16

Crossmodal learning for concept acquisition of human movements

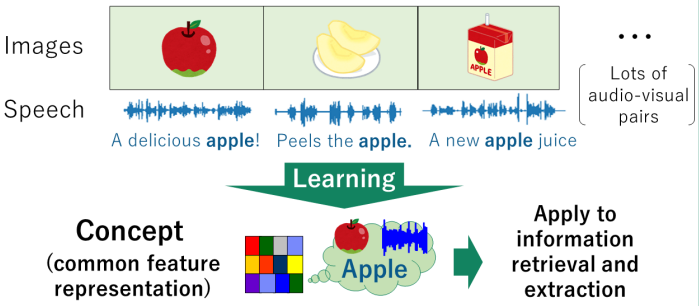
AI that acquires knowledge just by watching TV

Abstract

We developed a crossmodal learning method that can acquire "concepts" corresponding to specific objects and events on unlabeled audio and video signals. Achieving it in an unsupervised way is particularly important, since it is generally difficult to manually label all the objects and events appearing in audio-visual data for supervised learning. Our main idea was identifying concepts by looking at them from different modalities, just like looking at objects from different angles. To efficiently detect and utilize temporal co-occurrences of audio and video information, we employed a guided attention scheme. Experiments using real TV broadcasts of sumo wrestling with live commentaries show that our method can automatically associate specific athlete techniques and its spoken descriptions without any manual annotations. We are aiming for a future in which AI can acquire knowledge autonomously by just watching and listening to everyday scenes, or watching TV.

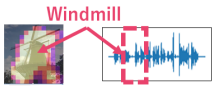
Crossmodal concept acquisition

Acquire “concepts” by learning semantic associations based on co-occurrences across different modalities in unsupervised manner



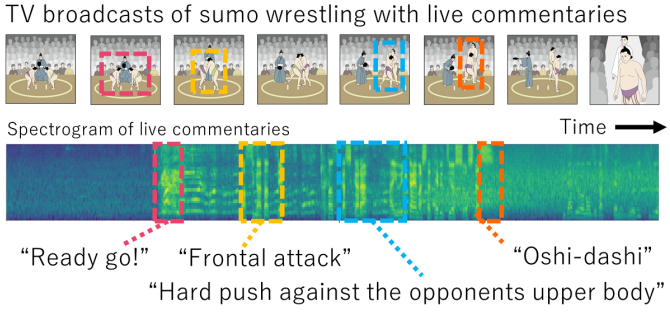
AI that acquires knowledge just by watching TV

Demonstrated feasibility of concept acquisition by associating visual objects with spoken words [1][2]



Concept acquisition of human movements

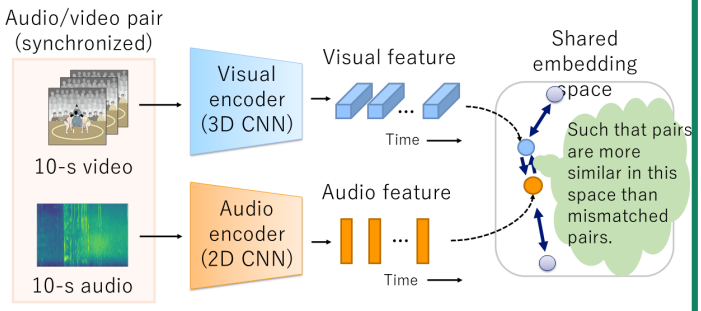
Utilize temporal proximity of spoken words appearing close to human movements in time



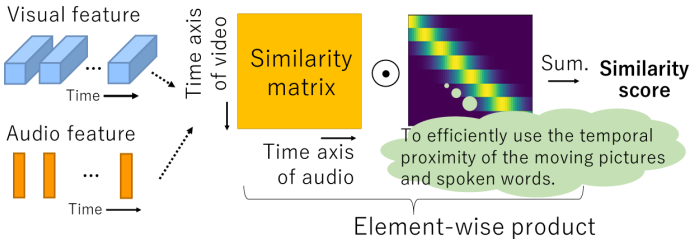
Introduce “guided attention” scheme along time to capture spatio-temporal correspondence [3]

Learning method and results

Method: train parameters of audio/visual encoders to optimize a ranking-based criterion

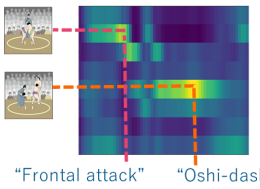


Point: compute similarity matrix between visual and audio features so that matrix becomes nearly diagonal



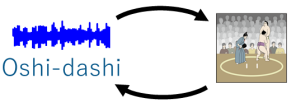
Results: implicitly learned underlying concepts of athlete movements in sumo bouts

Similarity matrix between visual and audio features



Concept retrieval (Recall score)

Audio-to-video retrieval:
Improved from 0.08 to 0.36



Video-to-audio retrieval:
Improved from 0.04 to 0.29

*Experimental conditions:
- 1,218 matches of NHK broadcast video of grand sumo tournaments in which the winners were determined by nine frequent winning techniques (1,128 for training and 90 for validation)

References

[1] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, J. Glass, “Trilingual Semantic Embeddings of Visually Grounded Speech with Self-attention Mechanisms,” in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*.
[2] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, J. Glass, “Pair Expansion for Learning Multilingual Semantic Embeddings using Disjoint Visually-grounded Speech Audio Datasets,” in *Proc. Interspeech 2020*.
[3] Y. Ohishi, Y. Tanaka, K. Kashino, “Unsupervised Co-Segmentation for Athlete Movements and Live Commentaries Using Crossmodal Temporal Proximity,” in *Proc. International Conference on Pattern Recognition (ICPR) 2020*.

Contact

Yasunori Ohishi / Recognition Research Group, Media Information Laboratory
Email: cs-openhouse-ml@hco.ntt.co.jp