BlazePose-Based Action Recognition with Feature Selection Using Stochastic Fractal Search Guided Whale Optimization

Motasem S. Alsawadi
Department of Electronic and Electrical Engineering
University College London
London, UK
motasem.alsawadi.18@ucl.ac.uk, malswadi@kacst.edu.sa

Marcial Sandoval-Gastelum
International Institute for Applied Systems Analysis (IIASA)
Laxenburg, Austria
sandovalgaste@iiasa.ac.at

Irfan Danish
Department of Computer Science
FAST - National University of Computer and Emerging Sciences
Islamabad, Pakistan
i202215@nu.edu.pk

Miguel Rio
Department of Electronic and Electrical Engineering
University College London
London, UK
miguel.rio@ucl.ac.uk

***Abstract*—The BlazePose, which models human body skeletons as spatiotemporal graphs, has achieved fantastic performance in skeleton-based action identification. A Spatial-Temporal Graph Convolutional Network can then be used to forecast the actions. This architecture performance can be improved by simply replacing the skeleton input data with a different set of joints that provide more information about the activity of interest. On the other hand, existing approaches require the user to manually set the graph's topology and then fix it across all input layers and samples. This research shows how to use Stochastic Fractal Search - Guided Whale Optimization Algorithm in conjunction with the BlazePose skeletal data to construct a novel implementation of this topology for action recognition. We utilized the NTU-RGB+D and the Kinetics datasets as benchmarks in our experiments.**

***Index Terms*—BlazePose, metaheuristics, convolutional networks, feature selection, action recognition**

I. Introduction

BlazePose is an architecture for human posture prediction using a lightweight convolutional neural network optimized for inference on mobile devices. During the inference process, the neural network generates 33 main body landmarks for a single person [1]. The standard method for action recognition systems generates heatmaps for every joint while simultaneously adjusting offsets for every position. Nevertheless, it makes the model for a single individual significantly more complex than what is needed for real-time inference on mobile phones.

In contrast, we have utilized an encoder-decoder architecture that regresses directly to the coordinates of all joints. Regression-based approaches attempt to forecast the mean coordinate values despite being less computationally intensive and more scalable.

II. Literature Review

We have arranged the previous works found in literature into two categories: action recognition and feature selection.

A. Convolutional Networks for Action Recognition

In deep learning, the term "geometric deep learning" refers to all developing techniques that generalize deep learning models to non-Euclidean domains like graphs. The concept of a Graph Neural Network, abbreviated as GNN, was first described in [2]. The hunch that underpins GNNs (Graph Neural Networks) is that the edges of a graph indicate the links between items or concepts, while the nodes represent the objects or concepts themselves. The Spatial-Temporal Graph Convolutional Network (ST-GCN) [3] is a sub-class of GNN that is specifically designed to handle graph-structured data with dynamic relationships over time. This architecture is able to model the interactions between nodes in a graph, considering both the spatial and temporal dimensions. The author [4] is credited with the original formulation of Convolutional Neural Networks (CNNs) on graphs. In his work, he adapted convolution to signals by employing a spectral construction.

B. Metaheuristics Algorithms for Feature Selection

Determining the best combination of characteristics is difficult and time-consuming to compute. Recently, metaheuristics have been helpful and dependable methods for tackling various optimization issues [5]. Metaheuristics offer superior performance compared to precise search methods since they do not have to look through the total search space. For instance, the Grey Wolf Optmization algorithm (GWO) was hybridized with the sine cosine algorithm to improve the power system's stability [6]. These hybrid strategies try to share the strengths of both of their components to increase the capability of

