

Video Transformer を用いた圧縮可能暗号化による プライバシー保護行動認識

Compressible Encryption-Based Privacy-Preserving Action Recognition Using
Video Transformer

Haiwei Lin[†]

今泉祥子[‡]

貴家仁志^{*}

[†] 千葉大学大学院情報・データサイエンス学部

[‡] 千葉大学大学院情報学研究院

^{*} 東京都立大学システムデザイン学部

Haiwei LIN[†]

Shoko IMAIZUMI[‡]

Hitoshi KIYA^{*}

[†] Graduate School of Informatics, Chiba University

[‡] Graduate School of Informatics, Chiba University

^{*} Faculty of System Design, Tokyo Metropolitan University

アブストラクト 本研究は、クラウド環境での映像の行動認識におけるプライバシー保護手法を提案する。近年、映像を暗号化したまま認識を実現する手法が注目されている。しかし、これらの暗号化された映像は従来の映像圧縮方式と互換性がなく、圧縮された場合、大幅な認識精度の劣化を生じさせるという課題がある。このことは、帯域幅に制約のある実環境での利用を困難にさせる。そこで本研究では、圧縮可能な暗号化映像を生成し、モデルの再学習を必要とせず高精度を実現する新しい手法を提案する。提案法により生成される暗号化映像は、Motion-JPEG や H.264 で効果的に圧縮可能であり、また提案法で暗号化されたパラメータをもつモデルは、暗号化映像に対して認識精度を一切低下させることなく維持できる。さらに圧縮率の高い条件下においても性能劣化を大幅に抑制できることを確認するとともに、暗号化映像の攻撃耐性についても評価した。

1 はじめに

行動認識は、映像または画像系列から人の行動を識別・分類することを目的とする、コンピュータビジョンにおける基本的なタスクの一つである。近年、深層学習の急速な発展により、この分野は大きく発展している。行動認識システムは、クラウドサーバ上に展開されることで、ローカルデバイスにかかる計算負荷を軽減している。しかし、映像データを外部サーバへ送信することは、プライバシー上の深刻な懸念を生じさせる。

これに対処するため、プライバシー保護行動認識 (privacy-preserving action recognition, PPAR) が活発

に研究されているいくつかの手法 [1]–[3] では、クエリ映像の品質を下げることにより、センシティブな情報を秘匿することができる。しかし、これらの手法は行動に関連する特徴も同時に失ってしまうため、認識性能の低下を引き起こす一方、暗号化に基づく手法 [4], [5] は、暗号化された映像を直接処理するため、わずかな精度低下のみで高精度を維持できる。さらに、近年の手法 [5] では、推論遅延や計算負荷を回避しながら、高精度な認識を可能にしている。

しかし、暗号化に基づく手法には重要な制約が生じている。暗号化された映像は一般に、国際標準の圧縮方式との互換性がない。これにより、ファイルサイズの効果的な削減が困難となるとともに、暗号化された映像が圧縮された場合、認識精度が著しく低下する。その結果、圧縮が不可欠な帯域制約の条件下において、PPAR の実用化は大きな課題となっている。

本研究では、圧縮可能な暗号化映像を生成することにより、圧縮の影響を抑制しながら高い精度で認識可能な PPAR 手法を提案する。提案法は、映像暗号化とモデル暗号化の二つの要素から構成される。映像暗号化では、圧縮可能な暗号化映像を生成する。一方、モデル暗号化では、事前学習済みモデルの特定のパラメータを暗号化し、再学習を行うことなく、圧縮可能暗号化映像における行動認識を可能にする。実験では、圧縮可能暗号化映像の認識精度について、暗号化されたパラメータをもつモデルが、プレーン映像を分類するプレーンモデルと同等の精度を達成することを示す。また、提案法により生成された圧縮可能暗号化映像は、Motion-JPEG および H.264

