# OpenPose: Optimizing the Military Physical Fitness Training Assistance System with a Single Image

**Jia-Hao Liu, Rui-Zhe Lu, Jine-Lun Peng, Cheng-Shun Lee**

**Department of Computer Information Science, Republic of China Military Academy, Kaohsiung, Taiwan**

## ABSTRACT

Currently, the military places a strong emphasis on scientifically training physical fitness, conducting three physical fitness tests annually. Among these tests, the most fundamental exercise is the push-up. The traditional testing method involves manual assistance, and recently, there has been the introduction of a push-up testing machine incorporating technology. However, this testing machine exhibits lower accuracy, longer testing times, and is constrained by limited space. Therefore, we aim to enhance the testing machine by integrating a camera and a computer to enable precise testing. We aim to establish a scientific definition of proper posture, which involves maintaining a straight line from shoulders to legs during the preparatory phase and defining the angles of shoulder, waist, hip, and leg axes. This will be achieved using the YOLOv7 model for accurate pose recognition.

Key words：Push-up, YOLOv7, Kettlebell Lateral Raises, Squat

## 1. Introduction

Due to the military's emphasis on scientific training, our initial idea was the push-up, as it requires no equipment and is an excellent physical exercise for improving the fitness of the troops. It is also one of the three physical fitness test items in the military.

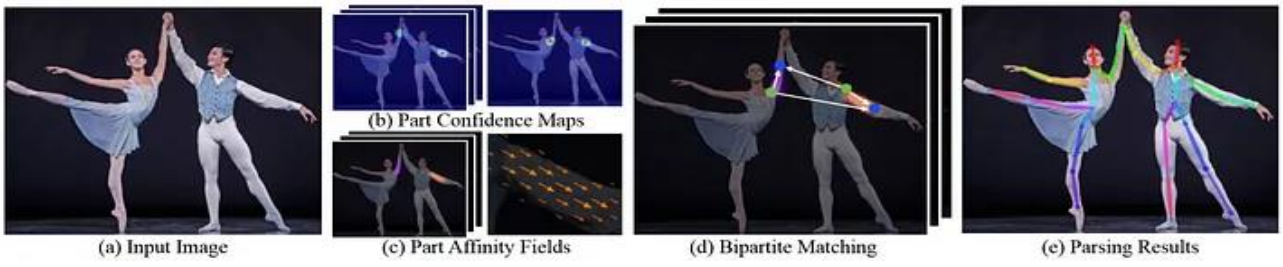


Fig.1. Overall pipeline. (a) Our method takes the entire image as the input for a CNN to jointly predict (b) confidence maps for body part detection and (c) PAFs for part association. (d) The parsing step performs a set of bipartite matchings to associate body part candidates. (e) We finally assemble them into full body poses for all people in the image

However, relying solely on manual assistance for testing cannot ensure that each push-up meets the required standard. Therefore, utilizing instrument-assisted measurements would be a more effective choice, enhancing the fairness of the tests.

We propose utilizing the Convolutional Neural Network (CNN) features of OpenPose [1] along with observing joint movements through YOLOv7[2], such as the angle of

elbow flexion, alignment of shoulders, waist, and hips, to clearly define the standards for push-ups.

"YOLOv7 can perform three tasks." One application is object detection,
which involves identifying objects in a scene, outlining them with bounding boxes, and displaying the predicted object names and probabilities within those boxes. Another task that YOLOv7 can handle is instance segmentation, which involves coloring the identified objects, also known as masks, providing more detailed segmentation than object detection by simply outlining the objects with bounding boxes. The third task performed by YOLOv7 is joint point detection, which is the most crucial feature in our project. The model can
identify and connect joint points recognized in the scene, resembling a stick figure, and can be used to predict movements and postures.

