# Skeleton-based Visual Recognition of Diver's Gesture

Bilal Ghader, Claire Dune, Eric Watelain, Vincent Hugel

# Skeleton-based Visual Recognition of Diver's Gesture

Bilal Ghader[1,2], Claire Dune[1], Eric Watelain[2], and Vincent Hugel[1]

*Abstract*— **Divers have developed a specific gesture language for efficient communication underwater. As companion robot drones are increasingly used to help divers, it is essential that these robots can understand basic commands. This paper focuses on diver gestures classification using geometrical features extracted from the diver's upper limb movements on underwater RGB-video. The extraction of upper limb skeleton points relies on pre-existing skeleton detection algorithms. Three geometric features are examined: angles, joint-line projections, and joint-joint distances. The classification is conducted using a 3-layer Bi-LSTM neural network for 11 different gestures. The database is acquired in a pool with 4 subjects and contains a total of 200 hand-cut videos. The recognition results show an accuracy of about 50% for the anangles and joint-lines features, this validates the potential of the method based on 2D skeleton extraction in videos.This study shows that the classification results are strongly related to the type of geometric feature used and to the characteristics of the gesture.**

## I. Introduction

Underwater missions are conducted by either divers or underwater robots, depending on the depth, duration, and complexity of the operations. While joint operations between humans and robots are limited, human-robot interaction is crucial for tasks that require the judgment of human operators. Divers communicate with each other using standardized gestures that have been developed over time and are easily recognizable. To enable communication with human divers, companion robots should be capable of recognizing these standard gestures.

Vision-based interaction between divers and companion underwater drones remains a challenge, mainly due to the optical properties of the marine environment that absorbs colors and reflects light, and due to the limited computing power that can be embedded in Autonomous Underwater Vehicles (AUV). While most of the existing studies focus on hand gesture recognition to classify diver commands [1]–[4], this work exploits the upper limbs movements of the diver. The use of upper limbs' movements shall make it possible to distinguish gestures from further away and in turbid water. Furthermore, by incorporating arm movements, one can effectively utilize the temporal dimension of the data, which was previously unexplored. The classification is therefore performed on a video sequence rather than images. This paper makes several contributions, including a novel perspective on the problem by taking into account the time dimension, the creation of a unique dataset of underwater diver gestures, and a comparison between the performance of 3 different geometric features in 2D: angles, joint-line projections, and segment lengths.

[1]COSMER EA 7398, Université de Toulon, France
[2]IAPS UR N°201723207F, Université de Toulon, France

The paper starts with a brief review of gesture recognition using RGB-data in II, presents the methodology used in III, the data collection, the experimental setup and the results in IV.

## II. Literature Review

Action and gesture recognition/classification is a broad field of research with various real-world applications such as video surveillance, autonomous navigation, sport analysis [5]. While different modalities exist, including Motion Capture, RGB-D, PointCloud, Event-Stream, Audio, IMU data, this work focuses on monocular video RGB. While monocular RGB data is not as rich as other modalities, it is a very popular modality because of its accessibility and ease of use.

The different RGB approaches can be classified into three categories based on how they handle the time-dimension of the data. The first approach handles temporal and spatial dimensions simultaneously. The most common tool used is the extension of 2D Convolutional Neural Network (CNN) to 3D, as in [6] where the auhtors extend state-of-the-art 2D image classification CNNs to 3D and compare the performance of ResNet, pre-activation ResNet, WideResNet, ResNeXt-101, and DenseNet in 3D action recognition. The authors in [7] evaluate a similar extension method from 2D to 3D, but focusing on the network's shift-variability. However, 3D CNNs have several downsides, such as their large number of parameters compared to 2D CNNs, therefore requiring more data for training. Additionally, 3D CNNs are computationally more intensive due to the extra dimension.

The second common approach processes each dimension separately using a two-stream 2D CNN, one stream for the spatial domain and the second for the time-domain. The work presented in [8] dedicates one stream to RGB images, and the second to flow information. Another approach is presented in [9] where the action is modeled as the transformation of the environment before and after it happens. A Siamese network is thus fed, a video before and after the action is performed in different streams, the comparison between both allows the classification.

