TextGuide: 説明文に基づくゼロショット長期間行動解析システム

筒川 和樹[1,a)]　佐藤 文彬[1,b)]　八馬 遼[1,c)]　関井 大気[1,d)]

概要

数十分以上におよぶ長期間の映像から人物の行動を解析するシステムは，応用時に認識対象行動や解析内容を変更する際，実装が複雑であればあるほど高い開発コストがかかる．この問題を解決するため，本研究では，長期間の動画像を入力として，ユーザーが文章で指示した行動解析タスクを，事前学習を除く DNN の学習なしに解く新たなタスク汎用化の問題に取り組む．具体的には，TextGuide 機構と呼ぶシステムを提案する．TextGuide 機構は，はじめに，ユーザーが文章入力で認識対象行動を指示するゼロショット行動認識手法を入力の動画像に適用する．次に，得られた行動認識結果を認識結果文として文章化し，ユーザーが入力した解析内容の説明文とともに大規模言語モデルで解析する．これにより，認識対象行動を汎用化しつつ，各ステップで用いる DNN を長期間の動画像なしに既存のデータセットで個別に学習することが可能となり，システム全体を多数の行動解析タスクで学習することを避ける．実験では，暴力行動認識の公開データセットである RWF-2000 を用いて，3 種類の行動解析タスクを対象として，先述の課題に対する提案法の有効性を検証した．

1. はじめに

Deep Learning 技術の勃興以降，カメラ映像から人物の行動を認識する技術の性能が飛躍的に向上 [6, 10, 16, 30] し，防犯カメラシステム [17] や店頭の購買行動解析 [14] といった多くの応用が実現されている．暴力行動 [31] や万引き [20] など短時間の間に取られる特定の行動を検出する場合とは異なり，統計的分析や映像検索など長期間に渡る行動解析結果を映像記録から得るには，より複雑なシステムが求められ，応用に合わせて高い開発コストがかかる．本稿では，数十分におよぶ比較的長期間の動画像を入力として，Deep Neural Networks（DNN）の学習時には未定義であった行動解析タスクをユーザーが文章入力で指示することにより解く高次なシステムの実現を目指す．

[1] コニカミノルタ株式会社
[a)] kazuki.tsutsukawa@konicaminolta.com
[b)] fumiaki.sato1@konicaminolta.com
[c)] ryo.hachiuma@konicaminolta.com
[d)] taiki.sekii@konicaminolta.com

1.1 関連研究と課題

長期間の動画像を DNN の入力としてあつかう長期間行動認識 [34, 38] に加え，意味索引付け [35]，質問応答 [36] といった行動認識に関する従来研究では，各シーンの平均動画長は数十秒〜数分程度であり，時間的に局所性のある情報と大域的な情報の両方を用いる [21]．近年では，時間的受容野がより広い注意機構 [15, 33] を DNN 構造として採用する手法 [2] が提案されている．しかしながら，デスクトップ PC など一般的な計算資源上で実行する場合，一次記憶に保持できる特徴量の量に制限があるため，数十分の動画像を DNN の入力としてあつかうことが難しい．そのため，長期間の動画像に対して部分的に DNN 処理（以下，認識ステップ）を適用し，結果を一次記憶以外の装置に保存した上で，それを入力として所望の解析結果を得る処理（以下，解析ステップ）をおこなう段階的システムが必要となる．

このような段階的システムとして，部分的に得られる認識ステップの結果をデータベースに保存し参照する行動認識手法 [34] が提案されている．一方，関連する研究として，映像要約技術 [22] は，認識ステップの結果を解析ステップでキーフレームの集合や合成フレームなどの形式に変換しユーザーに掲示する．映像検索技術 [1] は，インターネット上の大量の映像を認識ステップで処理し，検索クエリと各映像の意味の概念を解析ステップで関連付ける．現在では，スポーツ映像解析システム [9] など認識ステップの結果を解析ステップで統計的に分析する製品が提供されている．このような従来研究では，新たな応用に際して認識対象行動や解析ステップの変更が発生した場合，段階的システムが複雑であればあるほど高い開発コストがかかる．