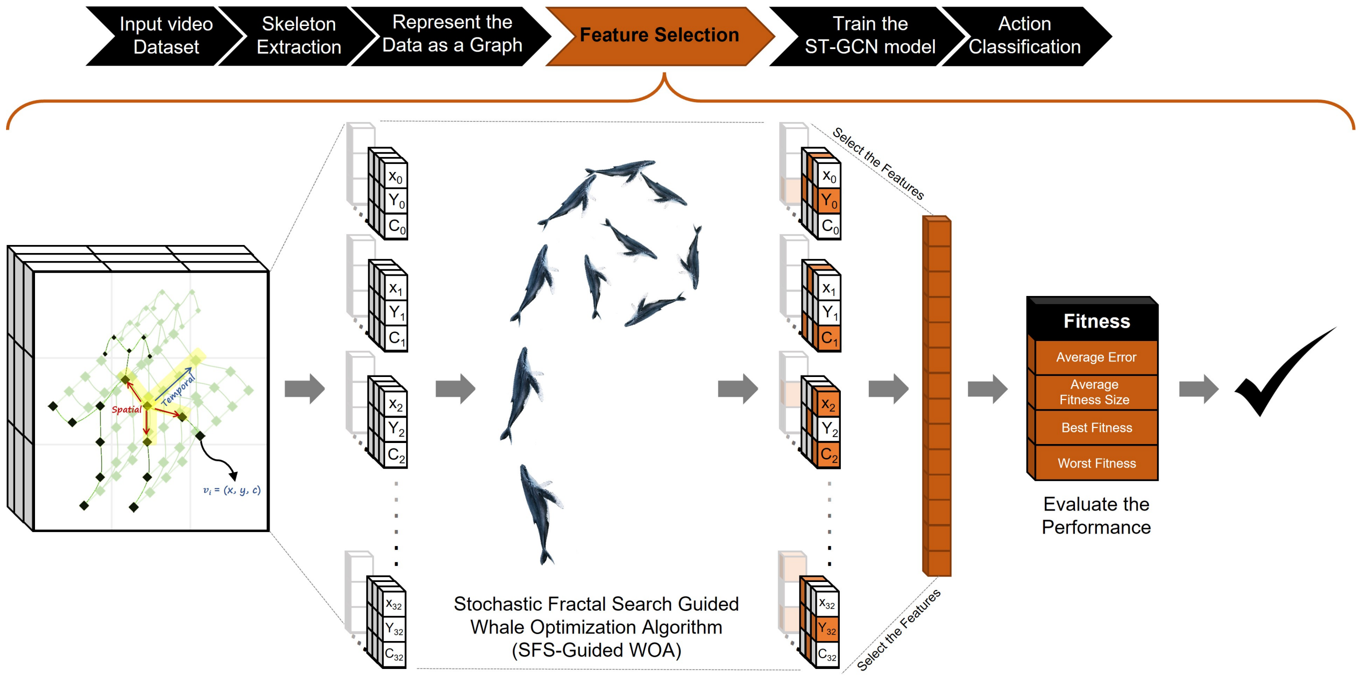

Fig. 1. The architecture of the proposed approach.

exploitation while simultaneously minimizing the likelihood of falling into an optimal local state. According to the findings of this research, the performance of the hybrid approaches was significantly superior to that of other global or local search methods.

Despite the excellent performance of the approaches described above, it is safe to say that none of them can handle all of the issues associated with the feature selection process. As a result, the solutions to problems involving feature selection can be improved by making enhancements to the approaches that are now in use.

III. The Proposed Methodology

From our perspective, the augmentation of the number of joints in the skeletal structure the BlazePose system's can provide a greater amount of data to facilitate the enhancement of the performance of the ST-GCN model compared to alternative topologies (e.g., OpenPose [7]). By adding feature selection layers, we hope to get a sense of how the shoulders and head move together during the activities for the Kinetics dataset and the NTU-RGB+D dataset. Fig. 1 shows the proposed pipeline definition for action recognition.

A. *SFS-Guided WOA for Feature Selection*

The Whale Optimization Algorithm (WOA) [8] as other Metaheuristic algorithms presented previously was inspired in nature. Specifically, the original WOA would drive the whales to swim in random circles around each other, just like the global search would. The Guided-WOA algorithm increases the number of whales during the exploration stage to improve its performance. This can push whales to explore additional territory without affecting their ability to hold leadership positions. This results in an enhancement in the Guided WOA's capability for exploration, and the diffusion process serves to find the best outcome.

The Guided-WOA algorithm is used to search for the optimal subset of features by optimizing a fitness function that represents the performance of the baseline model. For feature selection, each feature can be represented as a whale, and the objective is to find the main features that give the best performance on the ST-GCN.

B. *Fitness Function*

During the feature selection process, the metaheuristic algorithm selects which individual features from the population will be continuing onto the next generation. This can be achieved using the fitness function. Its value determines the probability of a feature being selected for further exploration and exploitation in the algorithm. Subsets of features with higher fitness scores are more likely to be selected for the next generation or iteration of the algorithm. We used the following equation to calculate the value of each solution:

$$Fitness = h_1 E(D) + h_2 \frac{|s|}{|f|} \tag{1}$$

Where s is the quantity of selected features in the iteration, f is the number of the complete population of features, $E(D)$ is the misclassification rate for each dimension and $h_1 \, \epsilon \, [0,1]$, $h_2 = 1 - h_1$ leverage the weight of the misclassification rate and the number of the chosen features.

