視覚言語モデルによる行動認識のための知識拡張テキストプロンプトチューニング

捧蓮　曲佳　宮本健　三輪祥太郎
三菱電機株式会社
sasage.ren@dr.mitsubishielectric.co.jp　kyoku.ka@dc.mitsubishielectric.co.jp
{miyamoto.ken, miwa.shotaro}@bc.mitsubishielectric.co.jp

概要

行動認識は映像から人の行動を分類する技術であり、製造現場の作業可視化や技術習得支援などへの応用が期待されている。近年の CLIP ベースの視覚言語モデルは、行動名のみをテキスト入力として多様な行動を軽量に認識できる一方、テキスト側に対象物や姿勢などの視覚的手掛かりが明示されず、意味の近い行動名（クラス）間で誤認識しやすい。本研究では、各行動に対して大規模言語モデルが自動生成したオブジェクト名や姿勢などの視覚語彙で行動ラベルを拡張する Knowledge-Augmented Text Prompt Tuning を提案する。ベンチマークデータセットでの few-shot 評価において、軽量かつ多様な行動の認識能力を維持しつつ、意味の近い行動名（クラス）間での誤認識を効果的に低減できることを示す。

1　はじめに

行動認識は、映像から人が行っている行動をクラス分類する技術であり、製造現場の作業可視化や技能伝承、教育支援、監視などへの応用が期待されている。深層学習 [1, 2, 3, 4, 5] と大規模データセット [6, 7] により高精度な行動認識モデルが提案されてきたが、アノテーションコストやドメインギャップの問題から、既存の大規模画像／動画モデルを利用した転移学習が重要となっている。

近年、CLIP [8] に代表される視覚言語モデル（VLM）を用いた few-shot 行動認識 [9] が注目されている。行動ラベルをテキストエンコーダに入力し、その埋め込みと動画特徴の類似度に基づいて分類することで、クラス数に対するパラメータ増加がなく推論も軽量である。しかし一般的な行動認識データセット [6, 7] では、多くの行動名が *wave*, *shake hands*, *cartwheel*, *somersault* のような動詞（句）で与えられており、対象物や接触様式、身体姿勢などの視覚的決定要因が明示されない。CLIP のテキストエンコーダは Web テキストから学習されている [8] ため、動詞と共起する物体概念を暗黙的に獲得していると考えられるが、先に挙げた意味的に近い行動名（クラス）はテキスト空間で近接しやすい。その結果、本来は接触の有無や姿勢、運動方向、回転軸といった視覚的差異で区別されるべきクラスで誤認識が多く、特に物体依存的な行動や抽象的な行動の識別が難しい。この不足を物体検出器や VLM の併用で補うことも可能だが、推論コスト増大や運用の複雑化を招く。

本研究ではこの課題に対し、CLIP ベース手法の「多様な行動を軽量に認識できる」利点を維持しつつ、テキスト表現に外部知識を注入する Knowledge-Augmented Text Prompt Tuning (KATP) を提案する。まず、大規模言語モデル（LLM）により各行動ラベルに対し、オブジェクト、身体姿勢、運動パターン、シーン要素などフレームから直接観測可能でクラス固有性の高い「視覚語彙」をオフライン生成し、「行動ラベル＋視覚語彙」から成る短いテキストプロンプトを作成する。次に、このプロンプトを CLIP のテキストエンコーダに入力してプロンプトチューニングを行い、各行動と相互作用する対象物や姿勢・動き方を内包したテキスト埋め込みを得る。これにより、従来テキスト側で区別しにくかった意味の近い行動名（クラス）間に視覚的決定要因に基づく分離を導入し、誤認識を効果的に減らす。

本研究の貢献は、以下のとおりである。

- CLIP ベースの few-shot 行動認識において、意味の近い行動名（クラス）間の誤認識を抑えるため、行動ラベルに視覚語彙を付与する KATP を提案する。

