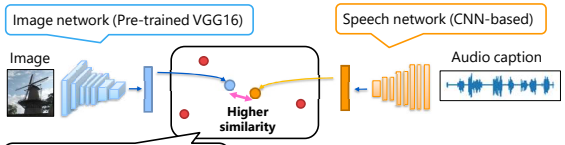
Unsupervised Co-Segmentation for Athlete Movements and Live Commentaries Using Crossmodal Temporal Proximity

Yasunori Ohishi, Yuki Tanaka, and Kunio Kashino
NTT Corporation, Japan

Related work

Embedding model (DAVEnet) that can directly associate visual objects with spoken words [Harwath+2016]

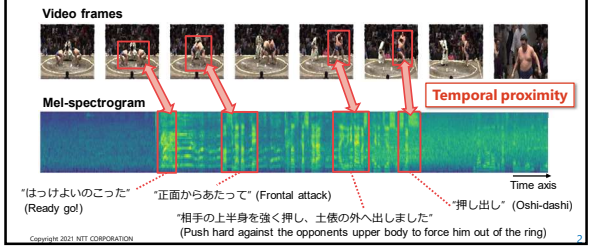


- Triplet loss function
- Margin softmax loss function
- Noise contrastive estimation

- 400K English captions [Harwath+2019]
- 100K Hindi captions [Harwath+2018]
- 100K Japanese captions [Ohishi+2020]

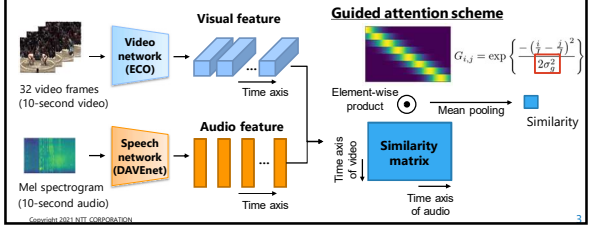
Our challenge

Co-segmentation of sports actions and live commentary



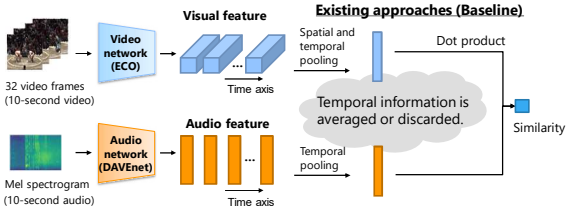
Model

Guided attention scheme to efficiently detect and utilize temporal co-occurrences of audio and video information



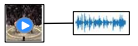
Model

Guided attention scheme to efficiently detect and utilize temporal co-occurrences of audio and video information



Dataset

- 170 hours of NHK broadcast of grand sumo tournaments
- 1,218 matches of nine frequent winning techniques
- 10-second video clips and their raw audio waveforms centered around labeled times as audio-visual pairs



Winning techniques	Training	Validation
Frontal push out	365	10
Frontal force out	362	10
Slap down	141	10
Thrust down	77	10
Over arm throw	45	10
Frontal thrust out	42	10
Frontal crush out	34	10
Rear push out	34	10
Frontal push down	28	10
	1,128	90

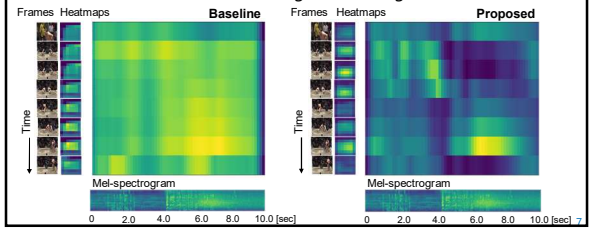
Crossmodal search results

Audio-visual retrieval recall scores when the correct result was defined as the clips with the same winning techniques as the query

	Audio to Video			Video to Audio		
σ_g	R@1	R@3	R@5	R@1	R@3	R@5
0.001	.289	.600	.739	.294	**.611**	.717
0.01	**.348**	**.656**	**.770**	.304	.604	**.785**
0.1	.304	.648	.763	**.307**	.581	.733
1	.289	.600	.711	.211	.511	.622
10	.211	.461	.611	.144	.389	.561
100	.122	.389	.511	.056	.211	.411
Baseline	.256	.422	.589	.233	.511	.633

Co-segmentation results

Our method better captures the correspondence between audio and visual information and the edges of the segments.



Co-segmentation results

Our method better captures the correspondence between audio and visual information and the edges of the segments.