KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer)
Volume 9, Nomor 1, November 2025
DOI: 10.30865/komik.v9i1.9567

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 28-35

Jaringan CNN 3D Berorientasi Konteks Untuk Mengenali Aksi Dengan Memanfaatkan Segmentasi Semantik (CARS)

Kevin H. Hutahaean[1,*], Sony Bahagia Sinaga[2], Chandra Frenki Sianturi[1]

[1]Fakultas Ilmu Komputer, Program Studi Teknologi Informasi, Universitas Budidarma, Medan
[2]Program Studi Manajemen Informatika, STMIK Mulia Darma, Labuhan Batu
Email : [1,*]kevinhutahaean@gmail.com, [2]sonybahagia@gmail.com, [3]chandrafrenki83sianturi@gmail.com
(* kevinhutahaean@gmail.com

Abstrak - Pengenalan aksi manusia menjadi topik penting dalam bidang visi komputer karena beragam aplikasinya, seperti pengawasan, interaksi manusia–komputer, dan sistem otonom. Walaupun metode CNN 3D terbaru mampu menangkap informasi spasial dan temporal dengan hasil yang cukup baik, pendekatan ini masih menghadapi kendala dalam memanfaatkan konteks lingkungan tempat aksi berlangsung. Keterbatasan tersebut mengurangi kemampuannya dalam membedakan aksi yang mirip serta mengidentifikasi skenario rumit secara lebih akurat. Untuk mengatasi permasalahan tersebut, penelitian ini mengusulkan pendekatan baru yang disebut Context-aware 3D CNN for Action Recognition based on Semantic Segmentation (CARS). Metode CARS mencakup modul pengenal adegan intermediari yang memanfaatkan model segmentasi semantik guna mengekstraksi petunjuk kontekstual dari rangkaian video. Informasi kontekstual tersebut kemudian direpresentasikan dan digabungkan dengan fitur yang dipelajari oleh model 3D CNN, sehingga terbentuk peta fitur global yang lebih kaya. Selain itu, CARS memasukkan Convolutional Block Attention Module (CBAM), yang menerapkan mekanisme atensi kanal dan spasial untuk menyoroti bagian paling penting dari peta fitur 3D CNN. Peneliti juga mengganti fungsi kerugian entropi silang konvensional dengan focal loss, yang lebih efektif dalam menangani kelas tindakan manusia yang jarang muncul dan sulit dibedakan. Serangkaian eksperimen pada berbagai dataset benchmark terkenal, seperti HMD51 dan UCF101, menunjukkan bahwa metode CARS yang diusulkan mampu melampaui kinerja pendekatan mutakhir berbasis 3D CNN. Selain itu, modul ekstraksi konteks dalam CARS bersifat generik dan plug-and-play, sehingga dapat meningkatkan akurasi klasifikasi pada berbagai arsitektur 3D CNN.
Kata Kunci : CNN 3D; Segmentasi Semantik; CARS; Computer Vision; CBAM; Context Aware

Abstract- Human action recognition has become an important topic in the field of computer vision due to its wide range of applications, such as surveillance, human–computer interaction, and autonomous systems. Although recent 3D CNN methods are able to capture spatial and temporal information with fairly good results, this approach still faces challenges in leveraging the environmental context in which an action occurs. These limitations reduce its ability to distinguish between similar actions and to identify complex scenarios more accurately. To address this problem, this study proposes a new approach called Context-aware 3D CNN for Action Recognition based on Semantic Segmentation (CARS). The CARS method includes an intermediate scene-recognition module that uses a semantic segmentation model to extract contextual cues from video sequences. This contextual information is then represented and fused with the features learned by the 3D CNN model, resulting in a richer global feature map. In addition, CARS incorporates the Convolutional Block Attention Module (CBAM), which applies channel and spatial attention mechanisms to emphasize the most important parts of the 3D CNN feature map. The researchers also replace the conventional cross-entropy loss function with focal loss, which is more effective in handling rare and hard-to-distinguish human action classes. A series of experiments on several well-known benchmark datasets, such as HMD51 and UCF101, show that the proposed CARS method outperforms state-of-the-art 3D CNN-based approaches. Moreover, the context extraction module in CARS is generic and plug-and-play, enabling it to improve classification accuracy across various 3D CNN architectures.
Keywords : 3D CNN; Semantic Segmentation; CARS; Computer Vision; CBAM; Context-Aware Learning.