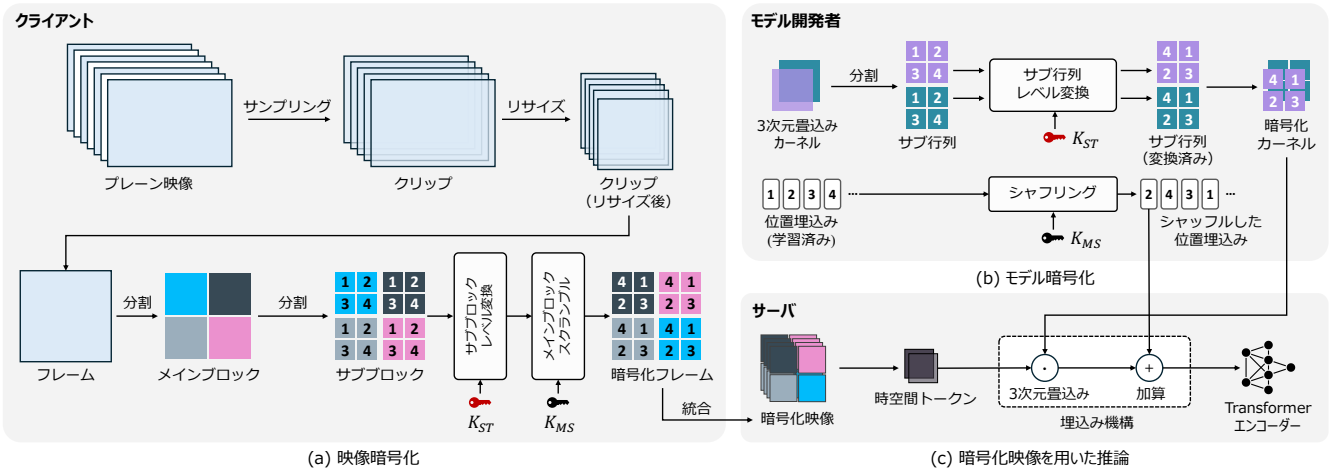


図 1: 提案法のフレームワーク

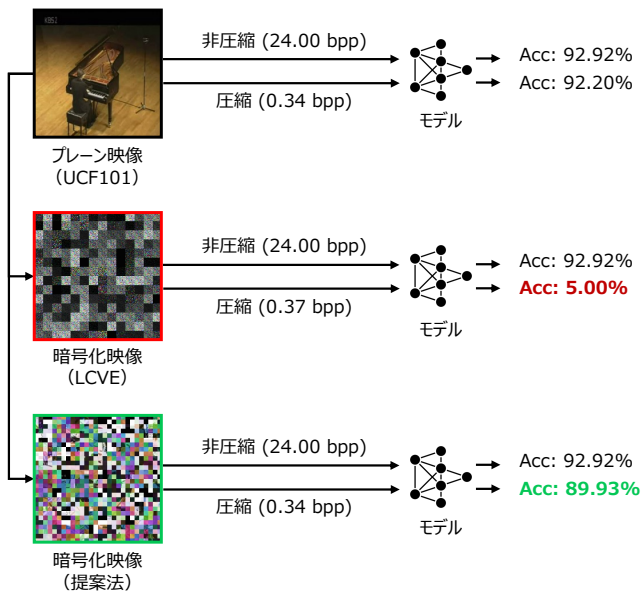


図 2: 提案法による圧縮後の暗号化映像に対する行動認識精度の向上

により効率的に圧縮可能であり、圧縮率が高い場合においても、提案法は性能低下を大幅に抑制することを確認した。さらに、提案法により生成された圧縮可能暗号化映像の攻撃耐性についても議論する。

2 準備

2.1 本研究の目的

本研究では、深層学習をサービスとして提供する (Machine Learning as a Service, MLaaS) 典型的なシナリオを想定する。このシナリオでは、クライアント、モデル開発者、信頼できないサーバ、および攻撃者の 4 者が互いに関係している。クライアントとモデル開発者は相互に

