（様式９）

学位論文審査の概要と結果

報告番号	東アジア博 甲 第 187 号	氏　名	張　慶琪（ZHANG QINGQI）
論文題目	Depth Camera-based Human Action Recognition with Object Detection and Graph Convolutional Network 物体検出とグラフ畳み込みネットワークを用いた深度カメラベースの人間行動認識		

（論文審査概要）

本学位論文は，深度カメラを用いた高精度かつ高効率な人物行動認識（HAR）フレームワークを提案する．このフレームワークは，まず物体検出アルゴリズムでシーン内の人物を特定し，正確な関心領域（RoI）を抽出する．次に，RoI 内の人物の骨格データを深度カメラデータから抽出し，これを時空間モデリングのためのロバストな入力として利用する．最後に，行動分類に特化して設計されたグラフ畳み込みネットワーク（GCN）を適用し，骨格シーケンスを分析して行動を分類する．RGB データと骨格データを含む複数のモダリティを効果的に組み合わせることで，本フレームワークは複雑なシナリオにおいても認識精度とロバスト性を大幅に向上させる．論文の構成は次の通りである．

第 1 章では，人物行動認識（HAR）およびそれに関するこれまで行われてきた先行研究の概略を紹介し，人物行動認識の研究における現状と課題を明らかにしたうえ，本研究の目的を示している．また，本学位論文の構成の説明を述べている．

第 2 章では，人物行動認識（HAR）に関連する既存の手法である物体検出，骨格抽出手法，および行動分類手法に焦点を当てて包括的なレビューを行っている．物体検出については，伝統的な手法を紹介した後に，深層学習に基づく物体検出手法の概要を述べている．骨格抽出手法については，2D 姿勢推定手法，3D 姿勢推定手法，そして深度カメラという 3 つの視点から提示している．行動分類手法については，CNN ベース，RNN ベース，GCN ベースの 3 つの主要なタイプに分類している．また，人物行動認識の応用例として，高齢者介護とスマート教育を取り上げると説明している．

第 3 章では，カスケードネットワークに基づく物体検出アルゴリズムを提案し，禁止品目検出と人物検出という 2 つの主要なタスクに適用する技術を提案している．これら 2 つのタスクは異なる対象物を扱うものの，根本的には同じ物体検出の範疇に属し，複雑な背景干渉やスケール変動といった類似の課題に直面する．そのため，本章では禁止品目検出を代表的な事例として取り上げ，物体検出における既存の問題と課題を分析することとなっている．また，提案手法が禁止品目検出と人物検出の両タスクにおいて有効であることを検証している．

第 4 章では，行動認識における問題と課題を分析するために，最も困難なケースである高齢者行動認識を代表例として取り上げている．骨格ベースの行動認識のために，高度な時間的特徴表現を効果的にモデル化する新しい骨格時系列融合グラフ畳み込みネットワーク（STF-GCN）を提案し，時間的特徴表現を強調する骨格時系列融合（STF）モジュールも導入している．最後に，提案手法の有効性を，高齢者行動認識と一般行動認識タスクの両方で広範な実験を通じて検証し，そのロバスト性と適応性を示している．

第 5 章では，第 3 章の物体検出アルゴリズムと第 4 章の行動分類アルゴリズムを基盤として，深度カメラに基づいて開発した行動認識システム，および，高齢者介護とスマート教育という 2 つの主要な領域への応用について述べている．高齢者介護の分野では，高齢者の安全対策の強化と自立した生活の支援に不可欠な転倒の検出と分析に焦点を当てており，スマート教育の領域では，学生の挙手動作のリアルタイム認識のために特別に設計されており，教室での参加度に関する貴重な洞察を提供し，教育者がそれに応じて指導方法等を調整することを可能にするとの説明をしている．

第 6 章では，本研究で得られた成果である，カスケードベースの物体検出手法（第 3 章），行動分類のための骨格時系列融合グラフ畳み込みネットワーク（STF-GCN）（第 4 章），および深度カメラベースの人物行動認識システム（第 5 章）をまとめている．また，これらの研究成果を踏まえながら，将来の研究展開について議論している．

以上の学位論文の内容から，審査委員会は次のように評価した．

1．創造性：人物行動認識（HAR）における現状と課題および関連研究を十分に理解し，独自に設計した物体検出アルゴリズムと行動分類アルゴリズムを基盤として，深度カメラの映像データに基づいた行動認識システムを自力で開発したことから，本論文は，創造性において極めて優れている．

2．論理性：先行研究を引用しつつ課題を明らかにしている．既存の手法の長所と限界を議論しながら，独自のアルゴリズムを設計し行動認識システムを開発している．このように，本論文は，課題の提示から解決法まで一貫性のある展開をしていることから，論理性において極めて優れている．

3．厳格性：関連する先行研究は幅広く網羅し，丁寧に渉猟している．また，提案するアルゴリズムは，その有効性とロバスト性について複数の実験結果に基づき厳密に検証しており，行動認識システムへの適用をしている．これらのことから，厳格性において優れている．

4．発展性：提案した手法および開発したシステムは，高齢者転倒や生徒挙手の動作検出シミュレーションにおいて確実に適応できている．これらの研究成果をさらに発展させていけば，福祉分野やスマート教育分野のみならば，人間の社会活動の様々な場面において活用が可能である．よって，発展性において極めて優れている．