1. PENDAHULUAN

Pengenalan aksi manusia merupakan topik penelitian yang penting dalam bidang visi komputer karena mencakup berbagai aplikasi bermanfaat, seperti interaksi manusia–komputer, analisis olahraga, sistem pemantauan kesehatan, serta sistem pengawasan [1]. Tujuan utamanya adalah mengidentifikasi dan mengklasifikasi aksi manusia dalam video secara tepat, sambil menghadapi beragam tantangan seperti pergerakan kamera, oklusi, perubahan pencahayaan, keterbatasan visibilitas tubuh, serta kerumitan kondisi lingkungan. Pendekatan-pendekatan awal dalam pengenalan aksi manusia umumnya bertumpu pada fitur yang dirancang secara manual serta algoritma pembelajaran dangkal [2]. Meskipun metode tersebut dapat bekerja dengan cukup baik dalam kondisi tertentu, pendekatan ini sering tidak mampu menangkap informasi spasial-temporal secara mendalam, terutama pada skenario yang lebih kompleks. Akibatnya, kemampuan generalisasi metode tradisional menjadi terbatas, khususnya pada dataset dan situasi yang menuntut pemahaman konteks lingkungan untuk membedakan aksi yang memiliki kemiripan. Kemunculan metode pembelajaran mendalam, khususnya Jaringan Saraf Konvolusional 2D (2D CNN), memberikan kemajuan besar melalui penerapan konvolusi dua dimensi untuk mengekstraksi

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 28-35

fitur spasial dari setiap frame video secara terpisah. Untuk merepresentasikan informasi temporal, jaringan berulang seperti Long Short-Term Memory sering diterapkan [3]. Namun demikian, model-model ini masih kesulitan menggabungkan fitur spasial dan temporal secara optimal, serta kurang mampu memodelkan ketergantungan jangka panjang maupun dinamika yang berubah-ubah pada rangkaian aksi yang kompleks. Sejalan dengan kemajuan pada CNN 2D, CNN 3D berkembang menjadi pendekatan yang lebih unggul untuk tugas pengenalan aksi. Model awal yang berpengaruh dalam kategori ini adalah Convolutional 3D Network (C3D) [4], disusul oleh arsitektur seperti Inflated 3D Convolutional Networks (I3D) [5] dan jaringan residual R(2+1)D, yang memisahkan konvolusi spasial 2D dan konvolusi temporal 1D [6]. Pendekatan-pendekatan tersebut mendorong kemajuan bidang ini dengan menggabungkan informasi spasial dan temporal secara lebih efektif dalam pemrosesan video. Selain itu, metode terkini seperti Grad-CAM + GRU, SegNet + BiGRU, dan AI-HAR juga menunjukkan performa yang kompetitif dalam tugas pengenalan aksi [7]. Namun, berbagai pendekatan tersebut belum secara eksplisit memanfaatkan informasi konteks di sekeliling tindakan, padahal hal ini sering kali penting untuk membedakan aksi yang tampak serupa secara visual. Kondisi tersebut menunjukkan kebutuhan akan model yang lebih peka terhadap konteks guna meningkatkan akurasi pengenalan dalam lingkungan yang kompleks.

Pendekatan yang peneliti ajukan, yaitu Context-aware 3D CNN for Action Recognition based on Semantic Segmentation (CARS), dirancang untuk meningkatkan kinerja metode mutakhir dalam mengenali aksi manusia pada berbagai kondisi dan beragam dataset dengan memanfaatkan informasi kontekstual serta mekanisme atensi yang lebih canggih. Secara khusus, peneliti berfokus pada pengembangan arsitektur berbasis 3D CNN guna meningkatkan tingkat akurasi dalam tugas pengenalan aksi. Peneliti juga mengusulkan pemanfaatan segmentasi semantik untuk mengekstraksi informasi kontekstual dan menggabungkannya ke dalam arsitektur 3D CNN. Pendekatan ini memungkinkan CARS tidak hanya memusatkan perhatian pada aksi yang diamati, tetapi juga mempertimbangkan keterkaitannya dengan lingkungan sekitar, sehingga meningkatkan akurasi pengenalan aksi pada beragam skenario yang kompleks. CARS merupakan arsitektur bertingkat yang memadukan ekstraksi konteks, mekanisme atensi, serta fungsi kerugian yang lebih efektif. Model segmentasi semantik One Former digunakan untuk melakukan segmentasi adegan dan memperoleh informasi kontekstual yang relevan [8].