## 2. Related Work

OpenPose utilizes a CNN [3] architecture to generate Confidence Maps (Figure b) for

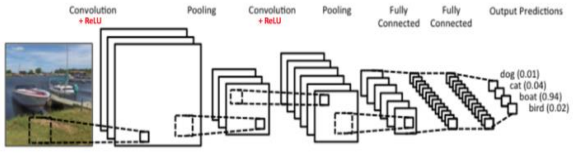


Fig.2. This diagram illustrates the concept of a Convolutional Neural Network (CNN)

each joint position and Part Affinity Fields (PAF) (Figure c), a new concept introduced by OpenPose. By integrating these two types of features, the model can more accurately predict the position of each limb. This description might seem similar to CPM [4], but CPM primarily predicts limbs in images that feature a single person. OpenPose, however, maintains a certain level of accuracy and speed in predicting human body joints and limbs in environments with multiple individuals and complex backgrounds.

### 2.1 Convolutional Neural Network, CNN

CNN mimics the cognitive process of the human brain, such as when we recognize an image. We first focus on distinct points, lines, and surfaces with vivid colors. Subsequently, these elements are abstracted and combined to form various shapes such as eyes, nose, and mouth. This abstraction process is how the CNN algorithm constructs its models. The Convolution Layer transforms the comparison from pointwise to local matching. By analyzing features in blocks and progressively stacking the comprehensive comparison results, better recognition outcomes can be achieved, as illustrated in Figure 2.

### 2.2 Convolution Layer

Taking each point in the image as the center, a neighborhood of N x N points is selected to form a region (N is referred to as the Kernel Size, and the N x N matrix of weights is called the 'Convolution Kernel'). Each point in the region is assigned a different weight. The weighted sum is calculated to measure the similarity between the feature and the local portion of the image. This is achieved by multiplying the values at each corresponding pixel, summing the products, and then dividing by the total number of pixels. The result serves as the output for that specific point. This process is repeated by moving to the next point in the image in the same manner until reaching the last point, constituting the Convolution Layer [5] of CNN, as depicted in Figures 3 and 4.

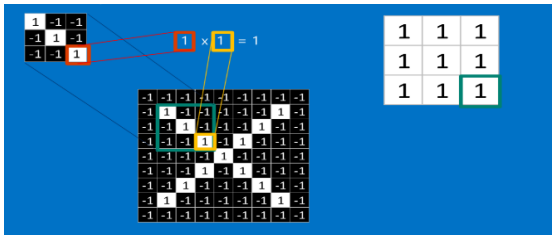

Fig.3.This diagram illustrates the multiplication of corresponding pixels

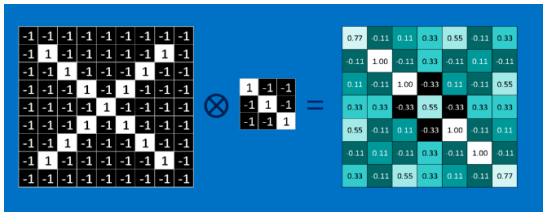

Fig. 4. This diagram illustrates convolution at different positions

2.3 Pooling Layer

Between convolutional layers, there is typically a pooling layer [6], which is a method for compressing images while retaining essential information. The sampling method also involves using a sliding window, but it typically employs max-pooling, where the maximum value is taken instead of a weighted sum. If the size of the sliding window is set to 2 and the 'stride' is also 2, the data volume is reduced to one-fourth of the original. However, since the maximum value is taken, it still retains the maximum possibility [7-9] of local range matching. In other words, the information after pooling is more focused on whether matching features exist in the image rather than precisely 'where' in the image these features exist. This helps the CNN determine if a particular feature is present in the image without needing to be concerned about the feature's specific location [10-12]. As a result, even with image shifts, the network can still recognize the features.

