Motivation
- Build an skeleton-based action recognition system
- Construct global information for action and select more related features
- Generate and utilize features to adapt for difference on action movements
- Extracted temporal feature representation

Challenges
- The importance of different channels varies in actions
- Some human actions only involve a small part of bodies
- Confusion on reversing actions

Contributions
- Extract the motion features from skeleton data and concatenating them with original spatial features
- Introduce a channel-wise attention module to emphasize channels with important features
- Use dense connection to ensure reuse of skeleton features and to generate a larger and sufficient features map
- Our model shows competitive performance with the state-of-the-art model on two large datasets, NTU-RGB-D and Kinetics

1. Introduction

2. Background

Skeleton Graph
- Skeleton graph G = (V,E)
- Given M frames and N joints of skeleton sequence
- Joints as vertices (V) and the connection between joints as edges (E)
- Each vertices contains three channels of information, a two-dimensional coordinate of corresponding joint and its estimation confidence
- Adjacency matrix A
- \( A_{ij} = 1 \) if i-th joint and j-th joint is connected, and 0 otherwise
- Joints are connected in one frame and adjacent frame

Baseline Model
- Spatial temporal graph convolutional networks (ST-GCN)
- Convolution operation on spatial and temporal dimension
- A mapping strategy to determine the size of convolution kernel and weight distribution of convolution on spatial dimension
- Temporal dimension convolution is similar to classical image convolution
- Two-stream adaptive graph convolutional networks (2s-AGCN)
- Based on ST-GCN

Introduce an adjacency matrix, the elements of it are parameterized and optimized together in the training process and can be arbitrary values

Introduce a similarity matrix, whose elements denotes the similarity of two vertices

Generate the connections and their importance between two vertices, which are not existed in the original graph

Data Preprocessing
- Vertex’s joint coordinate \( v_i = (x_i, y_i) \)
- Bone information, extracted from neighboring vertices, \( b_{ij} = (x_i - x_j, y_i - y_j) \)
- Motion information of joints and bones, extracted from consecutive frames of data, \( m_{ij} = (x_{ij+1} - x_{ij}, y_{ij+1} - y_{ij}) \)
- Concatenate the information of joints and their motion in the frame dimension. The same procedure was conducted with the bones.

Model Structure
- Two-stream fashion, each stream consists of 12 graph convolution blocks with late fusion
- A graph convolution block (GCB), consists of a spatial GCN, a temporal GCN and a channel-wise attention module (CAM), followed by a residual connection
- A DC-GCB group, consists of 4 GCBs with Dense connection implemented, followed by a transition layer

Channel-wise attention module (CAM)
- Encode the entire spatial and temporal feature on a channel as a global feature descriptor
- Analyze the interdependence between channels, generate a set of attention weights of corresponding channels
- A channel-wise multiplication is made to represent a global information based on feature channels
- Dense Connection
- Concatenation of all the preceding graph convolution block’s output features maps
- Transition layer between blocks to reduce the number of features

Training
- Training batch size 64 and Test batch size 64
- Stochastic gradient descent (SGD) as optimizer with an initial learning rate 0.1 and a cosine learning rate decay
- The weight decay of 0.0001 and Nesterov momentum of 0.9 are set
- The hyperparameter reduction ratio \( r \), used in channel-wise attention module, is set to 16

Dataset
- NTU-RGB-D
  - 60 different action classes including daily and health-related actions
  - 25 body joints collected by Microsoft Kinect v2
  - 40 distinct subjects recorded from 3 different horizontal angles
  - Cross-subject evaluation and cross-view evaluation
  - Kinetics
  - 400 action classes with at least 400 video clips
  - 18 body joints obtained by OpenPose toolbox

Ablation Study
- Channel-wise Attention Module
- Choose dataset NTU-RGB-D to test the Top-1 accuracy
- Valid effect on improving the performance of the model
- The module learns the non-linear relations between channels and the scale is not one-hot encoding
- Emphasize multiple channels with more importance
- Dense Connection
- Performance improvement shows that the network takes the advantage of Dense Connection
- Produces a larger and sufficient features map to make better results
- CAM, compared with DC, achieves higher accuracy improvements on cross-view benchmark, and vice versa, which can be explained by the relationship between modification modules and NTU-RGB-D setup

Comparison with the State-of-the-art methods
- Methods include hand-crafted methods [42][49], CNN-based methods [5][44][43][47], RNN-based methods [9][10][45][11] and GCN-based methods[14][13][15][48][46][44]

Comparison on Kinetics dataset
- Outperforms hand-crafted methods, CNN and RNN methods with a large margin
- A competitive result comparing with the state-of-the-art GCN-based methods

Comparison on NTU-RGB-D dataset
- Outperforms hand-crafted methods, CNN and RNN methods with a large margin
- A competitive result comparing with the state-of-the-art GCN-based methods

1. Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China
2. Department of Computer Science and Technology, Tsinghua University, Beijing, China

Michael Lao BanTeng, Zhiyong Wu

Channel-Wise Dense Connection Graph Convolutional Network
For Skeleton-Based Action Recognition