HUMAN ACTIVITY CLASSIFICATION USING DEEP LEARNING BASED ON 3D MOTION FEATURE

PAPER MLWA

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Human activity classification using deep learning based on 3D motion feature

Endang Sri Rahayu [a,b], Eko Mulyanto Yuniarno [a], I. Ketut Eddy Purnama [a], Mauridhi Hery Purnomo [a,c,*]

[a] Department of Electrical Engineering, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
[b] Department of Electrical Engineering, Jayabaya University, Jakarta, Indonesia
[c] University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS), Surabaya, Indonesia

ARTICLE INFO

Keywords:
Human activity
Motion feature
Deep learning
Classification

ABSTRACT

Human activity classification is needed to support various fields. The health sector, for example, requires the ability to monitor the activities of patients, the elderly, or people with special needs to provide services with fast response as needed. In the traditional classification model, the steps taken to start from the input of data and then proceed with feature extraction, representation, classifier and end with semantic labels. The classification stage uses Convolutional Neural Network (CNN) deep learning to data input, CNN, and semantic labels. This paper proposes a novel method of classifying nine activities based on the movement features of changes in joint distance using Euclidean on the order of frames in each activity segment as input to the CNN model. This study's motion feature extraction technique was tested using various window sizes to obtain the best classification accuracy. The experimental results show that the selection of window size 16 on the motion feature setting will produce an optimal model accuracy of 94.08% in classifying human activities.

1. Introduction

The Human Movement Analysis (HMA) research area is an interdisciplinary research area that attracts great interest from the computer vision, machine learning, multimedia, and medical research communities. The implementation of this research is utilized for human–computer interaction, security (intelligent surveillance), health (assisted clinical studies), information technology (content-based video capture), entertainment (special effects in somatosensory film and game production) for all aspects of our daily life (Seidenari et al., 2013). As an essential research series of HMA, Human Activity Recognition (HAR) forms the basis for all the applications mentioned above. Utilization has been widely used in health care applications such as elderly monitoring, exercise monitoring, and rehabilitation monitoring (Huang et al., 2020). HAR is also developing in applications in the fields of robotics, entertainment, biometrics, and multimedia (Bennamoun et al., 2020). An indoor emergency awareness alarm system was also developed using deep neural networks. The system uses mobile devices for people such as the elderly, people with special needs or children who may need help (Kim & Kim, 2020).

In recent years, the development of deep learning has resulted in significant advances in activity recognition. In various research topics, there are two main methods, framework-based activity recognition and sensor-based activity recognition (Goddard, 2021). One of the main problems of the existing HAR strategy is the relatively low classification accuracy, so it is necessary to increase the accuracy, which requires high computational overhead (Kong et al., 2021). Classification models using deep learning are continuously being developed to improve performance resulting from traditional video classification models.