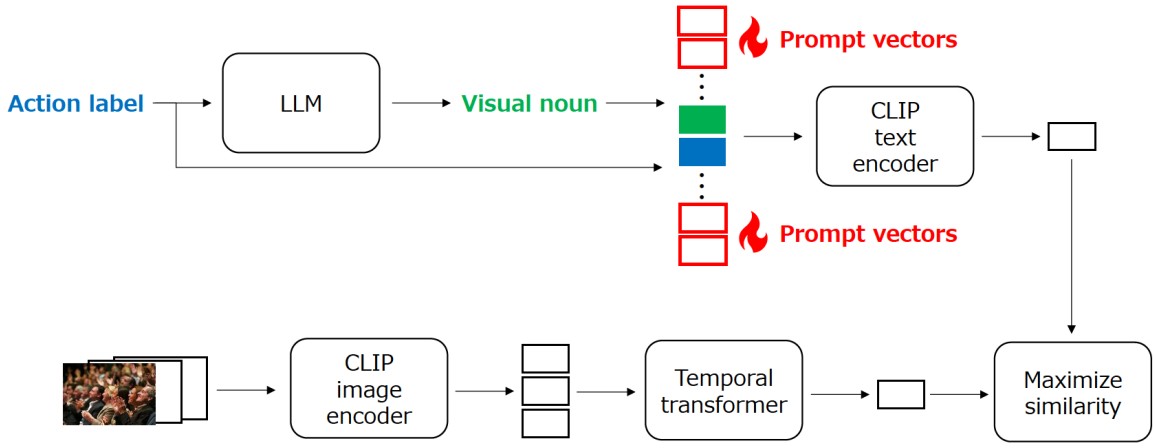

図 1 Knowledge-Augmented Text Prompt Tuning (KATP) の概要

- LLM が生成する候補語を視覚的に観測可能でクラス固有性の高い視覚語彙に限定して行動ラベルを拡張する語彙拡張手法を設計し、意味語彙・混合語彙との比較により、この設計がテキスト表現のクラス分離性を高めることを示す。
- HMDB-51[6] および Kinetics-400[7] を用いた few-shot 行動認識実験により、提案手法が既存手法を上回る性能を達成することを確認する。

2 関連研究

2.1 行動認識

従来の行動認識は教師あり学習に基づく動画分類モデルを中心に発展してきた。2D CNN＋時系列プーリング [10]、two-stream CNN[2]、3D CNN（C3D[1], I3D[3]）、Transformer ベースの TimeSformer[4] や Video Swin Transformer[5] などが高精度な行動認識を実現している。しかし動画ラベリングのコストや実環境とのドメインギャップにより、汎化性能に限界がある。

この問題に対し、少数サンプルで新規クラスを認識する few-shot 行動認識が提案され [11, 12]、メタラーニング [13] や距離学習 [14, 15] に基づく手法が検討されてきたが、多くは動画専用ネットワークを一から学習する必要があり、大規模事前学習モデルの活用という観点では制約が大きい。

2.2 Vision-Language Model を用いた動画理解

CLIP に代表される VLM は、画像とテキストのペアに対する対照学習により、両者を共通の埋め込み空間にマッピングし、クラス名を含むテキストプロンプトからの zero-/few-shot 分類を可能にする [8]。動画への拡張では、各フレームを CLIP で特徴抽出し、時間方向のプーリングや Transformer で動画表現を得る手法が提案されている [16, 17]。

画像分類では、テンプレート文の一部を連続ベクトルとして学習する CoOp [18, 19] が、テキスト側のみの微調整で精度向上を実現した。動画理解では Efficient-Prompt[9] が、CLIP のテキストプロンプトに学習可能なコンテキストトークンを付与し、軽量な時系列 Transformer とともに学習する枠組みを示したが、抽象的な行動名（クラス）では細かな視覚差をテキスト空間に明示的に埋め込めず、類似クラス間で誤認識が生じやすい。本研究はこの点に対し、LLM によって生成された視覚的に観測可能な視覚語彙による行動ラベルの拡張を提案する。

