Abstract

The field of artificial intelligence has witnessed remarkable progress in the last few years, but there still exists a disparity between the capabilities of intelligent systems and those of humans in terms of accuracy and efficiency. Humans possess the ability to perceive and interact with the world using multiple senses, enabling a deeper understanding beyond what can be achieved with a single 2D representation. Motivated by this, the focus of this thesis is on multi-modal learning to enhance model performance in terms of accuracy, robustness, and adaptability.

The thesis explores two interrelated research domains: Object Recognition (OR) and Egocentric Action Recognition (EAR). These domains present specific challenges that conventional uni-modal approaches, usually image-based models, struggle to address. The thesis explores two interconnected research domains: Object Recognition (OR) and Egocentric Action Recognition (EAR). These domains pose specific challenges that conventional uni-modal approaches, primarily image-based models, struggle to address. Image-based models tend to exhibit a bias toward texture and color information while encoding a limited amount of geometric and motion cues. Additionally, they are sensitive to variations in lighting and other environmental conditions, which compromises their generalization capability. Consequently, their performance may suffer when applied to diverse real-world scenarios. This research explores the integration of both visual and non-visual modalities, with the aim of leveraging their complementary characteristics and capturing the intricate nature of real-world information. The thesis proposes techniques to enhance model capabilities, enabling robustness and adaptability to different environments and tasks. Our research in multi-modal learning also includes the exploration of alternative modalities derived from innovative devices that go beyond traditional sensors, like event-based cameras.

Overall, this thesis contributes to the advancement of multi-modal learning for cross-domain analysis of EAR and OR, aiming to enhance the capabilities of AI systems and reduce the gap between human and machine perception.

Keywords: deep learning, multi-modal learning, domain adaptation, domain generalization, egocentric vision, event-based vision