1.2 提案法の概要

このような課題を解決することを目的として，ユーザーが文章で指示した人物行動を解析するタスクを，事前学習を除く DNN の学習なしに（ゼロショットで）解くシス

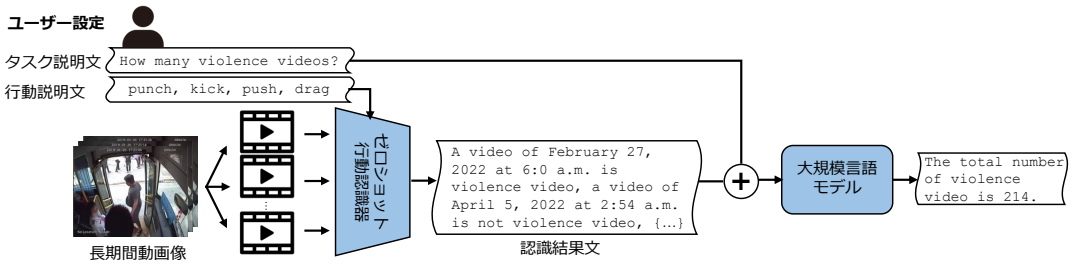

図 1: TextGuide 機構のシステム構成.

表 1: タスクごとの認識結果文（Recognition results），タスク説明文（Task description），LLM 入力文.

タスク	認識結果文	タスク説明文	LLM 入力文	
			回答方法を説明	回答例を例示
行動認識	A video of April 15, 2022 at 2:0 a.m. contains kick, push.	I define violence as any behavior that involves punch, kick, push, or drag. According to a video description, does the video contain violence?	{Task description} Answer 'A video of [Date and Time in the same format with the video description] is a violence video.' or 'A video of [Date and Time in the same format with the video description] is not a violence video.'. Video description: {Recognition results}	{Task description} {Recognition results} =>A video of February 27, 2022 at 6:0 a.m. is a violence video. {Recognition results} =>A video of April 5, 2022 at 2:54 a.m. is not a violence video. {Recognition results} =>
行動カウント	A video of February 27, 2022 at 6:0 a.m.is violence{...},a video of July 1, 2022 at 6:11 a.m. is not violence video.	How many violence videos?	{Recognition results} {Task description}Answer only with arabic numeral.	{Task description} {Recognition results} =>4 {Recognition results} =>
行動日時抽出		Extract the date and time the violence occurred.	{Task description} {Recognition results}	✗

テムを提案する．推論時にユーザーの意図に沿った任意のタスクを解く問題はタスク汎用化 [23] と呼ばれ，自然言語処理技術の発達にともない近年盛んに研究されている [11,27,37]．提案法は，新たな問題設定として数十分に渡る長期間の動画像を入力としたタスク汎用化に取り組む．

これを実現するため，本研究では，従来研究と同様の段階的システムを採用するとともに，TextGuide（Text-Guided Task-Generalizer）と呼ぶ，タスク汎用化のための新たなシステムを提案する．具体的には，認識ステップで得られる行動クラスが事前に定義されていると解析ステップで処理できるタスクが限定されるため，TextGuide 機構では，ユーザーが文章入力により認識対象行動を制御できるゼロショット行動認識手法を認識ステップに導入する．また，タスク汎用化システム全体を DNN によりモデル化する場合，多数のタスクを学習する必要がある．しかしながら，長期間の動画像を入力として用いる提案法がタスクの変化に対して汎化可能な量の学習データを準備するには，タスク汎用化の従来研究を超える高い人的コストが必要となる [11]．この問題に対して，認識・解析ステップをそれぞれ既存のデータセットのみで個別に学習できるよう，認識ステップの行動認識結果を文章化して解析ステップに入力するインターフェイスを導入する．解析ステップには大規模言語モデル（Large Language Model，LLM）を採用することで，ユーザーから文章を介して与えられた解析タスクを解く．

実験では，暴力行動認識の公開データセットである RWF-2000 [8] を用いて，文章表現に曖昧性のある暴力行動を対象として提案法を行動認識の State-of-the-Art 技術と比較することで，短期間の基本的な行動に対する提案法の有効性を確認した．次に，この結果を長時間に渡る解析対象として用いて，統計的分析とキーワード抽出を対象タスクとして，先述の課題に対する提案法の有効性を検証した．

以上を踏まえ，本研究の貢献は，(1) 認識ステップにゼロショット行動認識手法を導入することで対応可能な行動解析タスクを拡大し，(2) 認識ステップから解析ステップへの入力を文章化して既存のデータセットで学習可能なシステム構成を実現することにより，(3) 長期間の動画像を入力とした新たなタスク汎用化の問題を解くことである．

