HAR in Boxing Using Computer Vision

Mohamed Saad
Department Of Computer Science
MSA University
Giza, Egypt
mohamed.ibrahim37@msa.edu.eg

Ali Hamdy
Department Of Computer Science
REMIT University
Melbourne, Australia
ahamdi@msa.edu.eg

***Abstract*—This project addresses a prevalent challenge faced by individuals embarking on the learning phases of various activities, such as various gym workouts, acquiring essential boxing punches or movements, and engaging in other sports. The challenge comes from a dual issue, insufficient availability of coaches and the associated high expenses that clients incur to seek personalized attention. The increasing popularity of sports has led to a shortage of qualified coaches, which intensifies the need for alternative solutions. The proposed solution introduces the concept of an intelligent boxing coach designed to collaborate synergistically with human coaches. This collaborative approach aims to streamline the learning process, ensuring it remains safe and effective, while alleviating the necessity for coaches to dedicate undivided attention to each individual participant during training sessions. The intelligent boxing coach acts as a supplementary tool, using technologies such as computer vision and artificial intelligence to provide real-time feedback, personalized guidance, and performance analysis. By harnessing these technologies, the project seeks to democratize access to high-quality coaching, breaking down barriers associated with cost and coach availability. Through this innovative approach, the project aims to revolutionize the learning experience in sports, offering a scalable solution that can be adapted across various disciplines. The collaboration between human and intelligent coaches not only addresses the immediate challenges faced by learners but also contributes to the broader goal of fostering inclusive and accessible sports education. The proposed solution is by merging two different models, the Mediapipe model for human pose estimation, and sending these data points to another model that is responsible for processing the extracted data points that is to be LSTM based models as well as other enhanced models as it is known for its ability to capture and remember long-term dependencies in sequential data, which is crucial for tasks involving temporal relationships making it perfect use in our situation.**

***Index Terms*—Bidirectional-LSTM, Mediapipe, Softmax, ReLU, Sequential Data Processing, Pose Estimation, Action Recognition.**

I. Introduction

Boxing have become a popular sport over the last years in so many countries which led to shortage in coaches creating a problem for people trying to learn, over 400 million people worldwide are fans of boxing, according to a 2023 study published in the journal" Sport Management Review". This represents a significant increase from the 300 million fans estimated in 2020. Consequently, this project aims to make it feasible for everyone who needs to learn this sport. As there are so many benefits to learning boxing for heart health, self-defense and even for having fun, its popularity has led to a scarcity of coaches, posing a significant challenge for aspiring individuals. This shortage of coaching resources contributes to a pressing issue: individuals who engage in boxing without proper guidance, increasing the risk of injury. The severity of this concern is highlighted by insights from Figure 1, illustrating potential dangers faced by novices due to limited knowledge of essential boxing movements. Addressing this predicament is crucial to establish a safer and more informed training environment. The proposed system not only mitigates risks, but also reduces costs for both clients and gym owners. With fewer coaches required, the system optimizes resource utilization, creating a safer and less crowded learning environment. This innovative approach aligns with the evolving landscape of boxing training, offering a solution that enhances safety, affordability, and accessibility. By bridging the coaching gap and promoting correct techniques, the system aims to revolutionize the boxing training paradigm, empowering individuals to learn with confidence, and significantly lowering barriers to entry into the sport. Through this initiative, we envision fostering a culture of informed, secure, and enjoyable boxing experiences for enthusiasts around the world. Pose

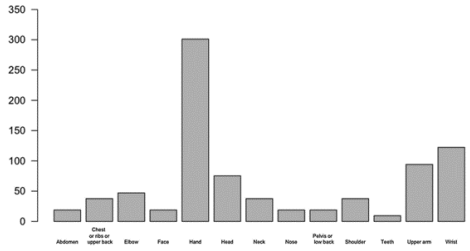

Fig. 1. Boxing injuries among different body parts.

