

ABSTRAK

Human Activity Recognition (HAR) adalah bidang penting dalam kecerdasan buatan yang bertujuan mendeteksi dan mengklasifikasikan aktivitas manusia secara otomatis. HAR memiliki banyak aplikasi, seperti keamanan, kesehatan, interaksi manusia-komputer, dan lingkungan cerdas. Pendekatan HAR bisa menggunakan kamera melalui visi komputer (CV). Kini, banyak penelitian HAR memanfaatkan *Deep Learning* (DL), terutama *Convolutional Neural Network* (CNN) berbasis 2D dan 3D, yang efektif menangkap data spasial (gambar) dan temporal (urutan gerakan) yang digabung menjadi spatiotemporal. Salah satu metode deteksi objek real-time yang banyak digunakan adalah YOLO (*You Only Look Once*), dengan versi Ultralytics di YOLOv8 dan YOLOv11. Untuk CV HAR spatiotemporal, arsitektur YOWOV3 menggabungkan deteksi spasial dari YOLO dengan analisis temporal dari CNN 3D. YOWOV3 adalah berbasis *dataloader* Torchvision dan PIL yang mana didesain untuk visualisasi data, tetapi kombinasi kedua modul memperlambat waktu training. Selain itu, YOLOv11 masih belum dibuktikan secara terperinci untuk dimasukkan ke YOWOV3. Dengan itu, proses pemuatan data diganti memakai modul Albumentations dan YOLOv8 diganti ke YOLOv11 sebagai perbandingan. Perbandingan ini berkontribusi terhadap pengembangan YOWOV3, salah satunya menuju efisiensi program. Evaluasi dilakukan dengan dataset MSCOCO untuk YOLO, dan AVAv2.2 untuk YOWOV3. Hasil modifikasi menunjukkan waktu pelatihan YOWOV3 turun drastis dari 1:10:00 menjadi 0:48:35. Namun, pergantian *dataloader* menghasilkan peningkatan *loss*, terutama saat menggantikan YOLOv8 ke YOLOv11. Dalam 9 epoch, YOWOV3 dengan YOLOv8 menghasilkan *loss* 7.933, sedangkan YOLOv11 8.0369. Ini menunjukkan tidak selamanya pergantian *dataloader* memenuhi efisiensi keseluruhan di mana terdapat tradeoff dari *dataloader* karena pendekatan optimisasi *dataloader* terhadap arsitektur dan *hardware* yang berbeda.

Kata kunci—computer vision, object detection, human activity recognition, HAR, spatiotemporal, Ultralytics.

ABSTRACT

Human Activity Recognition (HAR) is a vital field in artificial intelligence, supporting applications in security, healthcare, human-computer interaction, and smart environments. HAR aims to automatically identify and classify human activities using either sensor-based or computer vision-based methods. Recent progress highlights deep learning, particularly Convolutional Neural Networks (CNNs), 2D and 3D, which are the key for comprehensive spatiotemporal modeling of human actions. The Ultralytics YOLO series, especially YOLOv8 and YOLOv11, is widely used. Integrating YOLO's spatial detection with 3D CNN-based temporal analysis, the YOWOV3 architecture offers advanced capabilities for spatiotemporal HAR. However, performance-wise, YOWOV3 is still slow especially for the training time. The use of Torchvision +PIL dataloader causes performance bottlenecks that affect training time. On the other hand, YOLOv11 has not been thoroughly measured in YOWOV3. With that, this work proposes performance comparison to YOWOV3 by comparing YOLOv8 and YOLOv11 to YOWOV3 and changing dataloader between Torchvision+PIL dan Albumentations. This contributes to YOWOV3 development for performance efficiency. The evaluation uses the MSCOCO dataset for YOLO and the AVAv2.2 dataset for YOWOV3, ensuring thorough benchmarking across varied activity domains. Results show that the comparison has significant reduction in YOWOV3 training time with, from 1:10:00 to 0:48:35, using Albumentations and changing from YOLOv8 to YOLOv11. However, there is a tradeoff where in 9 epochs YOWOV3 using YOLOv11 has higher loss at 8.0369, compared with YOLOv8 at 7.933. This gives insight not every efficient dataloader can improve model performance, as each dataloader has different integration process between model architecture and hardware.

Keywords—Computer vision, object detection, human activity recognition, HAR, spatiotemporal, Ultralytics.