Computatio: Journal of Computer Science and Information Systems, 6/1 (2022), 51 - 60

PENDETEKSIAN AKTIVITAS MANUSIA DENGAN HUMAN POSE ESTIMATION DAN CONVOLUTIONAL NEURAL NETWORK

Andrean Lay[1], Lina[2]
[1,2] Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara, Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
E-mail: [1]andreanlay1@gmail.com, [2]lina@fti.untar.ac.id

Abstrak

Kamera pengawasan aktivitas yang memanfaatkan teknologi artificial intelligence semakin ramai digunakan akhir-akhir ini. Namun, aplikasi teknologi ini masih terpusat pada subjek orang dewasa, hal ini menyebabkan kurangnya penelitian pada kamera pengawasan aktivitas pada anak-anak yang menyebabkan kurangnya ketersediaan dataset dan referensi. Berdasarkan penjelasan singkat di atas, maka dilakukan penelitian yang bertemakan human activity recognition yang memiliki fokus pada anak-anak berusia 4 hingga 6 tahun. Metode yang akan dipakai dalam penelitian ini adalah Human Pose Estimation dan Convolutional Neural Network (CNN) dengan data yang dikumpulkan dari internet. Aplikasi yang dirancang akan terlebih dahulu mendeteksi skeleton dari objek manusia yang kemudian akan diklasifikasi oleh CNN. Aktivitas yang dapat diklasifikasi oleh aplikasi yang dirancang adalah belajar, berdiri, dan tidur. Hasil keluaran aplikasi berupa catatan aktivitas yang disesuaikan dengan waktu aktivitas tersebut terdeteksi. Hasil pengujian confusion matrix menunjukan bahwa model yang di-latih memiliki nilai akurasi sebesar 97.77%, presisi sebear 97.96%, recall sebesar 97.73%, dan F1-score sebesar 97.83%.

Kata kunci*—Aktivitas anak-anak, Convolutional Neural Network, Human Pose Estimation, Pendeteksian Aktivitas.*

Abstract

In the last few years, artificial intelligence-based human activity monitoring system becomes more popular. However, most proposed research focused on adult subject performing the action. Thus, leaving a gap in child activity recognition which causes low dataset availability and reference. Based on this situation, the focus of this study is to reduce the gap in child activity recognition. The proposed system in this study focused on 4 – 6 years old child. The methods used are Human Pose Estimation with BlazePose and Convolutional Neural Network (CNN) with images dataset gathered from the internet. First, the skeleton will be estimated using BlazePose, the resulting skeleton will be converted to matrix form and given to CNN to be classified. There are 3 activites which can be detected, they are studying, standing, and sleeping. Each activity will be recorded to a logbook with its timestamp when the activity detected. Confusion matrix testing shows that trained model has accuracy value of 97.77%, precision of 97.96%, recall of 97.13%, and F1-score of 97.83%.

Keywords—*Children activity, Convolutional Neural Network, Human Pose Estimation, Activity Detection.*

1. PENDAHULUAN

Perkembangan teknologi kecerdasan buatan pada beberapa tahun terakhir sangatlah pesat. Pesatnya perkembangan ini disebabkan oleh meningkatnya kekuatan komputasi pada perangkat keras komputer serta meningkatnya ketersediaan data yang membuat deep learning semakin populer. Dengan berkembangnya deep learning, cabang-cabang kecerdasan buatan seperti computer vision ataupun natural language processing yang dulunya sulit dilakukan, menjadi lebih mudah dilakukan.

Salah satu bidang ilmu computer vision yang aktif diteliti adalah Human Activity Recognition (HAR). Aktifnya penelitian pada bidang ini disebabkan oleh luasnya aplikasi HAR pada kehidupan sehari-hari. Beberapa aplikasi HAR untuk kamera pengawasan, militer, dan sistem kesehatan memiliki dampak yang besar karena dapat mengatasi kelemahan sistem jika dilakukan secara manual oleh manusia.

Dalam HAR, terdapat dua teknik utama, yaitu teknik berbasis visual dan teknik berbasis sensor [1]. Teknik berbasis sensor kurang praktis untuk digunakan karena membutuhkan perangkat tambahan untuk memonitor aktivitas. Selain membutuhkan perangkat tambahan, teknik ini juga mengurangi kenyamanan pengguna karena diperlukan perangkat tambahan yang perlu ditempelkan ke badan. HAR berbasis visual telah banyak digunakan sebagai alternatif HAR berbasis sensor. Pada HAR berbasis visual, pendeteksian akan dilakukan langsung melalui citra hasil tangkapan kamera, tanpa memerlukan perangkat yang ditempelkan ke badan.

