Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.050435

ARTICLE

Abnormal Action Recognition with Lightweight Pose Estimation Network in Electric Power Training Scene

Yunfeng Cai[1], Ran Qin[1], Jin Tang[1], Long Zhang[1], Xiaotian Bi[1] and Qing Yang[2,*]

[1]State Grid Jiangsu Electric Power Co., Ltd. Research Institute, Nanjing, 211103, China

[2]School of Computer Engineering, Nanjing Institute of Technology, Nanjing, 211167, China

*Corresponding Author: Qing Yang. Email: yangq@njit.edu.cn

Received: 06 February 2024 Accepted: 07 May 2024 Published: 20 June 2024

ABSTRACT

Electric power training is essential for ensuring the safety and reliability of the system. In this study, we introduce a novel Abnormal Action Recognition (AAR) system that utilizes a Lightweight Pose Estimation Network (LPEN) to efficiently and effectively detect abnormal fall-down and trespass incidents in electric power training scenarios. The LPEN network, comprising three stages—MobileNet, Initial Stage, and Refinement Stage—is employed to swiftly extract image features, detect human key points, and refine them for accurate analysis. Subsequently, a Pose-aware Action Analysis Module (PAAM) captures the positional coordinates of human skeletal points in each frame. Finally, an Abnormal Action Inference Module (AAIM) evaluates whether abnormal fall-down or unauthorized trespass behavior is occurring. For fall-down recognition, three criteria—falling speed, main angles of skeletal points, and the person's bounding box—are considered. To identify unauthorized trespass, emphasis is placed on the position of the ankles. Extensive experiments validate the effectiveness and efficiency of the proposed system in ensuring the safety and reliability of electric power training.

KEYWORDS

Abnormal action recognition; action recognition; lightweight pose estimation; electric power training

1 Introduction

Computer vision is a specialized field within Artificial Intelligence (AI) that focuses on enabling computers to interpret and extract information from images and videos. It has been applied to various fields including healthcare, transportation, industrial manufacturing, and electric power systems. The electric power system is one of the fundamental infrastructures that support the functioning of modern society. It provides a stable supply of electricity to households, businesses, medical institutions, schools, and other institutions, supporting various aspects of life and industrial activities. With the growing demand for electricity in society, higher requirements have been placed on the safety and reliability of the utilization of electric power energy. The application of computer vision technology to electric power systems could enhance safety, efficiency, and reliability in various aspects. By using computer vision technology to monitor power equipment, such as transformers, switchgear, and cables, potential faults or damages can be detected promptly [1]. Utilizing unmanned aerial vehicles equipped with

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

computer vision systems, conducting regular inspections of power lines enables the rapid detection and pinpointing of potential issues, such as line breaks or foreign object attachments [2]. In the data centers of the power system, computer vision technology can be exploited to monitor the operational status, temperature, and energy efficiency of equipment, aiding in enhancing operational efficiency [3]. It is also implemented to monitor the safety conditions of substations including foreign object trespass detection, fire detection, as well as the real-time surveillance of personnel while manipulating electrical power equipment.

Electric power training is essential to ensure the safety and reliable operation of the system, which enables employees to acquire the operational and maintenance skills necessary for electrical power equipment, playing a pivotal role in ensuring the regular functioning of equipment and timely maintenance. Electric power training typically includes theoretical knowledge sessions and practical hands-on training with electrical equipment, often in high-voltage and complex work environments. Therefore, operational safety is crucial in the electric power training scenario, particularly during the practical hands-on training part. Virtual Reality (VR) and Augmented Reality (AR) technologies have been employed in some practical hands-on training processes, where the trainer can undergo operational training for power equipment in a virtual environment [4], simulating real-world scenarios to enhance their proficiency and safety in actual work settings. However, some of the practical hands-on training still needs to be conducted in real-world work environments. Computer vision technology can be employed to monitor the actions of trainers in real time [5], providing immediate feedback to maintain order and safety during electric power training Abnormal fall-down is one of the most serious actions in electrical power training scenarios, particularly when dealing with complex or high-voltage electrical equipment. For safety training purposes, there are restricted zones where trainers are not permitted to enter, especially when power equipment is in operation. However, due to a shortage of teachers compared to trainers, there may be still abnormal trespass during the training process. Thus, it is of great significance to accurately and effectively detect abnormal falls-down and trespass using computer vision technology to ensure safety in electrical power training.