Proses ini menghasilkan vektor biner yang menandai keberadaan atau ketiadaan informasi tertentu yang kemudian diintegrasikan ke dalam arsitektur 3D CNN. Integrasi tersebut memperkaya pemahaman model terhadap konteks lingkungan serta urutan aksi. Selain itu, untuk memperkuat representasi fitur, kami turut memasukkan Convolutional Block Attention Module (CBAM) [9] ke dalam keseluruhan metode CARS, sehingga mekanisme atensi kanal dan spasial dapat dimanfaatkan secara optimal. Pendekatan ini membantu model memusatkan perhatian pada bagian data yang paling informatif, sehingga meningkatkan kemampuannya dalam membedakan tindakan yang tampak serupa secara visual (misalnya minum vs makan atau menendang vs. berjalan). Selain itu, kami menerapkan focal loss untuk menangani kelas-kelas yang sulit diklasifikasikan, terutama yang muncul akibat ketidakseimbangan kelas [10]. Dalam kasus tersebut, tindakan sering kali memiliki kemiripan tinggi dan hanya menunjukkan perbedaan kecil dalam konteks lingkungan maupun gerakan tubuh, sehingga menantang untuk dibedakan dengan tepat. Sistem yang kami ajukan diuji pada dua dataset acuan, yaitu HMDB51 [11] dan UCF101, untuk menilai efektivitas serta kemampuan generalisasinya pada berbagai arsitektur CNN seperti R(2+1)D [12], I3D, dan ResNeXt-101 [13]. Hasil evaluasi menunjukkan bahwa pendekatan CARS berhasil mencapai akurasi tertinggi pada dataset pengenalan aksi manusia dan memberikan peningkatan signifikan terhadap performa model CNN yang telah ada untuk tugas pengenalan tindakan. Selain itu, integrasi CARS ke dalam arsitektur yang sudah ada tidak hanya meningkatkan performanya, tetapi juga menghasilkan capaian terbaik pada dataset HMDB51 dan UCF101, melampaui pendekatan kuat terkini seperti Grad-CAM + GRU [14], SegNet + BiGRU, serta AI-HAR [15]. Temuan ini menegaskan peran krusial informasi konteks dan membuka peluang penelitian selanjutnya untuk mengeksplorasi serta memanfaatkan konteks yang lebih kaya guna meningkatkan efektivitas model.

2. METODE PENELITIAN

Jaringan saraf konvolusional tiga dimensi (3D CNN) telah memberikan kemajuan besar dalam pemahaman video dengan memadukan informasi spasial dan temporal yang diekstraksi dari data video. Pada tahap awal, 2D CNN digunakan untuk menganalisis fitur spasial dalam video, sedangkan fitur temporal diperoleh melalui metode seperti optical flow. Pendekatan ini dikenal sebagai arsitektur dua aliran, yang menggabungkan kedua jenis informasi tersebut untuk tugas pengenalan video. Aliran kedua merupakan aliran temporal yang memanfaatkan optical flow untuk menangkap informasi gerakan antar frame. Hasil dari kedua aliran tersebut kemudian digabungkan melalui teknik late fusion guna menyatukan informasi visual dan isyarat pergerakan, sehingga menghasilkan pemahaman tindakan yang lebih menyeluruh. Peningkatan kinerja yang diberikan oleh pendekatan ini menjadi dasar bagi berbagai metode lanjutan dalam pengenalan aksi [16]. Meski demikian, model C3D memerlukan biaya komputasi yang besar dan mudah mengalami overfitting. Selain itu, penggunaan jendela waktu yang tetap membuatnya kurang mampu menangkap ketergantungan temporal jangka panjang. Oleh karena

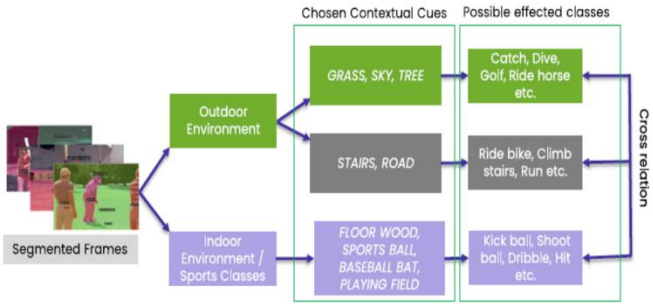
ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 28-35

