatz wurde ergänzend eingesetzt, da sich bei der ersten Analyse mit dem Va.Si.Li-Lab-Datensatz sehr niedrige Evaluationsmetriken zeigten. Um zu überprüfen, ob die Ursache dafür in der Qualität des Datensatzes oder im Modell selbst lag, wurde ein weiterer, anders strukturierter Datensatz zum Vergleich herangezogen.

Die Ergebnisse der Tests mit beiden Datensätzen sind nachfolgend dargestellt. Da jedoch nicht alle Videos und Analysedaten vollständig in diesem Dokument abgebildet werden konnten, ist eine vollständige Übersicht über die Materialien über den oben angegebenen GitHub-Link zugänglich.

Experiment und Ergebnisse

Ergebnisse binary classification

In diesem Vergleich zeigte das Modell eine signifikant bessere Erkennungsleistung. Die zuvor beobachteten Fehler traten nicht mehr auf. Die Klassifikation von "Geste" und "Nicht-Geste" erfolgte konsistent und mit hoher Genauigkeit.

3.1.1 Vi.Si.Li-Lab

(a) Labeled video (b) Tracked video

Abbildung 1: labeled & tracked Videos for Va.Si.Li-Lab

Tabelle 1: Durchschnittliche Evaluationsmetriken (aufgeteilt in 4 Gruppen)