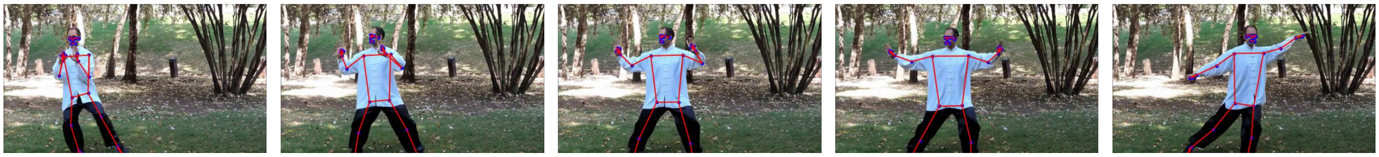

Fig. 2. A sequence of video frames corresponding to the ”tai chi” action from Kinetics

C. Evaluation Metrics

We evaluated the performance of our solution using the following evaluation metrics.

Average error: This metric describes the mean of the precision of the classifier by using the selected features of each iteration [9].

Average Fitness Size: The Fitness size can be described as the proportion of the original features size D considered in the features selection on each iteration. That is, to divide the size of the N selected features by D. Hence, this is metric describes the mean value of the proportions (i.e., $\frac{N}{D}$) calculated during the iterations of the optimization algorithm [9].

Best Fitness: This metric represents minimum value for the fitness function of a given optimization algorithm after completing all the tuning iterations [9].

Worst Fitness: Opposite to the *Best Fitness*, this metric describes the maximum value obtained by the fitness function of a a given optimization algorithm after completing all the tuning iterations [9].

IV. Experiment Settings

We resized each movie until it had the proportions 340 x 256 pixels. Consequently, we extracted the skeleton data using the framework provided by the BlazePose authors. This tool is very accurate to detect the joints on actions where there is a single performer. As a reference, we show the output obtained with this tool upon a ”tai chi” sample from the Kinetics dataset in Fig. 2. We did not consider any video frames in which the BlazePose system did not identify a skeleton as being present. Limiting the number of frames in the series of skeletons to just 300. As a result of this limitation, the majority of videos featured a limited amount of frames. Because of this, if a sequence contained fewer than 300 frames, we simply repeated the first few until we reached the appropriate length. On the other hand, if the sequence contained more than 300 frames, we arbitrarily removed some of the excess frames. We used spatial configuration partitioning to compute the convolution operation. Finally, we trained the model for 80 epochs.

Due to the nature of the Kinetics dataset which was extracted from Youtube with no quality checks, the BlazePose model presented difficulties in detecting the skeleton in many frames. Hence, we proposed to create a subset of samples for training. The sk-50 and sk-80 refer to two subsets consisting of samples using a minimum skeleton detection threshold indicated. For instance, the sk-50 st subset contains only samples on which it has been detected a skeleton at least in 50% of the frames on the videos.

V. Results and Discussion

The results are shown in two stages: First, the performance achieved in terms of SFS-Guided WOA with all feature selection performance metrics . The second, SFS-Guided WOA in conjunction with the BlazePose skeletal topology to construct a novel action recognition system.

A. Feature Selection Results

In this study, we provide a comparative analysis to assess the efficacy of the proposed feature selection algorithms with six other feature selection algorithms: bGWO [10], bPSO [11], bSFS [12], bWOA [13], bFA [14] and the bGA [15]. The