2. 提案法

2.1 システム構成

提案する TextGuide 機構のシステム構成を図 1 に示す．ユーザーは，解析したいタスクに必要な最小単位の行動・動作（以下，基本行動）を表す文章（以下，行動説明文）と，行動解析のタスクを表す文章（以下，タスク説明文）を入力する．行動説明文は，認識ステップにおいて 2.2 節で述べるゼロショット行動認識手法に入力され，行動説明文に合った行動が検知される．タスク説明文は，この行動認識結果を 2.3 節で述べるルールで文章化した内容（以下，認識結果文）とともに一文（以下，LLM 入力文）に統合され LLM に入力される．そのため，解析ステップには既存の LLM を特別な変更なく導入できる．結果として，認識・解析ステップ間の誤差逆伝搬なしに各処理で独立した学習が可能であり，複数の行動解析タスクを学習することが不要となる．

2.2 ゼロショット行動認識による基本行動の認識

本節では，はじめに提案法が採用するゼロショット行動認識手法 [28]（Zero-shot Anomaly Action Recognition，ZAAR）を概説し，続いて提案法による ZAAR の活用方法について説明する．

ZAAR は，入力の動画像に骨格検出 [29] を適用して得られる複数人物の骨格検出結果を入力として，視覚言語事前学習手法 CLIP [25] を応用した MotionCLIP [32] を拡張した手法である．具体的には，ZAAR は Permutation-Invariant な性質を持つ，点群深層学習に基づく DNN 構造 [7] を用いることで，3 次元の時空間上の関節点群から動画像全体を表現する特徴量（動画特徴量）を抽出する．事前学習では，これを比較的大規模な特定行動分類のデータセット（Kinetics-400 [6] など）を用いて，各動画サンプルから得られる動画特徴量を，対応する行動説明文を変換した特徴量（文章特徴量）と比較する対照学習 [25] をおこなう．これにより，動画特徴量は文章特徴量と共通の特徴空間で表現され，推論時には，動画特徴量とユーザーが入力した行動説明文の文章特徴量との距離から，認識対象行動が取られているかどうかを表すスコア [*1]（以下，行動スコア）を求めることができる．

提案する TextGuide 機構では，長期間に渡る動画像が DNN で処理できる長さに分割され，連続的に認識処理に入力されるものとする．ユーザーは解析処理に必要な基本行動の説明文を複数入力し，推論時には，分割された各動画は入力された基本行動のうち行動スコアが最大となるクラスに分類される．これにより，ユーザーの指示に基づき基本行動に応じた追加の DNN 学習をおこなうことなく基本行動をゼロショット（Text-Guided な設定）で認識し，解析ステップの入力情報として利用することができる．

2.3 大規模言語モデルを用いた行動解析

図 1 に示すように，解析ステップは，前節で述べた行動認識結果を文章化した認識結果文と，ユーザーから与えられたタスク説明文を統合し LLM に入力することで，指示されたタスクを特別な DNN の学習なしに実行し，結果を文章形式で出力する．認識結果文は，前節で述べた分割した動画像ごとに得られる行動認識結果を統合・変換することで得られる文章であり，長期間の動画像全体を表現した疎な情報として利用される．

2.3.1 大規模言語モデルの概要

LLM [3, 4, 26] は，近年その優れた汎用性能により，ゼロショットの自然言語処理タスクにおいて高い性能を示している．モデルパラメータと学習データの容量が大きいモデルの開発が盛んにおこなわれ，GPT-3 [4] では 300 億のトークン，1,750 億のパラメータが使用された．また，LLM のゼロショット推論精度を向上させる手法として指示チューニング [27] が提案されており，自然言語処理タスクの入出力の例が多数文章形式で定義された学習サンプルを用いて，LLM が更新される．本稿では指示チューニングにより学習された LLM（text-davinci-003 [24] など）を解析処理に用いる．

2.3.2 大規模言語モデルの入出力形式

各動画像の行動認識結果は，認識された基本行動のクラスと撮影時刻で構成される．このような特定の形式に沿った情報を LLM に入力する場合，表形式 [19] など文章以外の形式も提案されている一方，本稿では，最も基本的な入力形式で性能を検証することを目的として，文章入力を採用する．行動認識結果を LLM に文章入力する Text-Guided な設定によって，認識・解析各ステップが解く問題はそれぞれ既存の行動認識・自然言語処理に限定される．したがって，各ステップの DNN を既存のデータセットを用いて独立に学習することができる．そのため，タスク汎用化の従来研究 [11, 23, 27] とは異なり，複数のタスクを対象としてそれぞれ多数の画像を準備し，システム全体を End-to-End で学習してタスクの変化に汎化させる必要がない．

