Efficient Video Recognition with Transformers: Tackling the Redundancy in Video

January 2025

Tomoyuki Suzuki

別表５
(3)

(様式甲 3)

主　論　文　要　旨

No.1

報告番号	甲　　第　　号	氏　名	鈴木　智之

主 論 文 題 名：

Efficient Video Recognition with Transformers: Tackling the Redundancy in Video

(Transformer を用いた効率的動画認識：動画の冗長性への取り組み)

（内容の要旨）

動画は実世界の連続的な時間の流れの中で観測される視覚データである．コンピュータビジョン分野において，動画から解釈可能な情報を抽出する動画認識は，行動解析，危険予測，動画検索など，幅広い応用をもつ基盤技術として注目されている．一方で，その計算コストの大きさが実用や研究の障害となっている．我々は映写機の時代から，動画を画像の系列として表現・可視化してきた．そして現代の動画認識技術の主流は，同じようにコンピュータに動画を画像の系列として処理させている．しかし，動画の時間的連続性を考慮すると，動画の微小時間区間における変化は非常に小さく，これをすべて同等に処理することは非常に冗長であると考えられる．

本論文では，動画の冗長性を削減することで，動画認識の効率化を目指す．近年，動画認識において高い性能を示している Transformer をベースとするモデルを想定し，それらの内部で処理される冗長なトークンを検出・除外し，動画認識を効率化する手法を提案する．すべての時刻が観測可能なオフライン動画認識と，動画が刻一刻と観測されるオンライン動画認識の二つのシナリオに適した手法をそれぞれ提案した．いずれの設定においても，広範な実験を行い，提案手法により精度を維持しつつ，動画認識の速度を大幅に向上させることを実証した．

第 1 章では，動画や動画認識，動画の冗長性に関する問題を提起し，本論文の研究課題と貢献を示す．

第 2 章では，動画認識の歴史とその効率化の試みについて概説し，その歴史的な流れの中で本研究の位置付けを明らかにする．

第 3 章では，本研究の基盤となる Transformer とそれを用いた動画特徴抽出についての予備知識を説明する．

第 4 章では，オフライン動画認識において，動画の圧縮情報を活用して冗長トークンを検出・除外する手法を提案する．オフライン動画認識のベンチマークである行動認識において評価を行い，その効果を実験的に示す．

第 5 章では，圧縮情報が利用できないオンライン動画認識において，過去の情報を参照して現時刻フレームの冗長トークンを検出し，過去情報を再利用することでそれらの情報処理をスキップする手法を提案する．さらに削減対照を入力トークンだけでなく中間トークンに拡張する．オンライン動画認識における複数のタスクで評価を行い，その効果を実証する．

第 6 章では，本論文を総括し，今後の展望について議論する．

別表６
（３）

（様式甲 4）
Keio University

Thesis Abstract

No.

Registration Number	■ “KOU” □ “OTSU” No. *Office use only	Name	SUZUKI Tomoyuki

Thesis Title

Efficient Video Recognition with Transformers: Tackling the Redundancy in Video

Thesis Summary

Video is visual data observed in the continuous flow of time in the real world. The technology that extracts interpretable information from videos, video recognition, is attracting attention as a fundamental technology with a wide range of applications, such as action recognition, risk anticipation, video creative works, and video retrieval. Meanwhile, the computational cost of video recognition is a huge bottleneck for practical applications and development. Since the era of film projectors, videos have been represented and visualized as sequences of images. In today's mainstream video recognition approaches, computers also process videos in the same way as film projectors. However, given the temporal continuity of videos, the changes in videos over small time intervals are very small, and processing each frame equally is considered highly redundant.

This thesis aims to improve the efficiency of video recognition by reducing temporal redundancy. The core idea is to identify redundant tokens and exclude the computation associated with them in Transformer-based video models, which have recently demonstrated high performance in video recognition. Methods are developed for two scenarios: online video recognition, where videos are observed sequentially in time, and offline video recognition, where the entire video is observed at once. Extensive experiments in each setting demonstrate that the proposed methods significantly improve the speed of video recognition while maintaining accuracy.

Chapter 1 introduces videos and video recognition, raise the issue of video redundancy, and present the research questions and contributions of this study.

Chapter 2 reviews the history of video recognition and attempts to improve its efficiency, and confirm the position of this study in the historical context.

Chapter 3 introduces the Transformer and video feature encoding based on Transformer, which are the basis of this study.

Chapter 4 proposes a method to detect and exclude redundant tokens using video compression information in offline video recognition and experimentally demonstrate its effectiveness.

Chapter 5 updates the method proposed in Chapter4 for online video recognition, where compression information is not available. We detect redundant tokens by referring to past information and extend the method to intermediate feature tokens. We experimentally demonstrate the effectiveness of the proposed method in online video recognition.

Chapter 6 summarizes this thesis and discusses future prospects.