The last approach is the simplest, it handles the spatial dimension before the temporal dimension. In this approach, 2D features are extracted from the RGB-images using CNNs or classical image processing algorithms before they are fed into a tool meant to handle the temporal dimension. The most common tool used in this case is one of the types of Recurrent Neural Networks, such as LSTMs or GRUs. In the work of [10], [11] the feature extraction is done using

a CNN network, before an LSTM-based classification. The method employed in this article is similar to this approach.

As far as diver gestures classification is concerned, the current work deals with RGB images when it comes to classification. The common pattern is a first phase of hand-isolation and feature extraction, followed by a classification phase. The work in [2] isolates the hand of the diver based on a skin-color segmentation before classifying the gesture of the diver. Conversely, the study in [4] omits the hand-isolation process by using colored markers on the diver's hand to extract colored features, which are then fed into a random-forest classifier. In the works presented in [3], [12], two successive neural networks are employed, with the first network used to isolate the hand and the second used for classification. Until now, the work done on underwater did not try to exploit the temporal dimension of the data when it comes to RGB images.

## III. Methodology

We introduce a three-step method to recognize diver gestures: 1) extraction of diver's skeleton from RGB images, 2) computation of geometric features from the 2D joint positions, 3) training of a LSTM-based neural network to classify the gestures, by exploiting their temporal variation. The pipeline of this process is represented in Fig. 1.

### A. Skeleton detection

The diver is filmed performing the different gestures. Details of data acquisition are explained in sec. IV. The videos are manually segmented and labeled to isolate each gesture. The diver pose is estimated frame by frame using OpenPose [13]. A low-pass Butterworth filter of 5th order and cut-off frequency of 3 Hz is applied to the signal of the joint coordinates. The anatomical points that defined the Open Pose skeleton are presented in Fig. 3. They include shoulders (4 and 7), elbows (5 and 8), wrists (6 and 9) and hips (10 and 11) which are the main points used to compute the geometric feaures.

### B. Geometric feature extraction

The work presented in [14] compares the performance of multiple geometric features when it comes to action recognition on different databases. It focuses on 3D representation, notably Motion Capture and RGB-D data. Two different features are selected based on their performance on 3D data, namely Joint-to-Line distances, and angles. Here, a third feature, Joint-to-Joint distance, is also selected. While Joint-to-Joint features do not encode any real information in 3D, they can be useful in 2D data because they encode information based on 3D to 2D projection. For each arm, 4 different features are constructed.

The joint-to-joint features are the distances of segments (4,5), (5,6), (6,4) and (6,7) (Fig. 3(d)).

Figure 3(b) depicts the angle features. On the right side, point tuple (6,5,4) is used to calculate the elbow angle. The point tuples (10,4,5), (7,4,5) and (7,4,6) are used to get three angles at the shoulder. The left features are defined symmetrically. Given three 2D coordinates, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, let us defined the vectors $\mathbf{x} = \mathbf{a} - \mathbf{b}$, $\mathbf{y} = \mathbf{c} - \mathbf{b}$, $\mathbf{z} = \mathbf{c} - \mathbf{a}$. The angle relative to point $\mathbf{b}$ is defined as follows:

$$\theta = atan2(\|\mathbf{x} \times \mathbf{y}\|, \mathbf{x} \cdot \mathbf{y}) \tag{1}$$

Joint-to-Line distances are calculated using the lengths of the 3 sides $l$, $m$, and $n$ of the triangle formed by the points $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ such that $l \geq m \geq n$. The surface area $S$ of the triangle is computed using the numerically stable Heron's formula :

$$S = \frac{1}{4}\sqrt{(l+m+n)(n-l+m)(n+l-m)(l+m-n)} \tag{2}$$

The tuples chosen for the joint-to-line features are (4,5,6), (6,4,7), (5,4,7), and (6,5,7) ( Fig. 3c(c)) The Joint-to-Line distance relative to joint $\mathbf{b}$ is:

$$JL_d = 2S/\|\mathbf{z}\| \tag{3}$$

To reduce diver dependency, the joint-to-joint distances and the joint-to-line distances are normalized to the distance between shoulders. To limit signal noise, the signal is tracked on the entire video sequence and only the maximum value is used for normalization. After the feature extraction phase, a second low-pass Butterworth filter with a 3 Hz cutoff filter is applied.