信頼され、安全な環境で動作する一方、サーバは攻撃の脅威にさらされる可能性がある。学習済みモデルはサーバに置かれており、クライアントからアップロードされるクエリ映像に対して行動認識を行う。しかし、これらの映像は伝送や保存の過程で攻撃者に漏洩する恐れがある。さらに、サーバは帯域幅を確保するためにアップロードされるファイルサイズを制限することが多く、映像は圧縮されることが想定される。このようなシナリオに対応するため、本研究は、クライアントが以下の条件を満たす暗号化映像をアップロードできるようにすることを目的としている。すなわち、(i) 知覚的情報を秘匿できること、(ii) 標準的な圧縮方式との互換性を維持できること、(iii) 圧縮後もモデルによる認識が可能であることである。

近年の最先端の暗号化手法として、Learnable Cube-based Video Encryption (LCVE) [5] が提案されており、これは推論遅延や計算的オーバーヘッドを生じさせることなく高精度な認識を実現している。本研究で提案するフレームワークを図 1 に示す。LCVE と同様、本フレームワークでは、固有の埋込み機構を有する Video Transformer (VT) モデルを基盤として採用する。本フレームワークにおいて、モデル開発者はまず秘密鍵を生成し、サーバとは独立した通信チャンネルを通じてクライアントに安全に共有する。また、これらの鍵を用いて、VT の埋込み段階におけるモデルパラメータを暗号化する。この過程はモデル暗号化と呼ばれ、VT の認識領域を変換することで、暗号化映像を追加学習なしで認識可能にするものである。各クライアントは、受け取った鍵を用いてプレーン映像を暗号化した後にサーバへアップロードする。

その結果、暗号化パラメータを導入した VT は、図 2 に示すように、暗号化映像に対して高精度な認識を達成する。さらに、映像が圧縮を受けた後においても、LCVE

と比較して暗号化映像の認識精度は高く維持される。

2.2 Video Transformer

本研究では, spatio-temporal attention [6]–[8] に基づく Video Transformer (VT) に着目する。VT は, トランスフォーマ構造を用いており, 映像処理に広く用いられているモデルである。このモデルでは, 入力映像をまず時空間トークンに分解し, これらを埋込みトークンに変換した後, transformer encoder に入力する。

トークンの構成には, 各映像フレームを $H_p \times W_p$ のサイズの非重複 2 次元パッチに分割する。連続する τ フレームについて, 空間的に同一位置にある 2 次元パッチを時間方向に積み重ね, 3 次元パッチ $\mathbf{x}_{t,(i,j)}$ を形成する。ここで, t は時間インデックス, (i, j) は 2 次元パッチの空間座標をそれぞれ表す。

各 3 次元パッチは時空間トークンとして扱われる。各トークンに対して学習可能な 3 次元畳込みカーネル E を適用し, 対応する埋込みトークンを生成する。さらに位置情報を付与するため, 各トークンに位置埋込み E_{pos} を加える。この 3 次元畳込みと位置埋込みの加算をあわせて埋込みステージと呼ぶ。ここで得られた埋込みトークンは, transformer encoder に入力され, 行動認識タスクにおけるクラスが出力される。

3 提案法

本稿で提案する手法は, 圧縮可能な暗号化映像を生成するとともに, これに適したモデル暗号化を設計することで, 映像圧縮が認識性能を低下させることを軽減する。本章では, まず, 圧縮可能な暗号化映像を生成するための映像暗号化について述べる。次に, VT モデルが再学習を行うことなく暗号化映像を正確に認識できるようにするためのモデル暗号化について説明する。

3.1 映像暗号化

従来の多くの映像暗号化手法は, 画素をランダム性の高いパターンにスクランブルすることで視覚情報を秘匿する。しかし, 映像圧縮は映像内の冗長性を利用してデータを削減するため, これらの手法は圧縮に適しておらず, 圧縮を適用した場合, 低い圧縮効率となることが多い。

本節では, 視覚情報を秘匿しながら, 映像圧縮に必要な冗長性を保持することを目的とした Encryption-then-Compression (EtC) 法 [9] に基づく新しい映像暗号化手法を提案する。概要を図 1(a) に示す。暗号化はフレーム単位で行い, 各フレームに同一の変換処理を適用する。前処理として, プレーン映像はサンプリングおよびリサイズされ, VT の入力次元に合致させる。各フレームは $H_p \times W_p$ 画素のメインブロック (main block, 以降 MB) と呼ばれ