itu, sejumlah pengembangan lanjutan telah dilakukan. Salah satunya adalah memodifikasi arsitektur ResNet agar dapat beroperasi dalam ruang tiga dimensi untuk memperkuat representasi fitur melalui analisis spasio-temporal. Modifikasi ini melibatkan penggantian lapisan konvolusi 2D dengan konvolusi 3D, sehingga jaringan dapat memproses dan memahami rangkaian video, bukan hanya frame per frame. Dengan memperluas kedalaman dan kompleksitas ResNet ke ranah temporal, kemampuan ResNet 3D dalam mempelajari pola gerakan dan dinamika waktu meningkat secara signifikan, sebagaimana ditunjukkan pada model seperti ResNeXt-101 [17]. Sejumlah pendekatan terbaru terus mendorong peningkatan kinerja dengan menyempurnakan arsitektur 3D CNN. Multi-Scale Receptive Fields Convolutional Network memperkenalkan penggunaan kernel dengan beragam skala untuk memodelkan aksi secara lebih efektif pada berbagai resolusi spasial dan temporaL [18].

Metode yang dikenal sebagai ActionVLAD memanfaatkan kemampuan grafik relasional untuk memperdalam pemahaman terhadap aksi dengan mengekspresikan hubungan antara objek dan agen dalam sebuah adegan. Namun, karena pendekatan ini sangat bergantung pada pemodelan relasi, ActionVLAD berpotensi kurang sensitif terhadap perubahan gerakan berurutan yang biasanya ditangkap oleh 3D CNN, sehingga beberapa fitur penting terkait dinamika gerak dapat terlewatkan. Dibandingkan dengan 3D CNN, pendekatan berbasis konteks ini sering menghadapi beban komputasi yang lebih besar, kesulitan dalam menggabungkan berbagai jenis data, serta tantangan dalam mengatur parameter atensi secara optimal. Sementara itu, model multi-stream Inflated 3D ConvNet (I3D), yang semakin memperluas pemanfaatan informasi kontekstual, menawarkan alternatif lain dalam tugas pengenalan aksi [19].

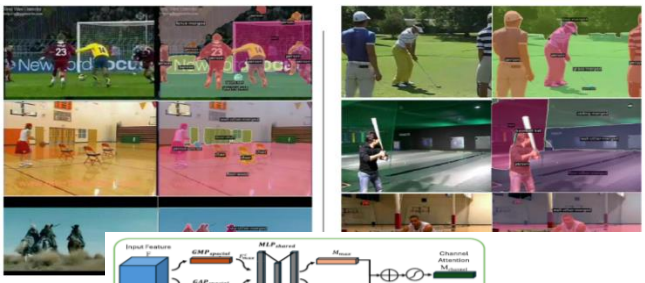
Informasi tersebut membantu mengidentifikasi jenis objek serta hubungan semantik antar-objek, yang merupakan aspek penting dalam memahami konteks untuk pengenalan aksi. Untuk melakukan segmentasi semantik, kami memanfaatkan OneFormer [20], sebuah model terdepan yang dikenal berkat kerangka terpadu dan kinerja unggulnya. Model ini sangat efektif dalam mengekstraksi informasi semantik pada level piksel, dan menunjukkan hasil yang lebih baik dibandingkan Mask R-CNN [21].

Selain itu, keterkaitan semantis antar kelas tindakan dapat muncul pada berbagai kondisi lingkungan. Sebagai contoh, seseorang dapat menggiring bola baik di area terbuka seperti taman bermain maupun di ruang tertutup seperti lapangan basket; demikian pula, aktivitas menaiki tangga dapat berlangsung di dalam maupun di luar ruangan. Dengan mengekstraksi petunjuk kontekstual yang berkaitan dengan lingkungan misalnya keberadaan RUMPUT, LANGIT, atau POHON pendekatan CARS kami menitikberatkan pada informasi lingkungan yang relevan, tanpa bergantung pada objek atau lokasi tertentu. Ketika seseorang menggiring bola di luar ruangan, CARS dapat mengenali keberadaan bola sekaligus menangkap petunjuk lingkungan seperti GRASS atau SKY melalui segmentasi semantik, meskipun tidak ada lapangan basket atau marka taman bermain yang tampak. Informasi kontekstual tersebut membantu model mengidentifikasi aksi yang berkaitan dengan aktivitas bola (kelas terkait) dan meningkatkan kemampuannya dalam membedakan dribbling dari tindakan lain yang mungkin serupa, seperti berlari atau berjalan. Secara umum, aktivitas tersebut tidak menampilkan petunjuk kontekstual khas lingkungan luar ruangan. Sebaliknya, aktivitas seperti bermain golf atau menyelam hampir selalu berlangsung di ruang terbuka, sehingga keberadaan isyarat lingkungan luar ruangan menjadi penanda yang lebih konsisten dan dapat diandalkan.