### C. Neural Network architecture

For the classification phase, an LSTM-based neural network is employed. The architecture consists of a 3-layer bidirectional LSTM, each followed by a dropout layer, and a final dense layer. Figure 4 provides a visualization of the proposed architecture. This family of network is popular in gesture classification [14]–[16]. The neural network is implemented using the Keras framework, with a Tensorflow back-end. Each of the LSTM layers consists of 128 units in each direction (256 in total) with a dropout rate set to 50%. As part of the training process, the data was augmented by generating mirrored versions of each signal.

## IV. Data collection

11 different recreational diving gestures were selected ('stabilize','go up', 'go down', 'panting', 'cold', 'assemble', 'ok', 'not well', 'stop', 'air reserve', 'no air'). They were randomly arranged to build 4 sequences of 25 gestures.

15 subjects participated in this study, ranging from beginner to professional divers. The experiment was conducted in a 2.5 m deep pool. The gestures were presented to the subjects in a briefing before recording the series. Divers are requested to start and end the gesture in their natural resting position. The divers are standing on their knees facing the camera and a tablet as seen in 2. This is not a natural position for divers in open water, however it was chosen to allow the divers to maintain stability in the pool. Each diver performs between 1 and 4 dives. A sequence of gestures out of 4 is chosen at every dive. Gestures labels are displayed one by one on the tablet facing the diver (Fig. 5). The diver does not know the order of the gestures before the beginning of

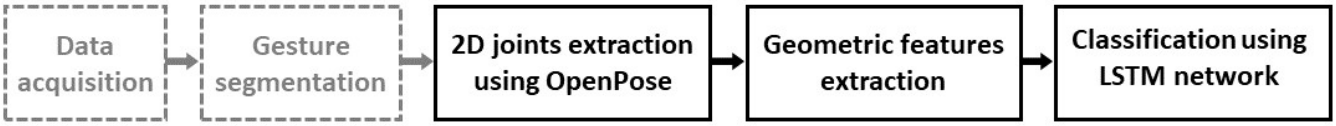



Fig. 1: Pipeline of the proposed classification methodology

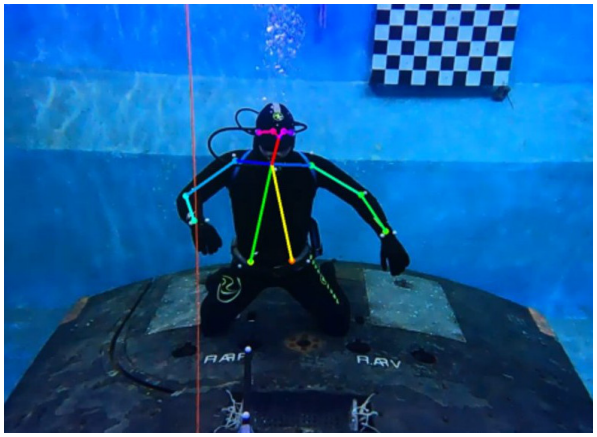
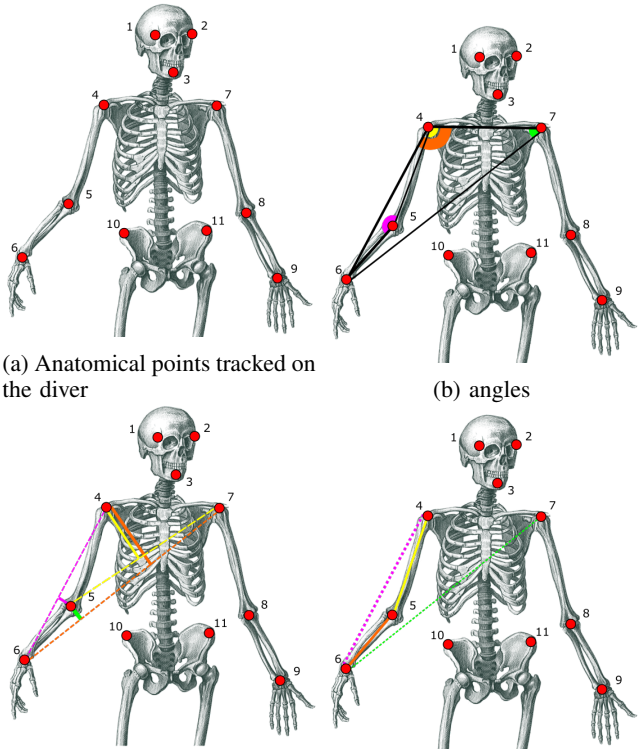


