DEEP LEARNING - BASED HUMAN POSE ESTIMATION METHODS FOR TRAINING ACTIVITIES: A COMPREHENSIVE REVIEW AND FRAMEWORK DESIGN FOR INTELLIGENT SHOOTING TRAINING SYSTEMS

KHẢO SÁT CÁC PHƯƠNG PHÁP ƯỚC TÍNH TƯ THẾ NGƯỜI TRONG CÁC HOẠT ĐỘNG HUẤN LUYỆN DỰA TRÊN HỌC SÂU, THIẾT KẾ FRAMEWORK CHO HỆ THỐNG HUẤN LUYỆN BẮN SÚNG THÔNG MINH

Vu Minh Hoang[1], Truong Quoc Hung[1], Nguyen Thi Lan[1,*]
Tran Thi Hai Anh[1], Truong Khanh Nghia[1]

DOI: http://doi.org/10.57001/huih5804.2024.207

ABSTRACT

Human Pose Estimation (HPE) has witnessed significant advancements in recent years, largely propelled by the breakthroughs in deep learning techniques. This paper presents a comprehensive review of Deep Learning-Based HPE Methods in the context of trainning activities. This review begins by introducing the fundamental concepts of human pose estimation and its significance in sports and physical training. Subsequently, the paper delves into the landscape of deep learning methodologies, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants, which have played a pivotal role in revolutionizing pose estimation. The core of the review lies in the analysis of state-of-the-art deep learning-based HPE methods tailored to training activities. Emphasis is placed on their accuracy, robustness to varying conditions, real-time processing capabilities, and integration with training environments. The challenges and limitations of deep learning-based HPE methods are addressed, along with ongoing research efforts to mitigate these issues. Finally, this paper proposes a shooting training model based on integrating of knowledge distillation into the context of HPE, highlighting its potential to enhance training assistance systems.

Keywords: *3D Human Pose Estimation, Deep Learning, action correction, physical and sports training assistance.*

TÓM TẮT

Ước tính tư thế con người (HPE) đã có nhiều tiến bộ đáng kể trong những năm gần đây bởi những đột phá trong kỹ thuật học sâu. Bài viết này nhằm đánh giá toàn diện về phương pháp HPE dựa trên học sâu trong các hệ thống huấn luyện thông minh. Đánh giá bắt đầu bằng việc giới thiệu các khái niệm cơ bản về ước tính tư thế con người và tầm quan trọng của nó trong các hệ thống hỗ trợ huấn luyện thể chất. Sau đó, bài báo đi sâu vào các phương pháp học sâu, bao gồm mạng CNN, RNN và các biến thể của chúng. Cốt lõi của nghiên cứu nằm ở việc phân tích các phương pháp HPE dựa trên học sâu hiện đại, phù hợp với các hoạt động huấn luyện; nhấn mạnh vào tính chính xác của chúng với các điều kiện khác nhau, khả năng xử lý thời gian thực và tích hợp với môi trường đào tạo; đồng thời chỉ ra những thách thức và xu hướng nghiên cứu đối với các vấn đề còn tồn tại. Cuối cùng, bài báo đề xuất mô huấn luyện bắn súng dựa trên việc tích hợp chắt lọc kiến thức vào bối cảnh của HPE, nêu bật tiềm năng ứng dụng của mô hình này trong các hệ thống huấn luyện thông minh.

Từ khóa: *Ước tính tư thế con người 3D, học sâu, điều chỉnh hành động, hệ thống hỗ trợ huấn luyện thông minh.*

[1]Institute of Simulation Technology, Le Quy Don Technical University, Vietnam
[*]Email: lannt.simtech@mta.edu.vn
Ngày nhận bài: 04/4/2024
Ngày nhận bài sửa sau phản biện: 15/6/2024
Ngày chấp nhận đăng: 25/6/2024

1. INTRODUCTION

Human movement analysis, a cornerstone of training activities, has long been integral to enhancing athletic performance and preventing injuries. Since 2014, the advent of deep learning has brought about a paradigm shift in the field of HPE [1]. By leveraging the power of complex neural networks, deep learning methods have not only refined the accuracy of pose estimation but have also unlocked new dimensions in real-time feedback and personalized coaching. This paper embarks on a comprehensive review of Deep Learning-Based Human Pose Estimation Methods, focusing on their transformative role in physical and sports training assistance.

The ability to accurately capture and analyze human movement is a fundamental challenge in fields ranging from sports coaching to rehabilitation. Human Pose Estimation, which involves the determination of joint positions and angles in the human body from images or video data, offers a solution to this challenge [2]. Accurate pose

estimation enables coaches, athletes, and healthcare professionals to gain insights into body posture, alignment, and movement patterns. These insights facilitate targeted training regimens, personalized corrective interventions, and the optimization of athletic techniques. Furthermore, pose estimation serves as a cornerstone for biomechanical analysis, aiding in the refinement of movement efficiency and injury prevention strategies.

Traditional computer vision techniques for pose estimation often relied on handcrafted features and heuristic algorithms [3, 4]. However, the emergence of deep learning has propelled pose estimation to new heights. Convolutional Neural Networks (CNNs) [18], Recurrent Neural Networks (RNNs) [19], and their variants have showcased remarkable capabilities in capturing intricate spatial and temporal relationships within human movement data [5]. These networks, fueled by large datasets and advancements in hardware, have demonstrated the potential to learn intricate pose representations directly from raw input data.

The foundation of our exploration is built upon the rich tapestry of research and reviews that have come before. Previous reviews of the field have consistently highlighted the profound implications of Deep Learning in reshaping the landscape of computer vision and HPE. They have documented the rapid evolution of Deep Learning techniques, particularly CNNs and RNNs, and their profound influence on the accuracy and efficiency of human pose estimation. These insightful assessments have collectively highlighted the seismic shifts brought about by Deep Learning in the realm of computer vision [6, 7].

However, a critical scrutiny of these reviews reveals certain limitations. Most notably, their comprehensive scope often necessitated a high-level overview, which, while valuable, occasionally obscured the nuances specific to the application of Deep Learning-based HPE in physical and sports training assistance. These reviews, in their attempt to encapsulate the entirety of the field, may have inadvertently overlooked the intricate challenges and opportunities within this specialized domain. This paper endeavors to bridge this gap by not only synthesizing insights from previous reviews but also focusing with precision on the intricacies and nuances that define the symbiotic relationship between Deep Learning-based HPE and physical training.

Moreover, this review aims to provide a comprehensive examination of the landscape of Deep Learning-Based Human Pose Estimation Methods, particularly in the context of physical and sports training assistance. The paper proceeds as follows: Section 3 offers an overview of deep learning methodologies tailored to pose estimation. Section 4 delves into the application of these methods in specific physical activities, highlighting their contributions to coaching and training. Section 5 addresses challenges, limitations and identifying future research directions. Finally, Section 6 provides the framework design for two real applications by using the integration of knowledge distillation into pose estimation, emphasizing its potential to enhance real-time processing and model efficiency.

