Crossmodal learning for concept acquisition of human movements

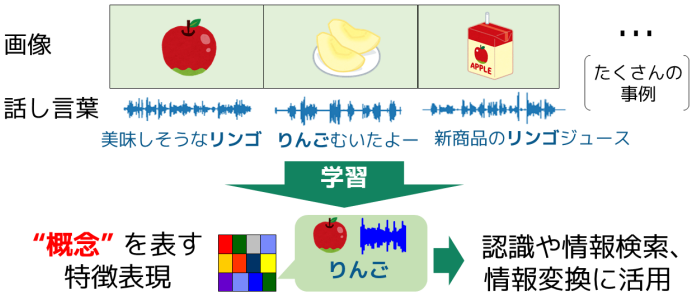
16 クロスモーダル学習による動作概念の獲得

TVを視聴するだけで賢くなるAI

どんな研究	大量データの準備には、手間がかかったり、データの入手自体が難しかったり、クラスラベルの付け方を事前に設計することが難しかったりします。この展示では、TV放送のようなメディアデータだけから、モノやコトの概念を自動獲得するAIを、より高度な認識や検索に活用する研究を紹介します。
どこが凄い	教師ラベルなしで映像における動作とそれを説明する話し言葉を時空間で対応付け、概念に相当する特徴表現を獲得する技術を考案しました。スポーツ実況の映像と音声データから、競技者の動作と実況の話し言葉の対応付けによる概念検索を実現しました。
めざす未来	TVを視聴するだけで、AIが音と映像を対応付けながら、知らないモノやコトを自ら学び、賢くなる未来をめざしています。音や映像、言語といったメディアの種類を横断する超大規模アーカイブ検索や自動アノテーションなどへの応用を検討しています。

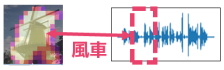
クロスモーダル概念獲得

異種のメディア情報の共起から、教師ラベルなしで相互の対応関係を発見し、メディア情報の種類に依存しない“概念”を獲得する



TVを視聴するだけで自ら賢くなるAIの実現

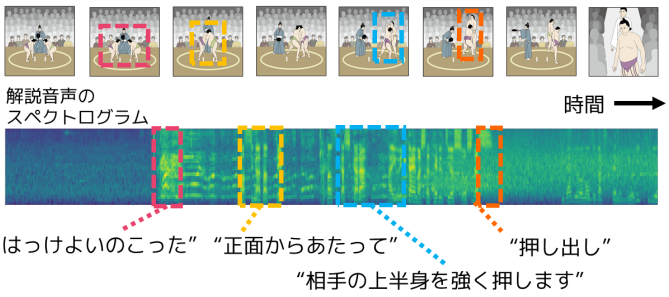
画像と話し言葉から、物体等に関する概念獲得・翻訳の実現可能性を示した[1][2]



動作概念の獲得の難しさ

映像における動作と解説の話し言葉の時間的なズレ

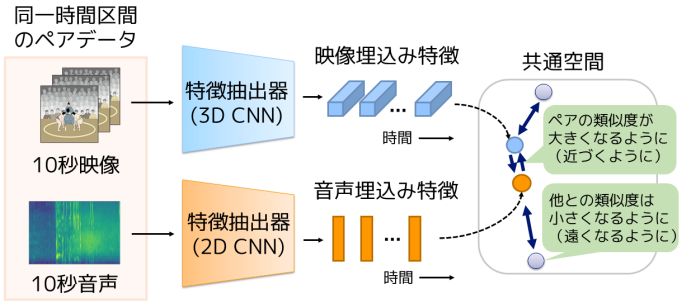
スポーツの実況映像（例：大相撲）



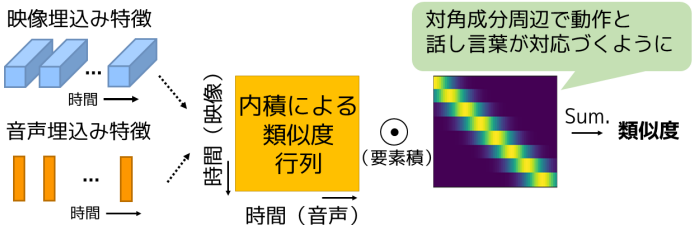
動作と話し言葉を時空間的に対応づけるために、時間的なズレを吸収するAttentionを考案した[3]

動作概念の学習法と結果

学習法：ペアの特徴の**類似度**が、他サンプルとの類似度よりも大きくなるよう特徴抽出器のパラメータを学習する

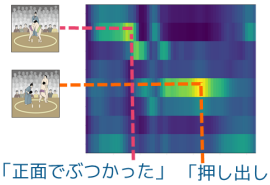


ポイント：類似度行列の対角成分周辺の重み付け (Guided Attention) により、動作と解説の時間的なズレを吸収する



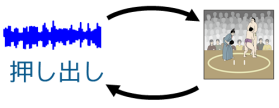
実験結果：動作と話し言葉の対応関係が抽出できること、”決まり手”に相当する概念検索が可能なことを確認した

類似度行列から分かる動作と話し言葉の対応関係



概念検索

音声から映像の検索精度
0.08 → 0.36 に向上



映像から音声の検索精度
0.04 → 0.29 に向上

※実験条件：
・NHKの大相撲中継から高頻度の9つの決まり手で勝敗が決まった1,128クリップ（各10秒）で学習
・決まり手毎に各10（計90）の未知クリップでテスト、評価尺度はRecall@1を利用

関連文献

[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*.

連絡先

大石 康智 （Yasunori Ohishi） メディア情報研究部 メディア認識研究グループ
Email: cs-openhouse-ml@hco.ntt.co.jp