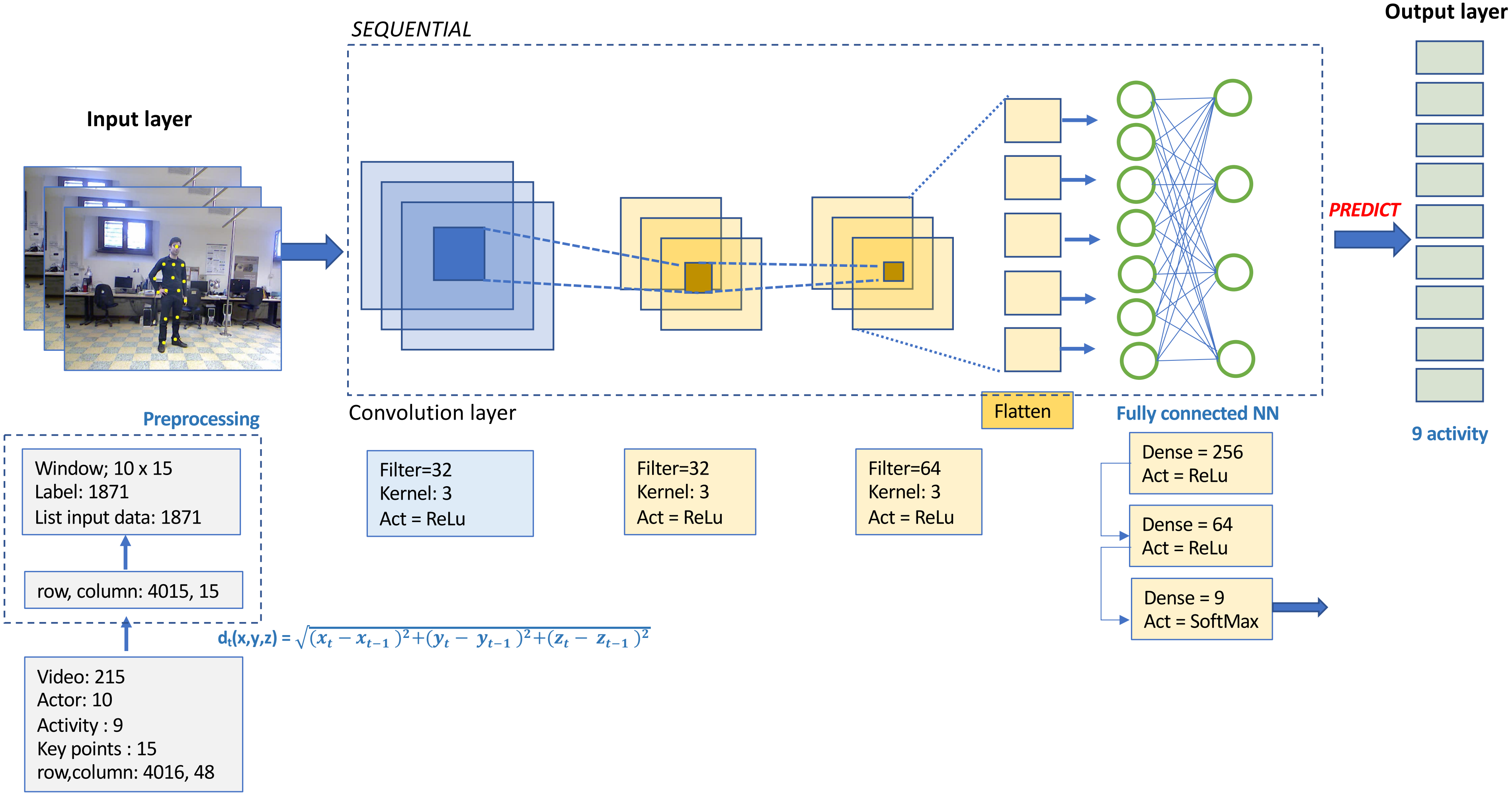
One of the Deep Learning algorithms to process image or sound data is a Convolutional Neural Network (CNN). CNN in this study was used to classify labelled data using the supervised learning method. Supervised learning work is the presence of target data for data training. Among all types of neural networks, CNN is known as the most successful and is widely used to solve problems of image recognition, object detection/localization, and even text processing (Alpaydin, 2021). CNN (convolutional kernels) combines some local filters with raw input data and generates local translation-invariant features in the convolutional layer. The successive pooling layer extracts fixed-length features via a sliding window from raw input data following some rules like mean, max., etc. (Zhao et al., 2019).

This paper proposes a model for recognizing human activities using deep learning to classify human activities (actions). The preparation of data as input data for the CNN model is the main concern in this paper. Meanwhile, we also pay attention to designing the CNN model that will be trained on the dataset. The source data of the 3D coordinate points of the joints' positions will be processed to detect changes in movement that occur by calculating the distance of the coordinates of the joints

[*] Corresponding author at: Department of Electrical Engineering, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia.
E-mail addresses: endangsr@jayabaya.ac.id (E.S. Rahayu), ekomulyanto@ee.its.ac.id (E.M. Yuniarno), ketut@te.its.ac.id (I.K.E. Purnama), hery@ee.its.ac.id (M.H. Purnomo).

https://doi.org/10.1016/j.mlwa.2023.100461
Received 25 January 2023; Received in revised form 6 March 2023; Accepted 8 March 2023
Available online 15 March 2023