To detect abnormal actions of fall-down and trespass in electric power training scenarios, we provide a novel abnormal action recognition system that utilizes the Lightweight Pose Estimation Network (LPEN). In specific, video frames are evenly extracted and input into the proposed LPEN network, which includes MobileNet, Initial Stage, and Refinement Stage for fast extraction image features, rapid detection of human key points, and further refinement of human key points, respectively. Then a Pose-aware Action Analysis Module (PAAM) is used to obtain the positional coordinates of human skeletal points in the frame. An Abnormal Action Inference Module (AAIM) is next employed to assess whether there is a current occurrence of abnormal fall-down or unauthorized trespass behavior. The assessment of abnormal fall-down considers three criteria: Falling speed, main angles of skeletal points, and the bounding box of humans. For unauthorized trespass, the major focus is on the position of the ankles. Finally, we experimentally validate the performance of the proposed Abnormal Action Recognition System (AARS) with respect to real captured video during the scene of electrical power training.

In summary, the main contributions of this article are as follows:

- We propose an Abnormal Action Recognition (AAR) system equipped with a lightweight pose estimation network to recognize the fall-down action and trespass action of people in electric power training scenarios.
- We introduce a new lightweight pose estimation network, named LPEN, to realize effective and efficient performance in terms of pose estimation.

- We design an Abnormal Action Inference Module (AAIM) to automatically detect the state of fall-down, as well as to automatically recognize the trespass action.

The rest of the paper is organized as follows. In Section 2, we briefly introduce the previous works of 2D pose estimation, along with a short review of OpenPose. Then, the overview of the proposed approach, including technical details, is described in Section 3. The experimental evaluations and analysis are presented in Section 4, and finally, we conclude the paper in Section 5.

2 Related Works

In this section, we review related works on pose estimation and abnormal action recognition.

2.1 Pose Estimation

Human pose estimation aims to accurately infer the positions of key joints of the human body from images or videos, thereby building a representation of the human body [6]. It involves detecting and locating the positions of key joints such as the head, shoulders, elbows, wrists, knees, and ankles. The spatial relationships of these key joints are critical for accurate pose estimation. The input data for human body pose estimation can be either two-dimensional or three-dimensional. Consequently, pose estimation methods are divided into two-dimensional and three-dimensional categories [7]. Since this paper focuses on electrical power training scenarios, the images captured by the sensor camera are in a two-dimensional format. Therefore, this paper primarily discusses two-dimensional human body pose estimation methods and their subsequent application in human body action recognition.

During the early stages, the predominant approaches of human pose estimation are normally the combination of manually designed feature images [8] with graph-structured models [9]. Grounding on this traditional framework, researchers have continuously worked to enhance the accuracy of feature descriptions and the efficiency of searching for body parts. However, due to the high flexibility of human poses, challenges arise in performance when faced with real-world scenarios. Simultaneously, there is a growing demand for more detailed feature descriptions of body parts to improve the accuracy of human pose estimation. With the development of deep learning technologies, especially the tremendous success of Convolutional Neural Networks (CNN) [10], a new dawn has emerged for addressing the aforementioned issues. Pose estimation methods based on deep learning [11] involve constructing neural network models to fit large amounts of training data, implicitly learning the mapping relationship from input images to the coordinates of human key points. In comparison to traditional methods that rely on manually designed representations, deep convolutional neural networks can automatically learn feature representations from data, thus avoiding the drawbacks associated with manual features.