2.4 Deep Learning and Athletic Performance

Z. Zhao, W. Chai Propose Deep learning has the potential to revolutionize sports performance, [13] with applications ranging from perception and comprehension to decision. This paper presents a comprehensive survey of deep learning in sports performance, focusing on three main aspects: algorithms, datasets and virtual environments, and challenges. Firstly, we discuss the hierarchical structure of deep learning algorithms in sports performance which includes perception, comprehension and decision while comparing their strengths and weaknesses. During the feature constructing stage, our method makes use of human landmarks to obtain the angles and distances between the joints. According the results, the proposed method provide comparable improvements for convolutional networks.

## 3.Experimental Parameters

In Figure 7, each number corresponds to a specific body part; for example, 13 corresponds to the left hand, 14 corresponds to the right-hand axis, and so on. In our program, the required body parts for kettlebell lateral raises and push-ups are indicated by the joint points of the shoulders (11, 12) and hand axes (13, 14). For squats, the joints of the waist (23, 24), knees (27, 28).

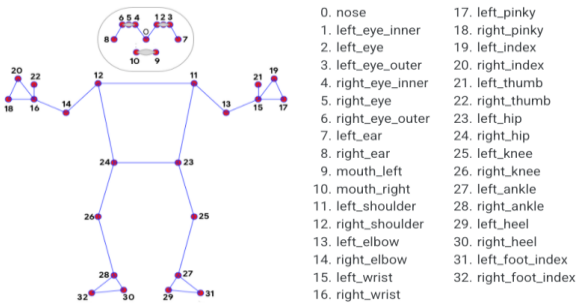


Fig.7. This diagram illustrates the joint points of various body parts

Based on the “112th Year Military Physical Fitness Diversified Training” our push-up test is conducted according to rigorous standards. Test subjects are required to place their hands flat on the ground with fingers naturally extended forward, arms shoulder-width apart, elbows fully extended, maintaining a straight body posture without lifting the hips or touching the knees to the ground, with feet together or spread apart at shoulder width. During elbow flexion, males' chins should be approximately 15 centimeters from the ground, while females' chins should be about 20 centimeters away. Our program conditions are based on the criterion that the joint point of the shoulders equals or is less than the joint point of the elbows to ensure the accuracy and consistency of movements. This enhancement will make the push-up test more accurate, efficient, and scientifically driven. We use the condition where the shoulder joint point equals or is less than the elbow joint point as the program's criteria.


Fig.8. This diagram depicts the preparatory posture for push-ups, with joint points highlighted in red to indicate that the condition is not met


Fig.9. This diagram represents a fulfilled condition for push-ups, with the shoulder joint points highlighted in green to indicate that the condition is met

From Figure 8, we can see that we have currently marked the joint points of the subject's shoulders and elbows. When the program condition is not met, with the shoulder joint point being greater than the elbow joint point, the joint point is highlighted in red. When the program condition is met, with the shoulder joint point greater than or equal to the elbow joint point, as shown in Figure 9, the shoulder joint point turns green to indicate the fulfilled condition, and the count is recorded. However, we still cannot overcome the issue of counting repetitions when the hips are either too high or too low (shoulders, waist, hips, and legs not forming a straight line).

Regarding the push-up shooting angle, we capture it from the front, and counting repetitions from a side view is not feasible. As the conditions become more complex, it becomes more challenging to capture human movements. Therefore, for the standard posture where the shoulders, waist, hips, and legs form a straight line, our initial concept was that during the preparatory movement, the hips should be positioned between the shoulders and elbows. However, this condition was challenging to detect, possibly due to its complexity or the difficulty for individuals to achieve such a precise standard. This is a challenge we plan to overcome in the future.

