Attention-Driven Body Pose Encoding for Human Activity Recognition

Bappaditya Debnath*, Mary O’Brien†, Swagat Kumar* and Ardhendu Behera**
Department of Computer Science, †Faculty of Health, Social Care & Medicine
Edge Hill University, Ormskirk, UK, L394QP
Email:{debnathb, obrienm, kumars, beheraa}@edgehill.ac.uk
• Aim: Detect human activity from videos

• Potential Applications:
  • Home-Based Rehabilitation,
  • Surveillance,
  • Human-Robot Interaction
Proposed Approach

• Multimodal:
  • Monocular video data
  • Human body-pose data

• Philosophy:
  • Learn the spatial and temporal relationships between various body-joints for enhanced body-pose representation
Pre-trained Inception ResNet V2

Spatial stream

SEU

1D Convs

Temporal stream

TEU

1D Convs

Multi-Head Attention

Bi-LSTM

GAP and Dense Layer

Network Diagram
Spatial Encoding Unit

T = Timesteps, N = number of joints, F = Number of Filters

Normal 1D Conv = T * F
Temporal Encoding Unit

T = Timesteps, N = number of joints, F = Number of Filters

Normal 1D Conv
SEU and TEU Analysis

- Normal 1D convolution
- SEU
- TEU

- Further SEU and TEU analysis
- Extensive ablation studies
QUESTIONS?

THANK YOU