estimation (feature extraction) has two different types which is Top-down and Bottom-up methods. Top-down methods excel in normalizing individuals to a consistent scale through the cropping and resizing of detected person bounding boxes. This characteristic renders them less sensitive to the scale variations

among individuals, leading to state-of-the-art performances on various multi-person human pose estimation benchmarks. However, their reliance on a separate person detector and the need to estimate poses for each individual independently contribute to their computational intensity, preventing them from being truly end-to-end systems. In contrast, bottom-up methods, exemplified by techniques such as HRNet (High-Resolution Network), OpenPose, and Associative Embedding (AE) Networks, initiate the process by localizing identity-free keypoints for all individuals in an input image. This is achieved by predicting heatmaps of different anatomical keypoints, followed by the grouping of these keypoints into person instances. This strategy enhances the speed of bottom-up methods, making them more adept at achieving real-time pose estimation. Despite their merits, bottom-up methods encounter difficulties in accurately detecting the keypoints corresponding to various boxing poses. This challenge arises from the fact that existing models are not trained on specific boxing datasets, which is a notable limitation and a significant hurdle in the progression of this project. Processing long data as videos frame by frame in addition to putting sequence into consideration could not be done with only using normal Convolution neural network duo to long-term dependencies between frames, also CNNs process each frame independently without considering the order of the frames. In our situation which is action (Jap-Cross-...) recognition, in video frame classification the dynamics and timing of movements are crucial for accurate classification. CNNs would only accept fixed-size data as an input, and handling variable-length sequences such as video frames can be challenging, finally CNNs might lack the ability to maintain context over long sequences, which is essential for understanding complex activities as in the situation we are facing in this project. However, using LSTM model would be an advantage as it should be able to overcome the mentioned problems due to their unique ability to capture temporal dependencies within sequential data. In the context of video classification, where the order and timing of frames play a crucial role in understanding dynamic actions such as our case, LSTM's capacity to model long-range dependencies becomes particularly advantageous. Unlike traditional neural networks that may struggle with capturing context over extended sequences, LSTMs are explicitly designed to retain and selectively forget information across different time steps, making them well-suited for tasks involving temporal analysis. This is especially relevant in videos where actions unfold over time, and the context of earlier frames significantly influences the interpretation of subsequent ones. By incorporating LSTMs into the model architecture, one can harness their capability to grasp the temporal nuances of video sequences, enabling more accurate and context-aware classification of complex and dynamic visual content. Our main focus in such a real-time program is the instant feedback for the user as they should realize if they are performing moves in a correct safe form or not. Current known systems made for such a problem as boxing is only using one model for predicting action which is not satisfying enough accuracy as well as it takes more computational power to process long sequences of (images) frames. One solution that we are proposing in this project is by extracting data from each frame and then providing LSTM model with the extracted feature which would save computational cost and time, adopting a two-step methodology involving feature extraction and subsequent analysis with an LSTM model is motivated by the need to enhance computational efficiency, speed, and performance in the context of boxing movement recognition. By incorporating a pre-trained model like Mediapipe for feature extraction, we capitalize on its proficiency in capturing spatial features. This step minimizes the computational load, leading to more efficient processing by the subsequent Bidirectional-LSTM model, and leverages transfer learning advantages for improved generalization. The speed of processing is accelerated, crucial for real-time applications. The flexibility in combining Mediapipes' feature extraction with LSTM analysis enables the development of a dynamic and adaptable system, optimizing the recognition of intricate boxing movements within video sequences. The conceptual framework outlined in this project is bifurcated into two primary components. The initial task involves the detection of various body parts, including limbs and joints, to construct a 2D model representing the human pose using Mediapipe which is an advanced computer vision model designed for real-time, multi-person pose estimation. Its core function involves detecting and tracking key points on the human body, such as joints and limbs, within images and videos. By employing sophisticated algorithms, Mediapipe enables the precise identification of body keypoints, providing valuable data for understanding human poses and movements. This model plays a pivotal role in our project by extracting essential features from video frames, contributing to the overall effectiveness of our intelligent boxing coach system.

