IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC-2021)
Singapore University of Technology and Design, Singapore
December 20-23, 2021

Light-weight Enhanced Semantics-Guided Neural Networks for Skeleton-Based Human Action Recognition

Hongbo Chen[1,2] **Lei Jing**[1,2]

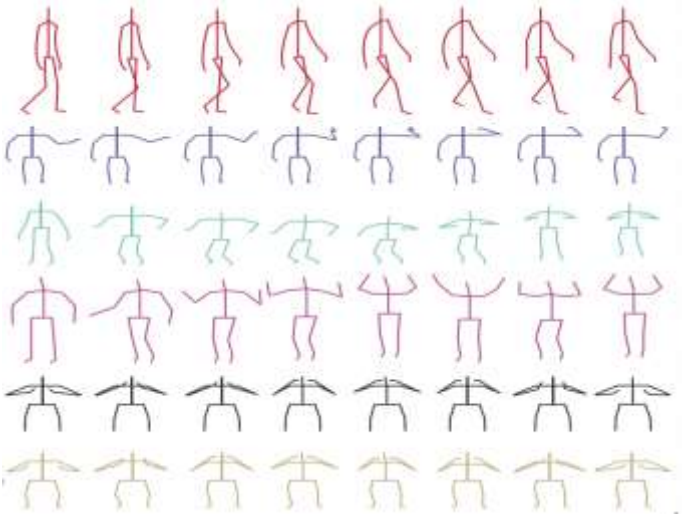
[1]The University of Aizu
[2]The CNLab of Aizu

Outline

- Introduction
  - Background
  - Purpose
  - Issue
  - Solution
- Methods
- Experiment
- Results
- Conclusion

Background

Human action recognition has many application scenarios in the real world, skeleton-based representations have been very popular for human action recognition, as human skeletons provide a compact data form to depict dynamic changes in human body movements.

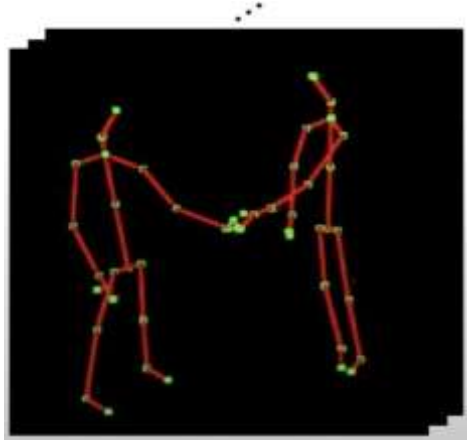


Purpose

1.Human actions can be efficiently represented by skeletons
2.Free of background clutter/lighting conditions/clothing variations

Input Video

Estimated 2D/3D Poses



Skeletons

Issue

1.Most of the previous methods had a very large number of parameters.
2.Low-parameter method (SGN) performs poorly in terms of accuracy.

Method	Year	Params	CS	CV	GCN-based
ST-GCN	2018	3.1	81.5	88.3	yes
SR-TSL	2018	19.07	84.8	92.4	no
RA-GCN	2019	6.21	85.9	93.5	yes
AS-GCN	2019	6.99	86.8	94.2	yes
2s-AGCN	2019	6.94	88.5	95.1	yes
ACG-LSTM	2019	22.89	89.2	95.1	no
DGNN	2019	26.24	89.9	96.1	yes
VA-fusion	2019	24.6	89.4	95.1	no
PL-GCN	2020	20.7	89.2	95	yes
NAS-GCN	2020	6.57	89.4	95.7	yes
4s-Shift-GCN	2020	2.76	90.7	96.5	yes
SGN	2020	0.64	89	94.5	yes

Params 0-1.0
Params 1.0-5.0
Params 5.0-10.0
Params 10.0-20.0
Params >20.0

Solution

1.Maintaining a light-weight baseline(<1.0M parameters)