るグリッドに分割され, これらを 2 次元パッチに対応させる。したがって, 3 次元パッチの時空間トークンについても, 時間軸方向に積み重ねられた MB の集合とみなすことができる。

各 MB は, さらに $H_s \times W_s$ 画素のサブブロック (SB) に分割される。各フレームに対して, まず MB ごとに, 回転, 反転, ネガポジ反転, チャンネルシャフリング, およびサブブロックスクランブルの 5 種類のサブブロックレベルの変換を行う。回転角度などの変換パラメータは鍵 K_{ST} から導出される。しかしながら, 先行研究 [10], [11] より, すべての MB に同一の変換パラメータを適用した場合, ジグソーパズル解読攻撃のリスクが生じることが示されている。この脆弱性を緩和するため, 提案手法では K_{ST} から生成される変換パラメータを MB ごとに別々に割り当てる。サブブロックレベルの暗号化後, 各フレーム内の MB の位置を空間的に攪拌するメインブロックスクランブルを行う。これは K_{ST} とは異なる鍵 K_{MS} により制御される。

最後に, 暗号化されたフレームをもとの時間順に積み重ねることで, 暗号化映像を生成する。

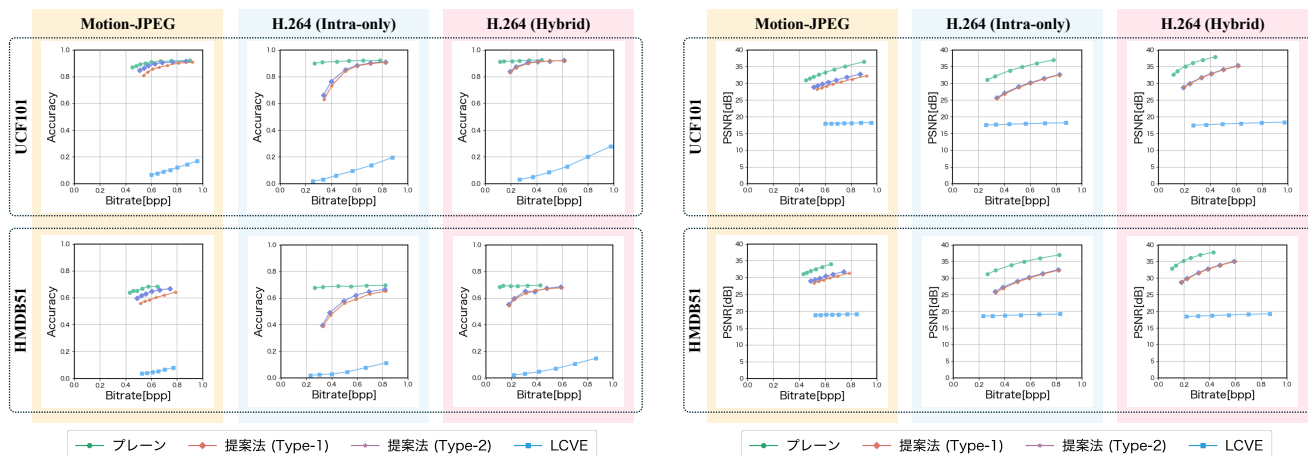
3.2 モデル暗号化

モデル暗号化は, プレーン映像で学習された VT モデルが, 3.1 によって生成された暗号化映像に対して再学習を必要とせず直接適用可能であることを保証する。図 1(b) に示すように, モデル暗号化は, 学習済み VT の埋込み機構で用いられるパラメータに対して, 一連の変換を施すことにより実行される。これらの変換は, 映像暗号化の際と同一の鍵を用いて行われる。

MB に対して行われる処理に対応して, まず 3 次元畳込みカーネル E をサブ行列 (sub matrix, 以降 SM) に分割する。各 SM の大きさは, SB と同一である。各 SM に対して, 共通鍵 K_{ST} を用いて, 回転, 反転, 符号反転, チャンネルシャフリング, およびサブ行列スクランブルの 5 種類のサブ行列レベル変換を適用する。