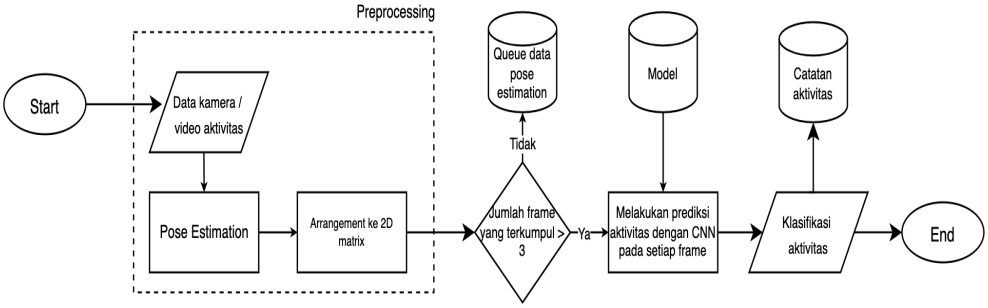
Penelitian-penelitian HAR berbasis visual seperti pada [2] menggunakan citra yang merupakan keluaran langsung dari perangkat kamera. Citra ini kemudian akan diproses oleh Convolutional Neural Network (CNN) untuk mengklasifikasikan aktivitas ke kelas-kelas tertentu. Sementara itu, beberapa riset lainnya seperti pada [3][4][5] menggunakan kamera RGB-D dengan *Depth Sensor* yang digunakan untuk menangkap data 3D *skeleton* objek manusia. Selain menggunakan titik-titik *skeleton* secara langsung, [6] melakukan transformasi hasil *pose estimation* ke dalam bentuk matriks dua dimensi untuk memanfaatkan hubungan spasial antar datanya.

Sebagian besar penelitian dalam HAR masih terfokus pada objek orang dewasa. Hal ini menyebabkan sulitnya HAR anak-anak yang dikarenakan beberapa hal seperti dataset yang terpusat pada aktivitas orang dewasa, serta kurangnya referensi yang ada. Oleh karena itu, sistem yang akan dirancang ini dibuat dengan fokus melakukan kontribusi pada penelitian HAR anak-anak.

Sistem yang dibuat ini akan menggunakan masukan citra dari kamera RGB yang kemudian akan dideteksi *skeleton* dua dimensinya. Koordinat-koordinat *skeleton* dua dimensi ini akan dikonversikan ke dalam bentuk matriks dua dimensi seperti pada [6] yang kemudian akan menjadi masukan model CNN yang telah dilatih. Sistem yang akan dirancang ini menargetkan tiga buah kelas aktivitas yaitu berdiri, belajar, dan tidur.

2. METODE PENELITIAN

Sistem yang dirancang menerima masukan berupa citra langsung dari kamera atau video. Citra yang didapat diproses oleh BlazePose untuk didapatkan *skeleton*nya yang kemudian akan dikonversikan ke dalam bentuk matriks dua dimensi untuk diprediksi oleh model CNN yang telah dilatih. Pada sistem ini, prediksi aktivitas akan dilakukan dengan merata-ratakan hasil keluaran CNN 3 *frame* terakhir dengan sistem pemilihan *frame* adalah *frame* pertama setiap detik. Alur kerja sistem dapat dilihat pada Gambar 1 di bawah ini.



Gambar 1 Diagram kerja sistem pendeteksian aktivitas

Metode penelitian yang dipakai akan dibahas secara berurutan sesuai dengan tahapan penelitian yang dilakukan. Tahapan penelitian yang dilakukan terdiri atas pengumpulan dataset aktivitas anak-anak, mengestimasi *skeleton* citra dengan Human Pose Estimation (HPE) BlazePose, dan melatih model CNN yang dirancang dengan data yang telah dibuat untuk mengklasifikasikan aktivitas.

