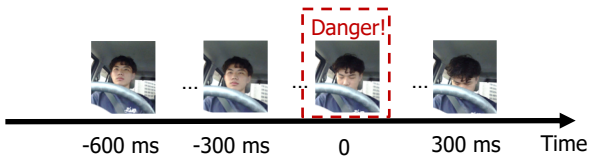
History Guided Pyramid Spatial-Temporal Neural Network with Displacement Based Refinement for Head Pose Prediction

許喆　池永研究室　修士課程修了

Background

Driver monitoring



Obtaining motion understanding before finishing the motion

Avoiding accidents in advance

Problems

No existing work for one-stage head pose prediction

- Unpowerful encoder
- Ignoring the information difference between features
- No information between frames

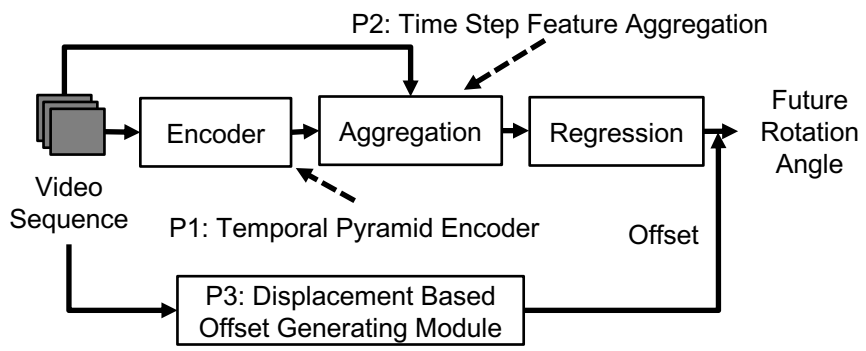
Solution

Hierarchically extracting temporal information for prediction

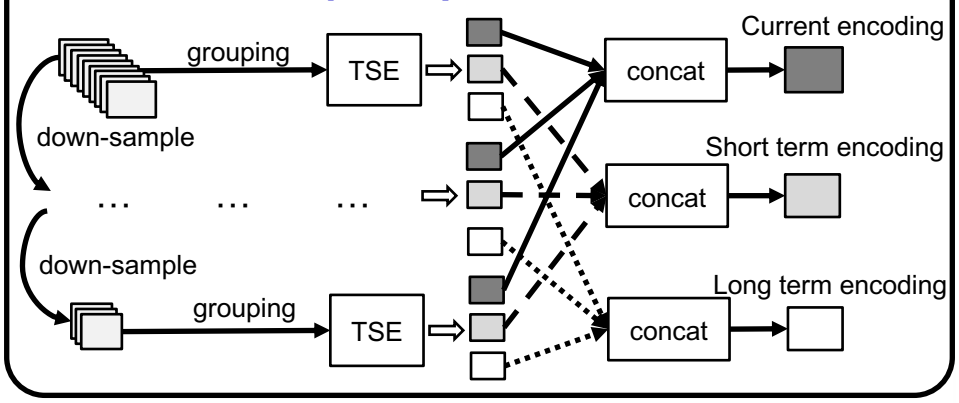
- **P1:** Temporal Pyramid Encoder
- **P2**: Time Step Feature Aggregation
- **P3**: Displacement Based Offset Generating Module

Proposed Methods

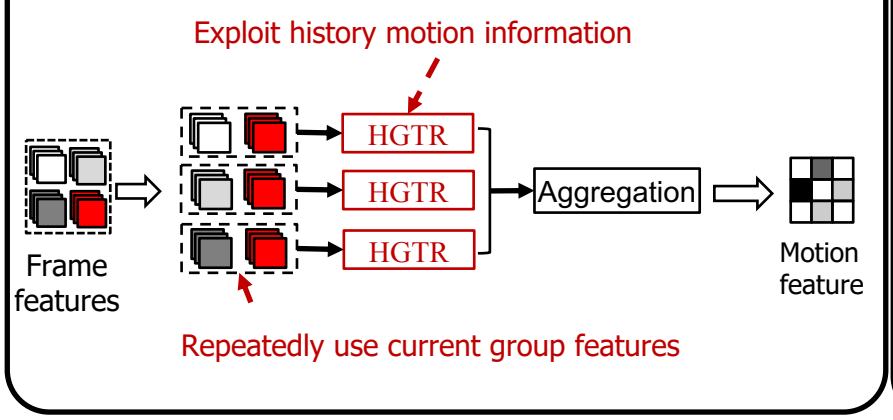
Overall framework



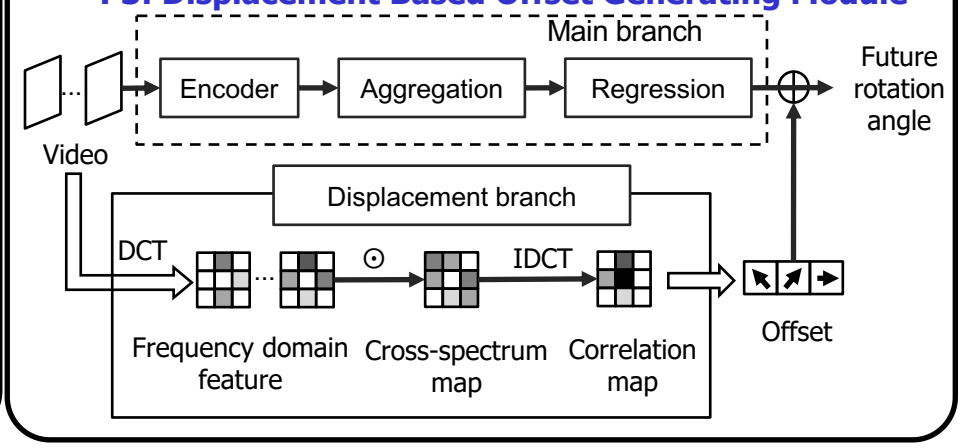
P1: Temporal Pyramid Encoder



P2: Time Step Feature Aggregation



P3: Displacement Based Offset Generating Module



Experimental Results

Motion	Method	100 ms prediction				300 ms prediction			
		Yaw	Pitch	Roll	Avg	Yaw	Pitch	Roll	Avg
Simple	TFSANet	10.8175	8.9473	6.6510	8.8052	12.6233	9.2746	7.5277	9.8085
	P1	7.9563	5.9475	4.7318	6.2118	10.1805	6.7803	5.2983	7.4179
	P1+P2	6.0852	4.6326	3.6802	4.7993	9.2322	5.9853	4.8352	6.6842
	P1+P2+P3	**5.0257**	**4.1531**	**3.1588**	**4.1125**	**8.3584**	**5.2371**	**4.0781**	**5.8912**
Common	TFSANet	10.3852	8.5550	6.7966	8.5789	13.8865	8.1460	7.0851	9.7058
	P1	8.2563	5.8257	4.8674	6.3164	11.0718	6.6950	5.4328	7.7332
	P1+P2	6.4729	4.2121	3.7444	4.8098	10.7265	**5.8934**	4.4459	7.0219
	P1+P2+P3	**5.2235**	**3.8424**	**3.6560**	**4.2463**	**8.9309**	5.9858	**4.1769**	**6.3645**

Conclusion

- **This work aims to achieve high accuracy head pose prediction based on monocular video.**
- **The proposed network outperforms TFSANet by relatively 50% improvement on 100 ms prediction and relatively 40% improvement on 300 ms prediction.**

Graduate School of Information, Production and Systems
Waseda University