Arsitektur Model CNN



SEQUENTIAL

Output layer

Input layer

PREDICT

Preprocessing

Convolution layer

Flatten

Fully connected NN

9 activity

Window; 10 x 15
Label: 1871
List input data: 1871

row, column: 4015, 15

Video: 215
Actor: 10
Activity : 9
Key points : 15
row,column: 4016, 48

$d_t(x,y,z) = \sqrt{(x_t - x_{t-1})^2+(y_t - y_{t-1})^2+(z_t - z_{t-1})^2}$

Filter=32
Kernel: 3
Act = ReLu

Filter=32
Kernel: 3
Act = ReLu

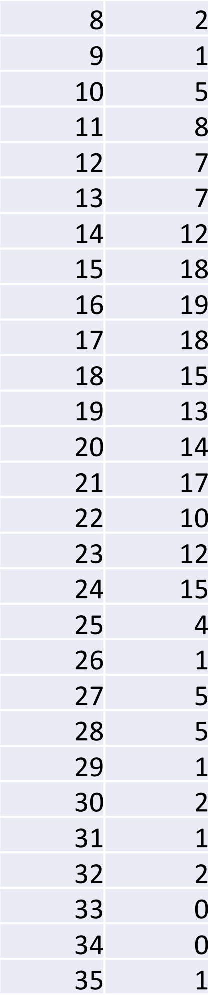
Filter=64
Kernel: 3
Act = ReLu

Dense = 256
Act = ReLu

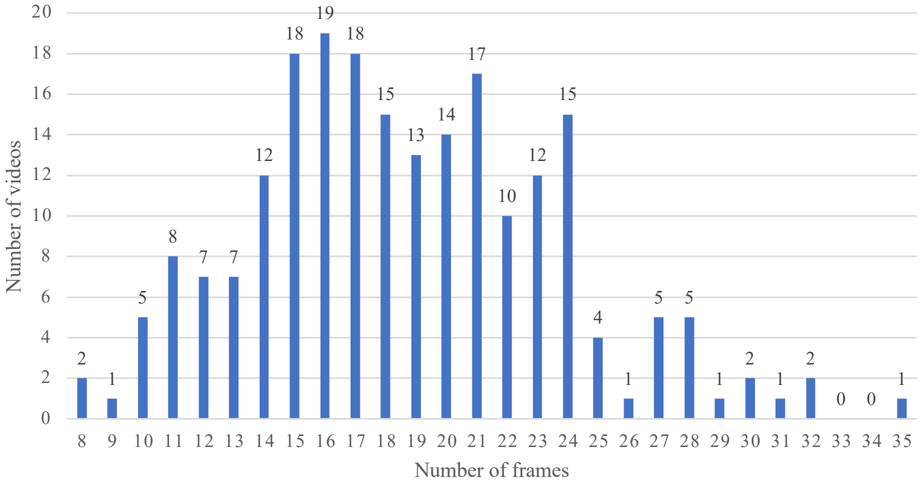
Dense = 64
Act = ReLu

Dense = 9
Act = SoftMax

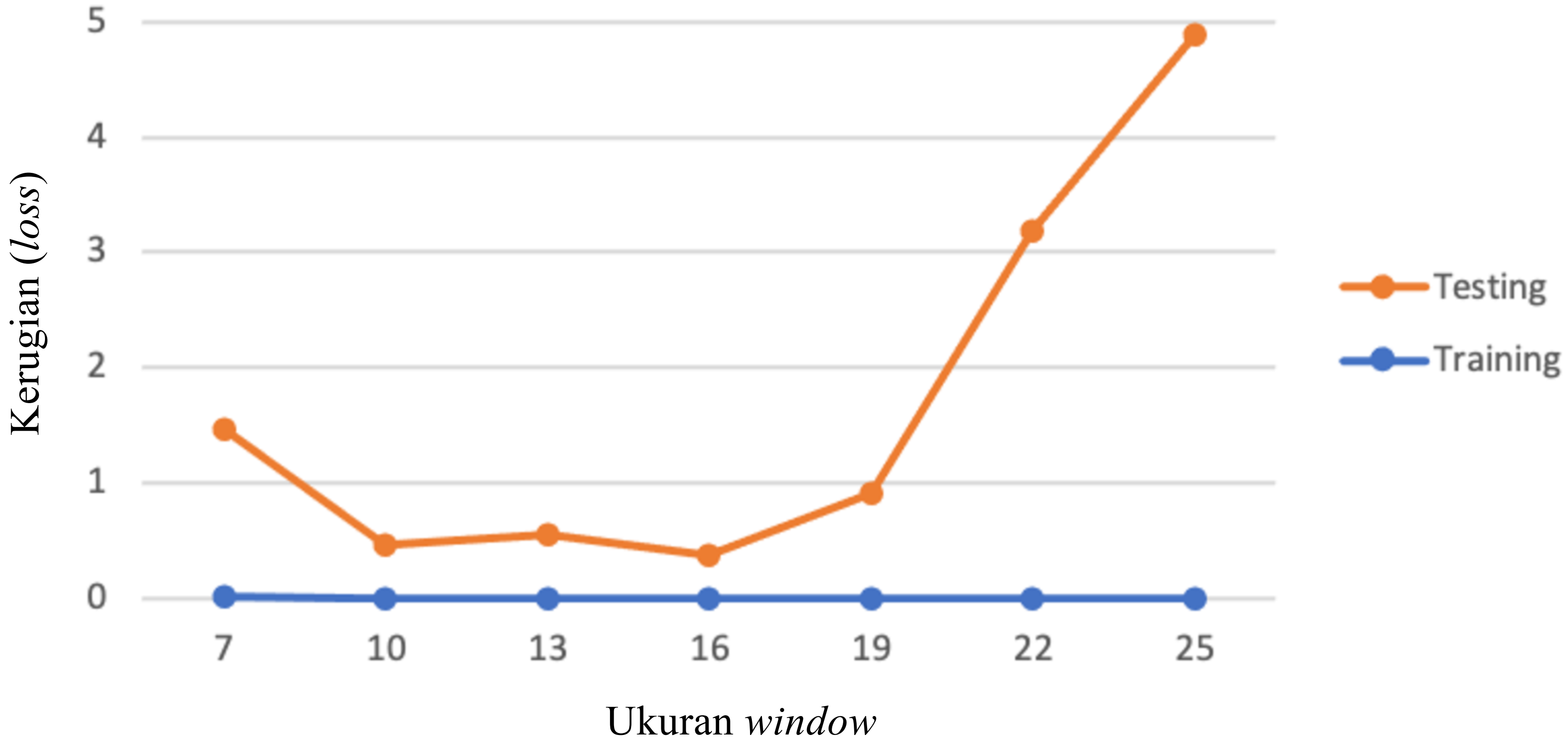
Count of video segments for each count of frames