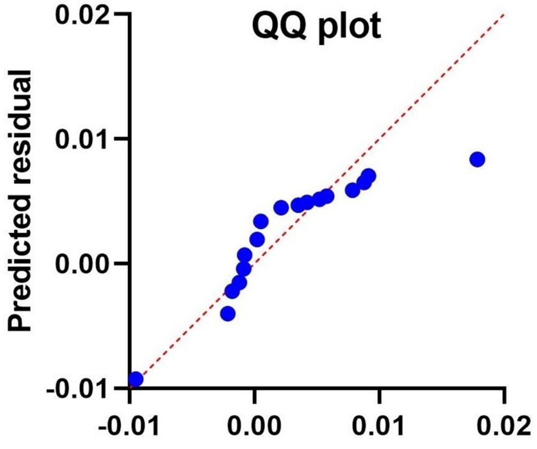

Fig. 3. Kinetics Performance

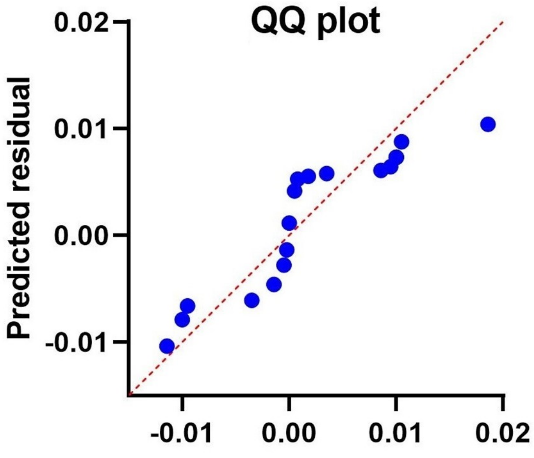

Fig. 4. NTU-RGB+D Performance

TABLE I
EVALUATION COMPARISON RESULTS

Average error							
Dataset	*SFS-GWOA*	*bGWO*	*bPSO*	*bSFS*	*bWAO*	*bFA*	*bGA*
Kinetics	0.27276	0.28648	0.2768	0.285465	0.27374	0.2806	0.27374
NTU-RGB+D	0.21014	0.22314	0.25091	0.230256	0.23596	0.24151	0.23083
Average fitness							
Dataset	*SFS-GWOA*	*bGWO*	*bPSO*	*bSFS*	*bWAO*	*bFA*	*bGA*
Kinetics	0.30049	0.33638	0.32279	0.3367	0.32376	0.33055	0.32376
NTU-RGB+D	0.22215	0.25712	0.28462	0.23	0.26981	0.27531	0.26473
Best fitness							
Dataset	*SFS-GWOA*	*bGWO*	*bPSO*	*bSFS*	*bWAO*	*bFA*	*bGA*
Kinetics	0.20535	0.22476	0.22476	0.285994	0.26358	0.26358	0.22476
NTU-RGB+D	0.17039	0.18731	0.22115	0.177115	0.19577	0.17039	0.20423
Worst fitness							
Dataset	*SFS-GWOA*	*bGWO*	*bPSO*	*bSFS*	*bWAO*	*bFA*	*bGA*
Kinetics	0.39988	0.41888	0.43829	0.360641	0.43829	0.4577	0.39946
NTU-RGB+D	0.32162	0.33115	0.33962	0.363322	0.35654	0.37346	0.38192

evaluation metrics measured based on the achieved results and recorded in Tab. I. In this table, six evaluation metrics are calculated and presented. The recorded values achieved by the proposed approach suggest that our approach outperforms the other alternative methods.

Another way to demonstrate the effectiveness of the proposed approach is through visualizing the achieved results. To achieve this, we utilized quartile-quartile (Q-Q) plots. A Q-Q plot is a graphical representation that compares two probability distributions by plotting their quantiles against each other. The purpose of a Q-Q plot is to determine if two sets of data come from populations with a similar distribution [16]. In Fig.3 and Fig.4, we show the plots for the obtained results based on the Kinetics and Ntu-RGB+D, respectively. From these figures, it can be noted that the performance of the proposed method is accurate in classifying the given actions.

On the other hand, the comparison of the time cost with previous works for feature selection is presented in Tab. II. As it can be noticed, we the time consumed by our proposed method is lower than the alternatives to select the needed important features for action recognition.

B. Action Recognition Results

The action recognition results achieved by the proposed approach is compared to the previous ST-GCN, Spatial-Temporal Graph Deconvolutional Networks (ST-GDN) [17], and BlazePose methods. The comparison results are presented in Tab.III in terms of the Kinetics dataset.

TABLE II
AVERAGE TIME CONSUMED FOR FEATURE SELECTION (IN SECONDS).

Method	Kinetics	NTU-RGB+D	Average Time
SFS-GWOA	**31.194**	**33.612**	**32.403**
bGWO	33.838	35.543	34.6905
bPSO	33.52	35.115	34.3175
bSFS	34.92	34.87	34.895
bWAO	33.327	34.448	33.8875
bFA	34.548	35.132	34.84
bGA	33.794	35.068	34.431

TABLE III
ACCURACY PERFORMANCE: KINETICS

Method	Top-1	Top-5
SFS-Guided WOA : BlazePose, sk-80	**56.87%**	**81.44%**
SFS-Guided WOA : BlazePose, sk-50	**51.79%**	**77.13%**
BlazePose, sk-80	37.38%	65.20%
BlazePose, sk-50	36.78%	61.69%
ST-GDN	37.30%	60.65%
ST-GCN	30.70%	52.80%

Regarding the NTU-RGB+D dataset, the evaluation was performed using the Cross-Subject (X-Sub) and Cross-View (X-View) criteria suggested by the dataset's authors. According to Tab. IV , The accuracy achieved by the proposed approach in case of sk-50 is 91.33% for X-View and 94.56% for the X-Sub, which are higher than the other methods included in the conducted experiments. In addition, the accuracy upon for the sk-80 subset is 93.13% and 96.74%, for X-View and X-Sub, respectively.

