

# Spatial-Temporal Attention Res-TCN for Skeleton-based Dynamic Hand Gesture Recognition

Jingxuan Hou<sup>1</sup>, Guijin Wang<sup>1</sup>, Xinghao Chen<sup>1</sup>, Jing-Hao Xue<sup>2</sup>, Rui Zhu<sup>3</sup>,  
and Huazhong Yang<sup>1</sup>

<sup>1</sup> Tsinghua University, Beijing, China  
{houjx14, chen-xh13}@mails.tsinghua.edu.cn  
{wangguijin, yanghz}@tsinghua.edu.cn

<sup>2</sup> University College London, London, UK  
jinghao.xue@ucl.ac.uk

<sup>3</sup> University of Kent, Kent, UK  
R.Zhu@kent.ac.uk

**Abstract.** Dynamic hand gesture recognition is a crucial yet challenging task in computer vision. The key of this task lies in an effective extraction of discriminative spatial and temporal features to model the evolutions of different gestures. In this paper, we propose an end-to-end Spatial-Temporal Attention Residual Temporal Convolutional Network (STA-Res-TCN) for skeleton-based dynamic hand gesture recognition, which learns different levels of attention and assigns them to each spatial-temporal feature extracted by the convolution filters at each time step. The proposed attention branch assists the networks to adaptively focus on the informative time frames and features while exclude the irrelevant ones that often bring in unnecessary noise. Moreover, our proposed STA-Res-TCN is a lightweight model that can be trained and tested in an extremely short time. Experiments on DHG-14/28 Dataset and SHREC'17 Track Dataset show that STA-Res-TCN outperforms state-of-the-art methods on both the 14 gestures setting and the more complicated 28 gestures setting.

**Keywords:** dynamic hand gesture recognition, spatial-temporal attention, temporal convolutional networks

## 1 Introduction

Dynamic hand gesture recognition has attracted increasing interests due to its potential relevance to a wide range of applications, such as touchless automotive user interfaces, gaming, robotics, etc [21, 3, 28]. However, it is still challenging to develop a highly precise hand gesture recognition system, owing to high intra-class variance derived from the various possibilities to perform the same gesture [30, 5, 3].

---

Corresponding Author





