3 種類の行動解析タスクにおける行動説明文，タスク説明文，LLM 入力文の例をそれぞれ表 1 に示す．LLM の出力文章として回答方法を説明する設定と，回答例を示す設定の 2 種類を提案する．以上に述べた文章形式によって LLM にタスクを指示することで，ユーザーの意図に応じた出力文章が生成される．

3. 評価実験

3.1 データセット

評価では共通して暴力行動分類データセット RWF-2000 [8] を用い，暴力行動を解析する 3 パターンの問題設定において提案法の有効性を評価する．RWF-2000 は暴力・非暴力 2 クラスの行動分類タスクを対象としており，評価データは行動ごとに 200 の動画像を含む計 400 の動画で構成される．各動画像は YouTube [*2]から収集・編集された 5 秒のクリップであり，400 の動画像の合計は 33.3 分となる．

認識ステップで用いる ZAAR では，比較的大規模な特定行動分類の Kinetics-400 [6] データセットを事前学習に用いた．事前学習と先述の RWF-2000 を用いた評価では，それぞれ動画像の各フレームに Pose Proposal Networks [29]（PPN）を適用して時系列の骨格情報を作成し，DNN の入力として用いた．PPN の DNN 構造には ResNet-101 [13] を採用し，骨格検出タスクの学習には MS-COCO データセット [18] を用いた．骨格の定義は OpenPose [5] と同様である．動画像はすべて 30 [FPS]・640×480 [px^2] に変換

[*1] 厳密には，ZAAR が算出する行動認識に関するスコアのうち，Prompt-Guided Action Score を指す．

[*2] youtube.com

表 2: 従来法と提案法の暴力行動分類精度の比較結果．提案法（Ours）では，回答方法として説明（w/ explanation）と例示（w/ sample）の 2 パターンを LLM に指示．

Method	Acc. (%)
PointNet++ [31]	78.2
DGCNN [31]	80.6
SPIL [31]	89.3
ZAAR w/ `violence`	78.0
ZAAR w/ `fighting`	81.5
Ours w/ explanation	84.5
Ours w/ sample	84.5

し用いた．

3.2 実装の詳細

前節で述べた ZAAR を高精度化するため，動画特徴量を抽出する DNN の構造に Structured Keypoint Pooling [12] を用いた．暴力行動を構成する基本行動（2.2 節参照）の行動説明文として，`punch`，`kick`，`push`，`drag` を入力した．行動認識結果それぞれに日時をランダムに付与し，認識結果文を作成した（表 1 参照）．基本行動を検出するしきい値は，次節で述べる分類精度を下に調整された．解析ステップで用いる LLM には，指示チューニングにより学習されたモデルである text-davinci-003 [24] を採用した．

3.3 暴力行動認識での評価実験

最初に，複数の基本行動で構成される暴力行動を対象として，提案法が認識結果文とタスク説明文を下に各動画像を暴力・非暴力行動に分類する短期間の行動解析をおこなえるかどうかを検証する．表 1 に示すように， 2.3.2 項で述べた 2 種類の設定で LLM 入力文を作成した．

暴力行動分類の従来法と提案法の分類精度を比較した結果を表 2 に示す．同表は，比較対象として，ZAAR に暴力行動を直接文章入力した場合の結果（w/ `violence`，w/ `fighting`）を含む．提案法は，LLM に対する 2 種類の回答方法の設定に依らず教師あり手法（PointNet++ [31]，DGCNN [31]）よりも高精度であり，State-of-the-Art 技術である SPIL と数 [%pt] の差に迫る頑健性を実現している．また，同表より，解析ステップなしに直接暴力行動を認識する ZAAR よりも提案法の方が高精度であることから，解析ステップが，文章表現に曖昧性のある暴力行動がいくつかの基本行動で構成されるという知識を活用していることがわかる．

