Deep Learning-Based Golf Swing Sequence Analysis

Amir Hajian
Department of Electrical Engineering
Chulalongkorn University
Bangkok, Thailand
amirhajian85@gmail.com

Karit Sookpreedee
Department of Electrical Engineering
Chulalongkorn University
Bangkok, Thailand
6230055921@student.chula.ac.th

Kingrak Phairoh
Department of Electrical Engineering
Chulalongkorn University
Bangkok, Thailand
6230040421@student.chula.ac.th

Watchara Ruangsang
Department of Electrical Engineering
Chulalongkorn University
Bangkok, Thailand
watchara.ru@student.chula.ac.th

Supavadee Aramvith
Multimedia Data Analytic and
Processing Research Unit
Chulalongkorn University
Bangkok, Thailand
supavadee.a@chula.ac.th

Abstract— **Golf is widely recognized as one of the most popular sports globally. However, one drawback of playing golf is the relatively high cost of equipment and coaching. While numerous training programs are available to assist players in their practice, there is currently no swing analysis program developed by Thai professionals. In this project, advanced deep learning models were employed: SwingNet, capable of predicting the sequence of eight golf swing events in videos and determining the confidence level of each swing, and MoveNet, designed to identify joint positions on the body and represent them as skeletons. These models were integrated into a customized template-matching algorithm that utilized angle-based measurements to analyze the sequence of golf swings. This analysis assessed the similarity score, represented as a percentage, between two individuals for each golf swing event. Furthermore, various techniques were implemented to enhance the efficiency of SwingNet. Through performance evaluation, it was observed that the efficiency of SwingNet surpassed by one percent compared to the pre-trained model.**

Keywords—action recognition, deep learning, cosine similarity score

I. INTRODUCTION

Golf is a popular sport worldwide, and according to research [1], the number of golfers globally has risen from 61 million in 2016 to 66.6 million in 2021. The demand for playing golf continues to grow each year. The most important limitation of playing golf is the significant expense of hiring a coach to learn the correct golf swing technique. Swinging a golf club involves intricate coordination of the entire body, making it a multifaceted motion that demands substantial synchronization.

Traditionally, amateur golfers used to rely on professional instructors who would give them real-time feedback based on naked-eye observation. However, because of the fast-paced characteristics of the golf swing, it was not always easy to identify the underlying issues immediately. As a result, slow-motion video analysis has gained popularity as a means of closely examining the complexities of the golf swing [2]. However, scrubbing through a slow-motion video to identify these eight golf swing sequences is impractical since only one event can be observed.

Nowadays, image/video-based convolutional neural networks (CNNs) have demonstrated impressive capabilities in various video recognition algorithms such as action localization and action detection in a sequence of frames.

This project aims to develop a program for analyzing golf swing sequences using deep learning techniques by inputting golf players' videos in a pose estimation model for swing analysis. It focuses on front-view video golf swings and calculates the similarity scores of each golf swing phase in percentage units. It involves training the model using a developed pose estimation technique to generate scores for swing analysis. Additionally, a deep learning model for detecting seventeen key points on the player's body. Subsequently, the program has been tested to evaluate real-world scenarios. The proposed model is anticipated to facilitate more accessibility to golf for amateur golfers and provide convenient training opportunities for beginners to improve their golf skills.

II. RELATED WORK

Utilizing an action recognition model is a practical solution to predict and analyze the complex tasks involved in the golf swing. The initial utilization for convolutional neural networks to recognize the action was explored by Karpathy et al. [3], who explored various techniques for integrating temporal information of a sequence of images. Simonyan et al. [4] further advanced this research by integrating optical flow into the action recognition architecture.

The 3D CNN is another technique to incorporating temporal information, which applies convolution across the video sequence.

By combining a 3D neural network and a two-stream architecture, Carreira et al. [5] developed the I3D architecture. Recurrent neural networks (RNNs), specifically those with long short-term memory (LSTM) cells, offer an alternative method for capturing long-term temporal dependencies in data [6]. Donahue et al. [7] were the first to apply this concept to action recognition, utilizing an LSTM network in a long-term recurrent convolutional network (LRCN), where the obtained 2D features were passed to the LSTM model.