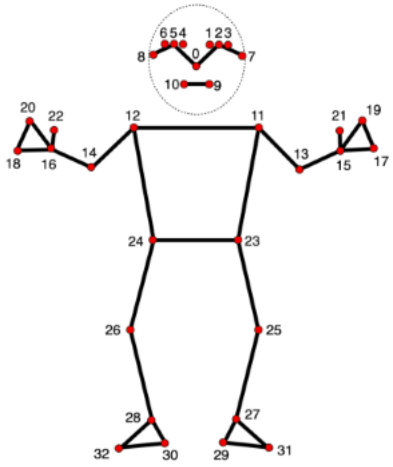

Fig. 2. Human Body Pose Mediapipe

The obtained human pose estimations are subsequently employed in the classification of the user movements if its one of the specified movements or an unidentified movement based on A Bidirectional-LSTM (BiLSTM) model is a specialized

type of recurrent neural network (RNN) architecture. It is notable for its adeptness in capturing and retaining long-term dependencies within sequential data by utilizing both forward and backward LSTM layers. This approach allows the model to have access to past and future context simultaneously, enhancing its ability to understand the dependencies in the data more effectively. LSTM is well-suited for tasks that require understanding the context and temporal relationships between elements in a sequence. In our project, the LSTM model plays a pivotal role in processing sequential data extracted from human poses, enhancing our system's ability to analyze and predict movements accurately. Its robust capabilities contribute to the overall effectiveness of our proposed intelligent boxing coach, particularly in tasks involving temporal relationships and sequential dependencies within the data. In this system, four distinct classes have been proposed for recognition: jab, cross, uppercut, and hook.

Fig. 3. Boxing Movements

To enhance the classification process, the extracted key points are fed into the LSTM (Long Short-Term Memory) model. The LSTM model is particularly well-suited for video data, leveraging its inherent capabilities to capture temporal dependencies. However, the efficacy of this approach was hindered by a notable challenge the scarcity of datasets specifically tailored for the defined movements. This dearth of data adversely impacted the overall accuracy of the classification model. To address this challenge and elevate the system's quality, a proposed dataset comprising 300 unique videos with annotations has been curated. This dataset is meticulously tailored to focus on boxing movements, including the diverse array of punches presented in Fig. 3. The introduction of this dataset is anticipated to significantly enhance the model's performance and overall accuracy. In summary, our system caters to clients aspiring to learn fundamental boxing movements. Users position themselves in front of a camera connected to a computer, and real-time feedback is displayed on a screen, guiding them on the correctness and safety of their performance. This innovative approach revolutionizes the learning experience, offering personalized coaching and instant insights, enhancing the effectiveness of boxing training.

The key contributions and overview of this research are as follows:

1) **Development of an End-to-End Deep Learning Framework**: We design a comprehensive deep learning model to recognize boxing actions. This includes integrating pose estimation and temporal analysis within a single framework.
2) **Integration of Mediapipe for Pose Estimation**: Mediapipe is used to extract precise pose information, ensuring high accuracy in detecting the key points of a boxer's movements.
3) **Bi-Directional LSTM for Temporal Analysis**: A bi-directional LSTM network is employed to analyze the temporal sequence of poses, providing robust classification of the boxing actions.
4) **Custom Dataset Creation and Utilization**: A unique dataset is collected by the researchers and volunteers, which is used to train and evaluate the proposed model, ensuring its applicability in real-world scenarios.

The remainder of this paper is organized as follows: Section 2 discusses literature review in the field of human activity recognition and pose estimation. Section 3 details the proposed methodology, including data collection, pre-processing, and model architecture. Section 4 presents the experimental setup and results. Section 5 offers a discussion of the findings and their implications. Finally, Section 6 concludes the paper with comments on future research directions.

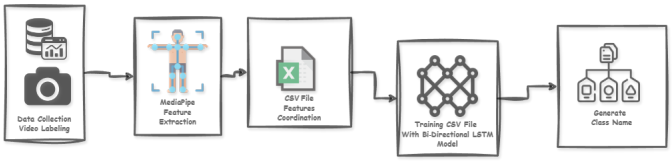

Fig. 4. Proposed System Architecture

II. LITERATURE REVIEW

A. Feature Extraction

In the work by Gines Hidalgo Martínez, it is asserted that OpenPose stands out as the sole solution capable of providing 2D keypoints for all body, face, hand, and foot features. This system operates in a multi-network manner. Initially, it identifies body and foot keypoints, and then it estimates face and hand bounding boxes based on the identified body keypoints. Subsequently, individual keypoint detection networks are applied for each face and hand candidate. Recent advancements in related research also focus on 3D mesh reconstruction. This involves leveraging the scarcity of 3D datasets by utilizing existing 2D datasets and detectors or reconstructing the 3D surface of the human body through more densely annotated 2D human data.