変換された SM は集約され, 暗号化カーネルが構成される。メインブロックスクランブルにより, 暗号化映像から抽出される各時空間トークンは空間的に攪拌された状態で配置されている。そのため, プレーン映像から学習された位置埋込み E_{pos} を直接加えた場合, 誤差が生じる。これを回避するため, E_{pos} の要素は共通鍵 K_{MS} により決定された順序に基づいて攪拌される。これにより, 埋込み機構において E_{pos} が攪拌されたトークンに対して適切に加えられることが保証される。

推論の際には, 暗号化カーネルおよび攪拌済みの位置埋込みが, 図 1(c) に示すように, サーバ上の VT に適用される。これによって, モデルは暗号化映像に対して直接



(a) 分類精度

(b) 画質

図 3: 映像圧縮による分類精度と画質の推移

推論することが可能となる。また、提案法は、パラメータの置換のみを行っているため、モデル暗号化による遅延は発生せず、リアルタイムに実行できる。

4 評価実験

ここでは、提案法の有効性について、認識性能、圧縮性能、および、攻撃耐性の観点から評価する。

4.1 実験条件

本実験では、行動認識モデルとして ViViT[6] を用いる。このモデルは、 224×224 画素の映像を入力とし、映像を $2 \times 16 \times 16$ 画素の 3 次元パッチに分割する。すなわち、 $\tau = 2$, $H_p = 16$, $W_p = 16$ となる。二つのモデルを Kinetics400 で事前学習されたパラメータで初期化し、UCF101 および HMDB51 の学習用映像を用いてそれぞれファインチューニングを行った。なお、ファインチューニングに用いられる映像は非圧縮とする。これにより、モデルが圧縮された映像に対して頑健になることを防ぎ、映像圧縮が認識性能に与える影響をより明確に評価できる。評価は UCF101 および HMDB51 のテスト用映像を用いて行った。これらのテスト映像は、均等なフレームサンプリングにより、32 フレームからなるクリップに変換され、各フレームは 224×224 画素にリサイズされる。リサイズ後のクリップは、LCVE および提案法を用いてそれぞれ暗号化される。提案法には Type-1 と Type-2 の 2 種類があり、Type-1 はすべての MB に同一の変換パラメータを適用し、Type-2 は各 MB に異なる変換パラメータを適用する。映像圧縮には、一般に広く用いられている 2 種類の映像圧縮規格 Motion-JPEG (MJPEG) と H.264 を用いる。両者のコーデックは FFmpeg を用いて実装した。

4.2 認識性能

ここでは、もとのパラメータを用いたプレーン映像上のモデル (プレーン設定)、LCVE[5]、および提案法の認識性能を比較し、評価する。後者二つの設定では、モデルのパラメータを暗号化し、対応する暗号化映像上で認識を行った。

まずプレーン映像に対しては、3 手法とも同等の認識精度を達成し、UCF101 で 0.93、HMDB51 で 0.70 の結果が得られた。この結果は、本手法がプレーン映像と暗号化映像を組み合わせた評価において、最先端 (state-of-the-art, 以降 SOTA) 手法と同等の認識性能を維持できることを確認している。

次に、圧縮映像における認識性能について確認する。H.264 で圧縮された映像については、標準的なハイブリッド符号化モード (hybrid モード) だけでなく、フレーム内符号化のみのモード (intra-only モード) も評価した。後者の intra-only モードでは、フレーム間予測を行わず、各フレームを独立に圧縮する MJPEG と同じ条件下での比較が可能となる。実験結果を図 3(a) に示す。この図から、LCVE はすべての圧縮条件下で極めて低い精度を示していることがわかる。このような著しい性能低下は、LCVE が映像圧縮に対して頑健性をもたないことを示している。

LCVE と対照的に、本手法は、MJPEG および H.264 の hybrid モード下で、プレーン設定に近い精度および性能低下傾向を維持する。ただし、H.264 の intra-only モードでは、hybrid モードと比較して認識精度が低下していることがわかる。したがって、本手法では、H.264 におけるフレーム間予測がフレーム内予測よりも親和性が高いことが示された。しかしながら、この低下は、提案法の有効性に影響を与えるものではない点に注意されたい。

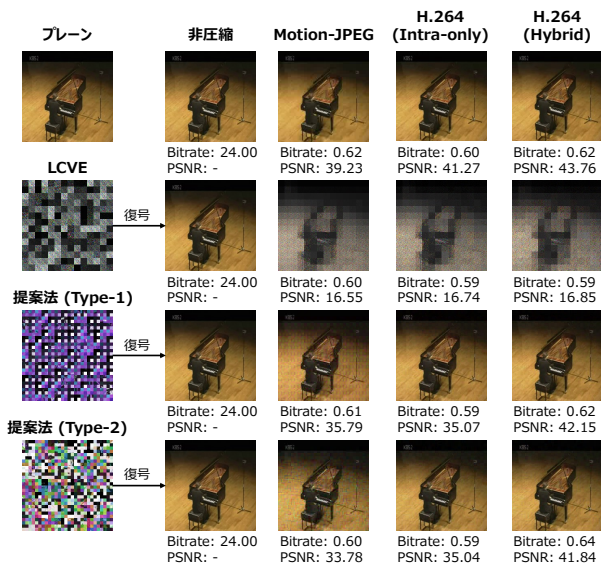


図 4: 暗号化領域において圧縮後、暗号解除された映像フレームの比較

H.264 の intra-only モードは実際にはほとんど使用されず、hybrid モードが一般的に利用されている。

4.3 Rate-distortion 解析

提案法で生成された暗号化映像の圧縮効果を評価するため、ビットレートと画質の関係を明らかにする rate-distortion 解析を行った。図 3(b) に、プレーン設定、LCVE、および提案法における RD 曲線を示す。同図より、LCVE による暗号化映像は、データセットや圧縮方式に関わらず、PSNR は常に 20dB を下回っていることがわかる。4.2 の結果と合わせると、この大幅な画質劣化が、圧縮された LCVE による暗号化映像の認識性能を低下させていることは明らかである。これに対して提案法により暗号化映像は、圧縮後も高い画質を保持している。

上記を視覚的に確認するため、図 4 に、各暗号化映像が 0.60bpp 程度に圧縮された後、暗号解除されたときの 1 フレームをそれぞれ示す。同図より、LCVE では、暗号化映像が圧縮後に暗号解除された場合、その内容が視認困難になる一方、本手法では深刻なアーティファクトを生じさせることなく、高い視認性が保たれている。

4.4 攻撃耐性

提案法によって生成された映像のセキュリティを評価するため、暗号文単独攻撃に対する耐性を検証する。提案法における大きなセキュリティ上の懸念の一つは、ジグソーパズル解法 (jigsaw puzzle solver, 以降 JPS) に基づく攻撃 [12], [13] に対する脆弱性であり、これは暗号化ブロック間の境界の相関性を利用して、もとの内容の復元を試みるものである。

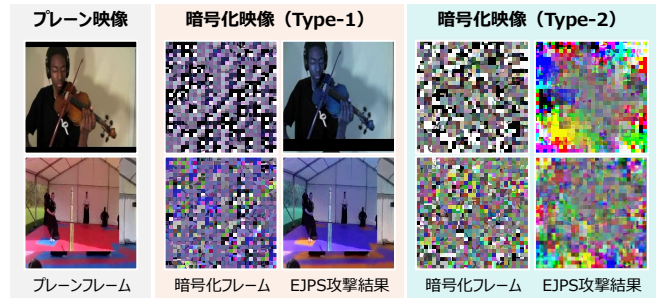


図 5: 提案法の各パターンで暗号化された映像の攻撃耐性の比較

本実験では、JPS の手法として、ブロック単位で画像暗号化された画像に対して効果的な拡張 JPS (Extended JPS, 以降 EJPS) [13] を採用した。この攻撃は、もとの内容を 2 段階で再構成するよう設計されており、第 1 段階では SB 変換の復元を、第 2 段階では MB の組立てを行う。具体的には、第 1 段階では各 MB 内で施された SB 変換の復元を、第 2 段階ではランダムに並べ替えられた MB 位置の再配置をそれぞれ試みる。ここでは、提案法の Type-1 および Type-2 で暗号化された映像の各フレームに対し、独立に EJPS を適用した。