3.4 長期間行動解析での評価実験

RWF-2000 データセットの評価データにおける 400 の動画像すべてを連結した計 33.3 分の映像を入力として用いて，提案法による長期間行動解析の精度を定量的に評価する．評価では，統計的分析タスクの一環として暴力行動カウントを，キーワード抽出タスクの一環として暴力行動発生日時抽出を，解析対象のタスクとする．表 1 に示した方法にしたがい，各タスクで LLM 入力文を作成した．LLM の 2 種類の回答方法のうち，回答方法を説明する設定がより高精度であったため，これを採用した．本実験で用いる LLM は，メモリ容量の制約上，長期間映像全体から得られる認識結果文を同時に処理できないため，認識結果文を分割して別々に解析ステップを適用した上でその結果を統合した．暴力行動カウントのタスクでは，一部の動画に対するカウント結果の総和を計算する質問応答を LLM で処理し，暴力行動の総数を得た．

表 3: 暴力行動カウント結果．GT は真値を表す．

Method	Action	# of Counts	Error Rate (%)
GT	GT	200	0.0
Ours w/ GT	GT	211	5.5
Ours	ZAAR	214	7.0

提案法による暴力行動カウントを定量的に評価した結果を表 3 に示す．真のカウント数である 200 に対して，提案法の出力は 214 であり，+7[%] のエラー率でカウントできている．また，評価に用いる行動認識結果を真値に設定した場合（Ours w/ GT）は，カウント数のエラー率は+5.5[%] に改善したため，本実験における認識ステップのカウント数への影響は限定的であることがわかる．なお，このときのカウント誤りは，解析ステップに用いた LLM の処理精度に起因しているため，今後の LLM 技術の進歩が期待される．一方，暴力行動発生日時抽出のタスクにおいて，抽出される日時を真の日時と比較した結果，再現率・適合率はそれぞれ 83.9[%]・81.4[%] であり，提案法は一定以上の検出精度でキーワード抽出が可能であることがわかる．以上の結果を踏まえると，提案法では，ユーザーが文章で長期間行動解析タスクを指示することにより，一定の問題設定においてタスク汎用化を実現できることがわかる．

4. まとめ

本研究では，長期間の動画像を入力として，ユーザーが文章で指示した行動解析タスクを事前学習を除く DNN の学習なしに解く問題に対して，TextGuide 機構と呼ぶ，タスク汎用化のための新たなシステムを提案した．TextGuide 機構では，ゼロショット行動認識手法と大規模言語モデルを文章を介して結合することにより，それぞれをユーザーが文章入力で制御できるタスク汎用化を実現するとともに，システム全体を多数の行動解析タスクで学習することを避ける．実験では，暴力行動認識の公開データセットである RWF-2000 を用いて，3 種類の行動解析タスクを対象として，先述の課題に対する提案法の有効性を検証した．

参考文献

[1] Awad, G., Curtis, K., Butt, A. A., Fiscus, J., Godil, A., Lee, Y., Delgado, A., Zhang, J., Godard, E., Chocot, B., Diduch, L., Liu, J., Graham, Y., and Qunot, G.: An overview on the evaluated video retrieval tasks at TRECVID 2022, *TRECVID* (2022).

[2] Beery, S., Wu, G., Rathod, V., Votel, R. and Huang, J.: Context R-CNN: Long Term Temporal Context for Per-Camera Object Detection, *CVPR* (2020).

[3] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., R, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K. and Liang, P.: On the Opportunities and Risks of Foundation Models, *arXiv:2108.07258* (2021).

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D.: Language Models are Few-Shot Learners, *NeurIPS* (2020).

[5] Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. and Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, *PAMI*, Vol. 43, No. 1, pp. 172–186 (2021).

[6] Carreira, J. and Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, *CVPR* (2017).

[7] Charles, R., Su, H., Kaichun, M. and Guibas, L. J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, *CVPR* (2017).

[8] Cheng, M., Cai, K. and Li, M.: RWF-2000: An Open Large Scale Video Database for Violence Detection, *ICPR* (2021).

[9] Eldridge, D., Pulling, C. and Robins, M.: Visual exploratory activity and resultant behavioural analysis of youth midfield soccer players, *Journal of Human Sport and Exercise*, Vol. 8, pp. S560–S577 (2013).

[10] Feichtenhofer, C., Fan, H., Malik, J. and He, K.: SlowFast Networks for Video Recognition, *ICCV* (2019).

[11] Gupta, T., Kamath, A., Kembhavi, A. and Hoiem, D.: Towards General Purpose Vision Systems: An End-to-End Task-Agnostic Vision-Language Architecture, *CVPR* (2022).