	Precision	Recall	F1-score
AVERAGE	0.354	0.431	0.358

(a) Gesture

	Precision	Recall	F1-score
AVERAGE	0.781	0.605	0.630

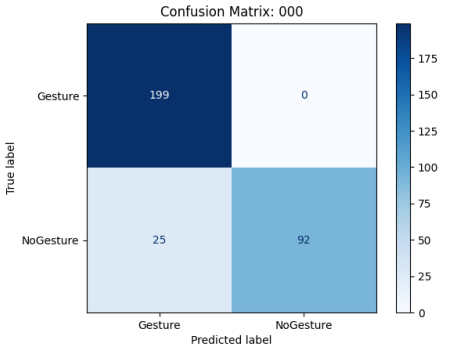
(b) NoGesture

	Precision	Recall	F1-score
AVERAGE	0.599	0.549	0.525

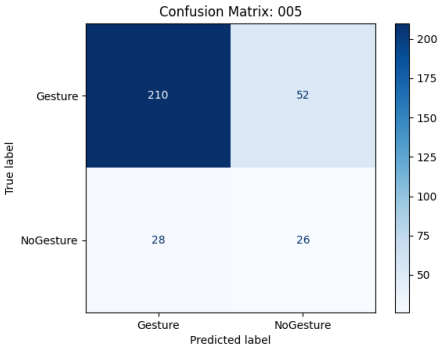
(c) Macro Average

	Precision	Recall	F1-score	Accuracy
AVERAGE	0.869	0.709	0.734	0.709

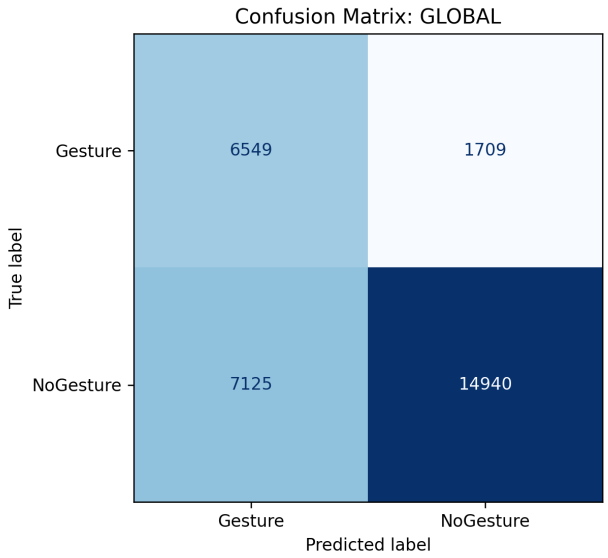
(d) Weighted Average + Accuracy



(a) bincary cm1



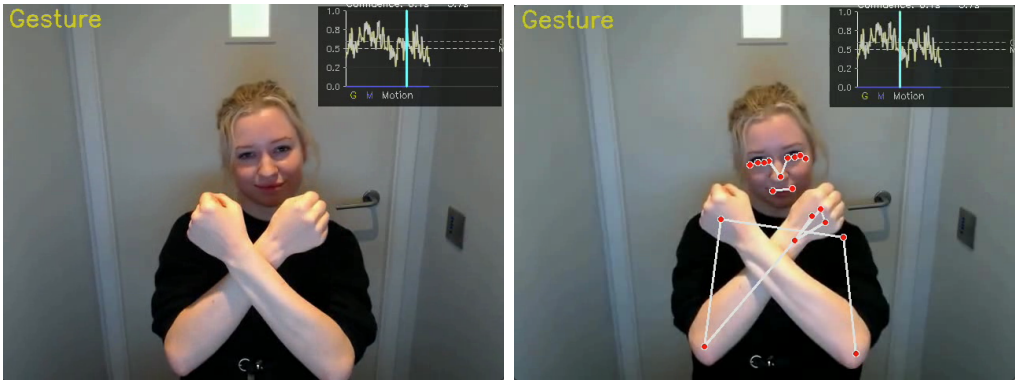
(b) binary cm2



(c) binary cm Avg

Abbildung 2: Confusion Matrices for Va.Si.Li-Lab

3.1.2 Univ. Edinburgh

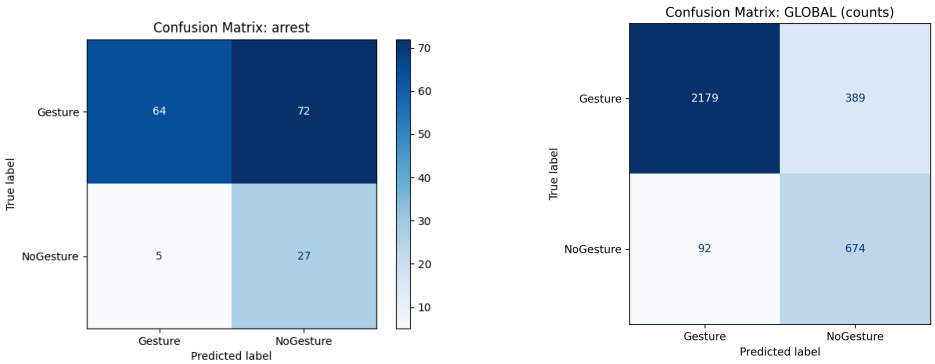


(a) Labeled video (b) Tracked video

Abbildung 3: labeled & tracked Videos for Univ. Edinburgh

Tabelle 2: Durchschnittliche Gesamtergebnisse

	Precision	Recall	F1-score	Accuracy
AVERAGE	0.955	0.845	0.890	0.853



(a) binary CM (b) binary CM Avg

Abbildung 4: Confusion matrices for Univ. Edinburgh

Ergebnisse BIO Schema

3.2.1 Va.Si.Li-Lab

Tabelle 3: Evaluationsergebnisse der BIO-Klassifikation - Va.Si.Li-Lab

	Precision	Recall	F1-score
AVERAGE	0.673	0.673	0.673

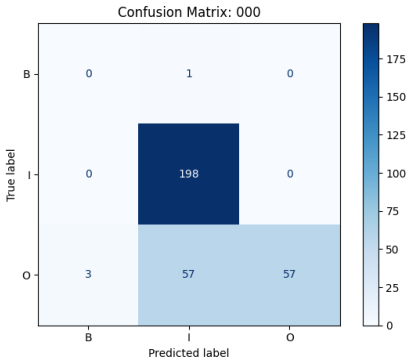
(a) Average

	Precision	Recall	F1-score
AVERAGE	0.410	0.406	0.370

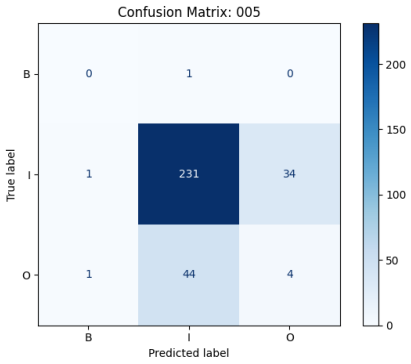
(b) Macro Average

	Precision	Recall	F1-score	Accuracy
AVERAGE	0.808	0.673	0.689	0.673

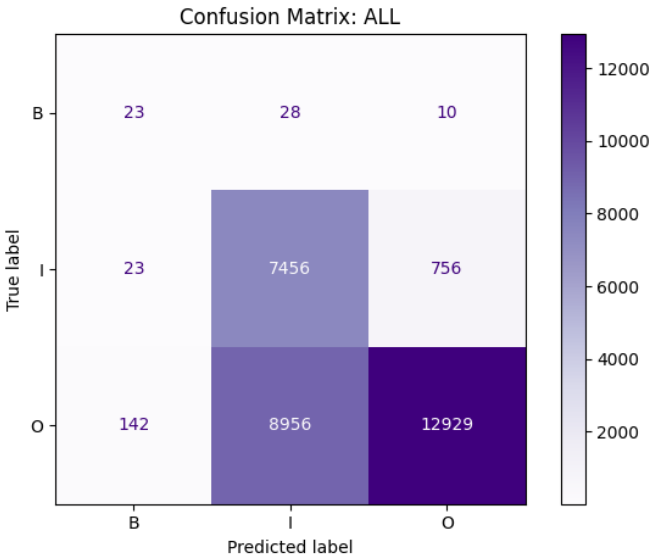
(c) Weighted Average + Accuracy



(a) BIO CM1



(b) BIO CM2



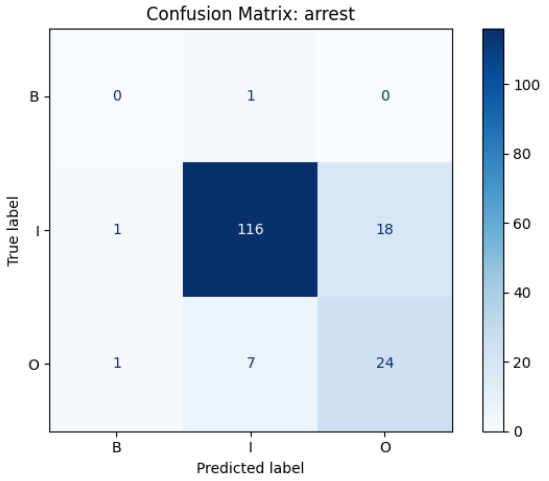
(c) BIO CM Avg

Abbildung 5: BIO Evaluation for Va.Si.Li-Lab

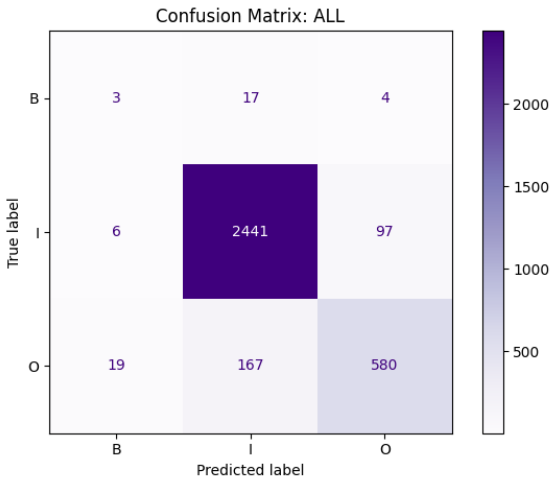
3.2.2 Univ. Edinburgh

Tabelle 4: Evaluationsergebnisse der BIO-Klassifikation - Univ. Edinburgh

	Precision	Recall	F1-score	Accuracy
AVERAGE	0.637	0.606	0.605	0.904



(a) BIO CM1



(b) BIO CM2

Abbildung 6: BIO Evaluation for Univ. Edinburgh

Schlussfolgerung

Diskussion

