2024 Spring Computer Vision VIVOTEK Final Project

CHIH-HAO LIAO[1,*], YI-HAN LEE[2,+], and HSIN-TZU LI[2,+]

[1]School of Forestry and Resource Conservation, National Taiwan University, Taipei, 106319, Taiwan
[2]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, 106319, Taiwan
[*]R11625015r@ntu.edu.tw
[+]these authors contributed equally to this work

ABSTRACT

This report serves as the final project for Prof. Shao-Yi Chien's spring course on Computer Vision: from Recognition to Geometry. The assignment involves developing a system for real-time monitoring of door status in public transit systems. In this project, we developed a fine-tuned Swin Transformer V2 model capable of accurately predicting the status of doors in public transit systems. The model can detect and localize doors using video from cameras, and it provides precise monitoring of door statuses, including Opening, and Closing statuses.

Introduction

This is the final project competition of the Computer Vision Course (Spring 2024, National Taiwan University) sponsored by Vivotek.

Motivation

An Automated Passenger Counter (APC) is an electronic device installed on transit vehicles, such as buses and trains, to record the times and locations of passengers boarding and disembarking. This data is crucial for analyzing travel patterns and enhancing the operational efficiency of transportation services. The APC relies on real-time signals from vehicle doors to determine when they open and close, initiating and finalizing the counting process accordingly. However, integrating APC systems with the door status signals of older public transit vehicles can be challenging due to difficulties in accessing and correctly connecting the necessary wiring. To overcome this, there is a need for a vision-based automatic door status monitoring technology that eliminates the dependency on external wiring for door status signals.

Methods

Swin Transformer

The Swin Transformer is a type of Vision Transformer designed for image classification and other vision tasks. It introduces a hierarchical architecture that processes images at multiple scales using shifted windows for efficient computation. This approach allows the model to handle high-resolution images and capture local context effectively.

In our study, the baseline model that we used is from Microsoft swin transformer v1[1], and the model structure that we used from PyTorch is swin transformer v2 base[2], and the pre-trained model that we used on Hugging Face[3] is inherited from the Microsoft swinv2-tiny-patch4-window8-256.[2].

References

1. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2021).
2. Liu, Z. *et al.* Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).
3. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, Online, 2020).