3 手法

本手法 (KATP) の概要を図 1 に示す。KATP は、CLIP に基づく Efficient-Prompt [9] を土台とし、(1) LLM によって各行動ラベルに関連する視覚語彙をオフライン生成し、「行動ラベル＋視覚語彙」から成る Knowledge-Augmented テキストプロンプトを構築し、(2) 拡張されたプロンプトを用いて、CLIP エンコーダを凍結したままテキスト側コンテキストトークンと時系列 Transformer を few-shot 学習する。

3.1 LLM による語彙生成

表 1 に示すように、行動ラベルのみを用いる Efficient-Prompt では、上半身ジェスチャ、姿勢・移動、体操・回転といった意味の近い行動クラス群で誤りが集中している。これらは本来、接触の有無や身体姿勢、運動方向、回転軸といった視覚的決定要因が、テキスト側に明示されていないためである。

表 1　Efficient-Prompt における誤認識の傾向（$a \to b$ は、真のクラス a がクラス b に高確率で誤分類されたことを示す。）

カテゴリ	誤認識例		
上半身ジェスチャ	wave	→	shake hands
	clap	→	push
姿勢・移動	jump	→	climb stairs
	walk	→	run
体操・回転	cartwheel	→	somersault
	flic flac	→	cartwheel

この問題に対し、本研究では LLM への指示段階で生成語彙を (1) 視覚的に観測可能な **視覚語彙** と (2) 高レベル意味カテゴリを表す **意味語彙** に分けて設計する。その概要を以下に示す。

- **視覚語彙**：1 フレームまたは短いクリップから直接観測できる物体、体の姿勢や手足の位置・向き、物や相手との接触の仕方、動きの方向・振り方のみを許可し、「非接触ジェスチャ vs 接触ジェスチャ」「回転 vs 並進移動」など、視覚的に類似した行動クラス間の微妙なモーション差・ポーズ差を補完することを狙う。
- **意味語彙**：コミュニケーション意図、社会的機能、イベント種別、相互作用様式といった抽象的な名詞句のみを許可し、「体操 vs 日常動作」「社会的ジェスチャー vs 移動動作」など、意味的に異なる行動クラスを離すことを狙う。ただし、意味語彙は複数の行動クラスで共有されやすく、テキスト空間のクラス間距離を縮めてしまう可能性もある。

提案手法 KATP では、初めに各行動ラベルを LLM に入力し、視覚語彙に限定して名詞句を生成させる。次に、得られた候補から視覚的で行動クラス固有性の高いもののみを選び、行動ラベル名と連結して Knowledge-Augmented テキストプロンプト s_c を構成する（表 2）。この処理を全クラスに対してオフラインで行い、接触様式や姿勢、運動パターンなどを含むテキスト表現を得る。

3.2　プロンプトチューニング

プロンプトチューニングでは、CLIP の画像・テキストエンコーダを凍結しつつ、動画側の時系列 Transformer とテキスト側のコンテキストトークンのみを学習する。

動画側では、各動画からサンプリングしたフレームを CLIP の画像エンコーダに入力し、得られたフレーム特徴列を軽量な時系列 Transformer に通して

表 2　Knowledge-Augmented テキストプロンプトの例

行動ラベル	拡張された行動ラベル
wave	wave with palm facing outward, side-to-side hand motion, wrist oscillation
shake hands	shake hands, thumb wrap and vertical hand pump in frontal stance
cartwheel	cartwheel with wheel-like sideways rotation, lateral handstand phase, two-hand plant
somersault	somersault, head over heel rotation, midair tuck posture

動画埋め込みベクトルを得る。テキスト側では、Knowledge-Augmented テキストプロンプト s_c をトークン化し、k 個の学習可能なコンテキストトークンを挿入したシーケンスをテキストエンコーダ ϕ_{text} に入力することで、各クラス c に対するテキスト埋め込み $\mathbf{c}_c$ を得る：

$$\mathbf{c}_c = \phi_{\text{text}}\big(a_1, \ldots, \text{TOKENIZER}(s_c), \ldots, a_k\big), \quad (1)$$