Fig. 2: Skeleton detection on underwater image using OpenPose [13]



(a) Anatomical points tracked on the diver

(b) angles

(c) Joint Lines Projections

(d) segments

Fig. 3: Different tracked geometric features

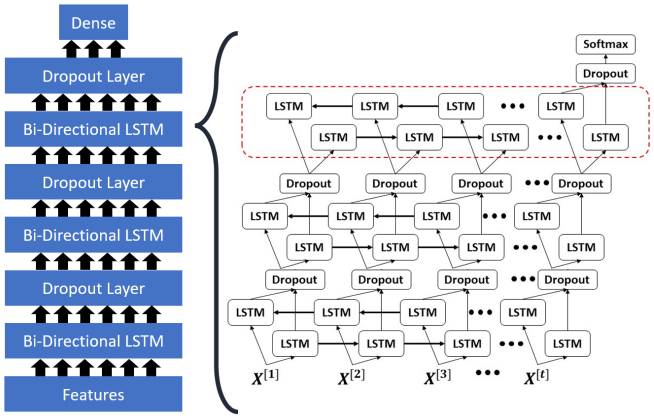



Fig. 4: Bi-directional LSTM neural network

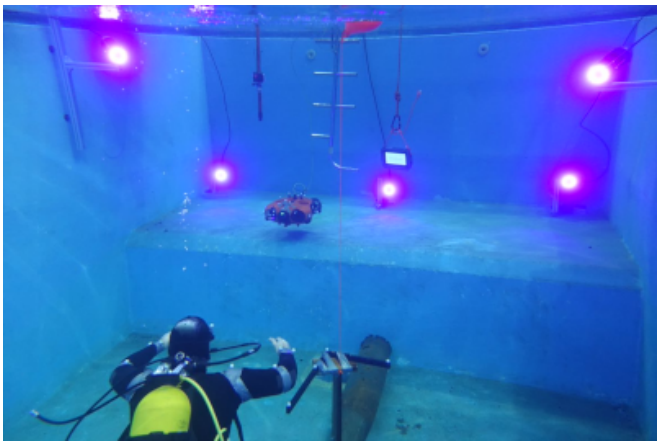


Fig. 5: Underwater setup for gesture data collection: a tablet displays a sequence of gesture labels. Motion capture data and images are simultaneously acquired.

the sequence. The divers are simultaneously recorded using an RGB camera and an underwater Motion Capture system.

## V. Experimental Results

We utilized a subset of the dataset to conduct a classification, which comprised 11 distinct gestures repeated by 4 divers, resulting in a total of 200 data samples. Multiple classifications were performed with different feature sets. The data was randomly divided into training and test data sets in an 80/20 ratio. Given the limited size of our dataset, each classification was conducted 5 times using different train/test splits.

The classification results are presented in Tab. I. The analysis reveals that the angles and joint-to-lines features