2. PREVIOUS RELATED WORK AND OUR CONTRIBUTIONS

Every year, many surveys related to the problem of human posture estimation are published. However, surveys conducted before 2015 mainly focused on conventional methods and rarely provided information on deep learning-based methods. For example, the work [3] refers to research on 3D HPE prior to 2012. This survey focused on conventional methods without addressing deep learning. A survey of both traditional and deep learning approaches related to HPE was presented in [4]. However, only a handful of deep learning-based approaches have been included.

Other recent papers have extensively reviewed models built from deep learning to estimate human posture with different scopes and approaches. Surveys [5, 8] cover 3D HPE methods, while surveys [11, 12] only consider 2D HPE methods. Survey [13] extensively evaluates 2D and 3D human pose estimation methods based on deep learning from monocular images or video footage as input. Besides, documents [14] present surveys of the most advanced methods that have been proposed to solve 2D and 3D pose estimation tasks. However, to date, there are few studies that provide information on the application of HPE in practice. Document [15] lists key studies in estimating human posture to aid training, from January 2011 to March 2021. However, the paper has not provided a general process for applying HPE in training support, and has not discussed the challenges as well as potential research directions in the future. Besides, the document [16] only focuses on analyzing yoga posture recognition systems using computer vision, machine learning and deep learning techniques without regard to other sports.

Therefore, our paper aims to investigate in depth a wide range of applications of HPE. Specifically, this article examines methods for applying HPE in posture correction and sports training support. In particular, each step in the process of building these applications will be thoroughly analyzed through a specific example. The new contributions in our paper compared to previous surveys are as follows:

- An overview of the latest deep learning-based HPE methods used in physical and sports training asisstance. In particular, the strengths and weaknesses of these HPE methods are pointed out.

- Presents the basic steps that need to be taken to apply HPE to support physical and sports practice.

- Discuss the main challenges of applying HPE in the sport sector and propose potential future research to improve the performance of these applications.

- Proposes an effective shooting training model and a welding robot training system based on integrating of knowledge distillation into the context of HPE.

3. HUMAN POSE ESTIMATION

Human pose estimation is a computer vision task that involves the localization and estimation of key points on a human body, known as keypoints or joints, in images or videos [17]. The primary objective is to determine the spatial positions of these keypoints, which correspond to specific anatomical landmarks on the human body, such as the head, shoulders, elbows, wrists, hips, knees, and ankles. These keypoints collectively represent the posture or pose of a person in a given image or frame of a video

3.1. Classification of Human Pose Estimation

The main kinds of human pose estimation can be classified based on various criteria, including dimensionality, methodology, and use cases. Here are some main kinds of human pose estimation:

- 2D pose estimation [20]: This method estimates the 2D coordinates of key points on the human body in an image or video. It is typically used when the camera is stationary, and the pose estimation is performed in a single 2D image. 2D pose estimation can be used for applications such as gesture recognition, activity recognition, and human-computer interaction. However, 2D pose estimation does not provide depth information, making it unsuitable for tasks requiring accurate spatial understanding. And it may not capture the full range of human motion and interactions in 3D space.

- 3D pose estimation [21, 22]: This method estimates the 3D coordinates of key points on the human body, which involves reconstructing the pose in 3D space. It is more challenging than 2D pose estimation as it requires more data and computational power. However, it can provide more accurate and detailed information about the pose and movement of the body. 3D pose estimation is typically used in applications such as virtual reality, augmented reality, robotics, and sports analysis.

- Single-person pose estimation [23]: Single-person pose estimation focuses on estimating the pose of a single individual within the image or video frame. Applications include sports analysis, gesture recognition, and single-person tracking. Traditional computer vision techniques, such as edge detection, and handcrafted feature extraction, were previously used for pose estimation but have become less common in favor of deep learning approaches due to their superior performance. Single-person pose estimation is a fundamental component in various applications that require an understanding of human body posture and movements. Advances in deep learning and the availability of large pose estimation datasets have significantly improved the accuracy and robustness of single-person pose estimation systems, making them valuable tools in computer vision and artificial intelligence applications.

- Multi-Person Pose Estimation [24]: Multi-person pose estimation is a computer vision task that involves detecting and locating key body joints or landmarks on multiple individuals within an image or a video frame. The goal is to accurately identify the positions of body parts such as the head, shoulders, elbows, wrists, hips, knees, and ankles for each person present in the scene. This task is challenging due to variations in human poses, occlusions between individuals, and differences in lighting conditions and camera viewpoints.

The output of a multi-person pose estimation algorithm typically includes keypoint locations and associated confidence scores for each detected person in the image. These keypoints can then be used for various applications such as human action recognition, gesture recognition, activity analysis, human-computer interaction, and sports analytics. To achieve multi-person pose estimation, deep learning-based approaches, particularly convolutional neural networks (CNNs), are commonly employed. These models are trained on large annotated datasets containing images with labeled keypoint coordinates. During inference, the trained model processes an input image and outputs the locations of keypoints for all detected individuals.

3.2. Deep learning-based HPE methods

Deep learning has revolutionized the field of human pose estimation by enabling accurate and robust models that can learn complex features from data. Some of the most popular deep learning-based methods for human posture estimation are:

Convolutional Neural Networks (CNNs): CNNs are a fundamental and widely used technique in HPE [25]. CNNs are particularly effective for this task because they excel at capturing spatial features in images, which is essential for detecting key body keypoints. They can be used to estimate human posture from images or video by training the network on a large dataset of annotated images or videos that capture different postures. The network can then identify specific body parts and their positions to estimate the overall posture of the individual. CNNs are primarily used in the initial stages of HPE architectures for feature extraction. The input image is processed through multiple convolutional layers, which apply a series of convolutional filters to extract features at various levels of abstraction. These features include edges, corners, and other visual patterns that help the network discern the locations of keypoints. CNNs used in HPE are often composed of multiple convolutional layers stacked on top of each other, creating a deep network. These deep architectures can capture complex and hierarchical patterns in the image, which is crucial for accurate keypoint localization. For 2D pose estimation, CNNs output heatmaps, where each heatmap corresponds to a particular keypoint. The peak or maximum value in each heatmap corresponds to the estimated keypoint location. By analyzing these heatmaps, the model can determine the 2D positions of keypoints. In 3D pose estimation, CNNs are often combined with depth information (e.g., from depth cameras or stereo vision). These networks learn to estimate not only the 2D positions of keypoints but also their depth or 3D coordinates in the real world. CNN-based HPE models have made significant progress in recent years, achieving high levels of accuracy and real-time performance [26]. These models have a wide range of

applications, including gesture recognition, action recognition, sports analysis, and human-computer interaction, and they continue to be an active area of research and development in computer vision.