ここで $a_i \in \mathbb{R}^D$ は i 番目の学習可能なプロンプトベクトル（コンテキストトークン）を表し、D はベクトル次元である。$\{a_i\}_{i=1}^{k}$ は全クラスで共有される。得られたテキスト埋め込みと動画埋め込みのコサイン類似度に基づき、クロスエントロピー損失で時系列 Transformer とコンテキストトークンを更新する。推論時には学習時と同じ s_c を用いてクラス埋め込みを事前計算し、各テスト動画との類似度により行動名（クラス）を予測する。

4　評価実験

4.1　データセット

評価には HMDB-51[6] と Kinetics-400（K-400）[7] の 2 つのデータセットを用いた。HMDB-51 は 51 クラスの日常的な行動を収めたデータセットで、ジェスチャや姿勢変化などのクラスを含む。K-400 は 400 クラスからなる大規模データセットで、物体依存クラスや抽象ラベルクラスが多数含まれる。いずれも公式の train/test 分割に従い、train split から各クラス 5 本のみを使用する few-shot 設定を構成した。

4.2　評価方法

5-shot-5-way と 5-shot-C-way の 2 設定で評価した。5-way 設定では、各試行でまず学習データから 5 クラスをランダムサンプリングし、それぞれのクラスから 5 本の動画を選んで学習用とし、同じ 5 クラスに属する残りの動画を評価用とした。この構成をランダムに 200 回繰り返し、top-1 accuracy の平均を求

表 3 HMDB-51 における 5-shot-C-way 設定での語彙種類別の行動認識性能（top-1 accuracy [%]）。視覚語彙を用いた場合に最大の改善が得られる。

Method	K-shot	N-way	HMDB-51
Efficient-Prompt	5	C_{ALL}	55.1
KATP (+visual noun)	5	C_{ALL}	**56.3**
KATP (+semantic noun)	5	C_{ALL}	55.5
KATP (+mixed noun)	5	C_{ALL}	55.4

めた。C-way 設定では、より難度の高いシナリオとして、データセットに含まれる全クラス数 C を同時に分類対象とした。train split から各クラス 5 本の動画をランダムにサンプリングして学習用とし、対応する test split 上で性能を測定した。これをランダムに 10 回繰り返し、top-1 accuracy の平均を求めた。

4.3 実験設定

画像・テキストエンコーダには事前学習済みの CLIP (ViT-B/16) を用いた。各動画から 16 フレームをサンプリングし、5-crop prediction により最終スコアを得た。時系列 Transformer とテキスト側コンテキストトークンのみを学習し、最適化には AdamW を用いた。語彙生成を行う LLM には GPT-5 を用いた。

5 実験結果

5.1 生成語彙の比較検討（視覚語彙 vs 意味語彙 vs 混合語彙）

KATP では、LLM が生成する名詞句の性質（視覚語彙や意味語彙）が、テキスト表現のクラス分離性を左右する重要な設計要素となる。3.1 節で述べたように、視覚語彙はフレームから視覚的に直接観測可能な要素に限定し、意味語彙はより抽象的な意味カテゴリを表す。

ここでは、それぞれの役割と効果を確認するため、生成語彙の種類別の行動認識性能を比較した。表 3 に示すように、視覚語彙で拡張した場合が 56.3% と最も高く、意味語彙あるいは混合語彙では 55% 台にとどまる。これは、抽象的・意味的な名詞句を用いた場合、多数の行動クラスで語彙が共通化し、クラス内よりもクラス間で共有されやすいためと解釈できる。一方、視覚語彙に限定すると、接触の有無や手足の向き、運動パターンといった視覚的差分がクラスごとに明確に埋め込まれ、CLIP のテキスト空間における細粒度なクラス分離が促進される。以上より、視覚語彙は、意味語彙や混合語彙より、本設定での誤認識低減に適した設計である。

