氏　　名　藤武 将人

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2329 号

学位授与の日付　2022 年 3 月 24 日

学位授与の要件　複合科学研究科 情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Feature Enhancement using Spatio-Temporal Information for Video Object Detection

論 文 審 査 委 員　主　査　杉本 晃宏
情報学専攻 教授
佐藤 いまり
情報学専攻 教授
児玉 和也
情報学専攻 准教授
池畑 諭
情報学専攻 助教
佐藤 真一
国立情報学研究所 コンテンツ科学研究系 教授

（様式３）

# 博士論文の要旨

氏　　名 藤武　将人

論文題目 Feature Enhancement using Spatio-Temporal Information for Video Object Detection

Video is an essential resource because of its ability to hold space and time information. Therefore, in the field of computer vision, a lot of research is conducted to extract various information, especially object detection in videos. It is expected to be applied to real-world applications such as surveillance cameras and robotics. Object detection consists of two processes: extracting the feature maps for detection from the video frames and detecting objects from them. In object detection in video, detectors for still images are sometimes applied to each frame. However, it is difficult to achieve stable detection due to apparent changes with time in the video, which leads to fluctuations of detection confidence score, and false-positive and false-negative detection results. Previous research tried to solve them by incorporating temporal information in the detection stage. However, the effect was limited since the feature maps obtained from the frames are deteriorated due to the changes in appearance, and it is difficult to detect objects from them. Therefore, it is essential to enhance feature maps with temporal information before the detection stage. Research on feature maps suitable for detection has been conducted mainly in offline methods that employ all future, current, and past information, and there have been few online methods, which do not rely on future information, aimed at real-world applications such as surveillance cameras and robots. In addition, for such applications, not only the accuracy but also the processing speed for real-time is essential. Previous works have proposed stabilizing the detection by propagating the past information from the last frame or a specific nearby keyframe for real-time processing in online settings. However, they have not yet achieved stable detection due to the limited use of temporal information. Therefore, this dissertation studies feature enhancement methods for real-time and online video object detection that utilizes more temporal and spatial information. To enhance feature maps for video object detection, we studied two aspects. One is to refine a feature map by aggregation, and the other is to enhance a feature map through prediction. First, we propose two new feature map aggregation methods: frame-level feature map aggregation and element-level feature map aggregation.  Feature map aggregation differs from previous real-time and online methods in that it directly exploits multiple past feature maps. It has been studied in offline methods and can provide stable feature maps; however, it requires more processing time due to the computation of the weight between detection and each surrounding frame. Therefore,

in frame-level feature map aggregation, we propose to refine the feature map by calculating which past frames should be focused on in a one-shot manner, which runs in real-time. To aggregate past features directly, we extend the detector with external memory. We experimentally show that frame-level feature map aggregation can suppress the issue of object confidence score fluctuations in time. At the element-level, the idea of the frame-level method is further extended. Each element of the feature map is refined considering local and global spatial information and short- and long-range temporal information; however, such dense aggregation takes much time to calculate in general. Therefore, we propose a novel sparse aggregation method to reduce computation processing time for feature aggregation. Furthermore, we also propose an adaptive feature update strategy in external memory to hold long-term information. Finally, we achieve state-of-the-art performance in an online detector that maintains real-time performance. We also show that the proposed method significantly reduces false-positive and false-negative detection results, which are challenges in video object detection. Next, we propose a novel feature map enhancement approach through prediction. The prediction-based approach differs from the feature map aggregation approach in that it does not utilize external memory but enhances the performance of the model itself. Therefore, it is suitable for conditions under strict GPU memory constraints, such as robotics. The prediction-based approach employs a future prediction task, which requires deep knowledge of objects, such as motion, to forecast the future clearly. The detector enhances the feature maps for stable object detection by learning features through prediction during the training phase. We leveraged the prediction from different perspectives: forecasts for the next and the next several frames. First, we propose a detector that jointly learns detecting objects and forecasting the next-frame feature map. This prediction approach is suitable for extending the recurrent neural network object detectors, and experiments show the effectiveness of learning features through the next frame forecast. Next, we propose a video object detection framework based on stochastic future prediction to leverage more extended time. The next several frames prediction is difficult to predict due to the future uncertainty; therefore, our model learns features by predicting the sampled and possible future. Experiments have shown the effectiveness of leveraging the stochastic long-term prediction for video object detection.