Recurrent Neural Networks (RNNs): RNNs are a class of neural networks that are often used in HPE, especially when dealing with temporal data, such as videos or sequences of images. RNNs are well-suited for modeling sequences because they can capture temporal dependencies and relationships between keypoints across frames [27]. RNN-based HPE models are used for real-time pose tracking in applications like sports analysis, where they can follow the movements of athletes over time to provide insights into their performance. Training RNNs for HPE typically requires labeled sequences of images or videos with annotated keypoint positions. This data is used to teach the network how to predict keypoints over time. RNNs are particularly valuable when the context of time is essential for accurate pose estimation. They can handle scenarios where the pose of a person evolves over a sequence of frames, such as tracking dance movements, analyzing sports plays, or monitoring rehabilitation exercises. However, it's worth noting that more recent approaches in HPE have started to incorporate hybrid models that combine the strengths of both CNNs and RNNs to achieve state-of-the-art performance in various applications.

Deep Convolutional Neural Networks (DCNNs): DCNNs have played a significant role in advancing HPE [28]. These networks are a class of deep learning architectures specifically designed for processing grid-like data, such as images, and they have proven to be highly effective in capturing spatial patterns and features in the human body for pose estimation. DCNNs are employed in HPE as feature extraction modules. They consist of multiple convolutional layers followed by pooling layers. These layers learn to extract hierarchical features from input images, capturing edges, textures, shapes, and higher-level representations relevant to the positions of keypoints on the human body. DCNN architectures used for HPE can vary widely, with some models designed specifically for this task. Researchers often experiment with modifications to the standard DCNN architectures to enhance their suitability for pose estimation. Specialized DCNN architectures and optimizations are applied to enable real-time or low-latency pose estimation, which is crucial for interactive applications and systems. The output layer of a DCNN for HPE typically consists of multiple neurons or nodes, each corresponding to a specific keypoint on the human body. The network's task is to predict the likelihood or coordinates of these keypoints. DCNNs have significantly improved the accuracy and efficiency of HPE systems, enabling their deployment in a wide range of applications, including sports analysis, healthcare, robotics, gaming, augmented reality, and security. Researchers continue to explore innovative DCNN architectures and training strategies to further enhance the capabilities of HPE models.

Long Short-Term Memory Networks (LSTMs): LSTMs [29] are a type of RNN architecture that are commonly used in deep learning for sequential data modeling. LSTMs are designed to overcome the vanishing and exploding gradient problem that can occur in traditional RNNs when dealing with long-term dependencies in a sequence of data. They have been used to estimate human posture from a sequence of video frames, similar to RNNs. LSTMs are designed to handle sequences of data, making them a natural choice for tracking the pose of a person as they move over time [30]. They can capture how the position of keypoints changes from one frame to the next and use this information to estimate the keypoints' positions in subsequent frames. LSTMs can be used to fuse information from multiple sensors or modalities, such as RGB images, depth maps, or infrared data. By combining information from different sources, LSTMs can provide a more comprehensive understanding of the pose in 3D space. LSTMs are a valuable tool for modeling the temporal aspects of human body pose, allowing HPE systems to track and predict keypoints' movements over time. However, they are often used in combination with Convolutional Neural Networks (CNNs) for initial feature extraction from images or video frames, creating hybrid architectures that leverage the strengths of both CNNs and LSTMs for accurate and robust HPE.

Graph Convolutional Networks (GCNs): GCNs [31] are a specialized deep learning technique used in Human Pose Estimation (HPE) for capturing spatial dependencies and relationships between keypoints by representing them as nodes in a graph structure. GCNs have gained popularity in 3D pose estimation tasks, as they can effectively model the connectivity of keypoints in the human body. In HPE, the human body can be represented as a graph, where keypoints (e.g., joints or body parts) are nodes in the graph, and edges represent possible connections between these keypoints. The graph's structure depends on the specific HPE task and the body model used but typically includes nodes for keypoints like the head, shoulders, elbows, wrists, hips, knees, and ankles. GCNs are designed to capture the spatial relationships between these keypoints. They learn how information flows between connected keypoints, allowing them to model dependencies effectively. For example, in a 3D pose estimation task, GCNs can capture how the position of the hip joint affects the position of the knee and ankle joints.

GCNs apply convolutional operations on the graph to process information. These operations are similar to the convolutional layers in Convolutional Neural Networks (CNNs), but they are adapted to work on graph-structured data. The convolutional operations involve aggregating information from neighboring keypoints and updating the features associated with each keypoint. In 3D pose estimation tasks, GCNs take into account not only the 2D positions of keypoints but also depth information, which can be obtained from depth sensors or stereo vision systems.

This allows GCNs to infer the 3D positions of keypoints in the real world [32]. GCNs have been employed in real-time HPE systems for applications like action recognition, human-computer interaction, and virtual reality, where capturing the spatial relationships between keypoints is crucial. GCNs can improve the robustness of HPE models by considering the graph-based structural information of the human body. This makes them especially useful in scenarios where pose estimation accuracy is critical. GCNs have shown promising results in HPE, particularly in 3D pose estimation tasks. By modeling keypoints as nodes in a graph and exploiting spatial relationships, GCNs can capture intricate pose information, making them a valuable addition to the toolbox of techniques used for accurate and robust human pose estimation.

4. HPE IN TRAINING ACTIVITIES

4.1. Applications of HPE in training activities

Human pose estimation has many applications in training activities. In physical training, HPE can be used to optimize performance, prevent injuries, and improve overall health and wellness [33]. By using HPE algorithms, trainers can monitor and analyze the movements of athletes or trainees and provide real-time feedback. For example, if a weightlifter's posture is not correct during a lift, the system can detect the incorrect posture and alert the trainer to provide corrective instructions. Furthermore, HPE can be used to develop customized training programs based on an individual's fitness level, goals, and injury history. These programs can be designed to improve an individual's strength, endurance, flexibility, and overall fitness while minimizing the risk of injury.

HPE can be extremely beneficial in professional sports training, where athletes are always looking for ways to improve their performance and gain a competitive edge. It can be used to analyze the movements of athletes during training or competition. Coaches and trainers can use the data to identify areas for improvement, correct technique, and prevent injuries. In addition, HPE can be used to evaluate an athlete's performance in real-time or post-competition. It can help assess the speed, agility, balance, and coordination of an athlete. Especially, HPE can be used to develop injury prevention strategies and rehabilitation programs for athletes recovering from injuries. This may involve analyzing an athlete's movement patterns, muscle imbalances, and areas of weakness, and developing a program to address these issues and prevent further injury.