以上より，全体的に優れていることから，論文審査を「合」と判定した．

論文審査結果
合・否

審査委員　主　査　（氏 名）萬 崎偉

（氏 名）中田 亮

（氏 名）北沢 千里

（氏 名）

（氏 名）

学 位 論 文 要 旨

学位論文題目 ： Depth Camera-based Human Action Recognition with Object Detection and Graph Convolutional Network（物体検出とグラフ畳み込みネットワークを用いた深度カメラベースの人間行動認識）

申 請 者 氏 名 ： 張 慶琪

Human Action Recognition (HAR) is a challenging and engaging research area in computer vision with diverse applications, including smart surveillance, robotics, industrial automation, healthcare, and education. Traditional methods relying on RGB video data often struggle to handle environmental noise, such as variations in lighting, viewpoint, background colors, and clothing, limiting their effectiveness in real-world scenarios. The rapid development of depth camera technology has opened up new opportunities for action recognition by enabling the use of three-dimensional human joint coordinates, which are highly effective for accurately modeling human actions. Additionally, depth cameras provide robustness in complex environments and offer privacy-preserving capabilities, making them particularly suited for sensitive applications.

This thesis proposes a depth camera-based framework that integrates object detection and Graph Convolutional Network (GCN) to achieve accurate and efficient HAR. First, the proposed framework employs an object detection algorithm to localize individuals in the scene and extract precise Regions of Interest (RoI). Next, depth camera data is utilized to extract skeleton data of individuals within the RoIs, providing robust input for spatiotemporal modeling. Finally, a GCN specifically designed for action classification is applied to analyze skeletal sequences and classify actions. By effectively combining multi-modal data sources, including RGB and skeletal data, the framework significantly improves recognition accuracy and robustness in complex scenarios.

As the first critical step in the proposed framework, object detection faces significant challenges, particularly the diversity of object scales and the overlapping of objects within complex scenes. To address the issue of scale diversity, this thesis proposes the Re-BiFPN feature fusion method, which incorporates a Coordinate Attention Atrous Spatial Pyramid Pooling (CA-ASPP) module and a recursive

connection. The CA-ASPP module extracts direction-aware and position-aware information from feature maps, enhancing the representation of multi-scale objects. The recursive connection further refines the feature maps by feeding the processed multi-scale features back to the backbone network for additional feature extraction. Additionally, to improve localization accuracy and robustness in cluttered scenes, a Rep-CIoU loss function is designed to mitigate the impact of object overlap.

To evaluate the effectiveness of the proposed methods, experiments were conducted on X-ray security inspection images, a highly challenging real-world application characterized by varying object scales and frequent object overlaps. The results demonstrate that the proposed methods significantly improve detection accuracy and robustness. Furthermore, to assess the generalizability of the methods, extensive experiments were also performed on the VOC and COCO datasets, which include human detection tasks. The experimental results show that the proposed methods achieve competitive performance, highlighting their broad applicability across diverse domains.

As another key step in the proposed framework, action classification is a challenging task, particularly when dealing with subtle motion patterns characterized by small amplitudes and prolonged durations. Two pivotal issues warrant further exploration: the development of enhanced temporal feature representations and the expansion of convolutional models' capacity to capture long-range temporal dependencies. This thesis proposes a novel Skeleton Temporal Fusion Graph Convolutional Network (STF-GCN) for action classification, which effectively models advanced temporal feature representations. Specifically, the STF-GCN employs three encoding strategies to integrate two types of temporal feature representations. These strategies are designed to capture the intricacies of motion dynamics and the subtleties in action variations, enabling more accurate and robust recognition of complex actions. Furthermore, a Skeleton Temporal Fusion (STF) module is proposed to highlight temporal feature representations, employing a structure that alternates between large and small kernel convolutions to achieve diverse effective receptive fields. The integration of large kernel convolutions allows our model to perceive an expanded temporal context, significantly enhancing its ability to understand action dynamics in depth.

To evaluate the effectiveness of the proposed methods, extensive experiments were conducted on both elderly action and general action datasets. The results demonstrate that the proposed methods achieve superior performance, not only in classifying elderly actions with subtle and prolonged motion patterns but also in general action classification tasks, highlighting its robustness and broad applicability.

Finally, this thesis explores the application of the proposed framework in two key areas: elderly care and intelligent education. In elderly care, fall detection serves as a case study. Fall detection is critical for enhancing caregiving safety and supporting the independence of elderly individuals. Experimental results demonstrate that the proposed framework achieves an impressive average accuracy of 96.3% in real-world scenarios, while maintaining low false positive and false negative rates. In intelligent education, hand-raising recognition serves as a case study. Recognizing hand-raising actions is essential for comprehensively evaluating student engagement and adapting teaching strategies accordingly. Experimental results show that the proposed framework achieves an average accuracy of 89.7% in real-world scenarios, also maintaining low false positive and false negative rates. These results demonstrate that the proposed depth camera-based action recognition framework is both accurate and practical for real-time applications in critical domains, and has strong potential for broader deployment in real-world scenarios.