2.Improving the accuracy of SGN to make it comparable to the previous method

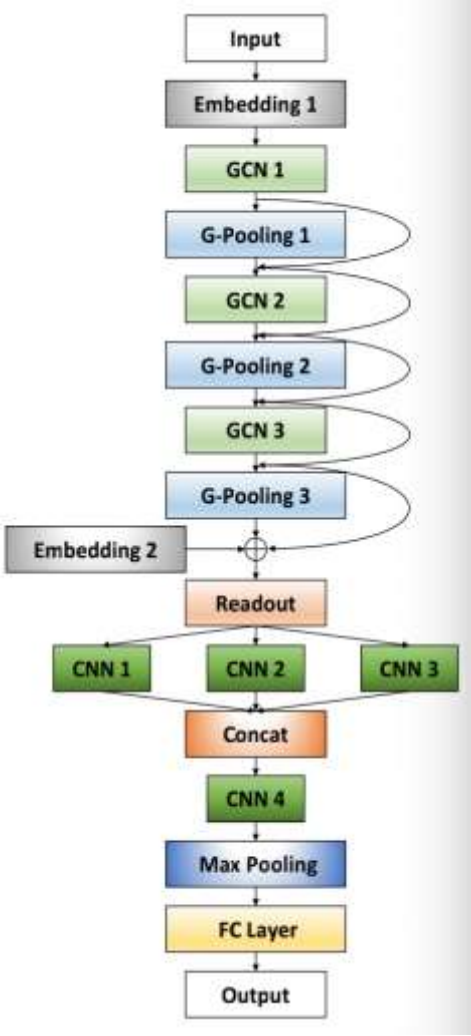
Methods

1.We increase the number of streams of the original information, and expand the single stream of data into four streams of fusion(Embedding part)

2.We propose a self-attention block based on pooling, which can effectively discover the most important joints.(G-Pooling part)

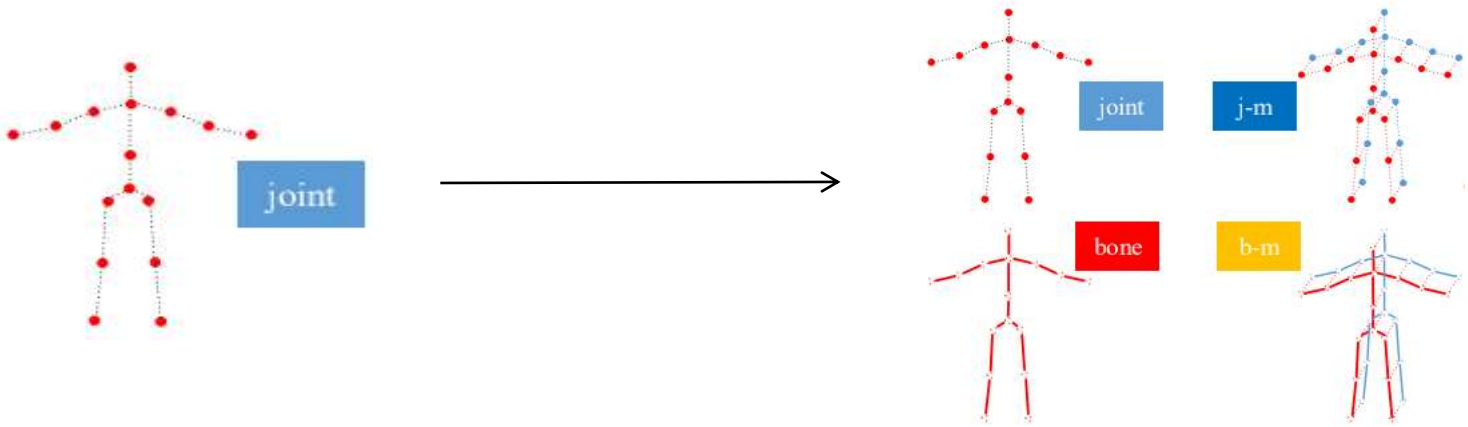
3.We propose a multi-scale CNN to extract temporal features of different scales.(CNN1,2,3 part)

4.In the loss part, we add a regularization term to improve the generalization ability of the model.



Method 1 - Enhanced Semantic

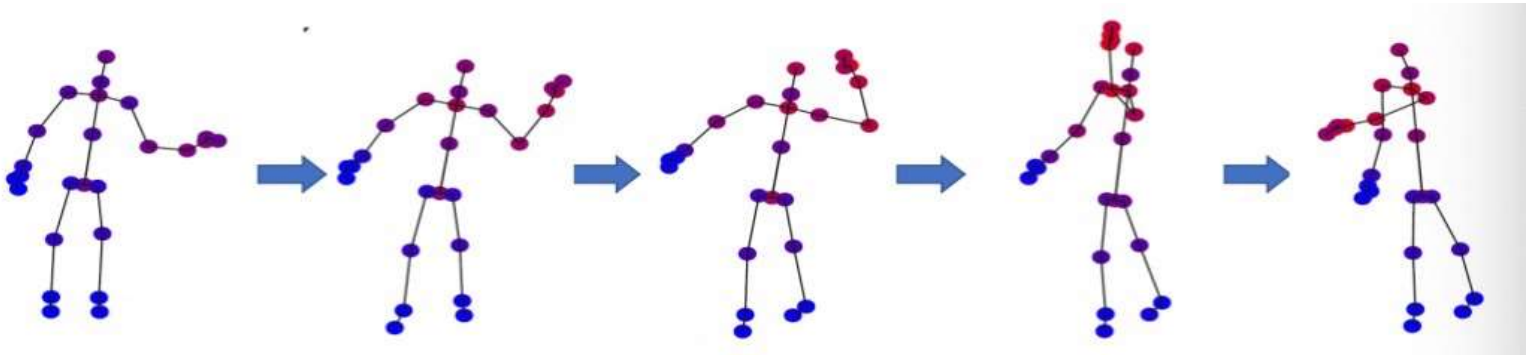
We enhance the original input, expand the original single stream of data to four streams, increase the accuracy, and the additional parameter burden is less than 0.1M.