Multimedia Data Analytics and Processing Research Unit, Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok 10330, Thailand.

Yeung and Russakovsky [8] constructed an end-to-end algorithm that combined extracted features with reinforcement techniques to gain the ability to identify the beginning and conclusion points of events. This method enables the consideration of only a subset of frames rather than the entire sequence.

Several studies have successfully combined 3D neural networks with object detection models. For instance, in [9], the 3D model was integrated with Faster R-CNN, while in [10], the 2D image regions model extended to the 3D video, resulting in the creation of Tube-CNN (T-CNN).

Mcnally et al. [11] adopted action recognition and temporal action detection to extract event frames in golf swing videos, enabling automated analysis of the swings. They introduced a novel video database specifically designed for golf swing sequencing. It addresses the need for a comprehensive collection of golf swing videos to aid in swing analysis. The database includes various phases or positions of the golf swing events to facilitate research and development in the field of golf swing analysis. These events consist of 8 positions [11], which are as follows:

(1) Address (A): The starting position of the golf swing before the club moves to the right side of the body.

(2) Toe-up (TU): The position where the clubhead points towards the right side of the body, with the club parallel to the ground.

(3) Mid-backswing (MB): The position where the club moves upward to the right side, with the hands at the level of the body and the arms parallel to the ground.

(4) Top (T): The position where the club reaches the highest point, with the clubhead facing towards the left side of the body.

(5) Mid-downswing (MD): The position where the club moves downward, with the hands at the level of the body and the arms parallel to the ground.

(6) Impact (I): The position where the clubhead contacts the golf ball.

(7) Mid-follow-through (MFT): The position where the club moves towards the left side of the body, with the club parallel to the ground.

(8) Finish (F): The ending position of the golf swing before the player returns to the starting position.

Additionally, this model determines video representation, action recognition, and temporal action detection tasks. The model output identifies and classifies different swing events in a graph containing eight. However, one limitation of this model is the absence of a similarity score (confidence level) typically represented as a percentage and a motion prediction capability.

III. PROPOSED MODEL

The architecture of the proposed Golf Swing model is shown in Figure 1. The proposed deep learning model aims to develop a golf swing analysis system. Two different front-view golf swing videos are used as input with a size of 160×160 pixels. The SwingNet model classifies all eight swing poses in the golf swing sequences. The MoveNet model detects various key points on the players' bodies, consisting of seventeen key points. The template-matching algorithm calculates the similarity scores based on the cosine similarity values for each swing pose. Finally, the similarity scores of each eight-swing pose and the total similarity score are calculated in percentage units.

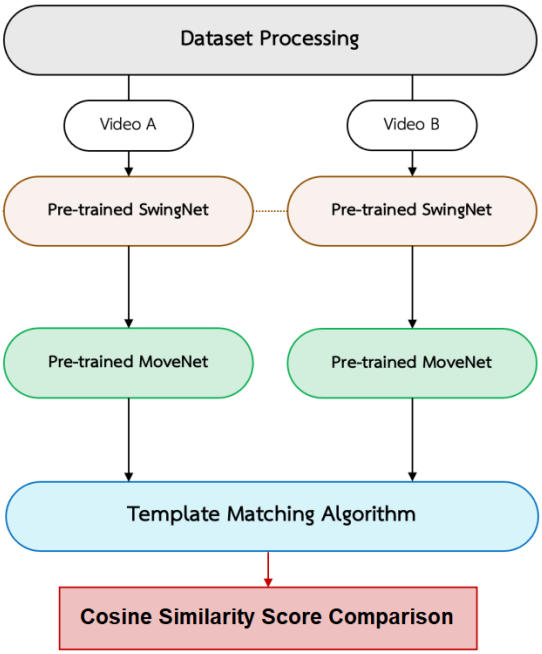

Fig. 1. The network architecture of the proposed golf swing analysis. Captures the similarity between the two videos and presents it as a percentage value.

A. SwingNet

SwingNet [11] is a model developed specifically to detect and classify golf swing motions, and its main architecture is illustrated in Figure 2.