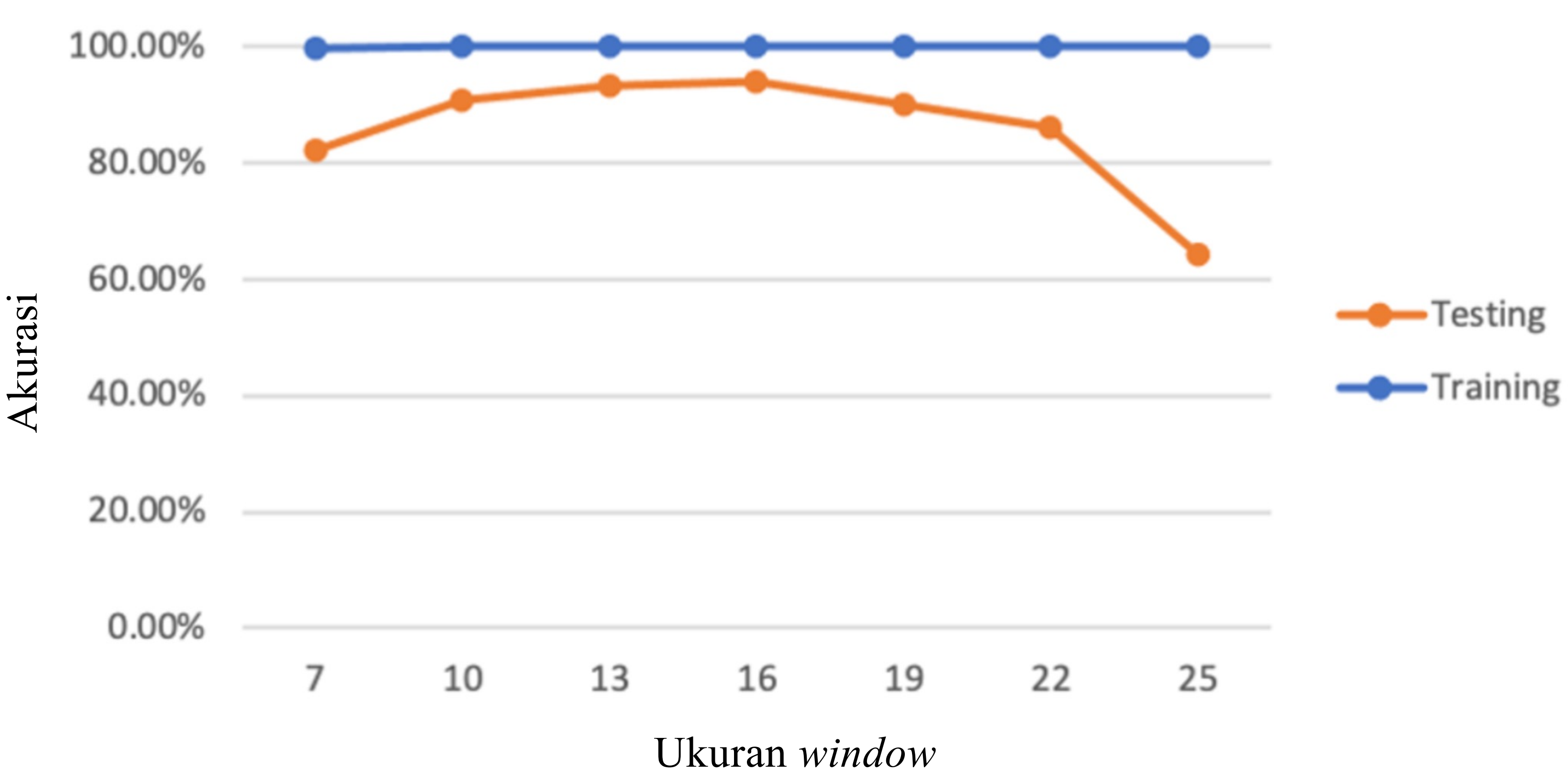


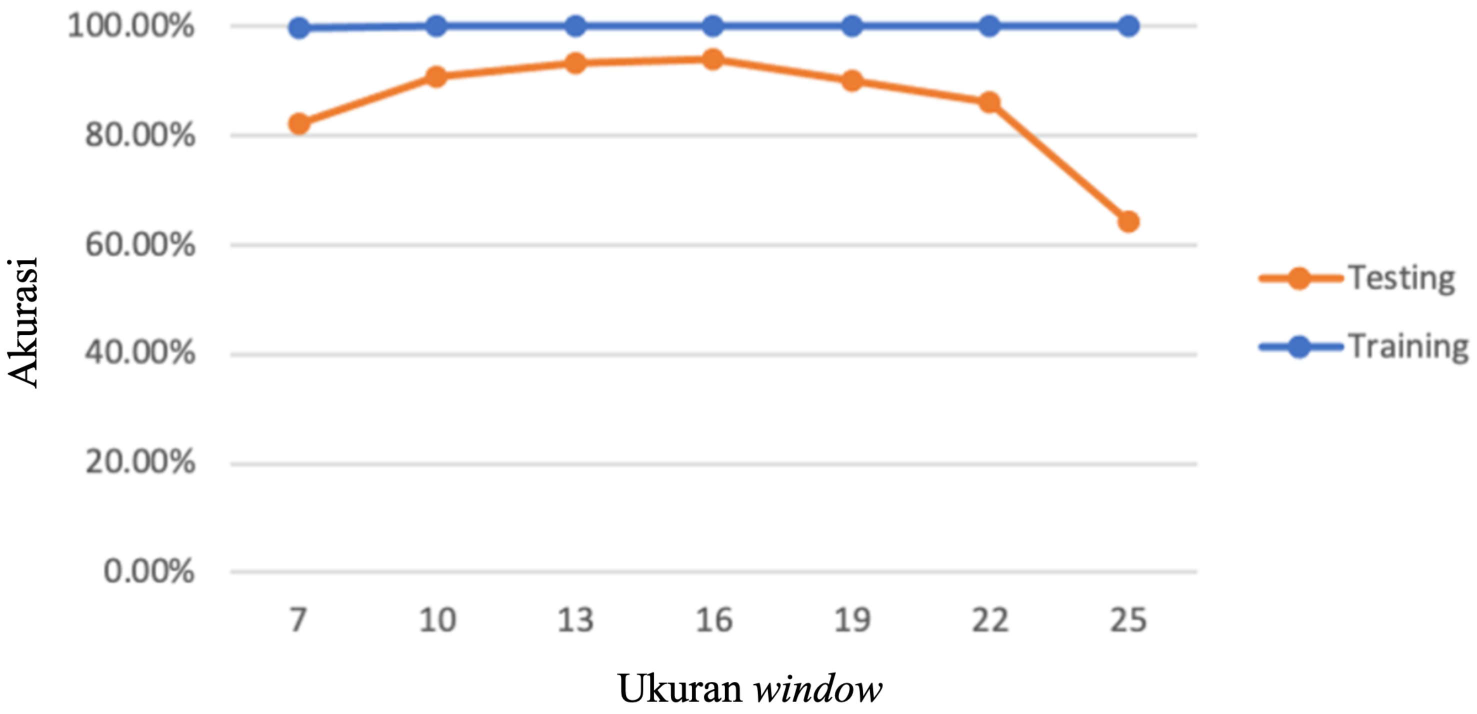
8	2
9	1
10	5
11	8
12	7
13	7
14	12
15	18
16	19
17	18
18	15
19	13
20	14
21	17
22	10
23	12
24	15
25	4
26	1
27	5
28	5
29	1
30	2
31	1
32	2
33	0
34	0
35	1