Pose estimation in videos can be divided into single-person and multi-person types, depending on the number of individuals present. The primary objective of single-person pose estimation is to identify the key points of a specific individual appearing in a captured image. In scenarios where multiple individuals are present in an image, it becomes necessary to segment the image into patches, ensuring that each patch contains only one person. This segmentation can be achieved using either an upper-body detector [12] or a full-body detector [13]. Single-person pose estimation methods can be categorized into two types based on the training data: Key point regression-based approaches and heatmap-based approaches [14]. Keypoint regression-based methods, also known as direct regression, directly capture the locations of key points from feature maps learned by an end-to-end framework. Toshev et al. [15] proposed a cascaded Deep Neural Network (DNN) regressor known as DeepPose

for pose estimation. This model uses a 7-layered generic convolutional DNN to regress the location of each body joint based on the full image input. DeepPose has shown better performance compared to traditional methods, paving the way for further advancements in learning-based pose estimation techniques. However, such methods, where the numerical location regression is directly derived from the end-to-end regressor, tend to result in a loss of spatial information of key points, which limits the model's spatial generalization ability. Subsequently, heatmap-based approaches are proposed to address this limitation. Tompson et al. [16] introduced a hybrid architecture for heatmap-based pose estimation, which combines a Deep Convolutional Network with a Markov Random Field. This method leverages the structural relationships between human key points and optimizes the prediction results using a Markov Random Field. As a result, it became one of the most advanced techniques for human pose estimation at the time. By transforming the problem of human pose estimation from coordinate regression to a detection problem based on heat map regression, this method maximizes the preservation of spatial information of key point coordinates. Consequently, it significantly enhances the spatial generalization capability of the learned pose estimation model and improves its accuracy. Wei et al. [17] proposed Convolutional Pose Machines (CPMs) for the task of articulated pose estimation by designing a sequential architecture composed of convolutional networks that directly operate on belief maps from previous stages. CPMs introduce intermediate supervision periodically through the network, thereby replenishing back-propagated gradients and conditioning the learning procedure. Newell et al. [18] proposed a Stacked Hourglass Network (SHN) that captures and consolidates information across different scales of an image. Based on this model, researchers have further developed methods for pose estimation, with a focus on improving accuracy by considering the structure of the human body. Chu et al. [19] proposed an end-to-end framework for human pose estimation that incorporates convolutional neural networks with a multi-context attention mechanism. SHNs are also adopted to generate attention maps from features at multiple resolutions with various semantics. Luo et al. [20] developed a technique called Scale-Adaptive Heatmap Regression (SAHR) to modify the standard deviation for each keypoint. This approach is more tolerant of various human scales and labeling ambiguities, but it aggravates the imbalance between fore-background samples. Huang et al. [21] proposed a principled Unbiased Data Processing (UDP) strategy. This method is equipped with unit length-based measurement and employs a combination of classification and regression for encoding-decoding processes. Additionally, there has been a focus on developing lightweight models, which are crucial for the practical implementation of single-person pose estimation applications [22].

Multi-person pose estimation, which deals with scenes containing several individuals, presents a more complex challenge than single-person estimation, compromising the top-down approach and bottom-up approaches [23]. The top-down approach transforms the original multi-person human pose estimation task into several single-person estimations. It first identifies each human body in the scene and then estimates the coordinates of key points for each identified individual. Chen et al. [24] presented a Cascaded Pyramid Network (CPN) for multi-person pose estimation, which includes two main components: GlobalNet and RefineNet. GlobalNet is responsible for localizing fundamental key points such as eyes and hands, while RefineNet handles key points that cannot be accurately estimated by integrating both global and local features. Xiao et al. [25] proposed a simple yet effective baseline scheme, which employs Faster R-CNN as the body boundary detector and ResNet as the backbone. Rodrigues et al. [26] presented a pose estimation method utilizing Time of Flight (ToF) cameras. DeepCut [27] and DeeperCut [28] are the earliest proposed bottom-up approaches for pose estimation. They initially detect all the key points of the human body from the image. These key points are then treated as nodes of an undirected graph, with the keypoint correlations serving as the weights

