SKELETON-BASED ACTION RECOGNITION WITH CONVOLUTIONAL NEURAL NETWORKS

CS 297

Under Guidance of :
Prof. Chris Pollett

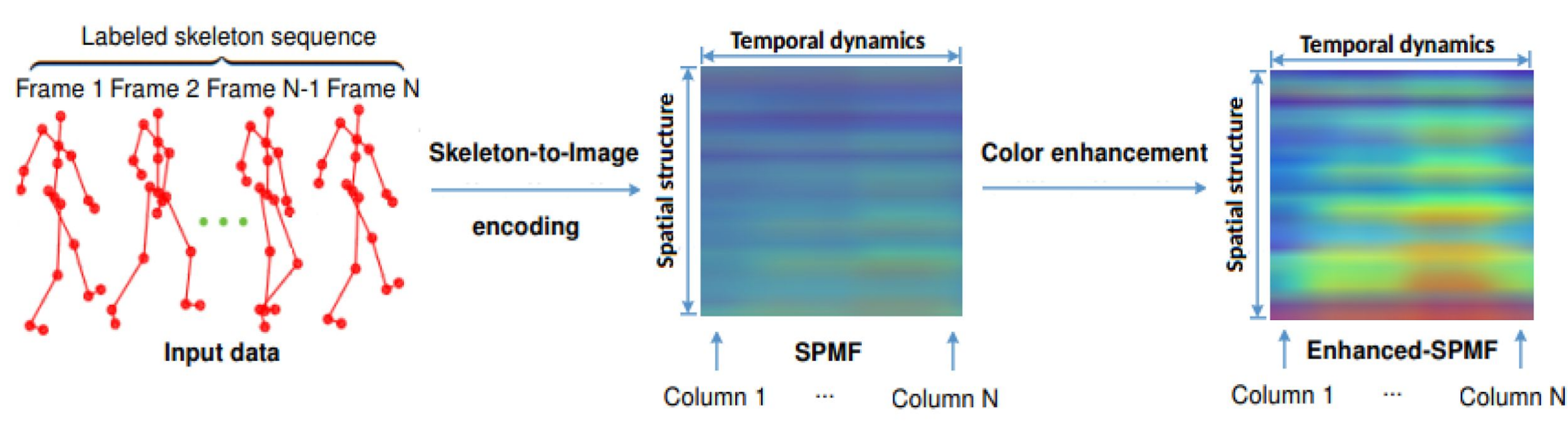
Presented By :
Charulata Lodha

FALL 2020

Introduction

- CNN based framework
- Action Classification and Detection
- 7-layer network
- 89.3% accuracy on validation set of the NTU RGB+D dataset

