IRSTI 28.23.15 https://doi.org/10.26577/JPCSIT.2023.v1.i1.05

Y.B. Beisen*, Y.S. Nurakhov

Astana IT University, Astana, Kazakhstan
*e-mail: beysenersultan@gmail.com

RECOGNITION OF HUMAN ACTIONS USING MACHINE LEARNING METHODS

Abstract. Human action recognition is a significant area of focus within computer vision, and it is intertwined with several other disciplines such as computer science, psychology, and healthcare. This is due to the increasing number of videos and the potential applications for automatic video analysis, such as video surveillance, human-machine interaction, sports analysis, and video search. In this research, we applied machine learning algorithms such as Random Forest, MLP Classifier, AdaBoost, and QDA to recognize human actions and compared the results. The results of the tests showed that the MLP Classifier had an accuracy of 97%, the Random Forest had an accuracy of 95%, the AdaBoost had an accuracy of 76%, and the QDA had an accuracy of 74%. In the training dataset, the MLP Classifier had an accuracy of 98%, the Random Forest had an accuracy of 99%, the AdaBoost had an accuracy of 76%, and the QDA had an accuracy of 74%. Out of all the algorithms, the MLP Classifier showed the best results.

Key words: Human action recognition, Machine learning, OpenPose, Algorithm, Dataset.

1 Introduction

The recognition of human actions has become a prominent topic in computer vision research. It involves detecting sequences of movements performed by individuals and can be categorized into three types based on the number of people involved: single-user, multi-user, and group actions. This recognition is accomplished through observations from various sensory devices [1-2].

The development of skeleton-based systems has made it easier to perform practical tasks in the field of human action recognition. Traditional sensors gather information about the coordinates of joints in the human body, which is used to create a parametric representation of the body [3]. This skeleton data is then processed using advanced algorithms to identify different actions.

The implementation of a skeleton-based action recognition system presents several challenges. Collecting data for this system is typically done in controlled laboratory conditions to minimize the impact of external factors, but to be more applicable in real-world scenarios, the system must be able to handle interruptions caused by changing background conditions. The system must also consider factors such as the size of the subject's body, its position, orientation, and changes in perspective, which can make recognizing actions difficult. Maintaining the correlation between joints in the skeleton structure during both data collection and analysis exacerbates this challenge. Furthermore, inaccuracies in tracking devices can lead to suboptimal results. Despite these challenges, researchers have worked to optimize the system to address these issues.

The hands and joints of the human skeletal system play a critical role in identifying specific actions. The authors divide the human skeleton system into five parts: two legs, two arms, and one trunk [4]. Many studies have highlighted the importance of these specific joints in action recognition. A review of skeleton-based models for classifying human actions, covering various preprocessing methods, action representation techniques, and classification methods, can be found in [5]. Several studies are also dedicated to action recognition models that use deep learning [6].

Figure 1 shows the general stages of the skeleton-based action recognition system.