Some specific applications of HPE in physical and sports training:

Action Correction: HPE can be used to correct the form and technique of athletes in real-time [34]. By comparing the athlete's pose with the correct pose, the system can provide feedback on the necessary adjustments to improve performance and reduce the risk of injury. This can be applied to various sports, such as weightlifting, yoga, and martial arts.

Sports Analysis: HPE can be used to analyze and improve sports performance. By tracking the movement of the human keypoints, the system can provide insights on the athlete's strength, balance, and coordination, among others. This can be applied to various sports [34]. For example, in soccer, a pose estimation model can be used to track the movements of players and analyze their positioning, passing accuracy, and other metrics. In basketball, a pose estimation model can be used to analyze shooting form and provide feedback on correct posture and technique.

Injury Prevention: HPE can be used to detect and prevent injuries in athletes [35]. By analyzing the biomechanics of the athlete's movements, the system can identify potential risk factors and provide recommendations for injury prevention. This can be applied to various sports, such as running, jumping, and throwing.

Rehabilitation: HPE can be used to monitor and assist with rehabilitation exercises for individuals with physical injuries or disabilities [36]. For example, a pose estimation model can be used to track the movements of a patient during physical therapy and provide feedback on correct posture and range of motion.

4.2. Steps for implementing HPE in training activities

The following are the general steps involved in applying HPE in training activities:

Step 1 - Data collection: Data collection is the first step in applying HPE in physical and sports training assistance. The goal of this step is to collect high-quality data of the athlete performing the desired movements or actions, which will serve as the input for subsequent analysis and processing.

There are several ways to collect data, including using cameras, motion capture systems, or wearable sensors. For example, a camera can be set up to record the athlete performing the movement or action from different angles. This data can then be used to estimate the athlete's joint positions and orientations using HPE algorithms.

To ensure high-quality data, it is important to consider factors such as lighting conditions, camera placement, and the range of motion of the athlete. The camera should be placed in a location that provides a clear view of the athlete's body, and the lighting conditions should be consistent across different recording sessions. It is also important to capture a wide range of movements or actions to ensure that the HPE algorithms are robust to variations in posture and movement patterns.

Step 2 - Data preprocessing: Data preprocessing is an important step in applying HPE in action correction and sports training assistance. The goal of this step is to prepare the collected data for analysis and processing using HPE algorithms.

There are several tasks involved in data preprocessing, including filtering, normalization, and data augmentation. Filtering involves removing noise and outliers from the data

to improve the accuracy of subsequent analysis. Normalization involves scaling the data to ensure consistency across different athletes and movements. This is important because the range of motion and body sizes can vary widely among individuals, and normalization helps to account for these differences. Data augmentation involves creating additional data samples by applying random transformations to the original data. This can help to increase the diversity of the data and improve the robustness of the HPE algorithms.

In addition to these tasks, it is important to ensure that the data is in a format that can be used by HPE algorithms. For example, the data may need to be converted to a specific file format or coordinate system to be compatible with the chosen HPE algorithm. It is also important to ensure that the data is properly labeled, with each joint and body part identified and annotated with a corresponding label or ID.

Step 3 - Human pose estimation: HPE is a critical step in applying HPE to physical and sports training assistance. The goal of this step is to detect and track the positions and orientations of the athlete's joints and body parts in the collected data.

There are several methods for performing HPE, including both 2D and 3D approaches. In 2D HPE, the goal is to estimate the positions of joints and body parts in the image or video frames captured by the camera. In 3D HPE, the goal is to estimate the positions of joints and body parts in 3D space. One popular approach for 2D HPE is using CNNs, which are deep learning models that can learn to detect and classify features in images. The CNN can be trained on large datasets of labeled images, such as the COCO dataset, to learn to detect and classify the positions of joints and body parts in the image. For 3D HPE, more advanced methods are required, such as using depth sensors, stereo cameras, or motion capture systems. These systems can provide more accurate 3D joint positions, but require more advanced hardware and setup.

Regardless of the specific method used, the output of the HPE step is a set of estimated joint positions and orientations, which can be used to analyze the athlete's movements and performance in subsequent steps.

Step 4 - Performance analysis: Performance analysis is a critical step in HPE to action correction and sports training assistance. The goal of this step is to analyze the athlete's movements and performance based on the estimated joint positions and orientations obtained from the HPE step.

There are several tasks involved in performance analysis, including joint angle calculation, motion analysis, and comparison to reference movements. Joint angle calculation involves using the estimated joint positions and orientations to calculate the angles between the athlete's joints, such as the angle between the elbow and wrist or the angle between the knee and ankle. This can provide insight into the athlete's technique and form, and identify areas for improvement. Motion analysis involves using the estimated joint positions and orientations to analyze the athlete's movements, such as the speed, acceleration, and trajectory of different body parts during the movement. This can provide insight into the athlete's overall movement patterns and identify areas for improvement in terms of efficiency and effectiveness. Comparison to reference movements involves comparing the athlete's movements and performance to reference movements, such as idealized movements or the movements of expert athletes. This can provide insight into areas where the athlete's movements differ from the ideal or expert movements, and identify areas for improvement.

Step 5 - Feedback and correction: The goal of this step is to provide feedback to the athlete based on the performance analysis obtained from the previous step, and to guide the athlete towards correct movements and techniques.

There are several ways to provide feedback and correction, including visual feedback, auditory feedback, and haptic feedback. Visual feedback involves displaying visual cues, such as diagrams or videos, to the athlete to illustrate correct movements and techniques. Auditory feedback involves providing auditory cues, such as spoken instructions or sound effects, to guide the athlete towards correct movements and techniques. Haptic feedback involves providing tactile cues, such as vibrations or pressure, to guide the athlete towards correct movements and techniques. In addition to providing feedback, correction is also important in this step. Correction involves guiding the athlete towards correct movements and techniques, such as adjusting the athlete's posture, adjusting the range of motion, or adjusting the tempo of the movement.

5. CHALLENGES FOR HPE IN TRAINING ACTIVITIES AND CURRENT RESEARCH TRENDS

While HPE has great potential for training activities, there are still several challenges that need to be addressed [37]:

Accuracy: Accuracy is a critical challenge when applying HPE in action correction and sports training assistance. HPE algorithms aim to detect and track the joints and body parts of athletes from video or sensor data. The accuracy of the algorithm refers to how well the algorithm is able to estimate the true position of each joint and body part in real-time, even when they are occluded, performing fast movements, or wearing equipment. Any inaccuracies in the pose estimation can result in incorrect feedback or training programs. The accuracy of HPE is affected by various factors such as occlusion, lighting conditions, camera angles, and the complexity of the human body movement. For example, when an athlete is performing a movement with one arm extended, the camera may not be able to capture the entire arm, resulting in an occlusion problem. In this case, the algorithm may need to estimate the position of the missing part based on other available information, such as the position of the other arm or body parts. However, this

estimation can result in inaccurate joint positions. Another factor that can affect accuracy is the variation in body shape and size among different athletes. For example, a pose estimation algorithm trained on a dataset of athletes of one specific body type may not work well on athletes with different body types. To improve accuracy [38], researchers have developed different algorithms such as CNN, pose estimation with graph convolutional networks, and optical flow-based methods. Additionally, researchers have collected large-scale datasets of human pose for training these algorithms, including datasets specific to certain sports or activities.