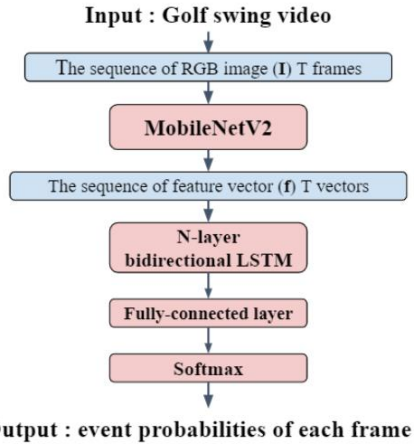

Fig. 2. The architecture of the swing recognition.

The SwingNet model takes video data as input, representing golf swing videos. The videos are then converted into an equal number of RGB images, corresponding to the number of frames in the video. These images are passed through the MobileNetV2 model, which performs the classification of golf swing motions and outputs a feature vector. The MobileNetV2 model is a low computational cost and a lightweight model due to utilizing additional depthwise separable convolutions, the Inverted Residual Bottleneck

Layer, and the removing ReLU activation in the last layer of the model. This model increases the classification performance compared to the MobileNetV1 while reducing the model size and gaining efficient computational cost.

The feature vector is then sent to a bidirectional Long Short-Term Memory (LSTM) [12] neural network with N layers (N-layer Bidirectional LSTM) that is developed from Recurrent Neural Networks (RNNs). The bidirectional approach has incorporated the weights from both directions and improves the learning capability of the model for robust prediction. Before the final prediction, a softmax function adjusts the output values to be within the range of 0 to 1, representing confidence levels. The output data is a probability distribution known as "confidence scores," which the SwingNet model can predict for all eight swing motions. The confidence score value determines a person's motion likelihood in each frame, assigning the graphical confidence profile to eight frames.

B. MoveNet

The MoveNet model is a bottom-up model that utilizes heatmaps to localize key points on the human body. The architecture consists of two main components: feature extraction and prediction, as shown in Figure 3. The feature extraction process combines the MobileNetV2 [13] model and the Feature Pyramid Network [14]. The reason for using this framework is to process images at multiple resolutions to improve accuracy while maintaining efficient computational costs. The concept of the Feature Pyramid Network is to create a top-down pathway to reconstruct different layers. In Figure 3, skip connections are added in the intermediate levels to mitigate the impact of both downsampling and upsampling while improving object localization accuracy. There are four predictions steps:

(1) Predict the center point heatmap of a golf player, by computing the mathematical average of all key points that belong to that person and selecting the region with the highest value.

(2) Predict the set of keypoint vectors (keypoint regression field) from the pixels associated with the golfer's center point.

(3) Predict all keypoint heatmaps of the player utilizing multiplication heatmap pixel by a weight value that is inversely proportional to the distance from the regressed keypoint.

(4) Predict the offset compensation by identifying the coordinates with the highest value in each keypoint channel of the heatmap.

C. Template Matching

Template Matching Algorithm [15] is a digital image processing technique used to identify the desired object by defining a template image, which is then matched with the corresponding region in the input image. This method involves using a convolutional neural network (CNN) to extract and match features from the image with the template. In the proposed model, the template matching detects the similarity between a reference scene and a comparison scene captured by a golf player user. The detected keypoint data represents the significant points on the human body that can be considered for pose estimation during a specific swing movement, depending on the player's orientation and varying physical characteristics in the scene.

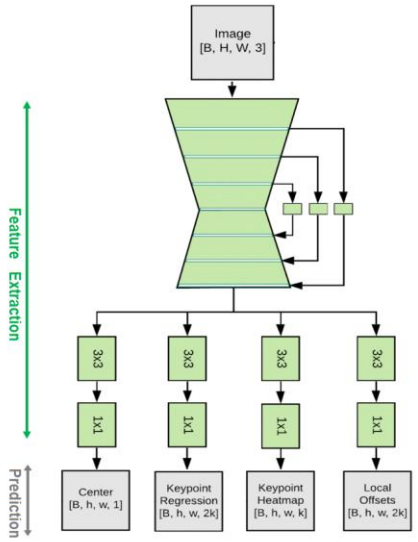

Fig. 3. MoveNet architecture. Feature extraction and prediction.