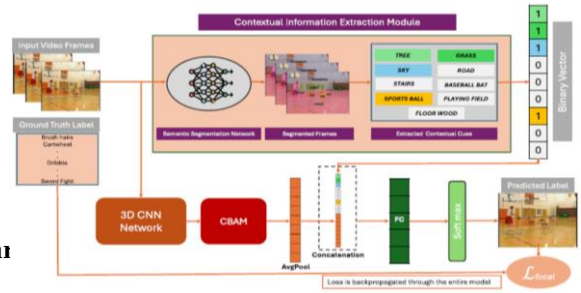
Gambar 1. Pengaruh informasi kontekstual terhadap klasifikasi aksi

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 28-35



Gambar 2. Hasil prediksi ataset HMDB51

Sebagaimana diperlihatkan pada Gam dialirkan ke Convolutional Block Attention Module (CBAM) ekanisme atensi kanal dan perhatian spasial [22]. Penyesuaian optimal pada data video, di mana pemahaman terhadap interaksi sangat diperlukan. CBAM berfungsi dengan melakukan rekalib sme atensi kanal dan atensi spasial.



Gambar g diajukan

Gambar 4. Struktur dari Modul Attention Berbasis Blok Konvolusional (CBAM)

Pada tahap akhir, sistem menghasilkan label tindakan yang diprediksi. Untuk menangani ketidakseimbangan kelas pada dataset evaluasi, peneliti menerapkan *focal loss*, yang sebelumnya terbukti efektif dalam tugas deteksi objek serta segmentasi instans [23]. Peneliti menemukan bahwa beberapa kelas sulit dibedakan karena kemiripan antar aksi dan distribusi kelas yang tidak merata. Dengan memanfaatkan *focal loss*, kontribusi dari sampel yang mudah diprediksi diperkecil, sehingga model terdorong memberi perhatian lebih pada kelas-kelas yang sulit. Pendekatan ini membantu menurunkan tingkat misklasifikasi secara keseluruhan.

3. HASIL DAN PEMBAHASAN

Kerangka CARS diimplementasikan menggunakan Python 11 dan memanfaatkan PyTorch sebagai pustaka pembelajaran mendalam sumber terbuka. Model dilatih selama 200 *epoch* dengan laju pembelajaran awal 1e-2 dan ukuran *batch* 8. Kami menggunakan pengoptimal AdamW dengan momentum 0,9 serta *weight decay* sebesar 1e-3. Ukuran iterasi ditetapkan pada 6 untuk membatasi penggunaan memori, dan progres pelatihan dicatat setiap 200 iterasi. Proses pelatihan dijalankan pada GPU NVIDIA GeForce RTX 2080i 11+11

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 28-35

yang tersebar di tiga sistem Alienware Aurora, masing-masing dilengkapi RAM 64 GB. Peneliti melakukan variasi jumlah bingkai yang digunakan untuk membentuk vektor konteks. Tabel 1 menunjukkan bahwa akurasi mulai mencapai titik jenuh pada rentang 20–30 bingkai. Pada dataset HMDB51, penggunaan 20 bingkai menghasilkan akurasi 81,70%, hampir sama dengan 81,72% ketika seluruh bingkai digunakan. Pada UCF101, akurasi tertinggi 97,17% juga dicapai dengan 20 bingkai. Jika jumlah bingkai kurang dari 15, performa menurun karena sejumlah isyarat penting tidak tertangkap. Sebaliknya, menambah bingkai lebih dari 30 tidak memberikan peningkatan akurasi, tetapi justru menambah biaya inferensi.

Tabel 1. Ablasi terhadap jumlah bingkai yang digunakan dalam segmentasi semantik untuk membentuk konteks 9-bit dilakukan menggunakan tulang punggung R(2+1)D dengan *cross-entropy loss*.

