Progressive Spatial Weighting and Temporal Aggregation Based Multi-Stage Regression for Video-Based Head Pose Estimation

李昱鴻　池永研究室　修士課程修了

Background

HCI　Attention detection　Behavior analysis

Head motion information from single image is limited.

Video-based head pose estimation is needed

Problems

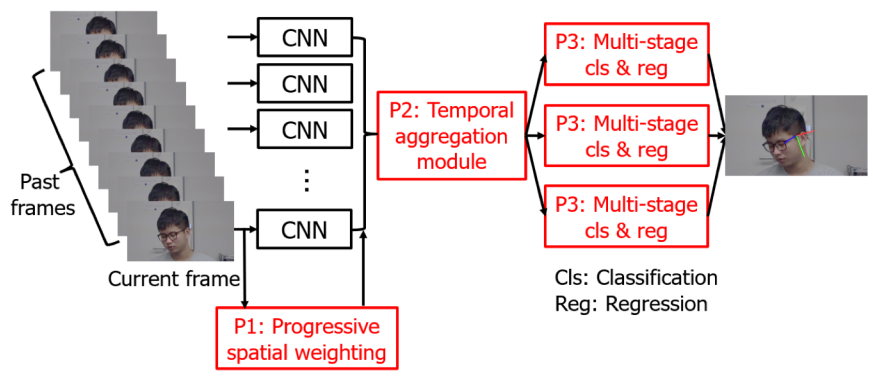
- Need pre-processing for input frames
- Simple operation and network to utilize temporal information
- No flexible and accurate for angle estimation

Solution

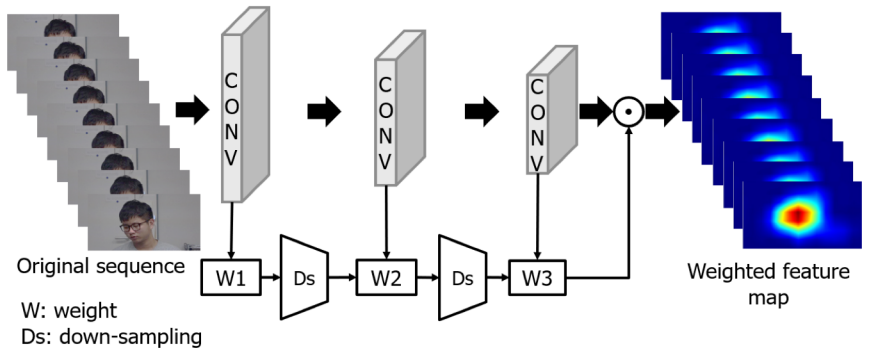
- **P1:** Progressive Spatial Weighting
- **P2:** Temporal Aggregation Module
- **P3:** Multi-Stage Classification and Regression

Proposed Methods

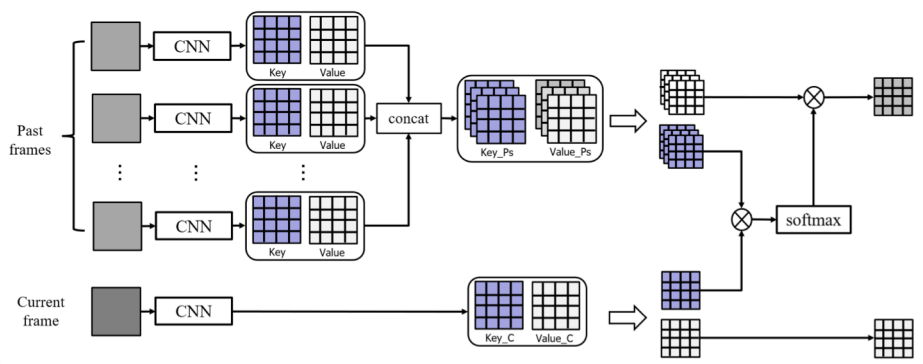
Overall framework



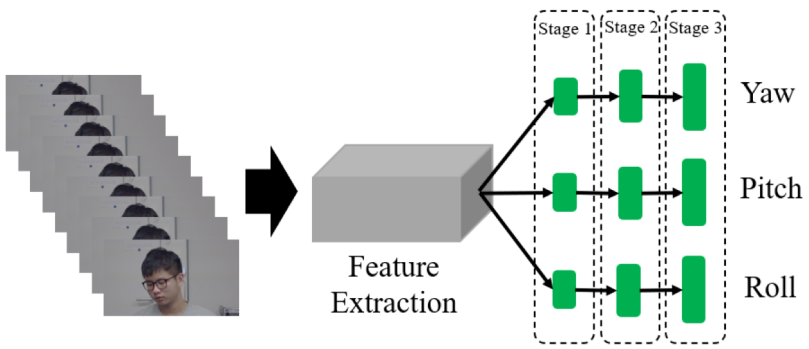
P1: Progressive Spatial Weighting



P2: Temporal Aggregation Module



P3: Multi-Stage Classification and Regression



Experimental Results

Private Dataset	Mean Absolute Error (MAE)					
	Simple			Complex		
	Yaw	Pitch	Roll	Yaw	Pitch	Roll
RW(1+2)	3.5137	3.0621	2.9073	6.5137	5.8621	4.9341
+P2	2.4504	2.8748	1.9806	4.3526	4.8748	3.9956
+P3	2.6569	4.1726	2.3078	5.6569	6.1681	4.3078
+P2+P3	2.4270	2.7688	1.9587	4.1333	4.2811	3.9457
+P1+P2+P3	**2.3562**	**2.7598**	**1.9521**	**3.7679**	**3.9564**	**3.1019**

BIWI Dataset	Mean Absolute Error (MAE)			
	Yaw	Pitch	Roll	Avg
RNN	3.14	3.48	**2.60**	3.07
RW(1+2)	4.33	4.42	4.09	4.28
Our	**2.78**	**3.38**	2.91	**3.02**

Conclusion

- **This work aims to achieve high accuracy head pose estimation based on monocular video.**
- **The proposed network achieves lowest MAE both on simple and complex backgrounds on the private dataset and outperforms RNN-based method by 0.05° on average MAE on BIWI dataset.**

Graduate School of Information, Production and Systems
Waseda University