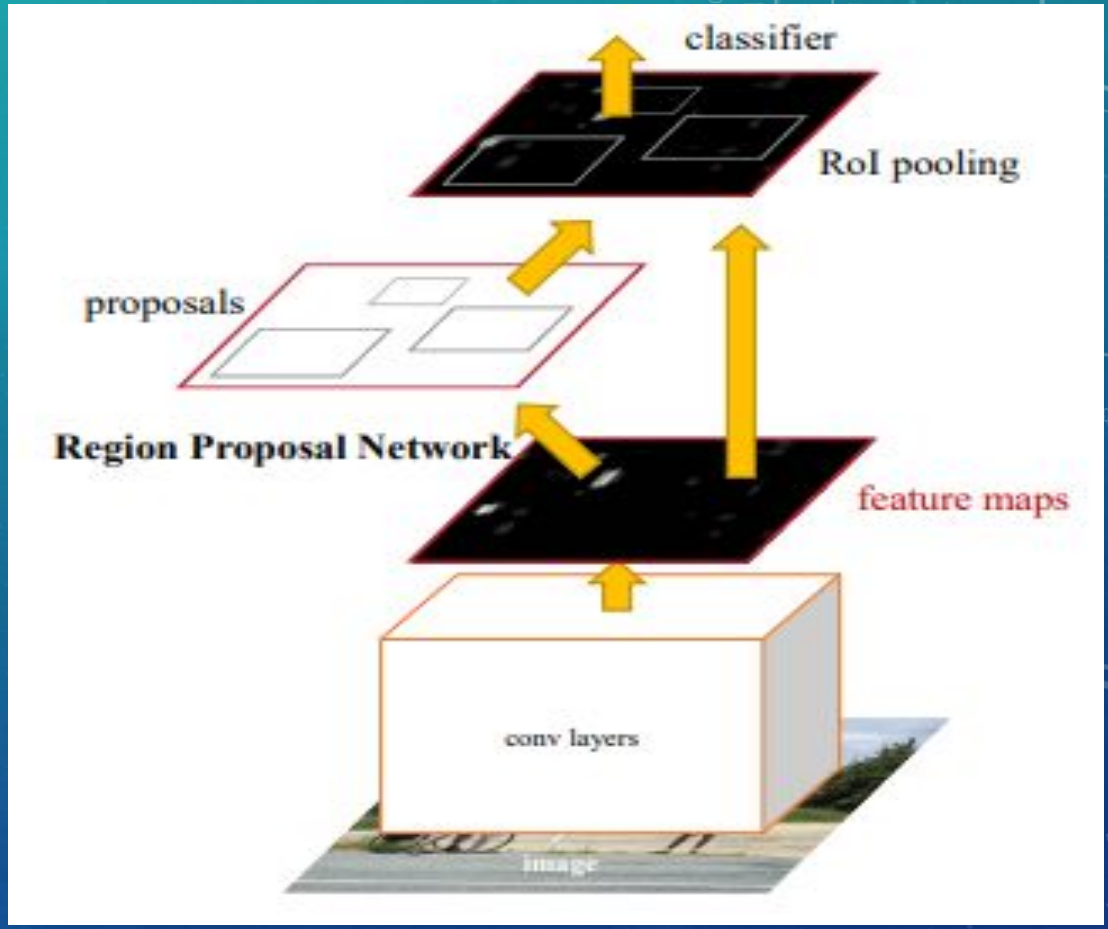
Introduction



ref: Convolutional Neural Networks ...mdpi.com

Faster R-CNN

1. RPN : putative region proposals
2. ROI Pooling: Feature extraction



RefL : Faster R-CNN | ML - GeeksforGeeks

Architecture

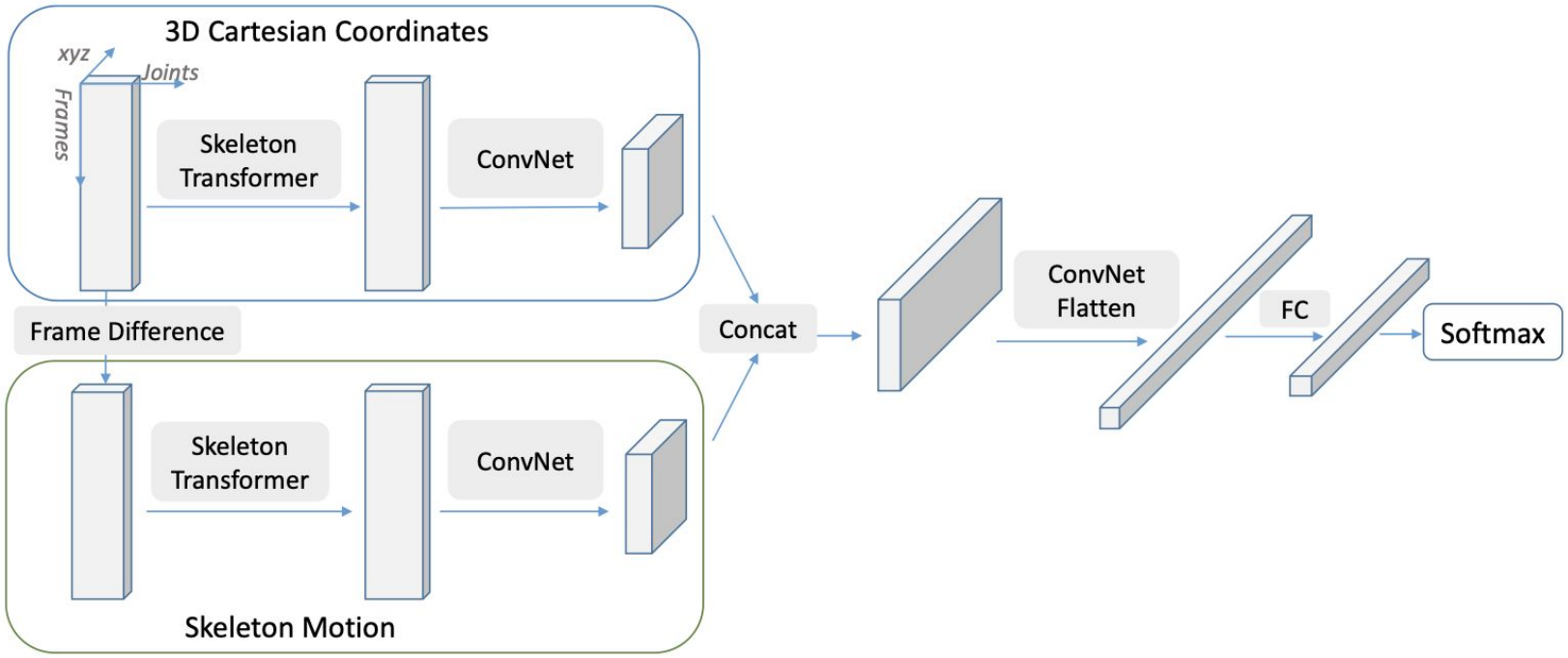


Fig. 1. CNN representation of skeleton sequences for action classification.