Form8 : Separate Sheet
（様式８・別紙１）

**Results of the doctoral thesis screening**

# 博士論文審査結果

Name in Full
氏　　名　藤武　将人

Title
論文題目　Feature Enhancement using Spatio-Temporal Information for Video Object Detection

博士論文は、「Feature Enhancement using Spatio-Temporal Information for Video Object Detection（動画物体検出のための時空間情報を用いた特徴量強化）」と題し、英文で書かれている。ライブストリーミング動画から実時間でそこに写っている物体を検出する技術は、交通状況の把握をはじめ多くの応用があり脚光を浴びている。本論文は、深層学習を使った動画物体検出をテーマとし、動画が有する冗長性を考慮して時系列情報を利用し、実時間処理を損なうことなく時々刻々見えが変化する物体を安定・高精度に検出する手法を開発し、その有効性を検証している。具体的には、過去のフレームから得られる画像特徴をその重要性に応じて集約し検出フレームにおける画像特徴を強化する手法、特徴量の重要性を時間・空間の要素レベルで評価し、大域的な観点でそれを集約し、検出のための画像特徴を強化する手法、将来フレームの予測を通して検出に必要な動画特徴表現を事前学習し、その特徴表現を検出に転移して学習させることで画像特徴を強化する手法、の３つの手法を提案し、それぞれの手法において、複数の標準的なデータセットを用いて、従来手法に対する優位性、導入した考え方の有効性を示している。

博士論文は６章で構成されている。まず、第１章で、ライブストリーミング動画からの物体検出の重要性を論じ、博士研究で取り組む問題の意義を議論している。そして、従来法でのアプローチの問題点をあげ、安定で高精度な検出とリアルタイム処理の必要性を論じて博士研究の位置づけを述べている。第２章では、それぞれの立場での従来研究の動向と問題点を示している。引き続く３つの章が本論文の主要部分となっている。第３章では、ライブストリーミング動画の過去フレームを外部メモリに格納し、格納した各フレームの画像特徴の重要性を評価して、検出フレームにおける画像特徴とともに集約することで、検出のための画像特徴を導出するネットワークモデルを提案している。外部メモリに格納する最適なフレーム数の検証を行った後、検出の安定性、精度、処理速度の観点から最先端の関連手法と比較し、提案手法の優位性を示している。第４章では、第３章を発展させ、時間・空間の要素単位で画像特徴の重要性を捉え、外部メモリに重要フレームを選択的に保持しつつ、長期的、大域的に画像特徴を時間・空間の要素レベルで疎に集約することで消費メモリを抑えつつ特徴量の強化を図っている。そして、実験によって、検出の安定性、精度、消費メモリ、処理速度の観点から提案手法の有効性を検証している。第５章では、現在までのフレームから将来フレームを予測するという異なるタスクを通して、将来まで含めた物体の見えの変化を捉える時系列画像特徴を事前学習させ、それを検出タスクに転移して学習させる手法を提案している。また、時系列画像特徴をリカレントネットワークの状態として記憶させることでメモリ消費を抑えている。検出の安定性、精度、消費メモ

リ、処理速度を最先端の関連手法と比較し、提案手法の優位性を示している。第６章では、まとめと今後の課題を議論している。

出願者による約４５分の発表もこの順で説明が行われ、その後、３０分程度の質疑応答があった。審査委員からは、提案した特徴量強化手法の他の動画解析タスクへの適用可能性、検出における将来フレーム予測の位置づけ、今後の展望などに質問とコメントが寄せられ、それらに対し出願者は適切に回答した。

質疑応答後に審査委員会を開催し、審査委員で議論を行った。審査委員会では、動画検出における問題を明確にし、それに対する提案手法の貢献が整理されていることを確認した。そして、出願者の博士研究がライブストリーミング動画における実時間物体検出に対して、時々刻々見えが変化する物体を捉えるために、時間的、空間的に画像特徴を活用して取り組むアプローチは独創的であることが評価されるとともに、研究成果として、情報学専攻が定めるトップカンファレンスに査読付き論文が１件採択され、また、他の国際会議において査読付き論文２編が採択されていることが確認された。以上の理由により、審査委員会全員一致で、博士論文として十分な水準にある研究であると認め、本論文が博士の学位請求論文として合格であり、学位の授与に値すると結論づけた。