between graph nodes. Finally, instance discrimination is modeled as an integer linear programming problem. However, this kind of problem belongs to the NP problems, and implementing integer linear programming on a complete graph incurs extremely high computational complexity. Newell et al. [29] introduced the Associative Embedding method, which assigns a token value to each keypoint by generating a heatmap. This method then uses clustering to group key points with similar token values, ensuring clear distinction between individuals. Cao et al. [30] developed an open-source real-time system, called OpenPose, for multi-person 2D pose detection, including body, foot, hand, and facial key points. OpenPose introduces Part Affinity Fields (PAFs) to achieve fast keypoint connections, which are a set of 2D vector fields used to encode the position and orientation of limbs in the image domain. Similarly, Kreiss et al. [31] designed the PifPaf network to predict the Part Intensity Field (PIF) that represents the positions of body key points, and the Part Association Field that represents the strength of associations between body key points. Finally, a greedy strategy is employed for instance matching. In contrast to methods that focus on partitioning and matching key points, some works are concerned with more accurately predicting instance-independent key points in the image. Cheng et al. [32] proposed a HigherHRNet network to address the scale variations challenges in multi-person pose estimation. Wang et al. [33] demonstrated that a high-resolution branch is unnecessary for a low-computation-region model through a progressive shrinking experiment. They proposed a fusion deconvolution head to eliminate redundant details and improve the performance of the bottom-up pose estimation model by utilizing large convolutional kernels. Although the bottom-up approach is generally faster than the top-down approach, requiring only one pose estimation step, it has its limitations. Notably, the network in bottom-up methods cannot directly obtain features from the frames, resulting in a lower average resolution per person during training compared to top-down methods, when using the same network and GPUs [34].

2.2 Abnormal Action Recognition

The key to identifying abnormal actions is to understand the definition of an exception, which may vary depending on the monitoring scenario. Additionally, an appropriate anomaly detection method should be selected by analyzing the characteristics of the data currently obtained, and the detection results are explained at last. At present, the most used methods for detecting anomalies in distributions include the deviation-based detection algorithm [35], the distance-based detection algorithm [36], etc. The deviation-based detection method compares the main features of the data, with a large deviation if the data in a certain section differs from other data objects. However, this method is not applicable to complex objects with multiple attributes since the requirement of obtaining the main features of the data first. The challenge of the distance-based detection algorithm lies in selecting the appropriate distance function. The basic distance function may not be sufficient for calculating distances of multi-dimensional features, due to the increase in feature dimension. Therefore, it is important to select an appropriate distance function based on the actual application scenario. The distribution-based detection method assumes that the data conforms to a certain probability distribution model and uses inconsistency to judge whether the data matches to determine the isolated points. The density-based detection algorithm utilizes the local anomaly factor to represent the anomaly degree of the data, which primarily relies on genuine anomalies and the correlation of local nearest neighbor densities of the objects to be detected.

The rapid growth of trajectory data holds significant potential for mining individual behavior patterns. Compared with the individual movement of a person at a certain moment, the pedestrian's trajectory can be represented in time and space. Clustering is an effective method to mine trajectory patterns. Based on human trajectory clustering, trajectory clusters are obtained, and behavior patterns

of specific people are mined. Kang et al. [37] used dynamic hierarchical clustering to describe a certain type of action pattern by using the central trajectory, and measured the similarity between the detection trajectory and the central trajectory to determine whether the anomaly was found. Lee et al. [38] proposed a trajectory segmentation detection framework and the trajectory clustering algorithm TRACLUS, which introduces the standard MDL (Minimum Description Length) widely adopted in information compression to extract the velocity feature points of the trajectory. Liu et al. [39] proposed a density-based Trajectory Outlier Detection method DBTOD (Density-based Trajectory Outlier Detection) based on TRADO, which can detect more anomaly points.