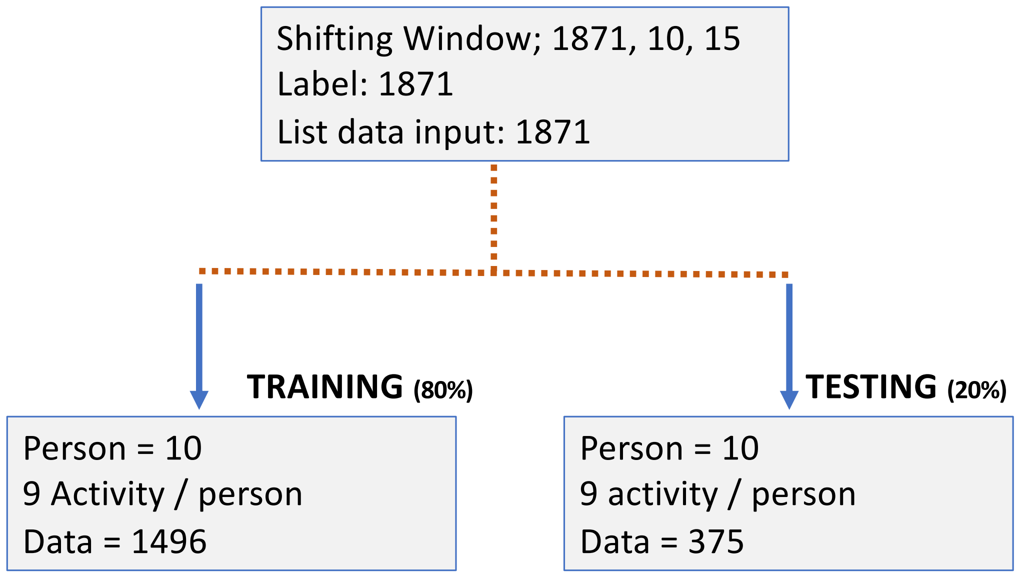
Count of actifity segments for some count of frames used during training


5
4
3
2
1
0
Kerugian (*loss*)
7
10
13
16
19
22
25
Ukuran *window*
Testing
Training


100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Akurasi
7
10
13
16
19
22
25
Ukuran *window*
Testing
Training


100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
Akurasi
7
10
13
16
19
22
25
Ukuran *window*
Testing
Training

