Master's Thesis at the Pattern Recognition Lab, FAU Erlangen-Nuremberg

Alignment-Stable Attention for Writer-Independent IMU Handwriting Recognition via Hybrid CTC–AR Training, Calibrated Decoding, and Pretrained LM Decoder Adaptation

Online handwriting recognition (HWR) from inertial measurement units (IMUs) enables pen-and-paper (or any-surface) writing without cameras, touchscreens, or specialized digitizers, making it attractive for privacy-preserving and low-infrastructure text input. In the writer-independent (WI) setting, recognition performance is limited by substantial inter-writer variability (e.g., stroke dynamics, writing speed, grip and sensor placement), sensor noise, and distribution shift between training and unseen writers. CNN-based encoders with CTC decoding provide a strong baseline for WI IMU-HWR, yet errors are still driven by imperfect alignment and decoding decisions, which can manifest as character-level substitutions as well as word-level mistakes. This thesis builds on the REWI baseline [1] and our current attention-based pipeline to improve both character error rate (CER) and word error rate (WER) across word-level and sentence-level datasets under explicit model-size and latency constraints. Building on our current AR pipeline, we observe a clear trade-off: autoregressive decoding substantially improves WER, while CER can degrade moderately due to imperfect character-level alignment. The primary objective of this thesis is therefore to achieve simultaneous improvements in both CER and WER by adopting a hybrid training regime that combines CTC and AR learning: CTC provides strong monotonic, character-level alignment supervision, while AR training leverages global context to reduce word-level errors. To further reduce residual errors, we will stabilize attention alignment and decoding behavior through alignment-aware regularization (e.g., skip/coverage penalties) and calibrated beam-search decoding (length normalization and EOS control), and we will optionally integrate a lightweight external language model for fusion/rescoring during beam search if sufficient transcripts are available. As a secondary objective (time permitting), we will explore a multimodal variant that retains the same CNN feature extractor while conditioning a pretrained language model decoder on the encoder tokens and fine-tuning it for IMU-HWR, in order to assess whether pretrained decoders can match or exceed the hybrid approach under comparable accuracy and efficiency constraints.

The thesis consists of the following milestones:

- **Baseline reproduction and protocol.** Reproduce the current AR pipeline on WI splits with consistent CER/WER reporting; document seeds, configurations, and evaluation scripts.
- **Hybrid multitask learning (CTC+AR).** Add a CTC auxiliary head on encoder outputs and train with a joint objective (AR cross-entropy + $\lambda \cdot$CTC) [2]. Perform a controlled sweep over λ and regularization to stabilize alignment.
- **Decoding study (calibrated beam search).** Implement beam search for AR decoding; tune length normalization and EOS control (EOS bias and/or minimum-length constraints). Add an alignment-aware *training-time* regularizer on decoder cross-attention (coverage/skip penalty) and evaluate *decoding-time* heuristics (coverage penalty and/or length-normalized scoring in beam search). Quantify improvements on WI validation with CER/WER and tail-risk metrics.
- **Evaluation and ablations.** Primary: CER/WER on WI test writers. Secondary: tail-risk metrics (distribution of normalized edit distance), writer-stratified performance, and confusion/collision

analysis. Ablate (i) loss regime (CTC/AR/joint), (ii) decoding settings, (iii) LM-decoder integration choices, and (iv) masked pretraining, where applicable.

- **Second-priority extension: pretrained LM decoder (multimodal).** After establishing a strong hybrid CTC+AR baseline, integrate a pretrained seq2seq LM decoder and condition it on the CNN encoder's downsampled IMU tokens via a lightweight projection/adapter. Fine-tune under a strict parameter/latency budget, and compare against the scratch AR decoder. Ground the LM decoding with the auxiliary CTC signal to mitigate fluent-but-incorrect outputs.
- **Generalization via masked modeling (optional if time allows).** Pretrain the existing CNN encoder using self-supervised masking (time-span/channel masking) and a reconstruction or latent-prediction objective [3]. Fine-tune on supervised HWR and measure WI robustness gains.
- **Encoder adaptor study (optional if time allows).** Evaluate a lightweight encoder adaptor by inserting a small Transformer encoder (2–4 layers) on the downsampled CNN tokens before the AR/LM decoder, under a fixed budget [4]. Run controlled ablations for AR-only vs hybrid training.

The implementation should be done in Python/C++

Supervisors:	Dr.-Ing. V. Christlein, Prof. Dr.-Ing. habil. A. Maier
Student:	Muhammad Ajlal
Start:	January 15, 2026
End:	July 15, 2026

References

[1] Jingdong Li, T. Hamann, J. Barth, et al. Robust and efficient writer-independent imu-based handwriting recognition. *arXiv preprint arXiv:2502.20954*, 2025.

[2] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[3] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.