Im Vergleich zeigte sich, dass die Evaluationsmetriken beim Datensatz des *Va.Si.Li-Lab* deutlich niedriger ausfielen als bei dem zweiten Datensatz.

Eine mögliche Erklärung hierfür ist, dass es sich bei den Videos des *Va.Si.Li-Lab* nicht um echte Aufnahmen von Personen handelt, sondern um multiperspektivische VR-Videos mit generierten Avataren. Dadurch kam es häufiger vor, dass der *Envisionhgdetector* auch in gestenfreien Sequenzen fälschlicherweise Gesten erkannte, da die Keypoint-Erkennung hier fehleranfälliger ist.

Zusätzlich wurden die Annotationen beim *Va.Si.Li-Lab* von Studierenden manuell vorgenommen, wobei Fehlbewegungen oder bedeutungslose Gesten oftmals nicht annotiert wurden. Das Modell hingegen klassifizierte solche Bewegungen möglicherweise als *Gesture*, was zu einer Diskrepanz zwischen Prediction und Ground Truth (*NoGesture*) führte.

Im Gegensatz dazu basiert der Datensatz der *Universität Edinburgh* auf echten Videoaufnahmen von Menschen, wobei jede Geste in separaten Clips dargestellt wird. Das trägt vermutlich zu einer höheren Präzision bei.

Fazit

Daraus lässt sich schließen, dass das Modell grundsätzlich zuverlässig arbeitet, jedoch die Wahl eines geeigneten Datensatzes entscheidend für die Ergebnisqualität ist. Diese Frage sollte daher in Rücksprache weiter diskutiert und entschieden werden.

Literatur

[1] W. Pouw, B. Yung, S. Shaikh u. a., *envisionhgdetector*, Computer software, PyPI, Version 0.0.5.0, 2024. Adresse: `https://pypi.org/project/envisionhgdetector/`.

[2] Y. Motamedi, M. Schouwstra, K. Smith, J. Culbertson und S. Kirby, *Evolving artificial sign languages in the lab: from improvised gesture to systematic sign (dataset), 2015–2018*, Dataset, University of Edinburgh, Centre for Language Evolution, 2018. DOI: `10.7488/ds/2447`. Adresse: `https://doi.org/10.7488/ds/2447`.

Erstellt von Jaehyun Shin am 17. Aug 2025