Equation (1) denotes the computation of angle a_i between detected key points of $z_1, z_2, and\ z_3$.

$$a_i = \cos^{-1} \frac{z_1 z_2 \ . \ z_1 z_3}{\|z_1 z_2\| \, \|z_1 z_3\|} \qquad (1)$$

D. Cosine Similarity

Similarity evaluation relies on the angle estimation of certain key points representing the player's body, inspired by previous research [16]. The similarity will be measured in percentage units. Finding the similarity between two vectors representing a certain aspect of two individuals can be obtained using (2), which measures cosine similarity.

$$CS(A,B) = \frac{A \ . \ B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} a_i \, b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}} \qquad (2)$$

n represents the number of features in the data, which is 18.
A represents a vector in an n-dimensional space with components $a1, a2, a3, \ \ldots, an$.
Brepresents a vector in n-dimensional space with components $b1, b2, b3, \ \ldots, bn$.
a_i and b_i represent angle values.
The cosine similarity value will be high when the angle between the two vectors is small, indicating a high similarity, and vice versa. The advantage of cosine similarity is that it does not consider the magnitude of the vectors. The resulting value ranges between 0 and 1, which can be interpreted accordingly.

IV. EXPERIMENTS

A. Experimental Settings

We were training our SwingNet model with the GolfDB dataset [11]. The dataset consists of a total of 1400 videos which are labeled with the name of professional golfers,

gender, different views, and the frame numbers in various shouts.

In our proposed model, we utilized the data augmentations technique. We investigate different augmentation approaches, including horizontal flip, random rotation between 30-70 degrees and random rotation between 10-180 degrees. Additionally, we crop with an angle of 5 degrees parallel to the x-axis and 5 degrees parallel to the y-axis in our re-trained SwingNet. We trained our swingNet model with Adam optimizer and set the learning rate to 0.001. The training batch size is 22 and 1800 iterations, utilizing 256 LSTM hidden size with 10 Frozen layers.

The performance of the SwingNet model is assessed using the Percentage of Correct Events (PCE). PCE evaluates the accuracy of key points by setting a threshold [17]. It involves comparing the percentage of correctly predicted events to the ground truth, determined by (3).

$$\delta = \max\left(\left[\frac{n}{f}\right], 1\right) \qquad (3)$$

Where δ is the acceptable difference between the frame numbers obtained from prediction and those obtained from the database. n represents the total number of frames in the video database, ranging from the start address to the end address. f denotes the frame rate of the video [frames/second].

For the MoveNet model, we utilized the original pre-trained model which has been trained with the COCO Keypoint Dataset Training Set 2017 [14, 18], which consists of 28,000 images with up to 2 people per image. Additionally, the model will utilize the Active Dataset Training Set, which contains 23,500 images with one person per image.

B. SwingNet Model Result

The performance of our modified SwingNet is shown in Table 1. PCE evaluations with different training configurations are compared. We were training our modified SwingNet model with an increased amount of data using the rotation augmentation technique with random angles between 10 and 180 degrees and cropping the edges parallel to the x-axis at 5 degrees and parallel to the y-axis at 5 degrees, resulting in a PCE total value of 72.5%. This is higher than the original pre-trained SwingNet model.

TABLE I. COMPARISON OF PCE VALUES FOR EACH EVENT AND PCE TOTAL IN PERCENTAGE BETWEEN THE DIFFERENT TRAINED MODEL AND THE ORIGINAL MODEL (SWINGNET)

Model	PCE of Each Swing Event								PCE total
	A	TU	MB	T	MD	I	MFT	F	
SwingNet	20.6	84.6	86.9	67.7	98.0	98.3	96.3	19.7	71.5
Horizontal Fleep	18.9	86.6	85.1	62.3	97.4	98.6	97.7	13.4	70
Rotate 30-70 Degree	24.3	81.4	86.9	66.6	98.0	97.4	95.1	23.4	71.3
Rotate 10-180 Degree	**27.7**	**80.3**	**88.0**	**75.1**	**97.4**	**94.6**	**96.9**	**20.3**	**72.5**

Figure 4 demonstrates expert swing golf video 584 of the GolfDB dataset. The results obtained from feeding these videos into the SwingNet model are as follows: Eight Frames of video show the detected swing golf actions, along with the frame sequence number, the name of the detected action, and the confidence score for each action.