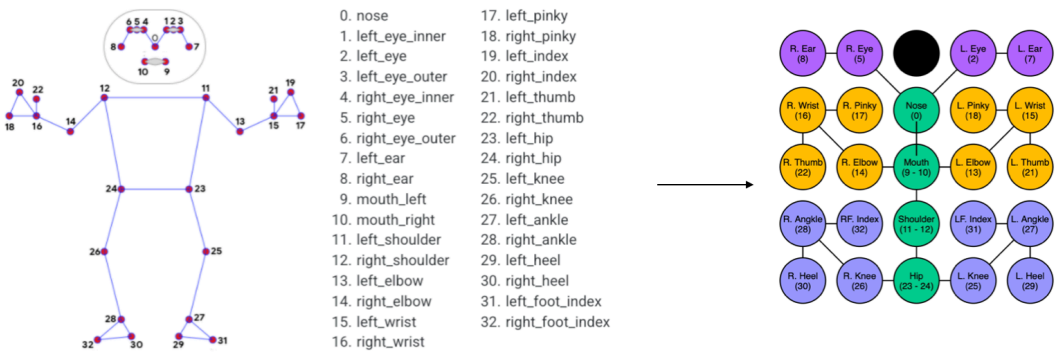
2.1 Pengumpulan Dataset

Dataset aktivitas anak-anak dikumpulkan dari internet dan diaugmentasi dengan operasi translasi, rotasi, dan *occlusion*. Aktivitas yang dikumpulkan adalah berdiri, belajar, dan tidur. Posisi duduk/bersila akan dianggap sebagai belajar walaupun sesungguhnya objek tidak sedang belajar. Dari data yang dikumpulkan, akan dibagi menjadi 75% data latih dan 25% data uji. Detail jumlah data per kelas dan pembagiannya dapat dilihat pada Tabel 1 di bawah ini.

Tabel 1 Pembagian data latih dan data uji

Jenis Kelas	Jumlah Data Latih	Jumlah Data Uji
Tidur	2861	521
Berdiri	5827	1005
Belajar	4826	859
Total	13514	2385

Setiap data citra aktivitas memiliki posisi dan pencahayaan yang bervariasi. Dari dataset yang telah didapat, akan diproses lebih lanjut dengan *Human Pose Estimation* (HPE) BlazePose untuk diekstrak data *skeleton*nya. Data skeleton yang terbentuk akan dikonversikan ke dalam bentuk matriks dua dimensi. Ilustrasi proses konversi ini dapat dilihat pada Gambar 2 di bawah ini.



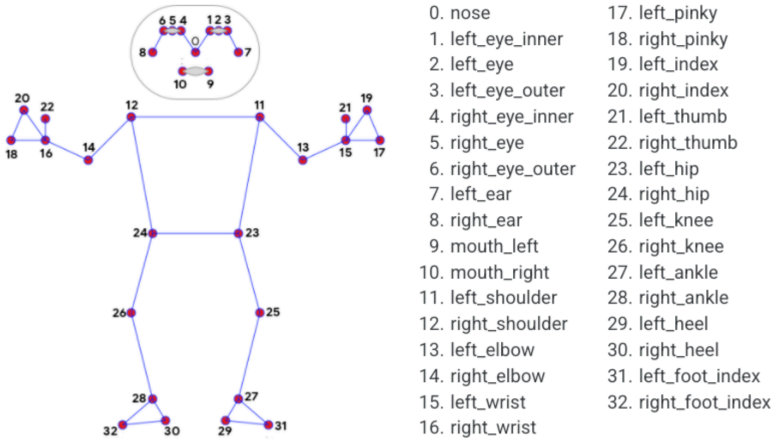
Gambar 2 Proses konversi data *skeleton* ke dalam bentuk matriks

2.2 Human Pose Estimation (HPE)

Metode HPE yang akan dipakai dalam sistem ini adalah BlazePose [7]. Arsitektur yang dipakai BlazePose untuk memprediksi *skeleton* dari object manusia adalah arsitektur berbasis *stacked hourglass* yang dimodifikasi. Sistem *encoder-decoder* dipakai untuk memprediksi *heatmap* dari semua sendi, yang kemudian diikuti oleh *encoder* untuk mengkalkulasi koordinat sendi tertentu.

BlazePose menggunakan sistem *detector-tracker* dimana setiap *frame* pertama, BlazePose akan mendeteksi *region-of-interest* (*ROI)* objek manusia, kemudian pada *frame* berikutnya BlazePose akan melakukan pelacakan objek manusia pada ROI sebelumnya. Jika objek manusia tidak terdeteksi, maka *detector* akan dijalankan kembali.

Detektor BlazePose mendeteksi objek manusia dengan cara mendeteksi bagian tubuh yang tetap, yaitu kepala. Sehingga, asumsi dasar BlazePose dalam mendeteksi manusia adalah bagian kepala harus terlihat pada *frame*. Jumlah titik yang akan dideteksi oleh BlazePose adalah 33 titik. Lokasi 33 titik dapat dilihat pada Gambar 3.