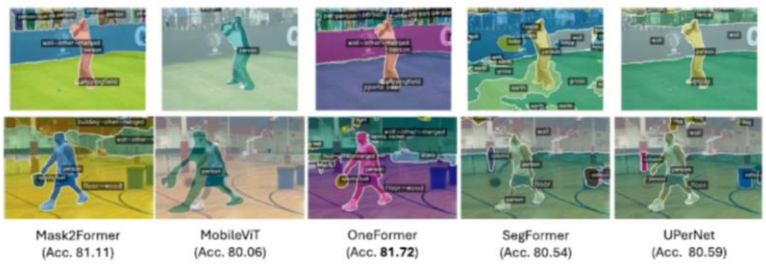
Frame used	Top-1 (%) HMDB51	Top-1 (%) UCF101	Seconds/video
All (64)	81.72	97.17	10.81
30	81.70	97.17	5.07
20	81.70	97.17	3.38
15	81.58	97.01	2.53
10	81.34	96.83	1.69

Dengan OneFormer sebagai penyedia fitur kontekstual dan R(2+1)D sebagai *backbone*, latensi per bingkai tercatat 169,0 ms atau setara 5,92 fps. Pemrosesan 20 bingkai memerlukan 3,38 detik per video, sedangkan 30 bingkai membutuhkan 5,07 detik peningkatan waktu proses sebesar 49% tanpa adanya keuntungan dalam akurasi. Temuan ini menegaskan bahwa penggunaan 20 frame memberikan keseimbangan paling optimal antara *throughput* dan akurasi pengenalan. Jumlah tersebut terbukti memadai untuk menangkap isyarat yang stabil sekaligus mempertahankan efisiensi waktu pemrosesan. Setelah itu, kami mengevaluasi hiperparameter *focal loss* dengan CBAM diaktifkan dan konteks 9-bit dipertahankan. Tabel 2 membandingkan tiga konfigurasi. Pengaturan S menggunakan atensi spasial, C menggunakan atensi kanal, sedangkan SC memadukan keduanya. Pada HMDB51, konfigurasi SC mencapai akurasi 83,66%, lebih tinggi dibandingkan C dengan 82,94% dan S dengan 82,61%. Pada UCF101, SC kembali memberikan performa terbaik dengan 97,17%, melampaui C yang memperoleh 96,91% dan S yang mencapai 96,67%. Temuan ini mengindikasikan bahwa S dan C menangkap informasi yang saling melengkapi: C menyesuaikan kembali respons fitur dengan menekankan filter yang paling penting, sementara S menekankan area informatif pada tiap bingkai.

Tabel 2. Ablasi CBAM dilakukan pada split-1 HMDB51 dan UCF101 dengan menggunakan R(2+1)D sebagai tulang punggung model.

CBAM Configuration	Top-1 (%) HMDB51	Top-1 (%) UCF101
All (64)	81.72	97.17
30	81.70	97.17
20	81.70	97.17
15	81.58	97.01
10	81.34	96.83

Terakhir, peneliti men k bersifat diskriminatif seperti {KURSI, PINT tika kategori {KURSI, PINTU, DINDING} di gan 20 frame dan *focal loss*, akurasi menurun menjadi 96,51% pada UCF101. Hasil ini me lerau dan menurunkan performa pengenalan, evan dengan tindakan. Pemanfaatan OneForm asilkan akurasi Top-1 tertinggi sebesar 81,72 Gambar 5.



ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 28-35

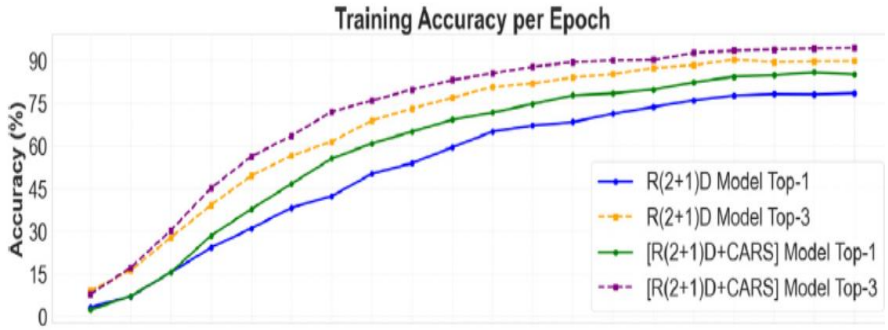
Gambar 5. Evaluasi segmentasi semantik SOTA terhadap akurasi Top-1 pengenalan tindakan manusia