[12] Hachiuma, R., Sato, F. and Sekii, T.: Unified Keypoint-based Action Recognition Framework via Structured Keypoint Pooling, *CVPR* (2023).

[13] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *CVPR* (2016).

[14] Herath, S., Harandi, M. and Porikli, F.: Going Deeper into Action Recognition, *Image Vision Computing*, Vol. 60, No. C, pp. 4–21 (2017).

[15] Hu, J., Shen, L. and Sun, G.: Squeeze-and-Excitation Networks, *CVPR* (2018).

[16] Huang, Z., Wan, C., Probst, T. and Van Gool, L.: Deep Learning on Lie Groups for Skeleton-Based Action Recognition, *CVPR* (2017).

[17] Jin, C.-B., Do, T. D., Liu, M. and Kim, H.: Real-Time Action Detection in Video Surveillance using a Sub-Action Descriptor with Multi-Convolutional Neural Networks, *Journal of Institute of Control, Robotics and Systems*, Vol. 24, pp. 298–308 (2018).

[18] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft COCO: Common Objects in Context, *ECCV* (2014).

[19] Liu, Q., Chen, B., Guo, J., Ziyadi, M., Lin, Z., Chen, W. and Lou, J.-G.: TAPEX: Table Pre-training via Learning a Neural SQL Executor, *ICLR* (2022).

[20] Martnez-Mascorro, G. A., Abreu-Pederzini, J. R., Ortiz-Bayliss, J. C., Garcia-Collantes, A. and Terashima-Marn, H.: Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks, *Computation*, Vol. 9, No. 2 (2021).

[21] Miech, A., Laptev, I. and Sivic, J.: Learnable pooling with Context Gating for video classification, *CVPRW* (2017).

[22] Narasimhan, M., Nagrani, A., Sun, C., Rubinstein, M., Darrell, T., Rohrbach, A. and Schmid, C.: TL;DW? Summarizing Instructional Videos with Task Relevance and Cross-Modal Saliency, *ECCV* (2022).

[23] Oh, J., Singh, S., Lee, H. and Kohli, P.: Zero-Shot Task Generalization with Multi-Task Deep Reinforcement Learning, *ICML* (2017).

[24] OpenAI: Model index for researchers, `https://platform.openai.com/docs/model-index-for-researchers`.

[25] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, *ICML* (2021).

[26] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I.: Language models are unsupervised multitask learners, *OpenAI blog*, Vol. 1, No. 8, p. 9 (2019).

[27] Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M. T.-J., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J. A., Teehan, R., Scao, T. L., Biderman, S., Gao, L., Wolf, T. and Rush, A. M.: Multitask Prompted Training Enables Zero-Shot Task Generalization, *ICLR* (2022).

[28] Sato, F., Hachiuma, R. and Sekii, T.: Prompt-Guided Zero-Shot Anomaly Action Recognition using Pre-Trained Deep Skeleton Features, *CVPR* (2023).
[29] Sekii, T.: Pose Proposal Networks, *ECCV* (2018).
[30] Simonyan, K. and Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos, *NeurIPS* (2014).
[31] Su, Y., Lin, G., Zhu, J. and Wu, Q.: Human Interaction Learning on 3D Skeleton Point Clouds for Video Violence Recognition, *ECCV* (2020).
[32] Tevet, G., Gordon, B., Hertz, A., Bermano, A. H. and Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space, *ECCV* (2022).
[33] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *NeurIPS* (2017).
[34] Wu, C.-Y., Feichtenhofer, C., Fan, H., He, K., Krhenbhl, P. and Girshick, R.: Long-Term Feature Banks for Detailed Video Understanding, *CVPR* (2019).
[35] Yang, A., Nagrani, A., Seo, P. H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J. and Schmid, C.: Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning, *CVPR* (2023).
[36] Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y. and Tao, D.: ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering, *AAAI* (2019).
[37] Zhong, W., Gao, Y., Ding, N., Liu, Z., Zhou, M., Wang, J., Yin, J. and Duan, N.: Improving Task Generalization via Unified Schema Prompt, *arXiv:2208.03229* (2022).
[38] Zhou, J., Lin, K.-Y., Li, H. and Zheng, W.-S.: Graph-Based High-Order Relation Modeling for Long-Term Action Recognition, *CVPR* (2021).