Multi-Person Pose Estimation: Multi-Person Pose Estimation is a significant challenge in the field of HPE, particularly in the context of physical and sports training assistance. This challenge arises when multiple individuals are close to each other or overlap in the image or video frame. The poses of individuals may partially occlude each other, making it difficult for a pose estimation model to distinguish body parts and joints. Addressing the challenge of Multi-Person Pose Estimation often involves using advanced computer vision techniques, such as object detection, keypoint association, and tracking algorithms [39]. Deep learning models, particularly CNNs, have significantly improved the accuracy of pose estimation in complex scenes. Some novel top-down convolutional networks are proposed to improve the robustness under complex field conditions in the wild. Additionally, combining pose estimation with identity tracking and using techniques like data association can help solve the identity-to-pose association problem. Continued research in this area is essential to improve the robustness, accuracy, and real-time capabilities of Multi-Person Pose Estimation systems, enabling a wide range of applications in sports, surveillance, entertainment, and beyond.

Limited Data for Specific Sports: The challenge of limited data for specific sports in HPE refers to the lack of sufficient labeled data for training and testing HPE models in specific sports. Collecting and annotating large-scale datasets for specific sports can be challenging and time-consuming. This challenge can significantly affect the performance of HPE models in specific sports or actions, leading to lower accuracy, lower precision, and reduced generalization ability of the models. For example, HPE models trained on general human pose datasets, such as COCO or MPII, may not perform well when applied to sports-specific movements, such as swimming strokes, golf swings, or tennis serves, as these movements have unique pose characteristics that are not well represented in the general datasets. Therefore, collecting and annotating large-scale datasets for specific sports is crucial to improving the performance of HPE models in these areas. To address this challenge, researchers can use various approaches, such as transfer learning, domain adaptation, and data augmentation [40]. Transfer learning involves using pre-trained models on a large dataset to initialize the weights of the HPE model and fine-tuning it on the specific sports dataset. Domain adaptation involves adapting a pre-trained HPE model to a specific domain or task, such as sports-specific movements. Data augmentation techniques, such as adding synthetic data to the training dataset to increase the diversity of the training data and improve the generalization ability of the model, extracting weak 3D information directly from 2D images or proposing a new self-supervised training method designed to train a 3D human pose estimation network using unlabeled multi-view images.

Real-time Performance: Real-time HPE is critical for physical and sports training assistance, as timely feedback is essential for correcting and improving performance. However, real-time HPE requires models that can process images or video frames quickly and accurately. This can be challenging due to the complexity of the models and the need for high-end hardware.

To address the challenge of real-time performance, researchers have developed various techniques such as deep neural network architectures, efficient inference algorithms, and hardware acceleration [41]. Deep neural networks can be optimized for fast inference by using lightweight architectures, reducing the number of parameters, and using efficient optimization algorithms. For example, MobileNet and ShuffleNet are efficient neural network architectures that have been used in real-time HPE systems. Efficient inference algorithms, such as pruning, filtering and knowledge distillation, can be used to reduce the computational cost of HPE models without sacrificing accuracy. Hardware acceleration techniques, such as using dedicated hardware like GPUs or TPUs, can further improve the real-time performance of HPE models [42]. These specialized hardware accelerators can perform computations much faster than traditional CPUs, making them well-suited for real-time applications.

Noisy and Dynamic Environments: The "noisy and dynamic environments" challenge refers to the difficulty of accurately detecting and tracking human pose in real-world environments that are noisy and dynamic, such as sports fields, gyms, or outdoor environments. These environments can pose significant challenges for HPE algorithms due to factors such as occlusions, lighting variations, camera angles, and rapid changes in the athlete's position or orientation. One of the main sources of noise in these environments is occlusion, which occurs when a body part is temporarily or partially obscured from the camera's view, making it difficult for HPE algorithms to estimate the position of the occluded joint [43]. Another source of noise is lighting variations, which can occur due to changes in ambient light or shadows cast by objects or other athletes. These variations can affect the accuracy of the algorithm by altering the appearance of the athlete's body and making it more difficult to detect the joints and body parts. Additionally, sports fields and gym environments are often dynamic, with athletes moving quickly and performing

complex movements that can challenge even the most robust HPE algorithms. The rapid changes in position and orientation of the athlete can result in missed or inaccurate joint detections.

To address these challenges, researchers have developed HPE algorithms that are designed to be robust to noise and dynamic environments [44]. These algorithms may incorporate additional sensors or modalities, such as accelerometers, to improve joint detection and tracking. Other strategies include incorporating prior knowledge of the athlete's body shape and movement patterns or using multiple cameras to capture different perspectives of the athlete's movement.

6. FRAMEWORK DESIGN FOR AN INTELLIGENT SHOOTING TRAINING SYSTEM

6.1. System description

6.1.1. The framework design of the proposed system

Designing an intelligent shooting training system based on human pose estimation involves combining various technologies and components to create a cohesive framework. Such a system aims to enhance shooting skills by analyzing and providing feedback on a shooter's posture, aiming, and shooting technique. Below is a description of the high-level framework design of our proposed system:

Sensors and Cameras: These capture data about the shooter's movements and actions. Multiple high-resolution cameras, possibly equipped with depth sensors are strategically placed around the shooting range or training area to track the shooter's body and firearm in real-time.

Pose Estimation Software: This software analyzes the data from sensors and cameras to estimate the shooter's body posture and firearm position. Pose estimation algorithms, often powered by deep learning models, process the image and depth data to generate 2D or 3D skeletal poses of the shooter.

Data Processing and Analysis: This component interprets the pose data, analyzes it, and extracts meaningful insights about the shooter's technique. Machine learning models and algorithms are used to process the pose data, identify posture errors, assess aiming accuracy, and track the shooter's movements over time.

Feedback Mechanism: This provides real-time feedback to the shooter to help them improve their skills. Feedback can be provided through various means, such as visual cues (overlaid graphics on camera feeds), auditory signals, or haptic feedback (vibrations or physical cues). The feedback is based on the analysis of the shooter's pose and performance.

User Interface (UI): The UI serves as the interaction point between the shooter and the system, displaying feedback and training modules. The UI can be displayed on a screen. It provides real-time feedback, training progress tracking, and customization options for the shooter.

Training Modules: These modules simulate various shooting scenarios and exercises to train and improve the shooter's skills. Training modules can include target practice, moving targets, timed shooting, and other scenarios relevant to the shooter's skill level and goals.