Gambar 3 Lokasi 33 *keypoints* BlazePose
(Sumber: https://google.github.io/mediapipe/solutions/pose.html)

Setiap titik yang dideteksi BlazePose akan memiliki 4 nilai yaitu koordinat x, y, z, dan nilai visibilitas. Nilai visibilitas didapat dari implementasi sistem klasifikasi visibilitas per titik yang akan mengindikasikan apakah sebuah titik terhalang ataupun yang memiliki tingkat akurasi rendah. Implementasi ini membuat BlazePose dapat mengestimasi titik-titik yang berada di luar *frame* ataupun terhalang sepenuhnya.

2.3 Convolutional Neural Network (CNN)

CNN atau yang sering disebut ConvNet merupakan salah satu jenis *deep neural network* yang menerima masukan berupa citra. CNN akan mempelajari bagian-bagian penting dari citra melalui *learnable weight* dan bias dari *kernel*. Yang membedakan ConvNet dari jaringan saraf tiruan biasa adalah ConvNet membutuhkan sedikit preprocessing pada data masukan, oleh karena itu ConvNet tidak memerlukan banyak fitur buatan karena kemampuan untuk mempelajari fitur-fitur kompleks jika dilengkapi dengan data yang cukup [8].

Secara umum, arsitektur CNN terdiri atas dua bagian yaitu *hidden layer* dan lapisan klasifikasi. Pada *hidden layer* akan dilakukan proses ekstraksi fitur-fitur. Fitur-fitur yang diekstrak ini akan dipakai oleh lapisan klasifikasi untuk diolah menjadi *output*. Cara CNN mengenali citra serupa dengan cara kerja penglihatan manusia, yaitu setiap lapisan Convolution

akan memproses citra dengan mengenali pola-pola yang ada pada citra mulai dari pola sederhana yang kemudian akan diteruskan ke lapisan yang lebih dalam untuk mengenali pola yang lebih kompleks. ConvNet juga dapat mengenali hubungan spasial dan temporal pada data. Arsitektur CNN yang akan dipakai pada sistem rancangan ini adalah arsitektur buatan sendiri. Detail arsitektur tersebut dapat dilihat pada Tabel 2.

Tabel 2 Arsitektur CNN yang dirancang

Jenis Lapisan	Filter/Unit	Size/Padding	Fungsi Aktivasi
Conv2D	32	(3, 3) / same	ReLU
Conv2D	32	(3, 3) / same	ReLU
Conv2D	64	(3, 3) / same	ReLU
Flatten	-	-	-
Dense	32	-	ReLU
Dense	3	-	Softmax

Model yang dirancang ini akan diuji dengan beberapa nilai *hyperparameter* dengan konfigurasi awal *batch size* sebesar 32, *epoch* sebesar 32, dan Adam optimizer dengan *learning rate* sebesar 0.0005. Tujuan pengujian dengan beberapa macam *hyperparameter* adalah untuk mencari *hyperparameter* terbaik yang dapat digunakan untuk pelatihan model CNN. Nilai-nilai *hyperparameter* yang akan diuji secara berurutan dapat dilihat pada Tabel 3.

Tabel 3 Nilai-nilai *hyperparameter* yang akan diuji

No	***Hyperparameter***	Nilai pengujian
1	Learning rate	0.01, 0.001, 0.0001
2	Batch size	16, 32, 64, 128
3	Epoch	32, 75, 100