Fig. 4. Eight swing events were obtained from video 584.

C. MoveNet Model Result

The golfer joints position from video 584 of the GolfDB dataset is shown in Figure 5.

As demonstrated, eighteen joint positions are detected in eight swing event frames. For convenience in using the results from the MoveNet model, the extracted joint data have been converted to angle values, which describe the flexibility.

D. Proposed Model Result

The Cosine Similarity between each pose calculated by (2), compared golf swing video 584 with 612 of the GolfDB dataset is shown in Table 2.

Figure 6 shows the heatmap of the coherence values, representing the Average Cosine Similarity of the nine expert-annotated videos from the dataset and one self-performed video.

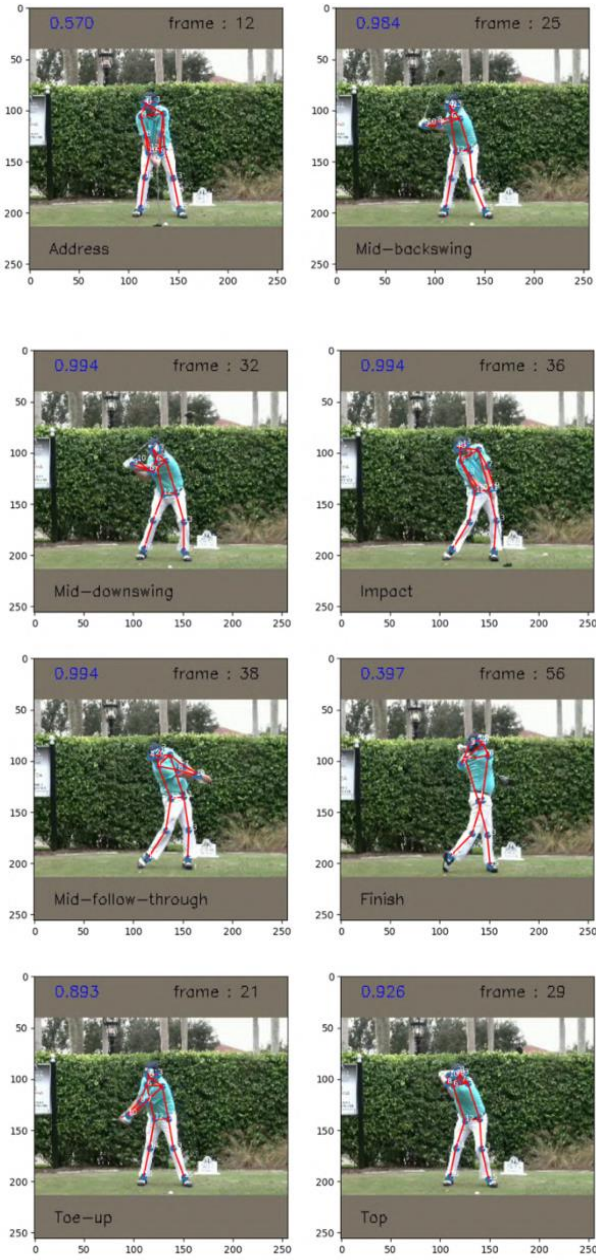

Fig. 5. The results MoveNet model to joint detection on video 584.

TABLE II. THE ACCURACY OF VIDEO 584 IN PERCENTAGE COMPARED WITH VIDEO 612 (CONSIDERING ALL EIGHT FRAMES).