Split data --> training, testing



Shifting Window; 1871, 10, 15
Label: 1871
List data input: 1871

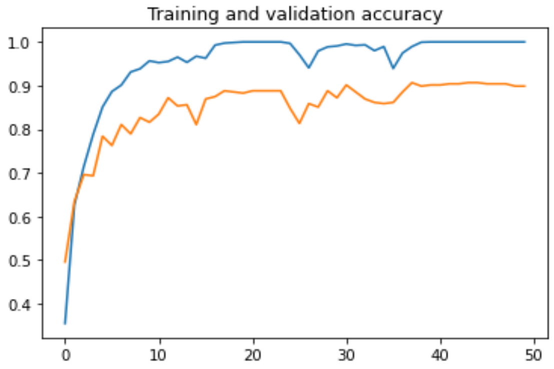
TRAINING **(80%)**

Person = 10
9 Activity / person
Data = 1496

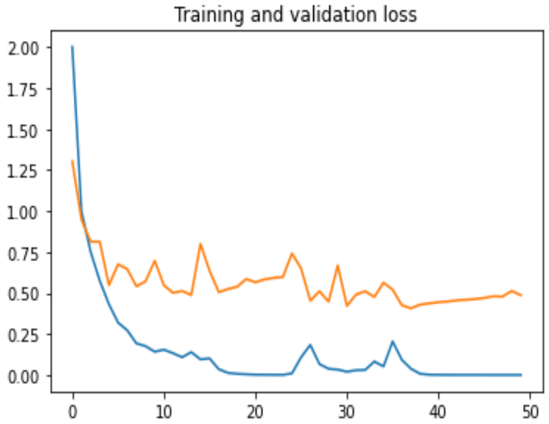
TESTING **(20%)**

Person = 10
9 activity / person
Data = 375

Hasil eksperimen: deteksi aktifitas (50 epoch)



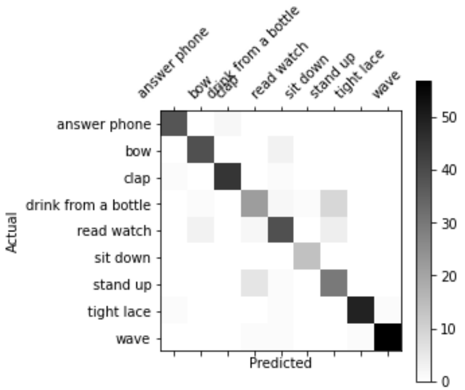
Training and validation accuracy

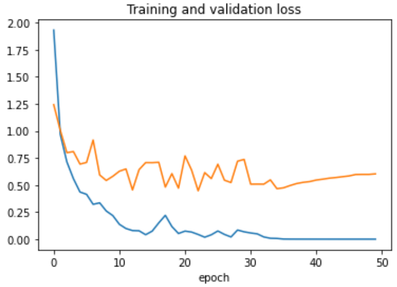


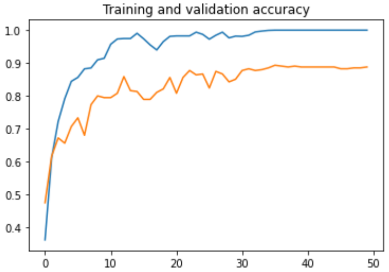
Training and validation loss

loss: 1.2694e-04 - accuracy: 1.0000 - val_loss: 0.6027 - val_accuracy: 0.8880

50 epoch



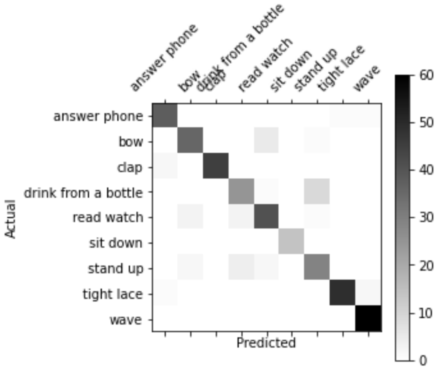


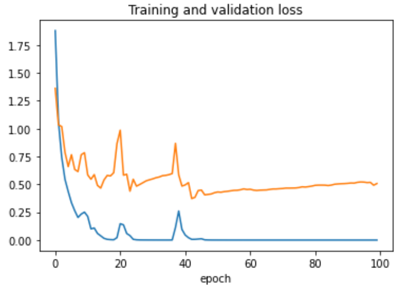


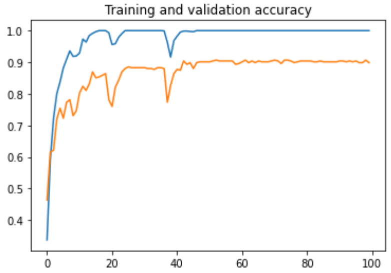
Actual	answer phone	bow	clap	drink from a bottle	read watch	sit down	stand up	tight lace	wave
answer phone	38	0	2	0	0	0	0	0	0
bow	0	39	0	0	3	0	0	0	0
clap	1	0	45	0	1	0	0	0	0
drink from a bottle	0	1	0	22	2	1	9	0	0
read watch	0	3	0	2	39	0	4	0	0
sit down	0	0	0	0	0	14	0	0	0
stand up	0	0	0	6	1	0	30	0	0
tight lace	1	0	0	0	1	0	0	49	1
wave	0	0	0	1	1	0	0	1	57