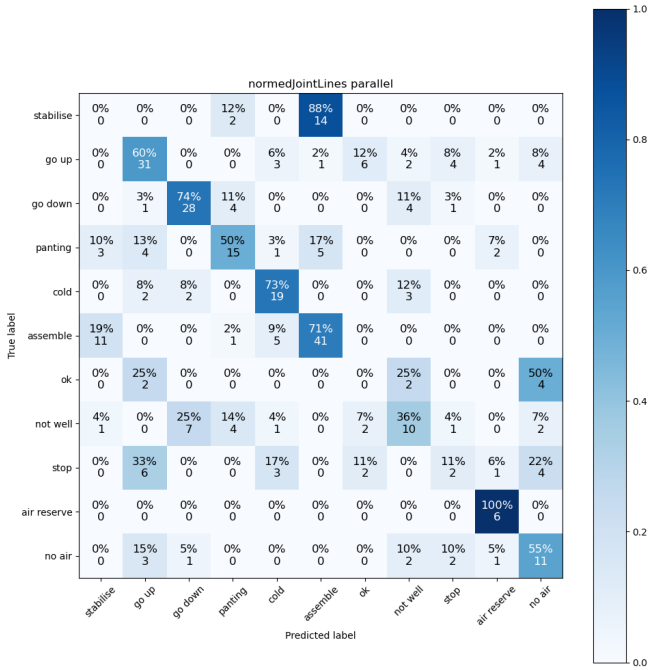



Fig. 6: Confusion matrix based on normalized Joint-to-line features for 5 classifications.

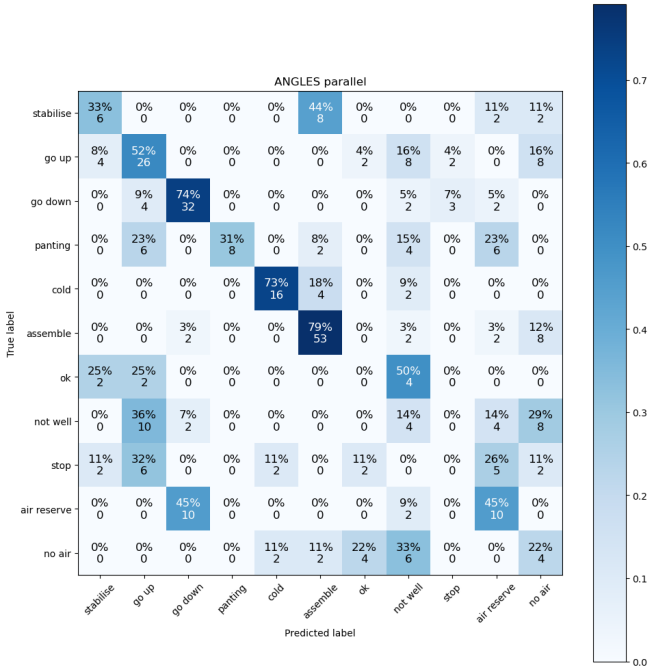



Fig. 7: Confusion matrix based on angles features for 5 classifications.

| Features | Average on 5 folds | best | worst |
|---|---|---|---|
| Angles | 0.5 | 0.65 | 0.4 |
| Joint-to-line distances | 0.54 | 0.64 | 0.43 |
| Joint-to-joint distances | 0.48 | 0.62 | 0.34 |

TABLE I: Results of classification across the different features, namely average performance across 5 classifications, best performing classification and weakest performing classification.

exhibit the best performance, with an average accuracy rate of 50% and 54%, respectively, while the highest classification rate reaches 65%. The Joint-to-joint distances give poor results, which can be explained by the fact that they are most sensitive to signal noise.

The confusion matrices presented in Figures 6 and 7 illustrate the performance of our model when using the joint-to-line and angle features, respectively. The matrix displays the actual label on the side, and the predicted label on the bottom. For each cell, two values are presented, one in percentage and the other in absolute value. In a perfect classification scenario, the confusion matrix would be a diagonal. We distinguish 3 behaviors when it comes to classifications. To begin with, we have classes that always offer satisfactory results such as *assemble* or *go down* regardless of the feature choice. We also have gestures that always perform poorly, such as *ok* or *not well*. And finally, we observed that some gestures showed different performance for different features, with *stabilise* performing better with the angle features, and *no air* performing better with the Joint-to-line features. Therefore, while the information contained in RGB videos is sufficient to classify certain gestures, this does not generalize across the board. This variation in performance between different geometric features was expected, as they encode different information, which is in line with the findings of [14] in their analysis of different geometric features for action recognition across different datasets.