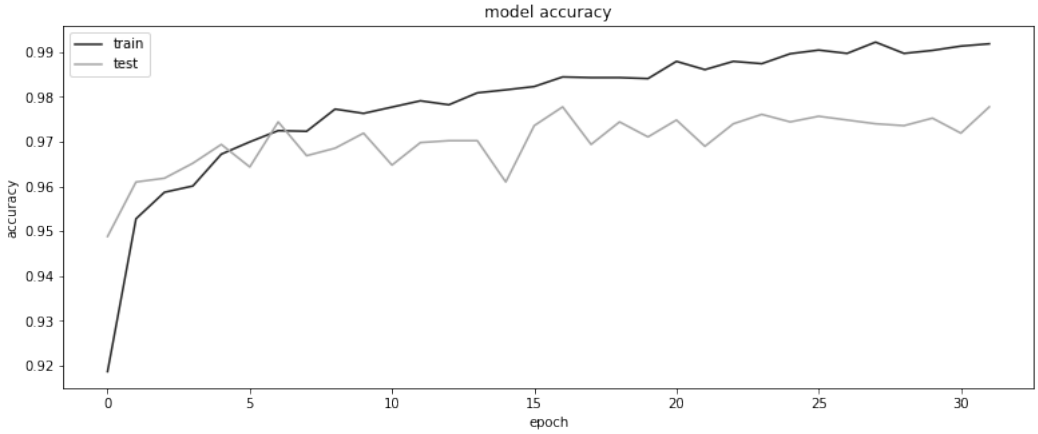
3. HASIL DAN PEMBAHASAN

Pengujian *hyperparameter* model dilakukan secara iteratif yaitu mulai dari *learning rate* (lr)*, batch size*, dan *epoch* dengan menggunakan pembagian data yang sama. Nilai akurasi validasi dan *loss* validasi terbaik dari pengujian *hyperparameter* pendahulu akan diambil dan dipakai pada pengujian nilai *hyperparameter* berikutnya. Hasil pengujian setiap *hyperparameter* dapat dilihat pada Tabel 4.

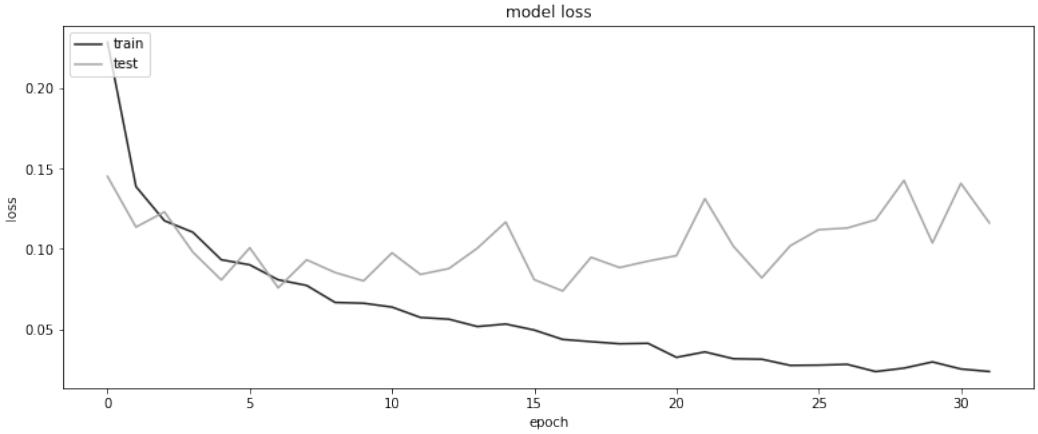
Tabel 4 Hasil pengujian nilai-nilai *hyperparameter*

Konfigurasi Hyperparameter			Akurasi Latih	Loss Latih	Akurasi Validasi	Loss Validasi
Learning rate	Batch size	Epoch				
0.01	32	32	96.60%	0.1047	95.22%	0.1519
0.001	32	32	99.14%	0.0251	97.65%	0.0935
0.0001	32	32	97.56%	0.0667	96.60%	0.0946
0.01	**16**	32	99.19%	0.0239	97.61%	0.1218
0.01	32	32	99.21%	0.0243	97.02%	0.1013
0.01	64	32	98.61%	0.0243	96.44%	0.0965
0.01	16	32	99.07%	0.0275	97.40%	0.0729
0.01	16	**32**	99.19%	0.0234	97.78%	0.1162
0.01	16	75	99.75%	0.0113	97.69%	0.1635
0.01	16	100	100%	1.67e-6	98.20%	0.1966