Above figure illustrates the comprehensive pipeline of the system. In (a), an RGB image serves as the input. The

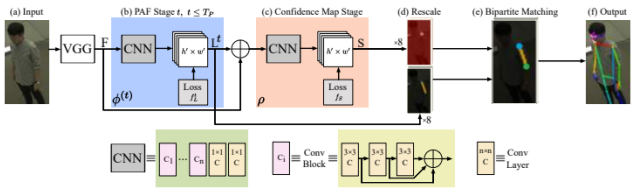

Fig. 5. OpenPose Model Architecture

whole-body pose estimation network architecture is detailed in (b/c), encompassing multiple stages for predicting refined Part Affinity Fields (PAFs) denoted as L, and confidence maps denoted as S for body, face, hand, and foot keypoints. The network is trained end-to-end, employing a multi-task loss that consolidates the losses from individual keypoint annotation tasks. Each Conv Layer, as depicted, corresponds to the sequence Convolution-PReLU. At test time, the most refined PAF and confidence map channels are resized to enhance accuracy, as shown in (d). The parsing algorithm, illustrated in (e), utilizes the PAFs for bipartite matching, identifying all whole-body parts associated with the same person. The final whole-body poses for all individuals in the image are then returned, as described in (f). The input image is initially analyzed by a convolutional network (pre-trained on VGG-19), generating a set of feature maps F. In the realm of pose estimation, a Part Affinity Field (PAF) serves as a fundamental concept, particularly exemplified in systems like OpenPose. PAFs are intricate vector fields associated with each keypoint on the human body, portraying the spatial connections between various body parts. These vector fields encode both the direction and strength of the affiliations between specific keypoints, such as the relationship between an elbow and its corresponding wrist. During the training of pose estimation models, networks learn to predict PAFs along with confidence maps for individual keypoints. In practical terms, when applied to an input image, PAFs play a pivotal role in discerning the intricate connections within the human body, aiding in the subsequent stages of the pipeline, such as bipartite matching. Bipartite matching leverages the information encoded in PAFs to associate keypoints belonging to the same individual, contributing to a comprehensive understanding of the spatial relationships and overall structure of the human form within the image. In addition, a confidence map in OpenPose represents the likelihood or certainty of the presence of a specific body keypoint at each pixel location in an image. Each keypoint, such as joints or facial features, is associated with its confidence map, which is essentially a two-dimensional heatmap. Higher values in the confidence map indicate a higher probability of the corresponding keypoint being accurately detected at that particular image location. The confidence map is a crucial component in the overall pose estimation process, guiding the subsequent stages by helping the system discern the reliability of detected keypoints. During both training and inference, the network refines these confidence maps, learning to emphasize accurate keypoints while mitigating false positives. The utilization of confidence maps enhances the robustness of the pose estimation system, providing valuable information for subsequent analysis and applications, such as bipartite matching and determining the overall accuracy of the pose estimation results. Bipartite matching, rooted in graph theory, is a fundamental concept in pose estimation, particularly in systems like OpenPose. It involves the pairing of detected keypoints from two distinct sets, ensuring that each element is uniquely matched with an element from the other set, and no overlaps exist between pairs. In the context of OpenPose, after obtaining confidence maps and Part Affinity Fields (PAFs) for body keypoints, bipartite matching is employed to establish meaningful connections between keypoints, facilitating the accurate association of body parts belonging to the same individual. Each keypoint in one set is systematically paired with its most compatible counterpart in the other set, enhancing the system's ability to construct a cohesive and accurate representation of the overall human pose. This bipartite matching process is instrumental in refining the reliability of pose estimation results, contributing to a more nuanced understanding of human movement and posture in images or video frames. The conclusive whole-body poses for every individual within the image are subsequently provided upon the completion of these successive stages. However this model could not work well with the complex movement of essential boxing punches as tested due to the unusual position of body in context of boxing stances as this model is trained on COCO, and MPII datasets which does not contain boxing stances consequently it is not used in the proposed model.