Method 2- Semantic-Pooling layer

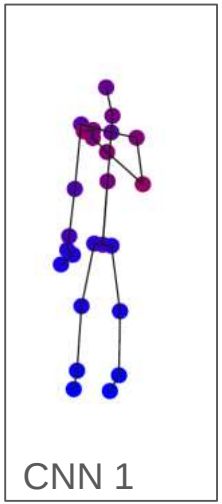
We visualize the Semantic-Pooling layer, and it is obvious that the model output the important joints.

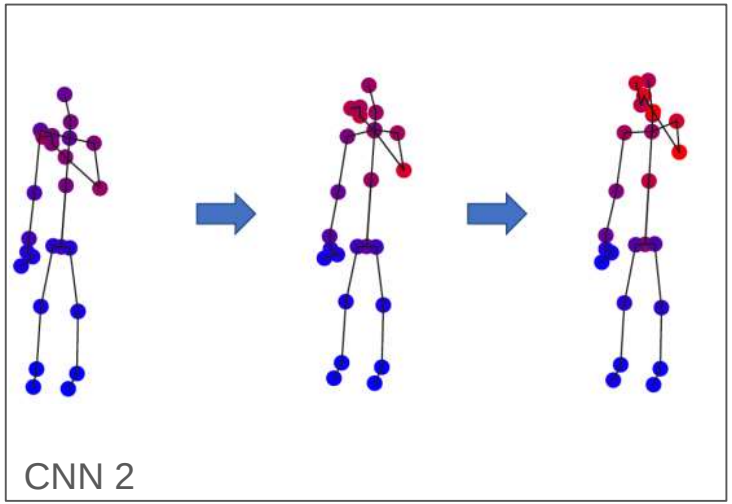
In the throwing action, the red circle part indicates the high importance joint in the movement, and the blue circle indicates the low importance joint.

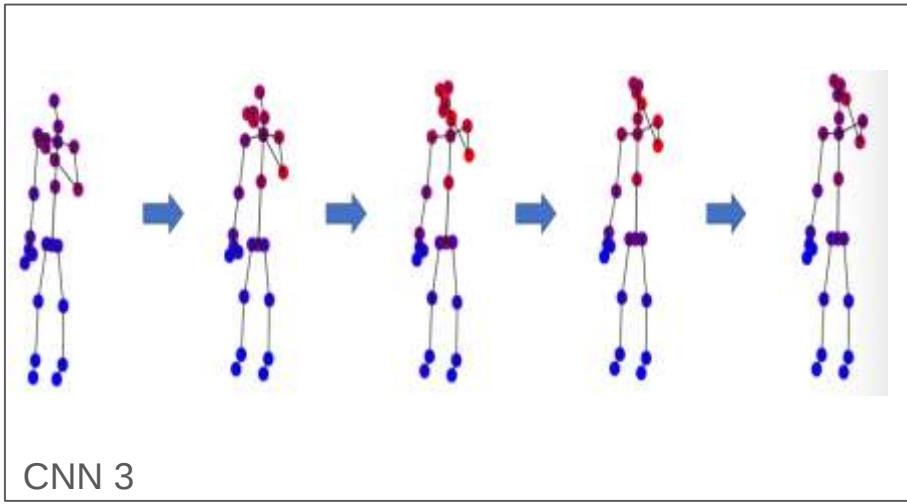


Method 3- Multi-scale CNN

We propose a multi-scale CNN to aggregate time information of different lengths of time, and it can be well matched with the Semantic-Pooling layer.

