Cognitive States Prediction based on Multimodal Analysis in Online Meeting with Coimagination Method

2310040 WEI JINGTONG

Mild Cognitive Impairment (MCI) represents a transitional stage between normal aging and dementia and is a key target for early intervention. Conventional clinical assessment typically relies on standardized scales and face-to-face interviews, which can be costly, infrequent, and limited in accessibility. With the rapid growth of online communication, developing non-invasive and scalable screening approaches based on natural conversational behavior has become increasingly meaningful in practice. However, in natural interactive settings, linguistic, acoustic, and facial behaviors are shaped not only by cognitive status but also by mood fluctuations, social strategies, and individual differences. These confounding factors can weaken feature discriminability and undermine model generalization. To address this challenge, this thesis investigates session-level multimodal MCI recognition in an online Coimagination meeting environment, with a particular focus on reducing the confounding effect of mood changes when incorporating facial Action Units (AUs), thereby improving robustness.

In this work, we formulate MCI screening at the session level and construct multimodal representations from language, speech, and video signals, together with facial AU features extracted by OpenFace. For AUs, utterance-level activation is determined using the proportion of activated frames within each utterance time window, and session-level AU descriptors are obtained by aggregating utterance-level binary activations across the session. Audio, text, and video modalities are represented using embeddings from pretrained models, where the video feature dimension is 1024. These modality-specific features are used to form session-level feature vectors for classification. To mitigate mood-related confounding in AU-based modeling, we use the change in Face Scale ratings before and after each session (ΔFace Scale) to quantify mood variation. Rather than coarse grouping, each distinct ΔFace Scale value is treated as a stratum. Within each stratum, we perform significance testing between the MCI group and the healthy control group on AU features, and select statistically significant AUs to form a reduced AU subset for subsequent classification experiments.

For classification, we implement a linear-kernel SVM as a conventional machine-learning baseline and a multi-branch DNN as the deep model. In the DNN, each modality (AU, audio, text, and video) is encoded by an independent MLP branch to learn compact latent representations, followed by feature-level fusion at an intermediate layer through concatenation and a final classification head. To ensure strict evaluation and fair comparability,

both SVM and DNN adopt Leave-One-Group-Out (LOGO) cross-validation, where each participant (all sessions of a subject) defines a group, enabling evaluation of generalization to unseen participants. Performance is reported using Accuracy and Macro F1, with systematic ablation over unimodal and multimodal settings. Since AU and video are both facial-information channels and may introduce redundancy or instability in fusion learning, AU and video are not combined within the same experimental setting.

Experimental results indicate that multimodal fusion benefits the deep model more substantially, and that mood-informed AU screening affects SVM and DNN differently. For configurations without AU, the best DNN performance is achieved by Audio+Text fusion (Acc = 0.725, Macro F1 = 0.683), outperforming the corresponding linear SVM and highlighting the advantage of deep feature fusion for capturing complementary speech–language cues. Using the full (unscreened) AU set, the DNN achieves the overall best performance under AU-only (Acc = 0.742, Macro F1 = 0.675), suggesting that the deep model can exploit nonlinear discriminative patterns from richer AU representations; AU-only is also the strongest setting for SVM under this condition, though with lower overall performance than DNN. After applying ΔFace Scale–guided screening, SVM improves under AU-only (Acc = 0.710, Macro F1 = 0.642), indicating that stratified significance-based selection effectively reduces mood confounds and weakly discriminative AUs for a linear classifier. In contrast, the best DNN setting with screened AUs shifts to AU+Text (Acc = 0.711, Macro F1 = 0.681), implying that a more compact AU subset can be more effectively complemented by linguistic information in deep multimodal learning.

In summary, this thesis presents a reproducible session-level multimodal framework for MCI recognition in interaction scenarios of online meeting scenarios and demonstrates that ΔFace Scale–guided stratified AU screening can mitigate mood-related confounding and improve robustness. The proposed methodology and empirical findings provide practical evidence for non-invasive cognitive screening in natural interaction settings and offer a foundation for future studies with larger-scale and multi-site online screening data.