Tabel 3 merangkum hasil akurasi Top-1. Pada HMDB51, empat konfigurasi pertama memanfaatkan *cross-entropy loss*, sedangkan konfigurasi terakhir menggunakan *focal loss* saja. Konfigurasi pertama menambahkan vektor biner-3 ke baseline [R(2+1)D] berisi isyarat kontekstual JALAN, TANGGA, dan PEMUKUL BISBOL dan menghasilkan peningkatan akurasi sebesar 0,59%, menegaskan efektivitas pendekatan ini. Konfigurasi kedua memperluasnya menjadi vektor biner-6 dengan menambahkan BOLA OLAHRAGA, LAPANGAN BERMAIN, dan LANTAI KAYU, yang memberikan peningkatan kinerja sebesar 0,79%. Konfigurasi ketiga, yaitu vektor biner-9, kembali meningkatkan performa sebesar 0,26% melalui penambahan isyarat kontekstual SKY, GRASS, dan TREE. Secara total, konfigurasi ini memberikan peningkatan kinerja kumulatif sebesar 1,64%

Tabel 3. Dampak segmentasi semantik SOTA pada akurasi pengenalan tindakan

#Model	Configuration	#Contextual cues	HMDB51 Top1	UCF 101 Top 1
R(2+1)D	Baseline	None	80.06%	96.06%
R(2+1)D	Config#1	Bin 3	80.65%	96.38%
R(2+1)D	Config#2	Bin 6	81.44%	96.56%
R(2+1)D	Config#3	Bin 9	81.70%	96.82%
R(2+1)D+CBAM	Config#4	Bin 9	83.01%	97.04%
R(2+1)D+CBAM	Config#5	Bin 9	83.66%	97.175

Pada gambar 6 memperlihatkan bahwa masukan diskalakan menjadi 224 × 224 piksel, dan klip video 64 bingkai digunakan sebagai input. Meskipun konfigurasi ini membutuhkan sumber daya komputasi lebih besar karena jumlah bingkai dan ukuran citra yang tinggi, pendekatan tersebut memungkinkan model menangkap fitur spasiotemporal dengan lebih kaya dan detail.



Gambar 6. Perbandingan pelatihan R(2+1)D dan R(2+1)D+CARS

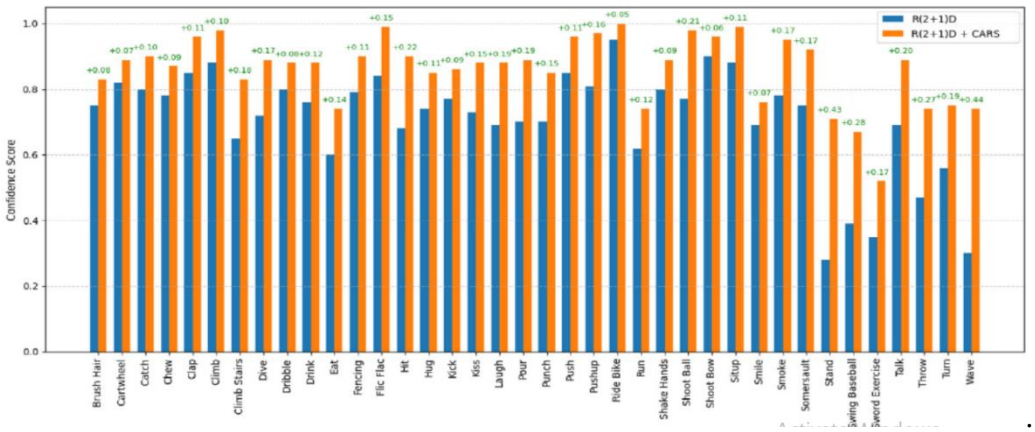
Hasil pada HMDB51 dan UCF101 dalam tabel 4 merupakan rata-rata menggunakan resolusi 112 × 112 dan klip video berdurasi 32 bingkai. Model ini lebih efisien secara komputasi dibandingkan I3D karena memanfaatkan jumlah bingkai yang lebih sedikit serta ukuran bingkai yang lebih kecil.

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 28-35

Tabel 4. Integrasi CARS meningkatkan kinerja model CNN 3D

Model	Pre-trained dataset	Frames	Params (M)	FLOPs (G/clip)	FPS	UCF101 acc.	HMDB51 acc.
ResNeXt-101	Kinetics	64	47.63	38.67	112.77	94.5%	70.2%
ResNeXt-101+CARS	Kinetics	64	48.15	38.68	110.10	95.57%	**74.44%**
Improvement							
I3D	Kinetics	64	12.34	111.42	59.90	95.4%	74.5%
I3D+CARS	Kinetics	64	12.47	111.42	59.31	96.62%	75.83%
Improvement							
R(2+1) D	Kinetics	32	63.52	153.15	49.67	96.8%	74.5%
R(2+1) D+CARS	Kinetics	32	63.55	153.15	49.01	**97.4%**	77.05%
Improvement							
R(2+1) D	IG65M	32	63.52	153.15	49.67	96.94%	80.05%
R(2+1) D+CARS	IG65M	32	63.55	153.15	49.01	97.65%	81.3%