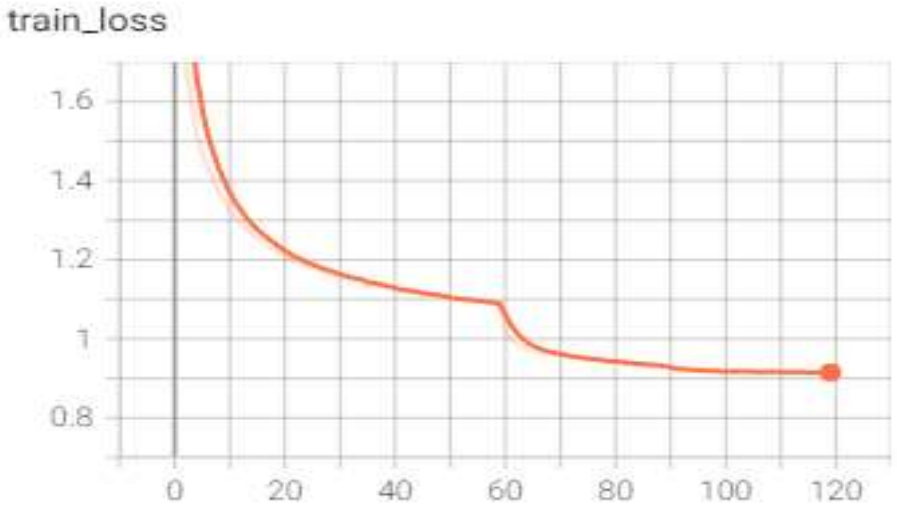






Method 4 - Regularized loss function

We have added a regularization term to the loss function,which will help improve the generalization ability of the model.At the same time, the training loss can be steadily and effectively dropped.



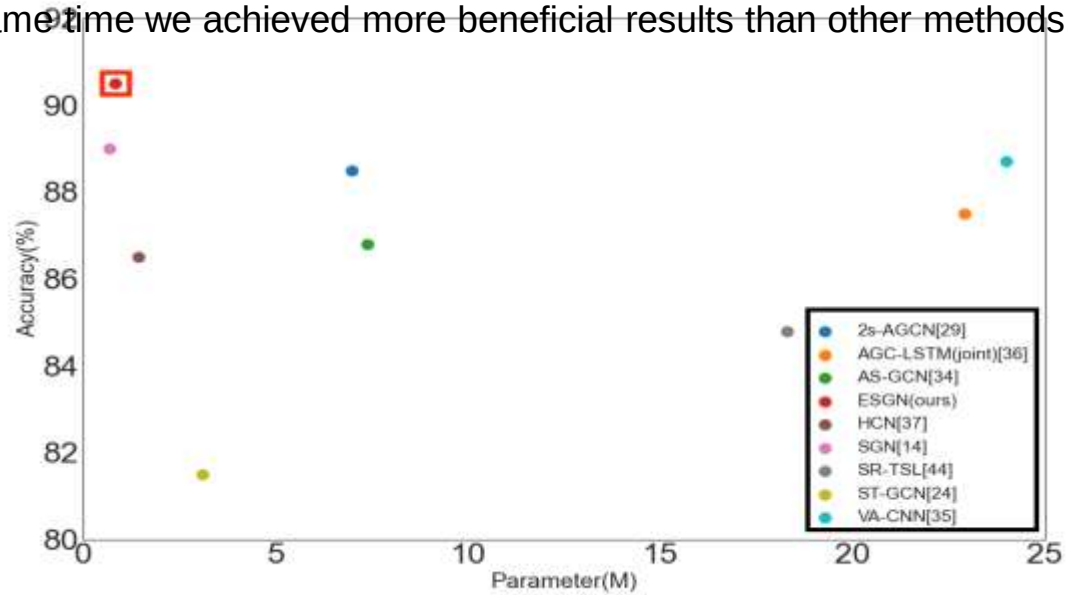
Experiments

- All experiments are conducted on the Pytorch platform with one 1070 GPU card.
- We use the Adam optimizer with the initial learning rate of 0.001.
- The learning rate decays by a factor of 10 at the 60th epoch, the 90th epoch, and the 110th epoch, respectively.
- We use two large data sets for evaluation, NTU60 and NTU120.
- The training is finished at the 120th epoch. We use a weight decay of 0.0001. The batch sizes for NTU60, NTU120, are set to 64.

Results

Although the method we proposed has a slight increase in the amount of parameters compared with SGN, the overall parameter scale is less than 1M, which is lightweight.

At the same time we achieved more beneficial results than other methods.



Conclusions

- In this paper, we propose an enhanced semantics-guided neural networks for the skeleton-based action recognition task.
- These all verify the reliability and accuracy of our model in the experimental part.
- In practical applications, our model can be applied to surveillance systems,autonomous driving, intelligent robot, and human-machine interaction, etc.
- The smaller parameter burden also means that the model has better potential to generalize to mobile devices.

Thank you for listening!