表 4 HMDB-51 および Kinetics-400 における 5-shot 行動認識性能（top-1 accuracy [%]）。提案手法 KATP は 既存手法 Efficient-Prompt を一貫して上回る。

Method	K-shot	N-way	HMDB-51	K-400
Efficient-Prompt	5	5	82.5	94.8
Efficient-Prompt	5	C_{ALL}	55.1	55.1
KATP (Ours)	5	5	**82.7**	**95.1**
KATP (Ours)	5	C_{ALL}	**56.3**	**57.0**

表 5 HMDB-51 における代表クラスの誤認識率（5-shot-C-way 設定、top-1 accuracy [%]）。KATP (+visual noun) は Efficient-Prompt と比較して、ジェスチャ系および体操系クラスで誤認識率を大きく低減している。

Action class	Efficient-Prompt	KATP
wave	89.3	74.7
cartwheel	86.7	76.0

5.2 提案手法の改善効果

表 4 に Efficient-Prompt と KATP (+visual noun) の比較を示す。いずれの設定においても一貫した改善が得られた。特に、C-way 設定では HMDB-51 で +1.2 %、K-400 で +1.9 %と、明確な向上が確認された。

精度改善は、HMDB-51 のジェスチャ系・姿勢／移動系・体操系や、K-400 の物体依存クラス・抽象的な行動名をもつクラスに集中している。代表例として、HMDB-51 におけるクラス別の誤認識率の変化を表 5 に示す。同表のとおり、ジェスチャ系の *wave*、体操系の *cartwheel* いずれにおいても誤認識率が大きく改善しており、近縁クラスとの混同が軽減されている。表 2 のように、KATP では、*wave*/*shake hands* に対して「掌の向き」「接触の有無」、*cartwheel*/*somersault* に対して「横方向回転」「縦方向回転」といった視覚的な違いを強調する名詞句が付与されている。その結果、行動名だけでは近接していたテキスト埋め込み間の距離が広がり、細粒度なクラス分離が促進されることで誤認識が低減したと解釈できる。

6 おわりに

本論文では、CLIP ベースの few-shot 行動認識において、行動ラベルのみでは視覚的決定要因が表現できず、意味の近い行動名（クラス）間で誤認識が生じやすい問題に対し、LLM によって視覚語彙を生成し、テキストプロンプトを拡張する Knowledge-Augmented Text Prompt Tuning を提案した。ベンチマークデータセットでの few-shot 評価により、既存手法を上回る性能を達成した。

参考文献

[1] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In **Proceedings of the IEEE international conference on computer vision**, pages 4489–4497, 2015.

[2] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. **Advances in neural information processing systems**, 27, 2014.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**, pages 6299–6308, 2017.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In **Proceedings of the International Conference on Machine Learning**, 2021.

[5] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pages 3202–3211, 2022.

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In **Proceedings of the IEEE International Conference on Computer Vision**, pages 2556–2563, 2011.

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. **arXiv preprint arXiv:1705.06950**, 2017.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **Proceedings of the International Conference on Machine Learning**, pages 8748–8763, 2021.

[9] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In **Proceedings of the European Conference on Computer Vision**, pages 105–124, 2022.

[10] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In **Proceedings of the European Conference on Computer Vision**, pages 20–36, 2016.

[11] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In **Proceedings of the European Conference on Computer Vision**, pages 782–797, 2018.

[12] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pages 475–484, 2021.

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In **Proceedings of the International Conference on Machine Learning**, pages 1126–1135, 2017.

[14] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. **Advances in neural information processing systems**, 30, 2017.

[15] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. **Advances in neural information processing systems**, 29, 2016.

[16] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. **arXiv preprint arXiv:2104.08860**, 2021.

[17] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In **Proceedings of the 30th ACM international conference on multimedia**, pages 638–647, 2022.

[18] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. **International Journal of Computer Vision**, 130(9):2337–2348, 2022.

[19] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pages 16816–16825, 2022.