Cross-Modal Contrastive Learning for Multimodal Yoga Pose Recognition

Name: Zolboo Damiran
Laboratory' s name: Ubiquitous Computing Systems Laboratory
Supervisor' s name: Keiichi Yasumoto

Abstract (should be within 1st page)

Human activity recognition (HAR) plays an important role in intelligent environments (IE), where heterogeneous sensing modalities, including visual and motion data, provide complementary information for understanding human behavior. Yoga pose recognition (YPR), a fine-grained HAR task, remains challenging due to subtle posture variations and individual differences in pose execution. However, multimodal learning remains difficult due to limited labeled data and modality discrepancies, which hinder cross-modal representation alignment. This dissertation investigates cross-modal contrastive learning to improve representation alignment and label efficiency in multimodal HAR under limited supervision. A preliminary study of representative self-supervised learning methods, including SimCLR, MoCo, and BYOL, reveals their limitations in learning discriminative representations for activity recognition. Motivated by these findings, this dissertation proposes CoRF-Yoga, a cross-modal contrastive learning framework that jointly learns aligned RGB and motion representations. To support this study, a synchronized multimodal benchmark dataset, DSZ-Yoga, is constructed using RGB and motion data for fine-grained YPR. Experimental results demonstrate improved performance in low-label regimes and enhanced cross-modal retrieval. Supervised cross-modal contrastive learning achieves up to 99.61% fusion accuracy using only 25% labeled data under Leave-One-Subject-Out evaluation. The results further demonstrate competitive performance with reduced labeling effort. Finally, the practical feasibility of the framework is examined through end-to-end analysis, including data processing and inference latency. These findings demonstrate computational efficiency and support practical deployment. Overall, this dissertation advances multimodal HAR by establishing a unified cross-modal learning framework, constructing a synchronized multimodal benchmark dataset, improving label efficiency, and providing empirical insights into cross-modal representation learning for IE.