LONG-TERM UNSUPERVISED MODEL FOR VIDEO MOTION SEGMENTATION

Author's: A. Malini[1], G. Thilipkumar[2], K. Sujith[3], N. Vanjulavalli[4]

Abstract

Video understanding requires the ability to decompose complex scenes into meaningful motion components without relying on extensive manual annotations. This work presents a long-term unsupervised model for segmenting motion components in video sequences, focusing on object-centric representation learning and temporal consistency. Unlike supervised segmentation approaches that depend on pixel-level labels, the proposed framework leverages motion cues derived from frame differences to discover independently moving objects and their dynamics in a self-supervised manner.The model integrates a convolutional motion encoder, a Slot Attention module for object-centric feature grouping, and a motion decoder for reconstructing dynamic components. Motion information is extracted by computing temporal frame differences, which highlight dynamic regions while suppressing static background information. These motion features are converted into spatial tokens and processed using Slot Attention, which clusters them into a fixed number of latent slots representing distinct motion components. To ensure long-term temporal coherence, a memory gating mechanism propagates slot representations across time steps, enabling stable object tracking and consistent segmentation over extended sequences.The decoder reconstructs motion maps for each slot and generates soft segmentation masks through normalized attention weights. The system is trained using unsupervised objectives, including motion reconstruction loss, mask entropy regularization, and temporal smoothness constraints. Reconstruction loss ensures that the segmented components collectively explain the observed motion, entropy regularization encourages confident slot assignments, and smoothness constraints reduce temporal flickering in segmentation masks.