Action Recognition from RGB-D Data: Comparison and Fusion of Spatio-temporal Handcrafted Features and Deep Strategies

Abstract

In this work, the multimodal fusion of RGB-D data is analyzed for action recognition by using scene flow as early fusion and integrating the results of all modalities in a late fusion fashion. The multimodal dense trajectory (MMDT) is proposed to describe RGB-D videos as handcrafted features. Multimodal 2D CNN (MM2DCNN) is proposed as the extension of 2D CNN by adding one more input stream (scene flow). The proposed methods are evaluated on two action datasets. Fusion of handcrafted and learning-based features achieved the state of the art results.

Introduction

- Action recognition is an active research area with potential applications of health-care monitoring, interactive gaming, surveillance, and robotics.
- Microsoft Kinect has facilitated capturing of low-cost depth images in real-time alongside color images (multimodal data).
- Late fusion of RGB, depth, and motion-based representations (like optical flow) is an effective method for action recognition.
- Scene flow [1] is the real 3D motion of objects that move completely or partially with respect to a camera.
- Considers as Early fusion of RGB and depth.

Multimodal Dense Trajectory (MMDT)

- Compute scene flow along the trajectories.
- Pruning dense trajectories.
- By the information achieved by scene flow in meters.
- Scene flow is invariant to the position of the subject relative to the camera.
- Scene flow has an additional dimension, which allows the measurement of motion through 2-axis.

HON descriptor

- New source of information: i.e., depth maps.
- Each normal is represented by two angles θ and φ:
  - \( 0 < \theta < \pi \) and \( -\pi < \phi < \pi \).
- 5 bins are considered, (size of \( \pi \) radians), total of 25 bins for sub-histogram.
- The final descriptor is the concatenation of 12 sub-histograms results in 300 dimensions.

Denoising and RGB-D Alignment

- Denoising missing pixels in depth images due to:
  - Limitations of the IR sensor,
  - Special reflectance materials,
  - Distance from the objects to the camera.
- Interpolating zero value pixels by its surrounding data,
- Hybrid median filter (HMF) to reduce pixel flickering,
- Compute medians for different spatial directions
  - Horizontal/vertical + diagonal
  - Compute the median of both of them

Video Summarization

- Deep methods mostly select a fixed number of frames with equal temporal spacing
  - Thus, some relevant information might be lost.
- Key frames selection
  - Select relevant visual information to discriminate actions,
  - Keeping the size of the data small.
- Sequential Distortion Minimization (SeDiM) [4]
  - The distortion between the original video and the synopsis video is minimized.
  - Computationally feasible and discriminative way to extract key frames.

Multimodal 2D CNN (MM2DCNN)

- Three streams with 2D CNN (VGG-16):
  - Spatial network (RGB)
    - Operating on key frames,
    - Using a pre-trained network on UCF-101.
  - Temporal network (Optical flow)
    - Using volumes of stacking optical flow fields between several consecutive frames,
    - Using a pre-trained network on UCF-101.
  - Temporal network (Scene flow)
    - Consider three dimensions of scene flow as three input channels,
    - Using a pre-trained model of its own RGB model.

Experimental Result

Denoising and RGB-D Alignment

- Without pruning
- With pruning

Multimodal Data

- A combination of multimodal and deep datasets.
- Handcrafted: powerful in describing motion information.
- Deep learning: good at describing appearance data.

References