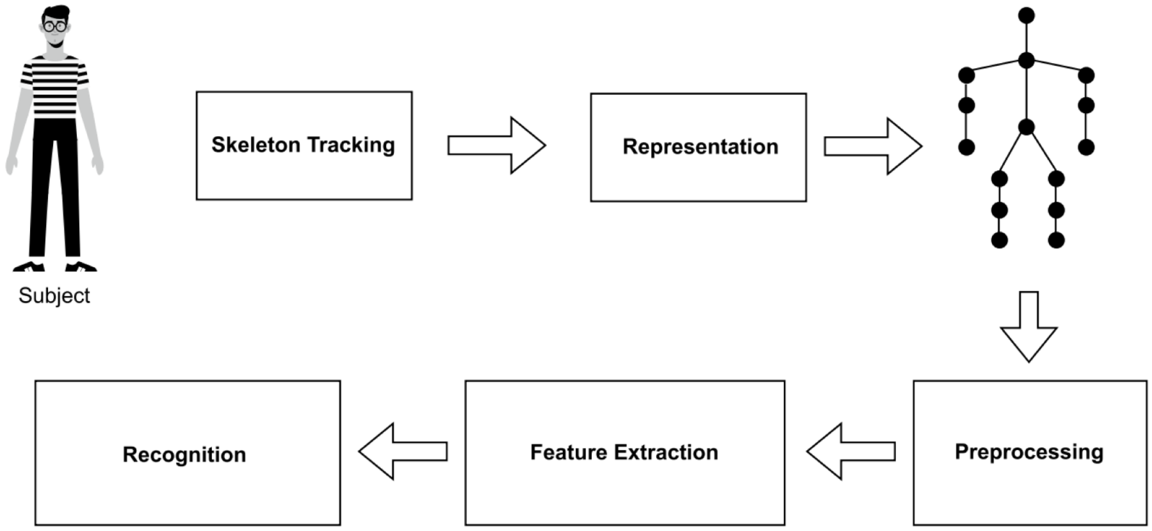


Figure 1 – General stages of the skeleton-based action recognition system

Skeleton tracking is the first and essential step in recognizing actions. Researchers have used various tracking methods to extract skeletal trajectories or postures, including depth sensors, the OpenPose toolset, and body markers. These techniques are considered some of the best methods for skeleton tracking.

In our research, we utilized the OpenPose toolset developed by researchers at Carnegie Mellon University. OpenPose is the first real-time system that can simultaneously detect the human body, face, hands, and feet (135 key points) in individual images [7-9]. It can determine the positions of 15, 18, or 25 key points on the joints of the human body or feet and provides a confidence score for each joint. In our work, we utilized 18 key points, as shown in Figure 2

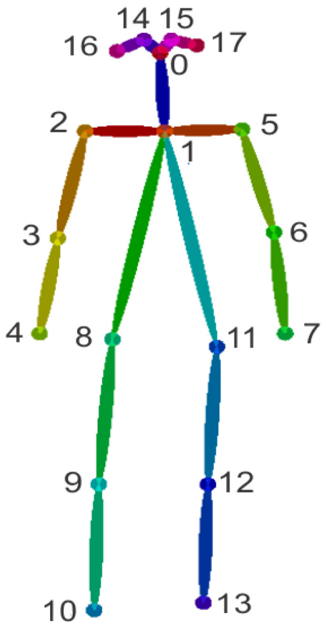


Figure 2 – Key points that can be detected using the OpenPose algorithm

In the field of human action recognition, we must consider two types of representation: the representation of the skeleton and the representation of the action. A frame with J joints can be represented as a set Xt = { X1t, X2t, ..., Xjt, ..., XJt } ∈ RD×J, where Xjt represents the joint coordinate j at time step t, and D represents the dimension of the skeleton (e.g. for a 3D skeleton, D=3). On the other hand, an action refers to the movement functions of the human body, often involving changing relative positions of body parts (such as walking, running, or jumping). The skeleton representation focuses on the location of the skeleton structure, while the action representation focuses on the movement functions of the human body.

The preprocessing steps involved in the skeleton-based activity recognition system are important for improving the accuracy of the system. These steps help to address issues such as differences in size and appearance, variations in body proportions, varying duration of skeleton sequences, and noise in the captured sequences [10]. Our preprocessing process is shown in Figure 3.