loss: 1.1070e-05 - accuracy: 1.0000 - val_loss: 0.5068 - val_accuracy: 0.8987

100 epoch







Actual	answer phone	bow	clap	drink from a bottle	read watch	sit down	stand up	tight lace	wave
answer phone	38	0	0	0	0	0	0	1	1
bow	0	36	0	0	5	0	1	0	0
clap	2	0	45	0	0	0	0	0	0
drink from a bottle	0	0	0	25	1	0	9	0	0
read watch	0	3	0	3	41	0	1	0	0
sit down	0	0	0	0	0	14	0	0	0
stand up	0	2	0	4	2	0	29	0	0
tight lace	1	0	0	0	0	0	0	49	2
wave	0	0	0	0	0	0	0	0	60

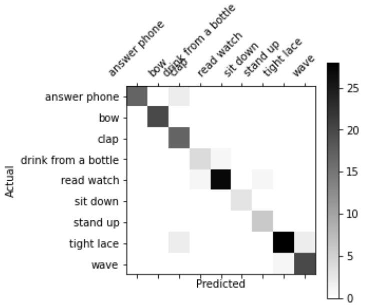
PROSENTASE

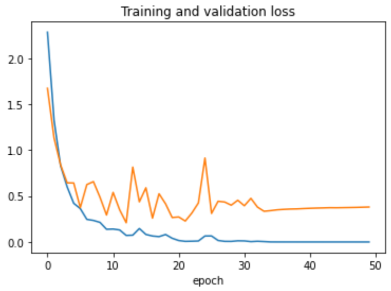
actual (total act)	prediction								
	1	2	3	4	5	6	7	8	9
1 (40)	95	0	5	0	0	0	0	0	0
2 (42)	0	93	0	0	7	0	0	0	0
3 (47)	2	0	96	0	2	0	0	0	0
4 (35)	0	3	0	63	6	3	26	0	0
5 (48)	0	6	0	4	81	0	4	0	0
6 (14)	0	0	0	0	0	100	0	0	0
7 (37)	0	0	0	16	3	0	81	0	0
8 (52)	2	0	0	0	2	0	0	94	2
9 (60)	0	0	0	2	2	0	0	2	95

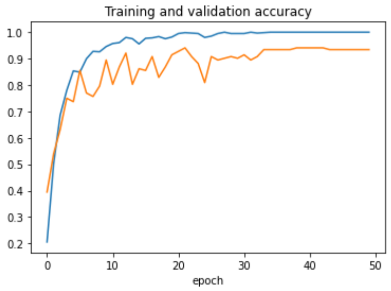
loss: 8.4611e-05 - accuracy: 1.0000 - val_loss: 0.3803 - val_accuracy: 0.9342

Window = 16

(760, 16, 15)







```
Model: "sequential"
_________________________________________________________________
Layer (type)             Output Shape            Param #
=================================================================
conv2d (Conv2D)           (None, 14, 13, 32)      320
_________________________________________________________________
conv2d_1 (Conv2D)          (None, 12, 11, 32)      9248
_________________________________________________________________
dropout (Dropout)          (None, 12, 11, 32)      0
_________________________________________________________________
conv2d_2 (Conv2D)          (None, 10, 9, 32)       9248
_________________________________________________________________
flatten (Flatten)         (None, 2880)            0
_________________________________________________________________
dense (Dense)             (None, 256)             737536
_________________________________________________________________
dense_1 (Dense)            (None, 256)             65792
_________________________________________________________________
dense_2 (Dense)            (None, 256)             65792
_________________________________________________________________
dense_3 (Dense)            (None, 64)              16448
_________________________________________________________________
dense_4 (Dense)            (None, 64)              4160
_________________________________________________________________
dense_5 (Dense)            (None, 9)               585
=================================================================
Total params: 909,129
Trainable params: 909,129
Non-trainable params: 0
```

Actual	answer phone	bow	clap	drink from a bottle	read watch	sit down	stand up	tight lace	wave
answer phone	17	0	2	0	0	0	0	0	0
bow	0	20	0	0	0	0	0	0	0
clap	0	0	17	0	0	0	0	0	0
drink from a bottle	0	0	0	4	1	0	0	0	0
read watch	0	0	0	1	27	0	1	0	0
sit down	0	0	0	0	0	3	0	0	0
stand up	0	0	0	0	0	0	6	0	0
tight lace	0	0	2	0	0	0	0	28	2
wave	0	0	0	0	0	0	0	1	20