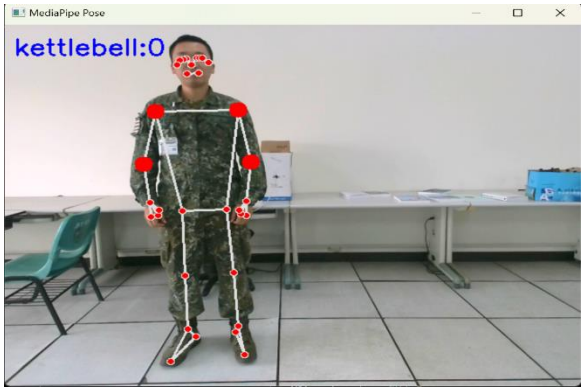


Fig.10. This diagram illustrates the preparatory movement for kettlebell lateral raises, with joint points highlighted in red to indicate that the condition is not met

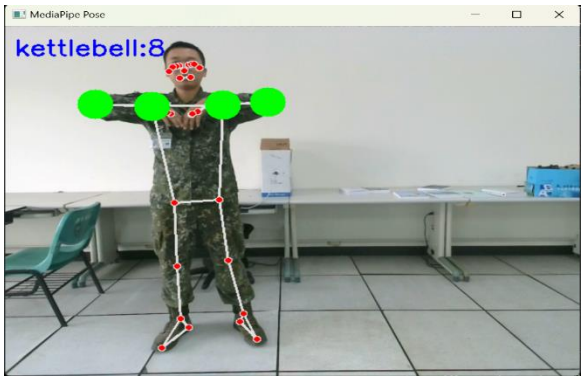


Fig.11. This diagram represents a fulfilled condition for kettlebell lateral raises, with the joint points highlighted in green to indicate that the condition is met

According to the “112th Year Military Physical Fitness Diversified Training” our kettlebell military press test adheres to strict standards. Test subjects are required to flex their elbows to lift the kettlebell upward to chest level while keeping their arms at approximately shoulder height. Subsequently, the kettlebell is lowered back down, returning to the starting position, completing one full repetition. Therefore, we use the condition where the angle at the elbow joint is greater than or equal to the angle at the shoulder joint as the criterion.

Program's criteria. This enhancement will make the kettlebell military press test more accurate, efficient, and scientifically driven. Through the use of standardized testing

methods, we can ensure that all test subjects are evaluated using the same criteria, thereby enhancing the fairness and reliability of the test. This contributes to the military's more effective assessment and enhancement of soldiers' physical fitness levels, thereby strengthening their combat capabilities and operational effectiveness. Therefore, we use the condition where the angle at the elbow joint is greater than or equal to the angle at the shoulder joint as the program's criteria.

From Figure 10, it is evident that we have currently marked the junction points of the shoulder and elbow. When the elbow is positioned lower than the shoulder, the joint point is marked in red, and the count is not registered. When the condition is met with the elbow greater than or equal to the shoulder, as shown in Figure 11, the joint point turns green to indicate the fulfilled condition, and the count is calculated.

Regarding the shooting angle, we capture it from the front as it is the most effective for counting repetitions, and the joint points of the elbows and shoulders can be clearly displayed.


Fig.12.This diagram illustrates the squatting posture, with joint points highlighted in green to indicate that the condition is met

Squats are one of the commonly used training methods in the military, effectively targeting lower body muscles, improving leg strength, and stability. This standard ensures the accuracy and safety of movements while effectively achieving training goals. Squat training helps soldiers enhance explosiveness and endurance, thereby improving combat capabilities. When performing squats, we use the hip and knee joint points as criteria. When the hip joint is lower than the knee joint, it indicates that the squat meets the standard. The condition is met, as shown in Figure 12. To ensure clear recording, we use a side-view angle. This angle of recording can present the movement more clearly, facilitating accurate assessment. If filmed from the front, it would be difficult to accurately capture the movement.