Swing Motion Event	Cosine Similarity (%)
Address (A)	99.69
Toe-up (TU)	99.88
Mid-backswing (MB)	98.04
Top (T)	97.69
Mid-downswing (MD)	98.62
Impact (I)	99.71
Mid-follow-through (MFT)	99.46
Finish (F)	97.63
Total Accuracy	**98.84**

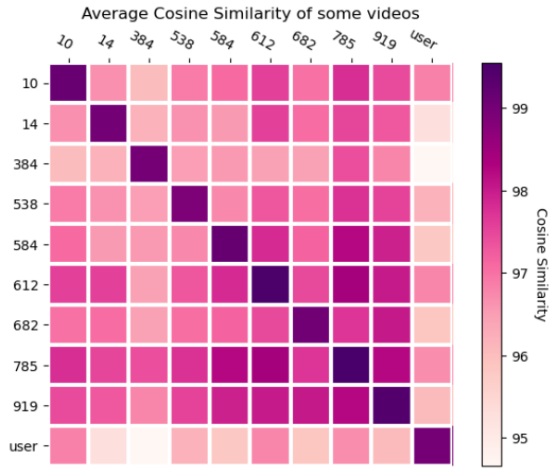

Fig. 6. Heatmap of the similarity values by Average Cosine Similarity.

From Figure 6, it can be observed that the darker shades indicate higher average similarity values. The average similarity values of self-performed video (user) compared to other expert-annotated videos are generally lower (lighter shades). This is because of the poor similarity of an amateur golf swing motion.

V. CONCLUSION

The model proposed in this study aims to classify and evaluate the similarity of golf swings in the swing sequences between two individuals and demonstrate the confidence score represented as a percentage. The model can accurately assess the similarity score between two individuals for each golf swing event by utilizing angle-based measurements and a customized template-matching algorithm.

Our model integrates two deep learning-based swing classifiers (The swingNet model and the MoveNet). Our model addresses the limitation of the previous model, which lacked a confidence level (similarity score) and improved motion prediction capability. The efficiency of the SwingNet model has improved (PCE=72.5) by adjusting hyperparameters and applying data augmentation techniques in training our model. Our model has demonstrated the robust ability to predict skeleton-shaped joint positions of golf swings and generate swing profiles for the sequence of eight golf swing events. The proposed model provides valuable insights into the performance of golfers, enabling coaches and players to understand better and improve their swings.

ACKNOWLEDGMENT

This research is funded by "Thailand Science Research and Innovation Fund Chulalongkorn University (CU_FRB65_ind (9)_157_21_23)" and The NSRF via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation [grant number B04G640053] and Thailand Science Research and Innovation Fund Chulalongkorn University (IND66210019). We would like to express our sincere gratitude to the Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University.

REFERENCES

[1] R&A, "Record Numbers Now Playing Golf Worldwide," 14 December 2021. [Online]. Available: https://www.randa.org/en/articles/record-numbers-now-playing-golfworldwide.

[2] Guadagnoli, M., Holcomb, W., & Davis, M. (2002). The efficacy of video feedback for learning the golf swing. *Journal of sports sciences*, *20*(8), 615-622.

[3] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

[4] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, *27*.

[5] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp.6299-6308).

[6] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[7] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp.2625-2634).

[8] Yeung, S., Russakovsky, O., Mori, G., & Fei-Fei, L. (2016). End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2678-2687).

[9] Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2018). A better baseline for ava. *arXiv preprint arXiv:1807.10066*.

[10] Hou, R., Chen, C., & Shah, M. (2017). Tube convolutional neural network (T-CNN) for action detection in videos. In *Proceedings of the IEEE international conference on computer vision* (pp. 5822-5831).

[11] McNally, W., Vats, K., Pinto, T., Dulhanty, C., McPhee, J., & Wong, A. (2019). Golfdb: A video database for golf swing sequencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 0-0).

[12] Lee, C. (2017). Understanding bidirectional rnn in pytorch. Towards Data Science (November 2017),

[13] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).

[14] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

[15] Abu-Bakar, S. A. (2019). Advances in human action recognition: an updated survey. *IET Image Processing*, *13*(13), 2381-2394.

[16] Borkar, P. K., Pulinthitha, M. M., & Pansare, A. (2019). Match Pose-A System for Comparing Poses. *International Journal of Engineering Research and Technology (IJERT)*, *8*(10).

[17] Munea, T. L., Jembre, Y. Z., Weldegebriel, H. T., Chen, L., Huang, C., & Yang, C. (2020). The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access*, *8*, 133330-133348.

[18] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing.