HOKKAIDO UNIVERSITY

Title	Reducing Annotation and Computation Costs for Efficient Compressed Video Action Recognition [an abstract of dissertation and a summary of dissertation review]
Author(s)	寺尾, 颯人
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第15999号
Issue Date	2024-03-25
Doc URL	https://hdl.handle.net/2115/91873
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	doctoral thesis
File Information	Hayato_Terao_review.pdf, 審査の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　寺尾　颯人

審査担当者　主　査　教　授　山本 雅人
副　査　教　授　野田 五十樹
副　査　教　授　川村 秀憲
副　査　特任教授　小野 哲雄

学位論文題名

Reducing Annotation and Computation Costs for Efficient Compressed Video Action Recognition
（効率的な圧縮動画分類に向けたアノテーションコストと計算コストの削減手法）

ソーシャルメディアや動画共有サイトの普及に伴い, オンライン上で利用可能な動画データの数は増加しており, それらの動画を活用するための動画解析技術の必要性は高まり続けている. また, 近年のディープラーニングの発展を受け, 深層学習を用いた動画解析技術が数多く提案されている. 特に, 入力として与えられた動画データを事前に決められたクラス群の中から最も適切なクラスに割り当てる動画分類問題は広く研究されている.

本論文は, 動画分類問題に対するアプローチのひとつである, 圧縮動画分類に取り組んでいる. 多くの動画分類手法では RGB フレーム列をディープネットワークの入力として扱う一方で, 圧縮動画分類は動画圧縮アルゴリズムを用いて圧縮された動画ファイルに保存されている I-frame, motion vector, residual と呼ばれる 3 種類の情報をディープネットワークに入力するという特徴がある. このアプローチの利点として, RGB フレームを取得するために必要であったデコード処理を介さずにディープネットワークへの入力を取得できる点が挙げられる. この利点により,RGB フレーム列を入力とする手法と比較してより効率的な動画分類が可能となることから, 圧縮動画分類はエッジデバイスやモバイルデバイスのような安価で省電力なデバイス上での運用などへの応用が期待されている.

本論文は, 圧縮動画分類の効率性を更に向上させることを目的とし, 特にアノテーションコストと計算コストを削減するための手法を提案している.

本論文は全 5 章で構成されている

第 1 章では研究背景と研究目的, および本論文の構成について説明している.

第 2 章では動画分類や圧縮動画分類に関する先行研究の詳細を述べている.

第 3 章では, 圧縮動画分類モデルの学習に必要なアノテーションコストを削減するための半教師あり学習手法が提案されている. 本論文では, 半教師あり学習手法で広く採用されるフレームワークである擬似ラベル法を圧縮動画分類に拡張した手法が提案される. ここで, 擬似ラベル法とは, 教師なしデータを学習中のディープネットワークに入力し, 得られた予測を基に生成した擬似ラベルを通常の教師ラベルの代わりとして用いる学習法である. この擬似ラベル法において, 教師なしデータに割り当てられる擬似ラベルの精度を高めることが最終的な精度を向上させる上で重要である. そこで, 本研究では I-frame, motion vector, residual を処理する 3 つのディープネットワークを同時に学習し, 擬似ラベルを生成する際に 3 つのディープネットワークの出力をアンサンブルす

ることで, より信頼のおける擬似ラベルを生成するアプローチを提案する. 実験により, このアプローチと擬似ラベル法を採用した最先端の半教師あり画像分類手法である Fixmatch を組み合わせた Compressed Video Ensemble based Pseudo Labeling (CoVEnPL) は,RGB フレームのみを入力とする半教師あり動画分類手法よりも少ないラベル数で同等の精度を達成できることが示されている. また, データの読み込み, 前処理に必要な時間についての比較をおこない, 提案手法が複数の入力を用いているにも関わらず RGB フレームよりも高速なデータ読み込み, 前処理が可能であることが示した. これらの結果から, 提案手法である CoVEnPL は RGB フレームを入力として扱う従来の半教師あり動画分類手法と比較して効率性と分類精度を両立させていることが示されている.

第 4 章では, 圧縮動画分類の計算コストの削減を目指した研究が行われている. 本論文では圧縮動画分類の入力を処理するネットワークの数自体を減らすことで計算コストを削減する multi-stream single network (MussNet) を提案している.MussNet は I-frame, motion vector, residual を単一ネットワークによる一度の順伝播計算で同時に分類をおこなうアプローチを採用する. これによって, 先行研究で採用された,3 つのネットワークで I-frame, motion vector, residual を個別に処理するアプローチと比較して少ないパラメータ数, 計算コストでの圧縮動画分類が可能となる. しかしながら, 単純に同じ動画から得られる I-frame, motion vector, residual を単一ネットワークで分類するように学習すると,3 つのネットワークを用いる手法と比較して精度が大きく劣化してしまうという課題がある. そこで, 本研究では学習時に異なる動画から得た I-frame, motion vector, residual を入力として与え, それぞれを独立に分類するように単一ネットワークを学習することを提案している. この学習手法によって, 単一ネットワーク内部に 3 つの独立したサブネットワークを構築し,3 つのネットワークを用いて圧縮動画分類をおこなう手法を単一ネットワークで近似することができる. 実験では動画分類データセットを用いたベースライン手法, 先行研究との比較をおこなうことで,MussNet が単一ネットワークと同等の計算コストで 3 つのネットワークを用いる手法と同等の分類精度を達成できることが示されている.

第 5 章では, 本論文の結論が述べられる. 特に, 本論文が取り組んだ問題と提案手法についてのまとめと, 結果全体を受けた今後の展望について述べる.

これを要するに, 著者は, 圧縮動画分類においてアノテーションコストおよび計算コストを削減するための研究をおこなった. アノテーションコストについては圧縮動画分類のための半教師あり学習手法を新たに提案し, 一般的な動画分類手法と比較して, より高い分類精度を達成した. また, 計算コストについては, 新しい圧縮動画分類モデルである MussNet を提案し, 単一ネットワークを用いながら複数ネットワークと同等の精度を達成可能であることを示している. これらの成果は圧縮動画分類の効率性をこれまで以上に高め, より安価に省電力なデバイス上での動画分類モデルの運用につながるものであり, 情報科学の分野に対して貢献するところ大なるものがある. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.