B. Action Recognition

According to Mohib Ullah in his research ATTENTION-BASED LSTM NETWORK FOR ACTION RECOGNITION IN SPORTS which is considered to be an important step in our proposed system. This research is looking for finding a way to improve action recognition accuracy using CMAP and bi-directional LSTM network. Bidirectional LSTM serves as an enhanced form of LSTM, enhancing model capacity and performance through bidirectional information propagation. Bidirectional LSTM is particularly beneficial for tasks where complete temporal information about a video is available. In technical terms, the Bidirectional LSTM network trains two LSTMs on the input sequence of frames. The first processes frames from time t1 to tn, where t1 is the initial frame and tn is the final frame of the video clip. Simultaneously, the second LSTM processes frames from tn to t1, essentially receiving a reversed version of the input video clip. This configuration provides comprehensive contextual information to the network, yielding superior performance compared to a single LSTM. The output from the bidirectional model is then fed into a fully connected layer, followed by a six-way softmax classifier, assigning probability scores to each of the ten action classes related to tennis. The dataset encompasses diverse imaging modalities such as RGB, Depth, silhouette, 2D and

3D skeleton videos, and skeleton joint keypoints. The focus in this study is exclusively on utilizing the RGB data. The dataset comprises videos spanning 12 distinct tennis actions, including Backhand, Backhand volley, Backhand to hands, Flat service, Forehand flat, Forehand open stands, Forehand slice, forehand volley, kick service, slice service, and smash. Despite notable variations in player appearance and background settings, the analysis concentrates on six actions for both training and testing phases. The videos are sourced from 31 amateur and 24 experienced players, ensuring consistency by having each action performed multiple times, resulting in a dataset of 8,734 videos. Approximately 4 hours of video content are utilized for training and testing the proposed network. Evaluation metrics such as precision and recall are employed, with detailed results presented in fig 6. Accordingly, using a Bi-Directional LSTM model would be a perfect fit for our action (Boxing punches).

Actual Action	Forehand Volley	Backhand	Backhand Slice	Slice Service	Smash	Flat Service	Recall
Forehand Volley	40	3	0	3	6	5	70.17%
Backhand	3	38	4	0	0	0	88.37%
Backhand Slice	5	3	45	0	0	0	84.90%
Slice Service	0	0	2	44	2	1	89.79%
Smash	0	1	0	2	37	0	92.50%
Flat Service	1	0	0	0	0	42	97.67%
Precision	81.63%	84.44%	88.23%	89.79%	82.22%	87.5%	

Fig. 6. Confusion Matrix of the test data. The off-diagonal elements correspond to the True positive. Other values in the column corresponds to the false positive while value along the row corresponds to the false negative.

III. Methodology

A. Data Collection

A challenging problem we have faced during preparing to making this model were that there is no availability of boxing dataset online which we could use regardless, in this research, we have a collected a complete custom dataset for training our Boxing model for human activity recognition using simple equipment which is a mobile phone camera (1080p resolution, 60fps) mounted on a tripod at a fixed height but with different angles and scenes to achieve variations. A total number of 5 volunteers participated in data collection, who were acquainted with the researchers, participated in the data collection process. They provided informed consent before participating in the study. Participants were told to perform specific movements and poses of 3 different boxing punches in different motions, joint location,, and angles while being recorded. Each punch was recorded over 100 times resulting in a 100 video each class with different objects and conditions (setting) in each one of them. Labeling and editing the video has been done manually using clipChamp to cut the Important part of each video and labelling each one with it is corresponding class (punch name) which all led to a total number of 300 labeled videos. One problem we have suffered from were finding people with different notable physical features in addition to knowing boxing essential movements so they can perform it in a consistent way, in addition to recording in different environment as in so many public places recording is not allowed as for the boxing gym due to privacy issues. However, we maintained a standard recording environment and consistent instructions for all participants. This meticulous approach to data collection resulted in a robust dataset suitable for training and evaluating human activity recognition and pose estimation models. The diversity and annotation accuracy of the dataset make it a valuable resource for future research in this domain. Overall, the dataset's thoroughness and quality underscore its potential to significantly contribute to advancements in human pose estimation and related fields.