Data Storage and Analytics: This element stores historical performance data and enables analysis of progress over time. Shooter data, including pose information, aiming accuracy, and training results, are stored in a database for later review and analysis. Analytics tools can help shooters track their improvement and identify areas that need more focus.

Privacy and Security Measures: Protect user data and privacy. Implement robust data protection measures to ensure user privacy and data security, especially if data is stored in the cloud or shared with other users.

These elements work together to create a comprehensive intelligent shooting training system based on human pose estimation, offering real-time feedback and performance analysis to help shooters enhance their skills and safety.

6.1.2. Why use HPE for this system

Applying HPE in an intelligent shooting training system offers several significant advantages and benefits:

Precise Posture Analysis: HPE provides accurate and detailed information about the shooter's body posture and movements. This precision is crucial for assessing shooting technique and identifying errors or areas for improvement in real-time.

Aiming Analysis: HPE can accurately track the alignment of the shooter's eye, firearm, and target. This enables the system to evaluate aiming accuracy and suggest adjustments to improve shot placement.

Real-time Feedback: HPE allows for the generation of real-time feedback based on the shooter's posture and aiming. Shooters can receive immediate cues on how to correct their technique, leading to more effective and efficient training.

Objective Assessment: HPE provides an objective assessment of the shooter's performance, reducing the reliance on subjective judgments. This is particularly important for competitive shooting or professional training where precision matters.

Progress Tracking: HPE enables the system to track the shooter's progress over time accurately. Shooters can see how their posture and aiming have improved or identify areas that still need work.

Personalized Training: HPE allows for personalized training programs. The system can identify specific weaknesses or errors in the shooter's technique and tailor training exercises to address those issues.

Safety Monitoring: HPE can be used to monitor and ensure safety. It can detect unsafe behaviors, such as improper firearm handling, and provide immediate warnings or intervention.

Accessibility and Inclusivity: HPE can be adapted to cater to shooters with different physical abilities and body types, making shooting sports more accessible and inclusive.

Continuous Improvement: HPE-driven systems can adapt and improve over time. As more data is collected and analyzed, the system can refine its recommendations and training modules.

In summary, Human Pose Estimation enhances an intelligent shooting training system by providing precise, real-time feedback and analysis of the shooter's posture and aiming. It contributes to more effective training, improved performance, and a safer shooting environment.

6.1.3. Objectives and requirements when designing an intelligent shooting HPE-based training system

a) Objectives

Designing an intelligent shooting HPE-based training system involves setting clear objectives to ensure that the system serves its intended purpose effectively. The objectives of our proposed system include:

Skill Improvement: The primary objective is to help shooters, whether novice or experienced, enhance their shooting skills, including posture, aiming, and firearm handling, through data-driven insights and training modules.

Real-time Feedback: Provide immediate and precise feedback to shooters during training sessions to correct errors, refine technique, and optimize their shooting performance in real-time.

Accuracy Enhancement: Assist shooters in improving the accuracy and precision of their shots by focusing on factors such as posture, grip, trigger control, and follow-through.

Aiming Proficiency: Help shooters develop better aiming skills by evaluating the alignment of their eye, firearm, and target and providing guidance on adjustments.

Progress Tracking: Enable shooters to monitor their performance over time, set performance goals, and measure their improvement accurately.

Customization: Tailor training programs to the individual needs, skill levels, and goals of shooters, allowing for personalized training experiences.

Integration: Ensure compatibility and integration with other shooting training equipment, such as firearms, targets, and range facilities, to create a seamless training ecosystem.

Privacy and Security: Implement robust data security and privacy measures to safeguard user data and ensure the confidentiality of training sessions.

Research and Development: Serve as a valuable tool for researchers and educators to study and enhance shooting technique.

By addressing these objectives, an intelligent shooting training system based on HPE can effectively assist shooters in improving their skills while promoting safety, engagement, and continuous development.

b) Requirements

Camera and Sensor Setup: Install high-quality cameras and, if needed, depth sensors, such as RGB-D cameras or LiDAR, at strategic positions around the shooting range to capture the shooter's movements accurately.

Pose Estimation Algorithms: Implement robust and accurate HPE algorithms capable of tracking the shooter's body posture and firearm position in real-time.

Real-time Processing: Ensure the system can process pose data in real-time with minimal latency to provide immediate feedback to the shooter during training sessions.

Feedback Mechanism: Develop a feedback mechanism that can convey information effectively, including visual cues, auditory signals, or haptic feedback.

Training Modules: Create a diverse range of training modules and scenarios that cater to different shooting disciplines and skill levels.

User Interface (UI): Design an intuitive and user-friendly UI that displays real-time feedback, training progress, customization options, and user profiles.

Data Storage and Privacy: Securely store user data and ensure compliance with privacy regulations to protect user privacy and data security.

Scalability and Maintenance: Design the system to be scalable, allowing it to accommodate growing user bases, and ensure ease of maintenance and updates.

Customization Options: Allow users to customize their training programs, set goals, and adjust training parameters to align with their preferences and skill development needs.

Continuous Improvement: Establish a feedback loop for users to provide input on their experiences, preferences, and suggestions for system enhancements.

User Training and Education: Provide instructional materials and training on how to use the system safely and effectively, especially for novice users.

6.2. Knowledge distillation

Knowledge distillation [45] is a machine learning technique that involves transferring knowledge from a larger, more complex model (often referred to as the "teacher model") to a smaller, simpler model (referred to as the "student model") (Fig. 1). The goal of this process is to make the student model mimic the behavior and predictions of the teacher model.

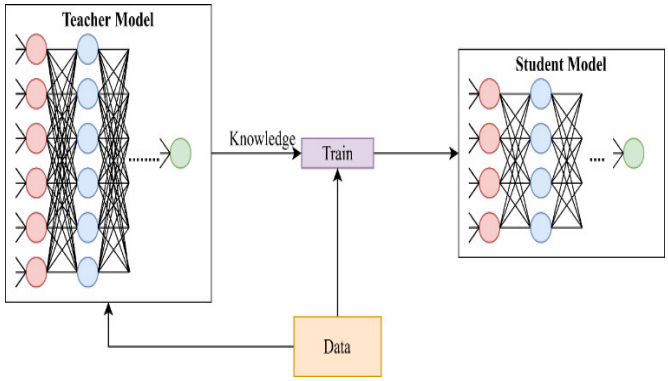

Fig. 1. A general illustration of knowledge distillation

The teacher model is typically a well-trained, high-capacity model that has learned to generalize from a large amount of data. The student model, on the other hand, might be a smaller neural network that is more computationally efficient and has fewer parameters. By training the student model to reproduce the outputs of the teacher model, the student model can often achieve similar performance to the teacher model but with a lower computational cost.