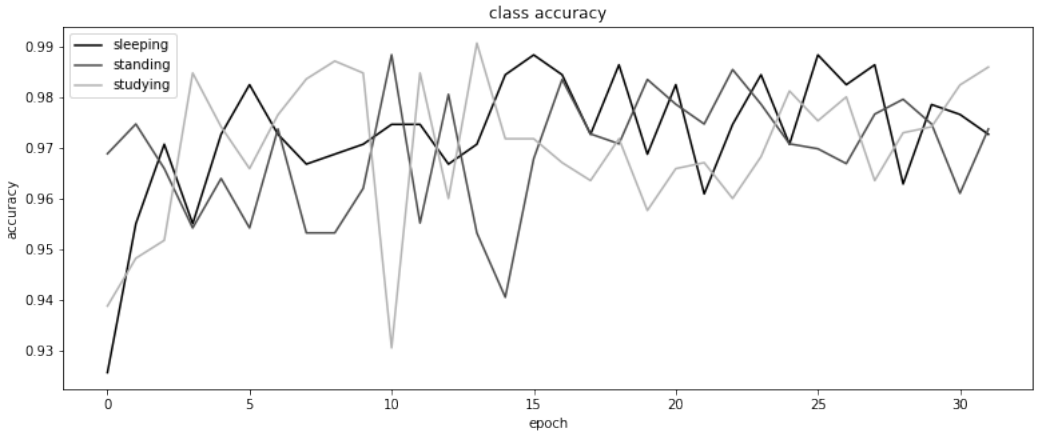
Berdasarkan hasil pengujian di atas dapat dilihat bahwa secara berurutan *hyperparameter* yang paling optimal adalah *learning rate* sebesar 0.001, *batch size* sebesar 16, dan *epoch* sebesar 32. Sehingga, konfigurasi *hyperparameter* ini akan dipakai untuk model akhir. Grafik akurasi, *loss*, serta *confusion matrix* dapat dilihat pada Gambar 4 hingga Gambar 8.



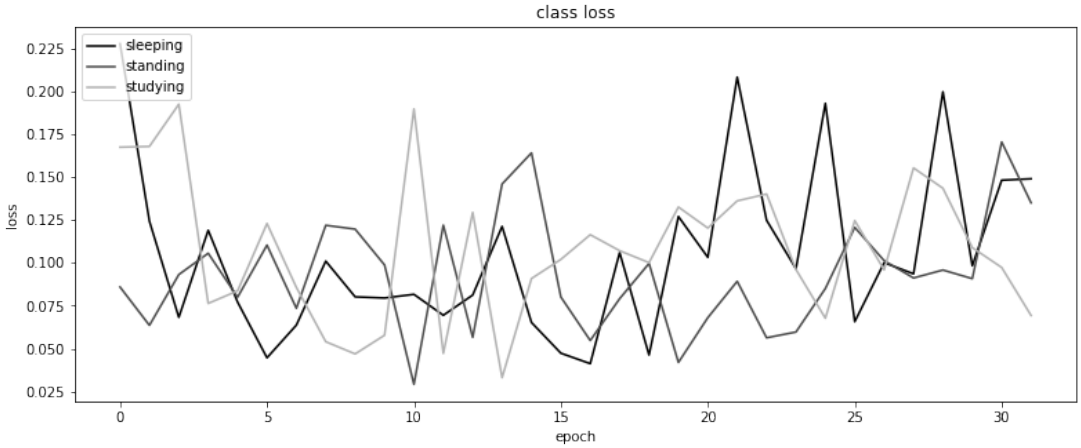
Gambar 4 Grafik akurasi model yang dirancang



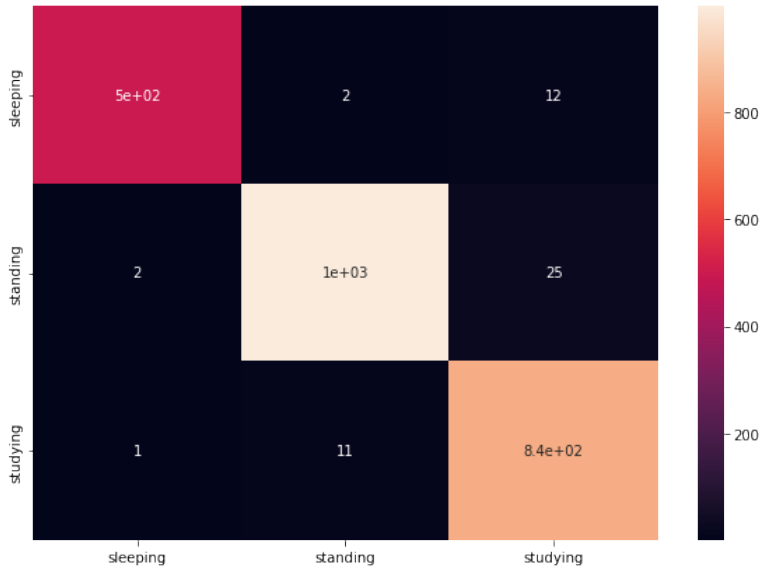
Gambar 5 Grafik *loss* model yang dirancang



Gambar 6 Grafik akurasi per kelas model yang dirancang



Gambar 7 Grafik *loss* per kelas model yang dirancang



Gambar 8 *Confusion matrix* terhadap data validasi pada model yang dirancang

Dari hasil pengujian *confusion matrix* terhadap data validasi, dapat dilihat bahwa kelas yang lebih sulit diklasifikasikan dengan benar dibandingkan kelas lainnya adalah kelas berdiri dan kelas belajar. Hal ini dikarenakan kesamaan fitur *skeleton* yang dimiliki kedua aktivitas tersebut. Selain itu, dari pengujian ini juga dihitung nilai akurasi, presisi, *recall*, dan F1-*score* terhadap data validasi. Hasil pengujian *confusion matrix* ini dapat dilihat pada Tabel 5.