## 4.Conclusion

Through a computer and a camera, we can assess exercises such as push-ups, kettlebell lateral raises, and squats. This setup enables operators to evaluate the standardization of their movements without being limited by space or requiring excessive equipment. During testing, we found that if the conditions are too complex, operators may struggle to meet them, or the computer may fail to detect them, resulting in uncounted repetitions. For example, achieving the standard of keeping the shoulders, waist, and hips in a straight line, with the hips positioned between the shoulders and elbows, proved challenging to detect. The reasons may align with what was previously explained. Regarding camera angles, we capture push-ups and kettlebell lateral raises from the front. While front-facing filming for kettlebell exercises poses no issues, attempts to film push-ups from the side have been unsuccessful in meeting the specified conditions. We speculate that the computer cannot accurately detect joint positions, leading to calculation errors. For squats, we opt for side-view filming, as it allows for a more accurate detection of the hip position. The hips should be at least parallel to the knees during the squat. Future corrections include addressing the filming angles for push-ups, adjusting the complexity of conditions, and ensuring operators can more accurately assess whether their postures are standard and calculate repetitions correctly.

**References:**

[1] Z. C. Tomas Simon, S. E. Wei, Yaser Sheikh “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields” arXiv:1812.08008v2, 2019.
[2] W.C.-Yao “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors” arXiv:2207.02696v 1, 2022.
[3] K. O’Shea, R. Nash, “An Introduction to Convolutional Neural Networks” arXiv:1511.08458v2, 2015.
[4] Z. Zhang, X. Han, H. Zhou, P. Ke, Y. Gu, D. Ye, Y. Qin, Y. Su, H. Ji, J. Guan, F. Qi, X. Wang, Y. Zheng, G. Zeng, H. Cao, S. Chen, D. Li, Z. Sun, Z. Liu , M. Huang , W. Han, J. Tang, J. Li, X. Zhu, M. Sun “CPM: A Large-scale Generative Chinese Pre-trained Language Model” arXiv:2012.00413v1, 2020.
[5] J. Redmon, S. Divvala, R. Girshick, “You Only Look Once: Unified, Real-Time Object Detection” arXiv:1506.02640v5, 2016.
[6] Hossein Gholamalinezhad1, “Hossein Khosravi, Pooling Methods in Deep Neural Networks, a Review”
[7] L. Ciamplconi, A. Elwood, M. L leonardi, A. M ohamed, and A. Rozza, A survey and taxonomy of loss functions in machine learning, arXiv:2301.05579v1, 2023.
[8] J. R. Terven, Diana M Cordova-Esparza, A.R. Pedraza, E. A. C. Urbiola, “loss functions and metrics in deep learning” arXiv:2307.02694v2, 2023.
[9] K. Janocha,W. M. Czarnecki, “On Loss Functions for Deep Neural Networks in Classification”arXiv:1702.05659v1, 2017.
[10] S. Kato, Meijo University, K. Hotta , Meijo University, “MSE Loss with Outlying Label for Imbalanced Classification” arXiv:2107.02393v1, 2021.
[11] J. Ren, M. Zhang, C. Yu, Z. Liu, “Balanced MSE for Imbalanced Visual Regression”
[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, R. Girshick,”Masked Autoencoders Are Scalable Vision Learners” arXiv:2111.06377v3, 2021.
[13] Z. Zhao, W. Chai,” A Survey of Deep Learning in Sports Applications: Perception, Comprehension, and Decision” arXiv:2307.03353v1, 2023

# 使用 OpenPose 以單一影像優化國軍體能訓練輔助系統

劉家豪 呂睿哲 彭敬倫 李政勳

陸軍軍官學校資訊系

## 摘要

現在國軍重視運動科學化訓練，每年都要進行三項體能測驗，其中最基本的莫過於伏地挺身。傳統的測驗方法為人工輔助測驗，近期結合科技也有了伏地挺身測驗機。但是此測驗機準確度較低，所需時間較久，場地受限。所以我們想改良測驗機，我們使用一台攝影機和電腦，就可以完成精確的測驗。我們要將姿勢做出科學化的定義，針對人體的各項關節點，肩、腰、臀、腿的標點及手軸彎曲角度等等，並利用 OpenPose 模型做出可視化的呈現，使電腦可以精確的判斷出姿勢的正確性。

**關鍵詞：**YOLOv7，OpenPose，伏地挺身，壺鈴平舉，深蹲。