The process of knowledge distillation involves training the student model on both the original training data and the "soft labels" produced by the teacher model. These soft labels are essentially the probability distributions over the classes instead of the hard labels (one-hot encoded vectors). The soft labels contain more nuanced information about the relationships between classes, which can help the student model learn more effectively.

Knowledge distillation has been applied to various computer vision tasks, including human pose estimation in several ways [46]. Researchers have been exploring various techniques to transfer knowledge from complex teacher models to simpler student models for more efficient and real-time pose estimation.

Overall, applying knowledge distillation to human pose estimation for a shooting training system offers several benefits and advantages:

Efficiency and Real-time Processing: Shooting coaching systems often require real-time or near-real-time feedback to users. Knowledge distillation allows us to create a more lightweight student model that can provide accurate pose estimations quickly, enabling instant feedback during workouts [47]. This means users can receive feedback without noticeable delays, enhancing the user experience in shooting coaching applications.

Reduced Computational Demands: Pose estimation models can be computationally intensive, especially when running on resource-constrained devices like smartphones or wearables. Knowledge distillation allows us to compress the knowledge from a complex teacher model into a smaller student model, making it feasible to run pose estimation on these devices without sacrificing accuracy significantly [48].

Noise Reduction: Knowledge distillation can help the student model reduce sensitivity to noise and outliers in the training data [49]. This is because the teacher model's knowledge serves as a form of regularization, discouraging the student model from fitting noise in the training data.

Reduced Overfitting: The teacher model's knowledge helps regularize the student model's training. This can prevent overfitting, where the model becomes too specific to the training data and performs poorly on new, unseen data. A more robust model is less prone to overfitting [50].

Customization and Personalization: By using knowledge distillation, we can create a student model that focuses on poses and movements relevant to shooting coaching, tailoring the model to the specific exercises and activities in our application.

Adaptability: Distilled student models can be more adaptable to varying lighting conditions, camera qualities, and user appearances due to the specific focus on relevant pose estimation features.

Quick Model Updates: As new exercise techniques or modifications emerge, updating a smaller student model with new knowledge is generally faster and less resource-intensive compared to updating a larger teacher model.

In general, knowledge distillation in HPE offers benefits in terms of model efficiency, real-time performance, reduced data footprint, energy efficiency, privacy preservation, scalability, and adaptability to edge devices. It enables the development of more practical and deployable HPE solutions that meet the constraints and requirements of various applications. This is the reason why we use knowledge distillation for human pose estimation in this system.

6.3. Applying knowledge distillation in human pose estimation for a shooting training system

Applying knowledge distillation in human pose estimation for a shooting training system (Fig. 2) involves several steps:

Teacher model selection: This step is a crucial part of applying knowledge distillation in human pose estimation for a shooting coaching system. The teacher model serves as the source of knowledge and expertise that will be transferred to the student model during training.

The teacher model is a complex and accurate pose estimation model that has been trained on a large and diverse dataset. It should demonstrate high accuracy in predicting human pose keypoints in various shooting stances and scenarios.

Choose a teacher model architecture that has demonstrated state-of-the-art performance in pose estimation tasks. This could be a deep convolutional neural network architecture specifically designed for pose estimation, such as Hourglass, OpenPose, or a stacked architecture.

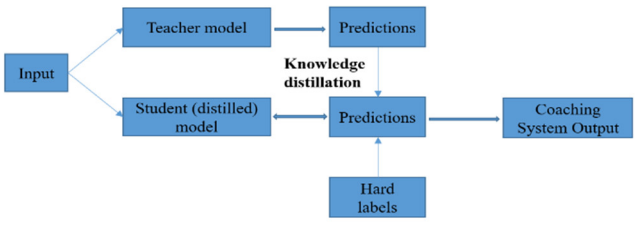

Fig. 2. Proposed model

Student model design: The student model is the lightweight network that will be trained to mimic the expertise of the more complex teacher model.

The student model should have a simplified architecture compared to the teacher model. It typically contains fewer

layers, fewer parameters, and is designed to be computationally efficient while maintaining reasonable accuracy.

Design the student model using lightweight components such as shallow convolutional layers, fewer filters, and smaller hidden layers. This helps reduce computational demands and memory requirements, making the model suitable for real-time processing.

Knowledge Distillation Strategy determines how the student model will learn from the teacher model's predictions. The distillation process involves transferring soft labels (confidence scores and logits) generated by the teacher model to guide the training of the student model.

Data Collection and Annotation: Gather a dataset of videos or images depicting shooters in various shooting stances and scenarios. Annotate the dataset with ground-truth keypoints representing key body joints.

Teacher Model Training: Train the teacher model on the shooting training dataset. Optimize the teacher model to achieve high accuracy in estimating shooter poses. Use traditional pose estimation loss functions, such as Mean Squared Error or Heatmap-based loss, to train the teacher model.

Student Model Initialization: Initialize the student model with random weights or pre-train it on a different dataset. The student model should start with no knowledge about shooting poses.

Knowledge Transfer and Training: During the knowledge transfer process, train the student model to mimic the behavior of the teacher model. This involves two key components:

- Pose Loss: The student model should be trained to minimize the traditional pose estimation loss against the ground-truth annotations. This encourages the student model to learn accurate shooter poses.

- Distillation Loss: Introduce a distillation loss that encourages the student model's predictions to match the softened predictions of the teacher model. The distillation loss can be based on functions like cross-entropy or Kullback-Leibler (KL) divergence, comparing the probability distributions of teacher and student predictions. Adjust the temperature parameter in the distillation loss function to control the softness of the teacher's predictions.

Hyperparameter Tuning: Fine-tune hyperparameters during training. This includes the weights assigned to the pose loss and distillation loss, as well as the temperature parameter for the distillation loss. Experiment with different values to achieve the desired balance between accuracy and model efficiency.

Evaluation: Evaluate the student model's performance on a separate validation or test dataset specific to shooting training. Assess its accuracy in estimating shooter poses, as well as its computational efficiency. Ensure that the model meets the real-time processing requirements of the shooting training system.

Integration into the Shooting Training System: Integrate the trained student model into the shooting training system. Ensure that the model can take input from cameras or sensors, estimate shooter poses in real-time, and provide immediate feedback or analysis to the shooter.

Continuous Improvement: Monitor the performance of the shooting training system and the student model in real-world scenarios. Collect additional data if necessary and fine-tune the model to further improve its accuracy and efficiency. Continuously update and refine the student model as new shooting techniques or improvements emerge. Knowledge distillation allows for efficient model updates.

In the near future, the authors will develop and refine a smart shooting training system based on the proposed model. Once completed, the system promises to significantly enhance shooting training at military schools.

REFERENCES

[1]. Toshev, et al., "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[2]. Song, et al., "Human pose estimation and its application to action recognition: A survey," *Journal of Visual Communication and Image Representation*, 76: 103055, 2021.