Meskipun peningkatan pada UCF101 relatif kecil, yaitu sebesar 0,36%, hasil tersebut tetap signifikan karena akurasi pada dataset ini sudah hampir mencapai titik jenuh, sehingga sulit untuk memperoleh kemajuan lebih lanjut. Dengan secara konsisten meraih performa terbaik pada kedua protokol evaluasi dan mengungguli berbagai model SOTA, metode yang kami usulkan membuktikan ketangguhan, skalabilitas, dan kemampuan adaptasinya dalam pengenalan tindakan spasio-temporal. Secara khusus, meskipun hanya menggunakan kerangka kerja aliran tunggal berbasis RGB, pendekatan kami mampu melampaui sejumlah metode multimodal dan arsitektur yang lebih kompleks, menegaskan efisiensi dan efektivitasnya



Gambar 7. Perbandingan skor keyakinan R(2+1)D dan CARS

4. KESIMPULAN

Berdasarkan metodologi menunjukkan bahwa dengan memadukan informasi kontekstual melalui segmentasi semantik serta menerapkan mekanisme atensi CBAM dan *focal loss* untuk menangani ketidakseimbangan kelas, CARS mampu meningkatkan performa model 3DCNN secara signifikan dalam tugas pengenalan aksi manusia (HAR). Berdasarkan pengetahuan terbaik peneliti, penelitian ini mencapai hasil terkini dan melampaui model-model 3D CNN berbasis RGB yang telah ada untuk HAR. Temuan tersebut menegaskan bahwa pemanfaatan informasi kontekstual merupakan arah yang menjanjikan untuk kemajuan pengenalan tindakan manusia. Dengan mengintegrasikan deteksi objek dan teknik leverage-aging seperti difusi stabil, data gambar-ke-teks dapat dihasilkan sebagai aliran input tambahan. Pendekatan multi-moda ini memungkinkan model untuk memanfaatkan isyarat visual dan semantik, sehingga menghasilkan representasi fitur yang lebih kaya. Dengan memusatkan perhatian pada *region of interest* (ROI), informasi yang diperoleh dapat menjadi lebih relevan. Dalam konteks pengenalan tindakan manusia, alih-alih memproses seluruh bingkai secara keseluruhan, memisahkan subjek manusia dari latar belakang akan jauh lebih efektif.

ISSN 2597-4645 (media online)
ISSN 2597-4610 (media cetak)
Page: 28-35

REFERENCES

[1] Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[2] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
[3] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[4] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., & Zisserman, A. (2018). A short note on the Kinetics-700 dataset. arXiv preprint arXiv:1812.02142.
[5] Girdhar, R., et al. (2019). Video action transformer network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[6] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
[7] Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? Proceedings of the International Conference on Machine Learning (ICML).
[8] Liu, Z., et al. (2022). Video Swin transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[9] Lin, T.-Y., et al. (2018). Feature pyramid networks for object detection (extended version). IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
[10] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLabv3+: Encoder–decoder with atrous separable convolution for semantic image segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[11] Zhao, H., et al. (2019). Pyramid scene parsing network (PSPNet) enhanced edition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
[12] Sun, P., et al. (2019). Optical flow guided feature: A motion representation for video action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[13] Kuehne, H., Arslan, A., & Serre, T. (2018). The language of actions: Encoding human action semantics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[14] Stroud, J. C., et al. (2020). D3D: Distilled 3D networks for video action recognition. Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV).
[15] Wu, C.-Y., et al. (2019). Long-term feature banks for detailed video understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[16] Piergiovanni, A., & Ryoo, M. S. (2018). Learning latent super-events to detect multiple activities in videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[17] Cheng, B., et al. (2020). Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[18] Zhu, X., et al. (2020). Deformable DETR: Deformable transformers for end-to-end object detection. Proceedings of the International Conference on Learning Representations (ICLR).
[19] Wang, J., et al. (2020). Deep high-resolution representation learning for visual recognition (HRNet). IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
[20] Li, Y., et al. (2021). Semantic segmentation with context encoding networks (expanded). IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
[21] Arnab, A., et al. (2018). Conditional similarity networks for video action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.
[22] Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI).
23] Xu, M., et al. (2022). Adaptive video recognition via spatiotemporal feature modulation. Proceedings of the European Conference on Computer Vision (ECCV).