Architecture

- Besides raw joint coordinates, motion of skeleton joints from two consecutive frames are fed as an extra input to the network.

- Max out Merge : Multi person setting

  Skeletons of different people go through the same network layers, and their feature maps are merged by an element-wise maximum operation after the last convolution layer.

Introduction

Given a 3D joint coordinate : **$J = (x, y, z)$**

Skeleton of one person is represented as a set of joint coordinates

- **$S = \{J^1, J^2, ..., J^N\}$**

Skeleton motion between two consecutive frames

- **$M = S^{t+1} - S^t$**

A skeleton sequence of T frames can be represented as : **$T \times N \times 3$ array,**

which is treated as a $T \times N$ sized 3-channel image

N → the number of joints per skeleton

Linear transformation

- Unrecognisable skeleton when points get shuffled
- order maintenance required
- $S' = (S^T \cdot W)^T$
- W : weight matrix

Window proposal network (WPN)

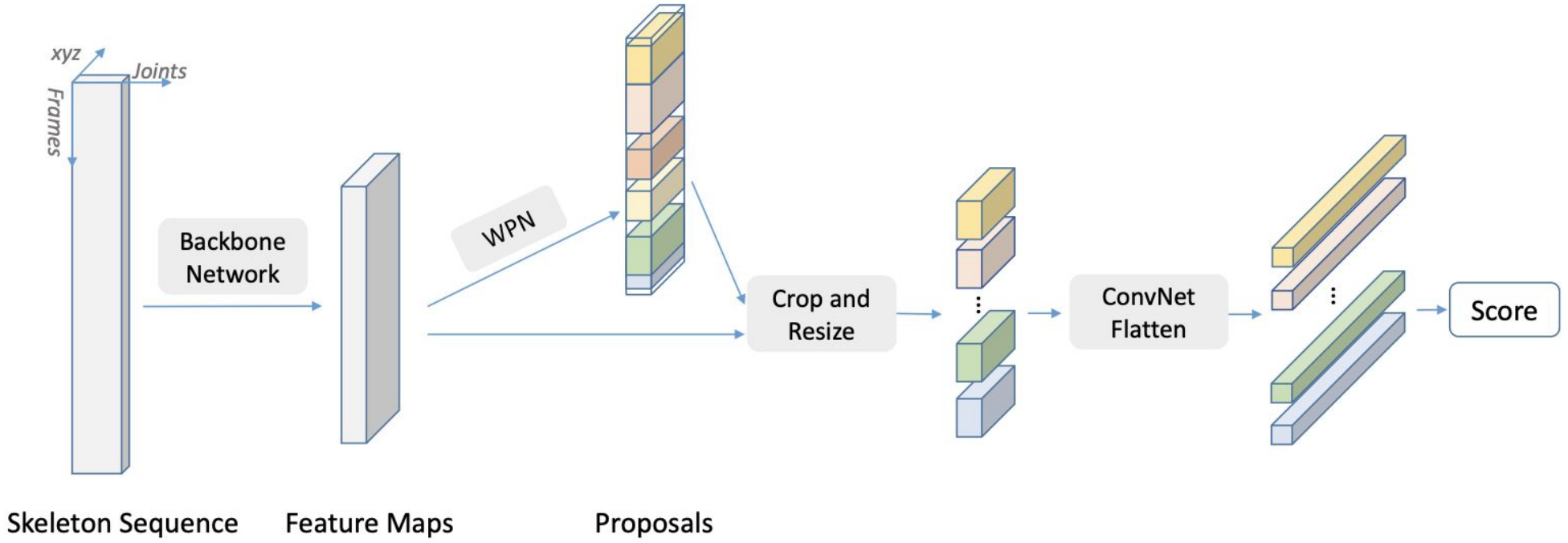


Fig. 2. Skeleton-based temporal action detection pipeline.

WPN Continued...

- Refine the temporal position of window proposals.
- After the proposals are ready, we pool features of each window
- Use shared feature maps with the crop-and-resize operation.
- Fed to the R-CNN subnetwork for classification and window regression.

References

- https://arxiv.org/pdf/1704.07595.pdf