TABLE IV
ACCURACY PERFORMANCE: NTU-RGB+D

Method	X-View	X-Sub
SFS-Guided WOA : BlazePose, sk-80	**93.14%**	**96.74%**
SFS-Guided WOA : BlazePose, sk-50	**91.33%**	**94.56%**
BlazePose, sk-80	87.62%	91.75%
BlazePose, sk-50	87.30%	90.34%
ST-GDN	89.70%	95.90%
ST-GCN	81.50%	88.30%

VI. CONCLUSION

This study introduces a new method for action recognition by building the BlazePose skeleton topology on top of the ST-GCN architecture and selecting features with SFS-Guided WOA. We have chosen the Kinetics and NTU-RGB+D benchmark datasets to give a reliable basis for comparison with the baseline model in. When the visual data has been acquired in unconstrained contexts, we advocated using alternative skeletal detection criteria to increase the model's performance. We've demonstrated that BlazePose's topology may be improved by selecting the appropriate features for feet and hands, resulting in more precise data about the motion being captured. In addition, the suggested topology in this research can improve performance even more.

Acknowledgment

The authors wish to express their gratitude for the support received from the King Abdulaziz City of Science and Technology (KACST), which played a vital role in the successful completion of this work.

References

[1] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," *arXiv:2006.10204*, 2020. [Online]. Available: https://arxiv.org/abs/2006.10204

[2] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph Neural Networks: A Review of Methods and Applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[3] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *arXiv*, 2018.

[4] M. S. Alsawadi and M. Rio, "Skeleton-Split Framework using Spatial Temporal Graph Convolutional Networks for Action Recognition," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*, Paris, France, 2021, pp. 1–5.

[5] M. Abdel-Basset, L. Abdel-Fatah, and A. K. Sangaiah, "Metaheuristic algorithms: A comprehensive review," *Computational intelligence for multimedia big data on the cloud with engineering applications*, pp. 185–231, 2018.

[6] D. S. Khafaga, A. A. Alhussan, E.-S. M. El-Kenawy, A. Ibrahim, M. M. Eid, and A. A. Abdelhamid, "Solving Optimization Problems of Metamaterial and Double T-Shape Antennas Using Advanced Meta-Heuristics Algorithms," *IEEE Access*, vol. 10, pp. 74 449–74 471, 2022.

[7] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[8] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Advances in engineering software*, vol. 95, pp. 51–67, 2016.

[9] E.-S. M. El-Kenawy, M. M. Eid, M. Saber, and A. Ibrahim, "MbGWO-SFS: Modified binary grey wolf optimizer based on stochastic fractal search for feature selection," *IEEE Access*, vol. 8, pp. 107 635–107 649, 2020.

[10] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371–381, 2016.

[11] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation*, vol. 5. IEEE, 1997, pp. 4104–4108.

[12] K. M. Hosny, M. A. Elaziz, I. M. Selim, and M. M. Darwish, "Classification of galaxy color images using quaternion polar complex exponential transform and binary Stochastic Fractal Search," *Astronomy and Computing*, vol. 31, p. 100383, 2020.

[13] G. I. Sayed, A. Darwish, and A. E. Hassanien, "Binary whale optimization algorithm and binary moth flame optimization with clustering algorithms for clinical breast cancer diagnoses," *Journal of Classification*, vol. 37, pp. 66–96, 2020.

[14] Y. K. Saheed, "A Binary Firefly Algorithm Based Feature Selection Method on High Dimensional Intrusion Detection Data," in *Illumination of Artificial Intelligence in Cybersecurity and Forensics*. Springer, 2022, pp. 273–288.

[15] "A new local search based hybrid genetic algorithm for feature selection," *Neurocomputing*, vol. 74, no. 17, pp. 2914–2928, 2011.

[16] D.-M. Tsai and C.-H. Yang, "A quantile–quantile plot based pattern matching for defect detection," *Pattern Recognition Letters*, vol. 26, no. 13, pp. 1948–1962, 2005.

[17] W. Peng, J. Shi, and G. Zhao, "Spatial Temporal Graph Deconvolutional Network for Skeleton-Based Human Action Recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 244–248, 2021.