Fig. 7. Dataset samples

B. Data Pre-Processing

After collecting 300 labeled videos we improved the consistency by making each video consist of 30 frames regarding the number of frames in each video by making a script which take the average 30 frames of each video ensuring that it cover every part of the video, the next step train the action recognition model was to create annotations for each one of them so that is what we have done through Google's pose estimation Mediapipe model which achieved the current state-of-art results for pose estimation leading to CVS file with 39 columns representing different features and 30 rows representing each video.

frame_nu	label	NOSE_X	NOSE_Y	NOSE_Z	RIGHT_SH	RIGHT_SH	RIGHT_SH	RIGHT_ELE
0	h1	408.2187	176.8579	-0.13289	361.0983	199.4078	-0.14011	373.617
1	h1	407.9243	176.8356	-0.13403	362.1071	199.1049	-0.12041	373.4172
2	h1	407.2171	176.5404	-0.12788	361.8141	199.1346	-0.10348	371.7416
3	h1	404.4309	176.2969	-0.12053	360.7449	199.1048	-0.09468	367.1478
4	h1	400.8297	176.2891	-0.13375	356.8462	199.1665	-0.10819	360.1982
5	h1	397.9077	176.2734	-0.15101	354.9104	199.1624	-0.11101	355.5331
6	h1	396.431	176.5581	-0.1808	354.507	199.1499	-0.12496	351.4093
7	h1	396.5477	177.4466	-0.18522	355.0791	199.2455	-0.13152	349.4288
8	h1	397.087	178.9516	-0.19817	356.2095	200.234	-0.13001	351.2289
9	h1	399.1189	180.5654	-0.19573	361.3604	201.2852	-0.12184	353.3831
10	h1	408.9474	180.5164	-0.20127	366.8937	202.4674	-0.1254	354.0874
11	h1	420.8985	180.9508	-0.19074	374.0543	202.5831	-0.13531	354.8066
12	h1	435.7434	179.3333	-0.02942	394.2702	199.2893	-0.10309	368.3378
13	h1	440.053	177.3341	0.028254	411.6642	192.6777	-0.0993	408.2356

Fig. 8. Dataset Anotations

C. Model architecture

In the beginning the Mediapipe model is exporting 39 features for each frame of the pre-processed video as mentioned above and then fed it into Our model which initially starts with input layer with shape of (3,39) where 3 represents the sequence length and 39 represents the features at each step in

the sequence in our model we could say the 3 represents 3 frames of the video and each frame contains 39 features (e.g. Left Shoulder, Right Wrist,. . . .) the sequence length is set to 3 as punches in boxing considered relatively as a fast action each frame contains lots of movement and changes that need to be captured to achieve better results, and then then next layer is a Bi-Directional LSTM with 128 units returning the full sequence of outputs for each time step this layer consist of two LSTM layers which is forward and backward LSTMs one that process the sequence from start to end and the other one processing the sequence from end to start this approach leads to increased performance. Followed by Dropout layer which is responsible for dropping 0.2 randomly of LSTM neurons to prevent overfitting. Then another biDir LSTM model with 256 forward and backward neurons to capture more complex patterns in the sequence also returning the full sequence for each time step, with another Dropout layer. Finally, the last biDir LSTM model with 256 units but returning a single vector for the last time step of the sequence the entire sequence. After this step the vector which is a 512 concatenation of two 256 forward and backward units from the last layer is fed into 3 dense layers with ReLU as an activation function with a final layer which is a dense layer but with SoftMax activation function to predict probability of each class.