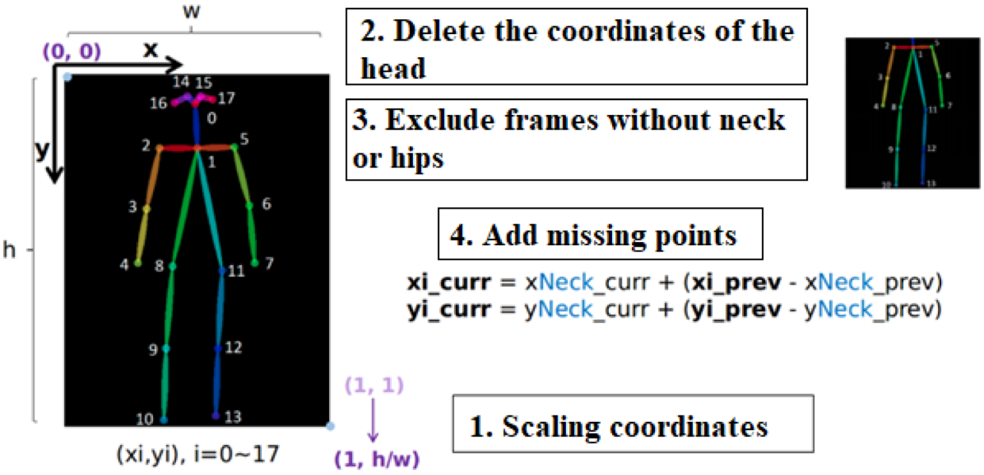

w
(0, 0)
x
y
h
14 15
16
17
0
2
5
1
3
6
4
8
11
7
9
12
10
13
(xi,yi), i=0~17
2. Delete the coordinates of the head
3. Exclude frames without neck or hips
4. Add missing points
xi_curr = xNeck_curr + (xi_prev - xNeck_prev)
yi_curr = yNeck_curr + (yi_prev - yNeck_prev)
(1, 1)
(1, h/w)
1. Scaling coordinates

Figure 3 – Features of preprocessing

Features play an important role in the recognition of actions since the effectiveness of the model depends on the extracted features. Table 1 shows our features.

Table 1 – List of extracted features

Features	Description
XS	Data collected from 5 frames (combined into one array)
H	The average height of the skeleton
V-body	Neck Length / H
X	Normalized data X = (XS – mean(XS)) / H
V-joints	Differences between frames

In the area of recognizing actions based on the skeleton, a variety of techniques have been suggested. Some researchers have proposed methods using machine learning, while others have employed neural network-based methods. In our study, we utilized various machine learning algorithms such as Random Forest, MLP Classifier, AdaBoost, and QDA.

2 Materials and Methods

In our research, we propose the recognition of human actions based on the skeleton using machine learning algorithms. We utilize the KTH Dataset as our dataset. This dataset comprises six types of human actions (walking, jogging, running, boxing, waving, and clapping hands) performed by 25 individuals in four different situations: on the street, on the street with a zoom change, on the street with different clothing, and indoors. Currently, the database contains 2391 sequences with a frame rate of 25 frames per second and a spatial resolution of 160x120 pixels. The sequences have an average duration of four seconds and were captured using a static camera on a uniform background [11]. In our task, we use only five actions (running, walking, boxing, waving, and clapping). An example from this dataset is shown in Figure 4.

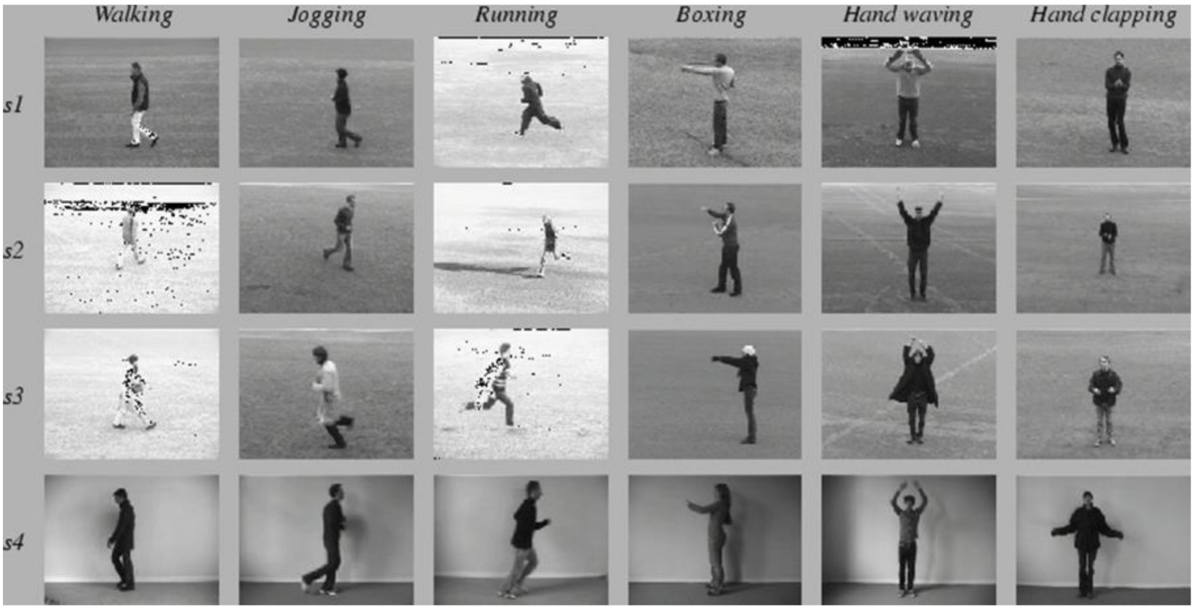


Figure 4 – Example of a KTH Dataset

In our research, we first extract human skeletons from 5 frames using the OpenPose system and assemble them into one array. The array then undergoes preprocessing (as shown in Figure 3), and features are extracted (as listed in Table 1). The total number of samples is 173,148. This set was split into 70% of the data (121,203 samples) for training and 30% of the data (51,945 samples) for testing. These features are then used as input data for various machine learning algorithms.

In this work, we utilized Python programming language within the Visual Studio Code environment and utilized the OpenPose and sklearn libraries to construct a machine learning-based model.

3 Literature review

In the realm of recognizing human actions, a range of methods has been proposed. These methods generally consist of three steps: preprocessing, feature extraction, and classification/detection. The current research has mainly concentrated on enhancing and implementing the latter two steps. This section presents a literature review of the various human action recognition techniques.

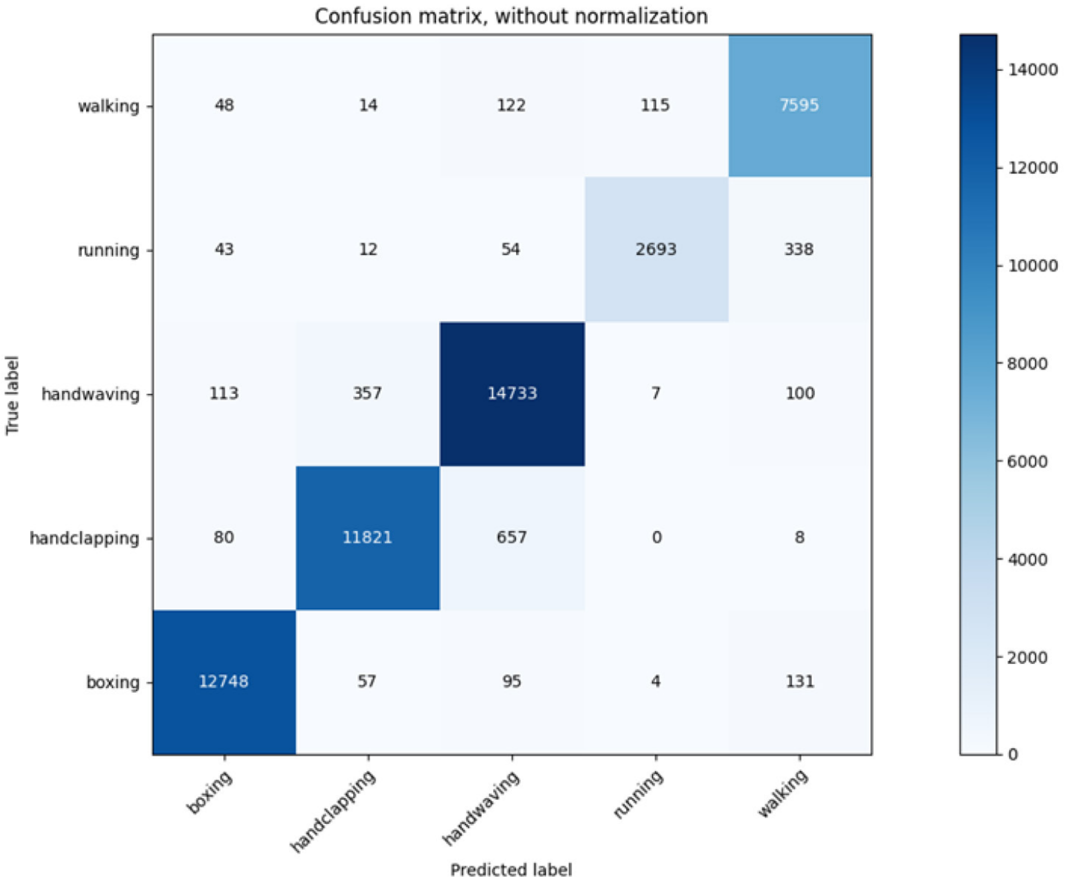
In recent years, the development of deep neural networks has increased the use of skeleton information for recognizing human actions [12][13]. The use of skeleton-based action recognition is becoming even more attractive due to the ability to extract skeleton data in real-time using just one RGB camera. Additionally, human activities can be directly identified using skeleton data [14].

Li et al. [15] developed a skeleton-based action identification system by analyzing important skeletal postures using a latent support vector machine. The study showed that distinguishing human actions takes only a few frames with key skeletal poses. Wang et al. [16] developed the model based on Naive-Bayes Nearest Neighbor (NBNN) algorithm considering Spatio-Temporal features to recognize actions. This study created a new model called the Spatio-Temporal NBNN by loosening the constraints of the NBNN algorithm and incorporating both spatial features and key temporal stages in the analysis of human actions. The bilinear classifier was used for classification purposes. Tang et al. [17] developed the model based on logistic regression and this model was constructed by estimating the parameters of the model using the logistic function and incorporating the depth and skeleton information from an RGB database.

4 Results and Discussion

During the study, five actions were selected to be recognized in human actions: running, walking, boxing, waving, and clapping. The experiment utilized machine learning algorithms such as Random Forest, MLP Classifier, AdaBoost, and QDA. The results from the testing dataset showed that MLP Classifier had an accuracy rate of 97%, Random Forest accuracy rate of 95%, AdaBoost accuracy rate of 76%, and QDA accuracy rate of 74%. The results from the training dataset showed that MLP Classifier had an accuracy rate of 98%, Random Forest accuracy rate of 99%, AdaBoost accuracy rate of 76%, and QDA accuracy rate of 74%. The highest accuracy was achieved by the MLP Classifier algorithm.

Figures 5-8 show the confusion matrices.



	boxing	handclapping	handwaving	running	walking
walking	48	14	122	115	7595
running	43	12	54	2693	338
handwaving	113	357	14733	7	100
handclapping	80	11821	657	0	8
boxing	12748	57	95	4	131

Figure 5 – Random Forest confusion matrix

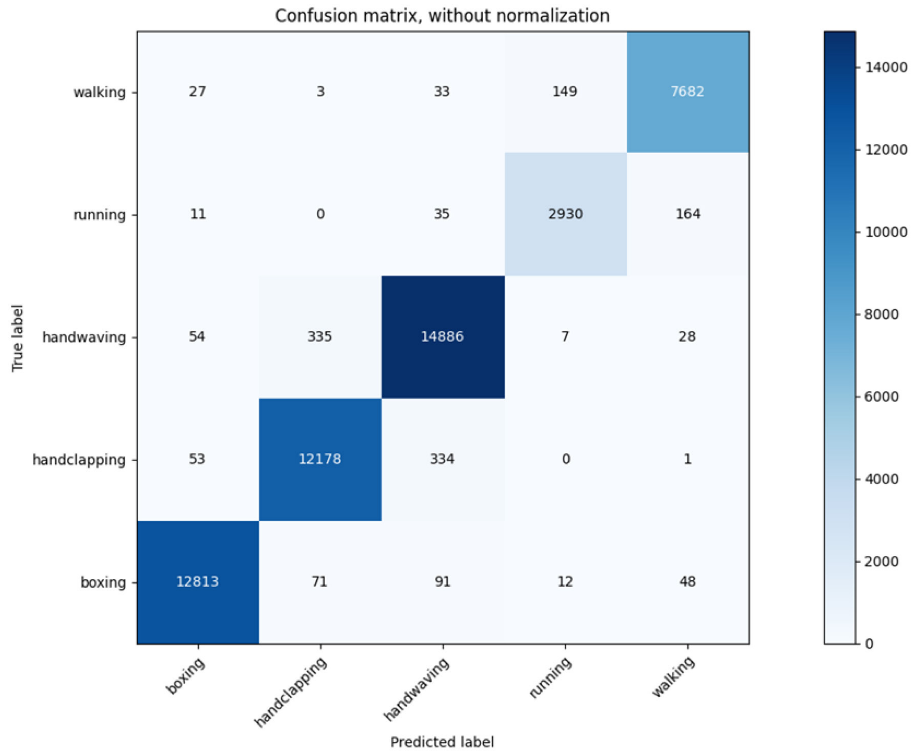


Figure 6 – MLPClassifier confusion matrix

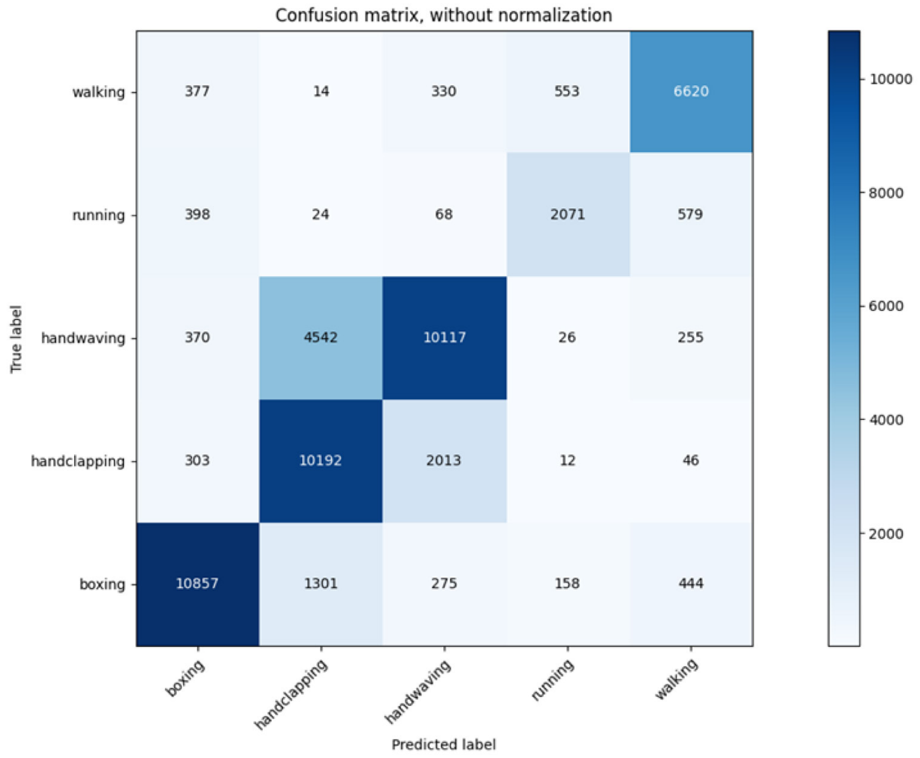


Figure 7 – AdaBoost confusion matrix

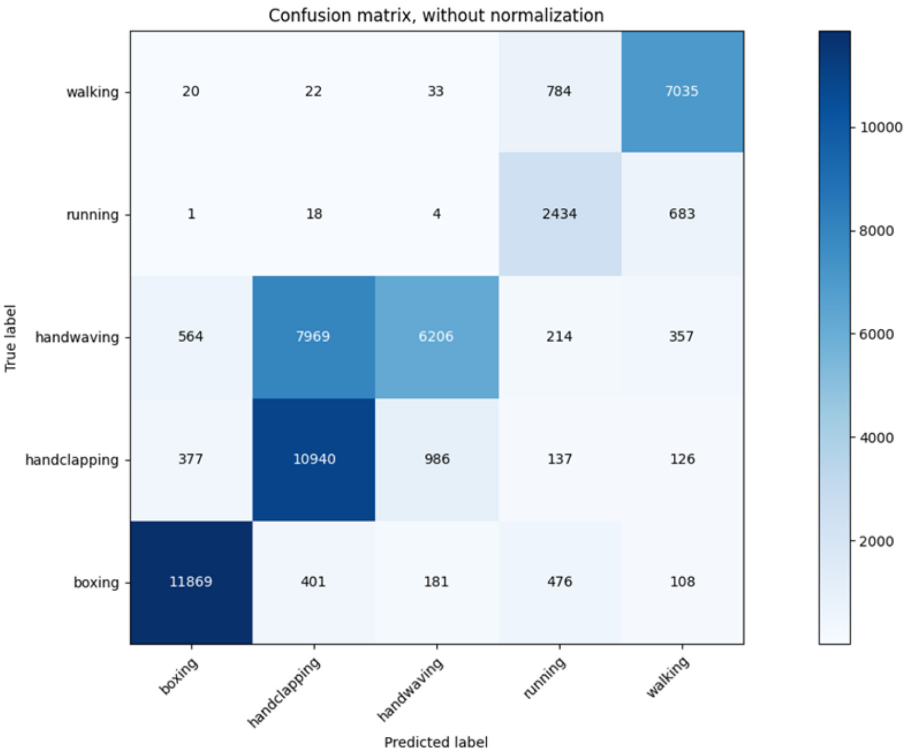


Figure 8 – QDA confusion matrix

The confusion matrices reveal that there are often mix-ups between the classes "clapping hands" and "waving hands," as well as between "walking" and "running" due to their similar style.

Recognizing human actions is challenging due to issues such as variations in viewpoint, body structure, camera movement, and background clutter. Our study presents a framework for recognizing human actions based on skeleton information using machine learning algorithms and the KTH dataset. Despite the high accuracy achieved, further research can be conducted in this field by considering more actions, using other datasets such as Weizmann, HAHA1, CMU MoBo, Human Eva, and NTU RGB+D 120 to expand the training data.

5 Conclusion

This study developed a system to recognize five types of human actions (running, walking, boxing, waving, and clapping) using the skeleton and machine learning algorithms. The experiments revealed that the MLP Classifier algorithm had the highest accuracy at 97%. Plans include expanding the range of human actions and utilizing different datasets.

References

1. Wang, J., Liu, Z., Ying, W., Yuan, J.: Learning action let ensemble for 3d human action recognition. IEEE Trans. Pattern Analy. Mach. Intell. 36(5), 914–927 (2013).

2. Wang, L., Gu, T., Tao, X., Lu, J.: Sensor-based human activity recognition in a multi-user scenario. In: European Conference on Ambient Intelligence, pp. 78–87. Springer (2009)

3. Batabyal, T., Chattopadhyay, T., Mukherjee, D.P.: Action recognition using joint coordinates of 3d skeleton data. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 4107–4111. IEEE (2015)

4. Pham, H.-H., Khoudour, L., Crouzil, A., Zegers, P., Velastin, S.A.: Exploiting deep residual networks for human action recognition from skeletal data. Comput. Vis. Image Underst. 170, 51–66 (2018)

5. Presti, L.L., Cascia, M.L.: 3d skeleton-based human action classification: a survey. Pattern Recogn. 53, 130–147 (2016)

6. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: a survey of deep learning-based methods. Comput. Vis. Image Underst. 192, 102897, 03 (2020)

7. Cao, Z., Simon, T., Wei, S.-E., Sheikh, S.-E.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)

8. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017)

9. Wei, S.-E., Ramakrishna, S.-E., Kanade, T., Sheikh. Y.: Convolutional pose machines. In: CVPR (2016)

10. Liliana [Lo Presti], Marco [La Cascia]: 3d skeleton-based human action classification: a survey. Pattern Recogn. 53, 130–147 (2016)

11. Sch, C., Barbara, L., . Recognizing human actions: A local SVM approach.

12. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1112–1121 (2020)

13. Huynh-The, T., Hua, C.-H., Tu, N.A., Kim, J.-W., Kim, S.-H., Kim, D.-S.: 3d action recognition exploiting hierarchical deep feature fusion model. In: 2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM), pp. 1–3. IEEE (2020)

14. J. Liu, N. Akhtar, and A. Mian, “Skepxels: Spatio-temporal Image Representation of Human Skeleton Joints for Action Recognition,” pp. 10–19, 2017, [Online]. Available: http://arxiv.org/abs/1711.05941

15. X. Li, Y. Zhang, and D. Liao, “Mining key skeleton poses with latent SVM for action recognition,” Applied Computational Intelligence and Soft Computing, vol. 2017, pp. 1–11, 2017, doi: 10.1155/2017/5861435.

16. J. Weng, C. Weng and J. Yuan, “Spatio-Temporal Naive-Bayes Nearest-Neighbor (ST-NBNN) for Skeleton-Based Action Recognition,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 445-454, doi: 10.1109/CVPR.2017.55.

17. Tang, N.C., Lin, Y.-Y., Hua, J.-H., Weng, M.-F., Mark Liao, H.-Y.: Human action recognition using associated depth and skeleton information. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4608–4612. IEEE (2014)