It is important to note that several factors impact the classification results obtained in our study compared to the work of [14]. First, our study was limited to using skeleton data extracted from RGB images, which inherently contains less information than the 3D modalities (RGB-D and Motion capture) used by [14]. Additionally, our pose detector was developed for images, resulting in extremely noisy data with joint jumps between successive frames. The position of the diver is also problematic, given that the legs are not shown in the picture. OpenPose frequently misclassifies the joints, classifying the knees as legs. This noisy behavior has a negative impact on the LSTM's classification capacity. The implementation of low-pass filters at multiple stages in our pipeline helped mitigate the impact of this noise, however it was not completely eliminated. And finally, due to the small size of our dataset, its ability to generalize is limited. This is evidenced by a difference in loss between the training and testing data during the training phase.

## VI. Conclusion & Future work

This paper presented an attempt to diver gesture classification using upper limbs' movement based on monocular RGB video. Three different type of geometric features were tested: angles, joint-to-line distances, and joint-to-joint distances. We note a better performance of angles and joint-to-Line

features compared to joint-to-joint distances. We also note a non-uniform performance across the different gestures and the different geometric features. To enhance the classification rate, there are several options that can be explored. First, increasing the amount of data would be necessary. Second, improving pose estimation is crucial, since it is currently noisy in the context of the video. This can be achieved by implementing better skeleton tracking methods, such as adding restrictions on the segments' elasticity and on joint velocities, which will remove the discontinuities between successive frames. Another possible approach would be to utilize a human body model to ensure fluid movements of the limbs. Lastly, a potential improvement lies in the feature selection, and combination. Given the variation in performance observed across different features for different gestures, it may be beneficial to combine multiple features in a single classifier or employ multiple classifiers with a result fusion scheme at the end to improve classification accuracy. A different features choice can also involve the use of the derivatives of our pre-built features, or the definition of a different type of features to better capture 2D motion information.

## REFERENCES


[1] Đ. Nađ, C. Walker, I. Kvasić, D. O. Antillon, N. Mišković, I. Anderson, and I. Lončar, "Towards advancing diver-robot interaction capabilities," *IFAC-PapersOnLine*, vol. 52, no. 21, pp. 199–204, 2019.

[2] M. J. Islam, M. Ho, and J. Sattar, "Understanding human motion and gestures for underwater human–robot collaboration," *Journal of Field Robotics*, vol. 36, no. 5, pp. 851–873, 2019.

[3] R. Codd-Downey and M. Jenkin, "Human robot interaction using diver hand signals," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '19. IEEE Press, 2019, p. 550–551.

[4] A. G. Chavez, C. A. Mueller, T. Doernbach, D. Chiarella, and A. Birk, "Robust gesture-based communication for underwater human-robot interaction in the context of search and rescue diver missions," *CoRR*, vol. abs/1810.07122, 2018.

[5] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3200–3225, 2023.

[6] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" 06 2018.

[7] Q. Shi, H.-B. Zhang, Z. Li, J.-X. Du, Q. Lei, and J.-H. Liu, "Shuffle-invariant network for action recognition in videos," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 18, no. 3, mar 2022. [Online]. Available: https://doi.org/10.1145/3485665

[8] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3218–3226.

[9] X. Wang, A. Farhadi, and A. Gupta, "Actions transformations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2658–2667.

[10] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[11] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.

[12] R. Codd-Downey and M. Jenkin, "Finding divers with scubanet," in *International Conference on Robotics and Automation (ICRA'2019)*, 2019, pp. 5746–5751.

[13] Z. Cao, G. Hidalgo, T. Simon, S. 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. 01, pp. 172–186, jan 2021.

[14] S. Zhang, X. Liu, and J. Xiao, "On geometric features for skeleton-based action recognition using multilayer lstm networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 148–157.

[15] M. Khokhlova, C. Migniot, A. Morozov, O. Sushkova, and A. Dipanda, "Normal and pathological gait classification LSTM model," *Artificial Intelligence in Medicine*, vol. 94, pp. 54–66, Mar. 2019.

[16] M. Asadi-Aghbolaghi, A. Clapes, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *12th IEEE international conference on automatic face & gesture recognition (FG'2017)*. IEEE, 2017, pp. 476–483.