Tabel 5 Skenario video percobaan yang dibuat

Akurasi	Presisi	Recall	F1-Score
97.77%	97.96%	97.73%	97.83%

Setelah mendapatkan konfigurasi *hyperparameter* yang cocok untuk dipakai, tim peneliti melakukan percobaan untuk menguji keberhasilan pendeteksian aktivitas. Percobaan ini dilakukan dengan cara membuat sendiri 4 buah data video yang mengandung skenario-skenario aktivitas anak secara kronologis. Setiap video memiliki resolusi tidak lebih dari 540p dan memiliki sudut pengambilan gambar yang berbeda-beda. Skenario-skenario setiap video dapat dilihat pada Tabel 6.

Tabel 6 Skenario video percobaan yang dibuat

Nama Data	Detik ke-	Aktivitas
Video 1	0 – 3	Berdiri
	4 – 6	Belajar
	7 – 9	Berdiri
	10 - 13	Tidur
Video 2	0 – 8	Berdiri
	9 – 11	Belajar
	12 – 19	Tidur
	20 – 22	Belajar
	23 – 31	Berdiri
	32 – 39	Belajar
Video 3	0 – 2	-
	3 – 8	Berdiri
	9 – 10	Belajar
	11 – 14	Berdiri
	15 – 18	Belajar
	19 - 21	Tidur
Video 4	0 – 8	Berdiri
	9 – 10	Belajar
	11 – 12	Tidur
	13	Belajar
	14	-
	15 - 17	Tidur

Frame yang akan diambil untuk melakukan prediksi aktivitas adalah *frame* pertama setiap detiknya. Pada tabel di atas, setiap detiknya memiliki *ground-truth* label yang menandakan aktivitas yang benarnya, label ini diperoleh dengan anotasi secara manual. Detik yang memiliki label strip (-) menandakan detik tersebut tidak mengandung aktivitas atau aktivitas yang ada tidak termasuk ke dalam target kelas aktivitas.

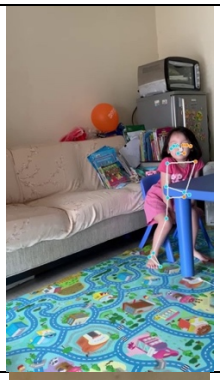
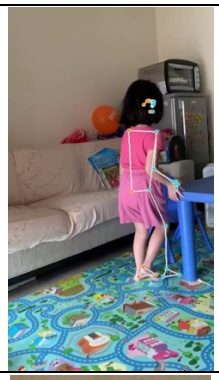
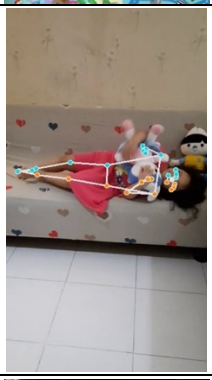
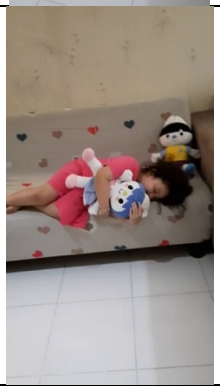
Terdapat dua jenis pengujian yang akan dilakukan pada pengujian dengan video skenario ini yaitu pengujian dengan prediksi setiap *frame* dan rata-rata 3 *frame* terakhir. Di mana pada prediksi setiap *frame*, prediki aktivitas pada detik saat ini akan diprediksi dengan menggunakan *frame* saat ini, sedangkan untuk rata-rata 3 *frame* terakhir akan dipakai akumulasi vektor keluaran CNN yang dirata-ratakan. Akurasi pengujian berbagai skenario untuk model yang telah dilatih dapat dilihat pada Tabel 7.