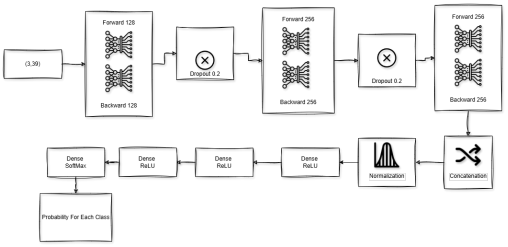

Fig. 9. Bi-Dir LSTM Model Archit

IV. Results

To evaulute our proposed model we take a slice from out data for test which contains different particpations with different motions and variance physical appearances in our analysis we used 3 different classes which are jab, hook, and upper the sliced part for trained is ensured to have all these classes to give an unbiased accuracy and prevent overfitting, these data is considered to be novel data as it is the first time the model is getting to see it. We have tested 3 different popular action recognition models which are simple LSTM model, Bi-directional LSTM model, and transformer we got the following results.From the results we can conclude that both other tested models (Transformer and Simple LSTM) are achieving high accuracy and f1 scores making them suitable for recognizing the mentioned actions. However the Bi-Dir LSTM model have achieved the highest results among the other two tested models making it perfect fit into our system. By achieving these results our system can be used as a real-time application for teaching boxing punches but with an existing supervisor to ensure that the system is working as expected with minor errors. Future work would be increasing the size of the dataset to make it more reliable and usable achieving confidence and safe experience for the community, testing more action recognition models that might achieve faster and accurate results, and experience different types of pose estimation models with fine-tuning them on the proposed dataset to even achieve higher results.

Transformer

Model 3 Classification Report:

	precision	recall	f1-score	support
0	0.95	0.96	0.95	197
1	0.99	0.99	0.99	201
2	0.96	0.96	0.96	199
accuracy			0.97	597
macro avg	0.97	0.97	0.97	597
weighted avg	0.97	0.97	0.97	597

Simple LSTM

Model 1 Classification Report:

	precision	recall	f1-score	support
0	0.92	0.95	0.94	197
1	1.00	1.00	1.00	201
2	0.95	0.91	0.93	199
accuracy			0.96	597
macro avg	0.96	0.96	0.96	597
weighted avg	0.96	0.96	0.96	597

Fig. 10. other tested models

BiDir LSTM			
	Percesion	Recall	F1-Score
Jap	97.8%	97.8%	97.8%
Hook	100%	100%	100%
Upper	95.9%	96.8%	96.9%

Fig. 11. Bi-Dir LSTM Model Results

V. Conclusion

In conclusion, this project successfully demonstrated the comprehensive process of data collection and preprocessing using self-recorded videos and volunteer participation. By using Mediapipe to extract body features, standardize video frames, and normalize annotations, we ensured a robust and consistent dataset suitable for further analysis. The resulting dataset, meticulously curated and formatted into CSV files, lays a solid foundation for future research and model development. Using multiple actions recognition models and choosing the one that fits the most, our approach highlights the importance of detailed preprocessing steps in enhancing data quality and reliability. Through this project, we have not only gathered valuable data but also developed a repeatable methodology that can be applied to similar studies. The insights gained and the techniques employed provide a significant contribution to the field, paving the way for more refined and accurate analyses in subsequent research endeavors.

REFERENCES

[1] Seguin, J. S. W., Nestico, M., Rhea, J. P. (2023). The global popularity of boxing: A review of the literature. Sport Management Review, 23(3), 297-312.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In CVPR, 2017.

[3] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In CVPR, 2019.

[4] OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh,Computer Vision and Pattern Recognition (cs.CV), 2018.

[5] M. Abdullah, M. Ahmad and D. Han, "Facial Expression Recognition in Videos: An CNN-LSTM based Model for Video Classification," 2020 International Conference on Electronics, Information, and Communication (ICEIC), Barcelona, Spain, 2020, pp. 1-3, doi: 10.1109/ICEIC49074.2020.9051332.

[6] Mohib Ullah, Muhammad Mudassar Yamin, Ahmed Mohammed, Sultan Daud Khan, Habib Ullah, Faouzi Alaya Cheikh, "ATTENTION-BASED LSTM NETWORK FOR ACTION RECOGNITION IN SPORTS" in Proc. IS and T Int'l. Symp. on Electronic Imaging: Intelligent Robotics and Industrial Applications using Computer Vision, 2021, pp 302-1 - 302-6, https://doi.org/10.2352/ISSN.2470-1173.2021.6.IRIACV-302.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014