図 5 は、提案法で暗号化された映像フレームが EJPS で攻撃された例を示している。同図より、Type-1 と Type-2 では、EJPS 攻撃に対する耐性に大きな差があることがわかる。各 MB に同一の変換パラメータを適用する Type-1 で暗号化された映像は、色の歪みが見られるものの、EJPS によりほぼ完全に復元されている。一方で、Type-2 で暗号化された映像からは、EJPS によって意味のある情報を復元することができなかった。したがって、攻撃耐性の観点から、提案法の中でも Type-2 が有効であることが明らかである。

5 まとめ

本稿では、圧縮に適した暗号化映像を入力として処理するための新しい PPAR 手法を提案した。提案法では、暗号化と圧縮の間のトレードオフに対応することで、暗号化映像が高い行動認識精度と高い圧縮効率をともに保持できる。まず、認識性能の観点から、提案法は圧縮されていない暗号化映像に対して、SOTA 手法と同等の認識性能を達成することが確認された。さらに、提案法は SOTA 手法と比較して、圧縮後の暗号化映像の認識精度および画質の低下を大幅に抑制できることが明らかとなった。また、提案法の Type-2 は、EJPS に対して高い耐性を有することが示された。今後、選択的暗号化などの暗号化手法を検討し、提案法の実用性をさらに高めていく予定である。

参考文献

- [1] M. Ryoo, K. Kim, and H. Yang, “Extreme low resolution activity recognition with multi-siamese embedding learning,” in Proc. AAAI Conference on Artificial Intelligence, vol.32, no.1, 2018.
- [2] S. Kumawat and H. Nagahara, “Privacy-preserving action recognition via motion difference quantization,” in Proc. European Conference on Computer Vision (ECCV), pp.518–534, 2022.
- [3] F. Ilic, H. Zhao, T. Pock, and R. P. Wildes, “Selective interpretable and motion consistent privacy attribute obfuscation for action recognition,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.18730–18739, 2024.
- [4] M. Kim, X. Jiang, K. Lauter, E. Ismayilzada, and S. Shams, “Secure human action recognition by encrypted neural network inference,” Nat. Commun., vol.13, no.1, p.4799, 2022.
- [5] Y. Ishikawa, M. Kondo, and H. Kataoka, “Learnable cube-based video encryption for privacy-preserving action recognition,” in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp.7003–7013, 2024.
- [6] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, “ViViT: A video vision transformer,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp.6836–6846, 2021.
- [7] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, “Video Swin transformer,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.3202–3211, 2022.
- [8] Z. Tong, Y. Song, J. Wang, and L. Wang, “Video-MAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol.35, pp.10078–10093, 2022.
- [9] H. Kiya, A. P. M. Maung, Y. Kinoshita, S. Imaizumi, and S. Shiota, “An overview of compressible and learnable image transformation with secret key and its applications,” APSIPA Trans. Signal Inf. Process., vol.11, no.1, 2022.
- [10] H. Lin, S. Imaizumi, and H. Kiya, “Privacy-preserving ConvMixer without any accuracy degradation using compressible encrypted images,” Information, vol.15, no.723, 2024.
- [11] T. Chuman, N. Ono, and H. Kiya, “Security evaluation of compressible and learnable image encryption against jigsaw puzzle solver attacks,” in Proc. IEEE Global Conference on Consumer Electronics (GCCE), pp.823–826, 2023.
- [12] T. Chuman, W. Sirichotedumrong, and H. Kiya, “Encryption-then-compression systems using grayscale-based image encryption for JPEG images,” IEEE Trans. Inf. Forensics., vol.14, no.6, pp.1515–1525, 2019.
- [13] T. Chuman and H. Kiya, “A jigsaw puzzle solver-based attack on image encryption using vision transformer for privacy-preserving DNNs,” Information, vol.14, no.311, 2023.