Tabel 7 Skenario video percobaan yang dibuat

Skenario 1		Skenario 2		Skenario 3		Skenario 4	
1 Frame	3 Frame	1 Frame	3 Frame	1 Frame	3 Frame	1 Frame	3 Frame
78.57%	83.33%	40%	57.14%	31.81%	36.84%	38.88%	20%

Pada skenario-skenario yang memiliki sudut pengambilan video yang sulit, tingkat akurasi cenderung berkurang karena model HPE gagal memprediksi *skeleton* objek manusia. Selain itu, jika bagian kepala dari objek manusia yang menjadi target terhalang maka model HPE juga akan gagal melakukan prediksi *skeleton*. Kedua hal ini membuat sistem yang dirancang gagal memprediksi aktivitas pada skenario yang sedang terjadi. Beberapa sampel *frame* video pengujian dapat dilihat pada Tabel 8.

Tabel 8 Sampel *frame* dan hasil prediksi pengujian skenario video

Citra HPE	Aktivitas		Citra HPE	Aktivitas	
	Ground-truth	**Prediksi**		**Ground-truth**	**Prediksi**
	Belajar	Belajar		Berdiri	Berdiri
	Belajar	Berdiri		Tidur	Berdiri
	Tidur	HPE Gagal		Belajar	HPE Gagal

4. KESIMPULAN

Berdasarkan hasil pengujian yang dilakukan, dapat ditarik beberapa kesimpulan yaitu sebagai berikut

1. Sistem yang dirancang berhasil melakukan pendeteksian aktivitas anak-anak dengan dataset citra yang dikumpulkan dari internet. Hasil evaluasi pelatihan model yang diperoleh sistem memiliki nilai akurasi 97.77%, presisi 97.96%, *recall* 97.73%, dan F1-*score* 97.83%.
2. Penggunaan HPE pada HAR membuat aktivitas-aktivitas yang memiliki bentuk *skeleton* yang sama dapat salah diklasifikasikan. Seperti aktivitas belajar dan berdiri pada anak-anak, hal ini dikarenakan tidak adanya konteks lingkungan yang dapat membantu pendeteksian.
3. Posisi kamera dalam pengambilan citra sangat berpengaruh, oleh karena itu dibutuhkan dataset yang memiliki sudut kamera yang bervariasi.

4. Pada kondisi di mana kepala dari objek manusia terhalang benda/tidak jelas, BlazePose gagal mendeteksi *skeleton*, sehingga membuat satu sistem gagal.
5. Secara umum, pendeteksian dengan menggunakan 3 *frame* terakhir menghasilkan hasil yang lebih baik. Hal ini dikarenakan hasil HPE yang tidak stabil (berubah-ubah).

Dalam penelitian selanjutnya, penggunaan BlazePose dapat diganti dengan *pre-trained* model HPE yang lebih akurat dalam mendeteksi *pose*. Selain itu juga diperlukan dataset aktivitas yang lebih bervariasi sehingga hasil pendeteksian dapat lebih akurat.

DAFTAR PUSTAKA

[1] Hussain, Z., Sheng, M., dan Zhang, W. E., 2019, Different Approaches for Human Activity Recognition – A Survey, *Journal of Network and Computer Applications*, No. 102738, Vol. 167.

[2] Gruosso, M., Capece, N., dan Erra, U., 2021, Human segmentation in surveillance video with deep learning, *Multimedia Tools and Applications*, Vol. 80, Hal. 1175-1199.

[3] Cippitelli, E., Gambi, E., dan Spinsante Susanna, 2017, *Human Action Recognition with RGB-D Sensors*, Diedit oleh Travieso-Gonzales, C., *Motion Tracking and Gesture Recognition*, IntechOpen, London.

[4] Liu, Y., Ma, R., Li, H., Wang, C., dan Tao, Y., 2021, RGB-D Human Action Recognition of Deep Feature Enchancement and Fusion Using Two-Stream ConvNet, *Journal of Sensors*, Vol. 2021, No. 8864870.

[5] Bagate, A., dan Shah, M., 2019, Human Activity Recognition using RGB-D Sensors, IEEE, *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, Madurai.

[6] Trascau, M., Nan, M., dan Florea, A.M., 2019, Spatio-Temporal Features in Action Recognition Using 3D Skeletal Joints, *Sensors (Basel) 2019*, Vol. 19, No. 243.

[7] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., dan Grundmann, M., 2020, BlazePose: On-device Real-time Body Pose tracking, *https://arxiv.org/abs/2006.10204*, Diakses pada 1 September 2021.

[8] Saha, S., 2018., A Comprehensive Guide to Convolutional Neural Network – the ELI5 way, https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53, Diakses pada 3 September 2021