[3]. Holte Michael B., et al., "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *IEEE Journal of selected topics in signal processing* 6.5: 538-552, 2012.

[4]. Gong, et al., "Human pose estimation from monocular images: A comprehensive survey," *Sensors* 16.12: 1966, 2016.

[5]. Sarafianos, et al., "3D human pose estimation: A review of the literature and analysis of covariates," *Computer Vision and Image Understanding,* 152: 1-20, 2016.

[6]. Desmarais, et al., "A review of 3D human pose estimation algorithms for markerless motion capture," *Computer Vision and Image Understanding*, 212: 103275, 2021.

[7]. Zheng, et al., "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, 2020.

[8]. Wang, et al., "Deep 3D human pose estimation: A review," *Computer Vision and Image Understanding*, 210: 103225, 2021.

[11]. Dang, et al., "Deep learning based 2d human pose estimation: A survey," *Tsinghua Science and Technology*, 24.6: 663-676, 2019.

[12]. Munea, et al., "The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation," *IEEE Access* 8: 133330-133348, 2020.

[13]. Chen, et al., "Monocular human pose estimation: A survey of deep learning-based methods," *Computer vision and image understanding* 192: 102897, 2020.

[14]. Gamra, et al., "A review of deep learning techniques for 2D and 3D human pose estimation," *Image and Vision Computing* 114: 104282, 2021.

[15]. Difini, et al., "Human pose estimation for training assistance: a systematic literature review," in *Proceedings of the Brazilian Symposium on Multimedia and the Web*, 2021.

[16]. Rajendran, et al., "A Survey on Yogic Posture Recognition," *IEEE Access* 11: 11183-11223, 2023.

[17]. Aftab, et al., "A boosting framework for human posture recognition using spatio-temporal features along with radon transform," *Multimedia Tools and Applications* 81.29: 42325-42351, 2022.

[18]. Sigal, et al., "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International journal of computer vision* 87.1-2: 4-27, 2010.

[19]. Medsker Larry R., L. C. Jain. "Recurrent neural networks," *Design and Applications* 5.64-67, 2001.

[20]. Cao, et al., "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[21] Liu, et al., "Feature boosting network for 3D pose estimation," *IEEE transactions on pattern analysis and machine intelligence* 42.2: 494-501, 2019.

[22]. Tang, et al., "MLP-JCG: Multi-Layer Perceptron with Joint-Coordinate Gating for Efficient 3D Human Pose Estimation," *IEEE Transactions on Multimedia*, 2023.

[23]. Kulkarni, et al., "PoseAnalyser: A Survey on Human Pose Estimation," *SN Computer Science* 4.2: 136, 2023.

[24]. Cai, et al., "Learning delicate local representations for multi-person pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Part III* 16. Springer International Publishing, 2020.

[25]. Zheng, et al., "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys* 56.1: 1-37, 2023.

[26]. Liao, et al., "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition* 98: 107069, 2020.

[27]. Chen Haoming, et al., "2D Human pose estimation: A survey," *Multimedia Systems*, 1-24, 2022.

[28]. Tulbure, et al., "A review on modern defect detection models using DCNNs–Deep convolutional neural networks," *Journal of Advanced Research* 35: 33-48, 2022.

[29]. Cheng, Jianpeng, et al., *Long short-term memory-networks for machine reading*. arXiv preprint arXiv:1601.06733, 2016.

[30]. Rani Challapalli Jhansi, et al., "An effectual classical dance pose estimation and classification system employing convolution neural network-long shortterm memory (CNN-LSTM) network for video sequences," *Microprocessors and Microsystems* 95: 104651, 2022.

[31]. Zhang Si, et al., "Graph convolutional networks: a comprehensive review," Computational Social Networks 6.1: 1-23, 2019.

[32]. Zhao, Long, et al., "Semantic graph convolutional networks for 3d human pose regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[33]. Aritz, et al., "A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise," *Sensors* 21.18: 5996, 2021.

[34]. Difini, et al., "Human pose estimation for training assistance: a systematic literature review," in *Proceedings of the Brazilian Symposium on Multimedia and the Web*, 2021.

[35]. Ludwig Katja, et al., "Self-supervised learning for human pose estimation in sports," in *2021 IEEE International Conference on Multimedia & Expo Workshops* (ICMEW), IEEE, 2021.

[36]. Stenum Jan, et al., "Applications of pose estimation in human health and performance across the lifespan," *Sensors* 21.21: 7315, 2021.

[37]. Zheng, Ce, et al., "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, 2020.

[38]. Shan Wenkang, et al. "Improving robustness and accuracy via relative information encoding in 3d human pose estimation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[39]. Zhang Yifu, et al., "Voxeltrack: Multi-person 3d human pose estimation and tracking in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.2: 2613-2626, 2022.

[40]. Taketsugu, et al., "Uncertainty Criteria in Active Transfer Learning for Efficient Video-Specific Human Pose Estimation," in *2023 18th International Conference on Machine Vision and Applications* (MVA), IEEE, 2023.

[41]. Garau Nicola, et al., "CapsulePose: A variational CapsNet for real-time end-to-end 3D human pose estimation," *Neurocomputing* 523: 81-91, 2023.

[42]. Zhong Fujin, et al., "DSPNet: A low computational-cost network for human pose estimation," *Neurocomputing* 423: 327-335, 2021.

[43]. Ran, Hang, et al., "3D human pose and shape estimation via de-occlusion multi-task learning," *Neurocomputing,* 126284, 2023

[44]. Zhang Zhongyang, et al., "Neuromorphic High-Frequency 3D Dancing Pose Estimation in Dynamic Environment," *Neurocomputing,* 126388, 2023.

[45]. Gou Jianping, et al., "Knowledge distillation: A survey," *International Journal of Computer Vision* 129: 1789-1819, 2021.

[46]. Li Zheng, et al., "Online knowledge distillation for efficient pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[47]. Bian Cunling, et al., "Structural knowledge distillation for efficient skeleton-based action recognition," *IEEE Transactions on Image Processing*, 30: 2963-2976, 2021.

[48]. Salimi Mohammadamin, José JM Machado, João Manuel RS Tavares, "Using deep neural networks for human fall detection based on pose estimation," *Sensors* 22.12: 4544, 2022.

[49]. Li Yuncheng, et al., "Learning from noisy labels with distillation," in *Proceedings of the IEEE international conference on computer vision,* 2017.

[50]. Mishra, et al., *Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy*. arXiv preprint arXiv:1711.05852, 2017.

THÔNG TIN TÁC GIẢ

Vũ Minh Hoàng, Trương Quốc Hùng, Nguyễn Thị Lan, Trần Thị Hải Anh, Trương Khánh Nghĩa

Viện Công nghệ Mô phỏng, Học viện Kỹ thuật Quân sự