SEMINAR ANNOUNCEMENT

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
COLLEGE OF DESIGN AND ENGINEERING
Website: https://cde.nus.edu.sg/ece

Area: Signal Analysis & Machine Intelligence

Host: Asst Prof. Shou, Zheng

TOPIC	**:**	**One-Stage Action Detection**
SPEAKER	**:**	**Mr. Chang Shuning** **Graduate Student, ECE Dept, NUS**
DATE	**:**	**Friday, 27 May 2022**
TIME	**:**	**4.30PM to 5.30PM**
WEBINAR	**:**	**Join Zoom Meeting** **https://nus-sg.zoom.us/j/86825312359?pwd=UkF3NjFMT0RsZ2NFMEFMYkxXdThqZz09** **Meeting ID: 868 2531 2359** **Passcode: 877566**

ABSTRACT

Localizing people and recognizing their actions from videos is a challenging task for media interpretation and high-level video understanding. Existing methods are mostly two-stage based, with one stage for person bounding box generation and the other stage for action recognition. However, such two-stage methods are generally with low efficiency. We observe that directly unifying detection and action recognition usually suffers from (i) inferior learning due to different desired properties of context representation for detection and action recognition; (ii) optimization difficulty with insufficient training data. In this work, we present a decoupled end-to-end one-stage network dubbed DOAD, to mitigate above issues and improve the efficiency for spatio-temporal action detection. Detection and action recognition are decoupled into two branches. Specifically, one branch focuses on detection representation for actor detection, and the other one for action recognition. For the detection branch, we utilize a temporal aggregation module to enhance the person features by aggregating temporal information. For the action branch, we design a transformer-based module (TransPC) to model pairwise relationships between people and contexts. Different from commonly used vector-based dot product in self-attention, it is built upon a novel matrix-based key and value for Hadamard attention when modeling person-context information. It not only exploits relationships between person pairs but also takes into account the context and relative position information. The results on AVA and UCF101-24 datasets show that our method is competitive with two-stage state-of-the-art methods, and the inference speed is improved by over 50% compared with typical two-stage methods.

BIOGRAPHY

Mr. Chang Shuning is a PhD candidate of ECE, NUS, supervised by Prof. Shou Zheng. My interested field is video understanding and efficient model design.