Some researchers combine the global and segmented difference measurement methods to test the abnormal trajectory. For instance, Wang et al. [40] reconstructed the trajectory and represented it as a symbolic sequence. Atluri et al. [41] used clustering to identify spatio-temporal anomalies by checking time-domain and space-domain information. Ullah et al. [42] proposed an LSTM network based on an attention mechanism, incorporating Channel Attention and Spatial Attention modules. These modules enhance the most useful information in videos. Farooq et al. [43] proposed a method based on motion shapes and deep learning to detect anomalous behaviors in high-density crowds, specifically focusing on crowd dispersal behaviors. Morris et al. [44] evaluated different similarity measures and clustering methods to find out their advantages and disadvantages in locus clustering. Current research indicates that pedestrian movement is often complex. By extracting the inherent features of the moving target, such as size, color, and posture, and combining them with the motion state information of the pedestrian's speed and direction, we can obtain a measurement method that represents richer trajectory information. This can then be used to identify and track the target, providing significant assistance in acquiring motion trajectory and constructing a normal trajectory model.

3 The Proposed System

To detect abnormal actions such as fall-down and trespass in electric power training scenarios, we have developed a new abnormal action recognition system that uses the Lightweight Pose Estimation Network (LPEN). The overall framework of the proposed technology is shown in Fig. 1. The framework comprises four main components: Surveillance in an Electric Power Training Scenario for frame capture; a Lightweight Pose Estimation Network (LPEN) for human body key points detection; a Pose-aware Action Analysis Module (PAAM) for obtaining the main factors utilized in abnormal action detection; and an Abnormal Action Inference Module (AAIM) for the detection of abnormal fall-down and trespass.

3.1 Lightweight Pose Estimation Network

An LPEN (Lightweight Pose Estimation Network) is employed to detect the key points of the human body. This network is constructed based on the OpenPose framework, with a lighter MobileNet V1 [45] replacing the VGG-19 backbone. This modification aims to maintain high accuracy while making the model more lightweight, improving recognition efficiency, and lowering hardware processing requirements. The proposed LPEN not only replaces its backbone network but also simplifies the multi-stage structure of OpenPose, including only an Initial Stage and a Refinement Stage. The architecture of LPEN is shown in Fig. 2.

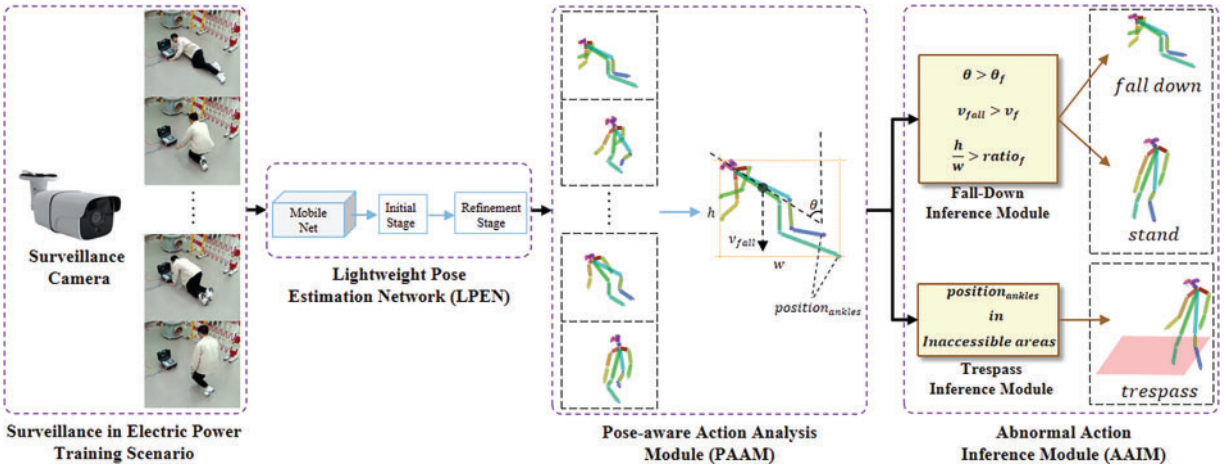

Figure 1: The framework of the proposed Abnormal Action Recognition (AAR) system. This framework consists of four modules, namely Surveillance in Electric Power Training Scenario, Lightweight Pose Estimation Network (LPEN), Pose-aware Action Analysis Module (PAAM), and Abnormal Action Inference Modules (AAIM)

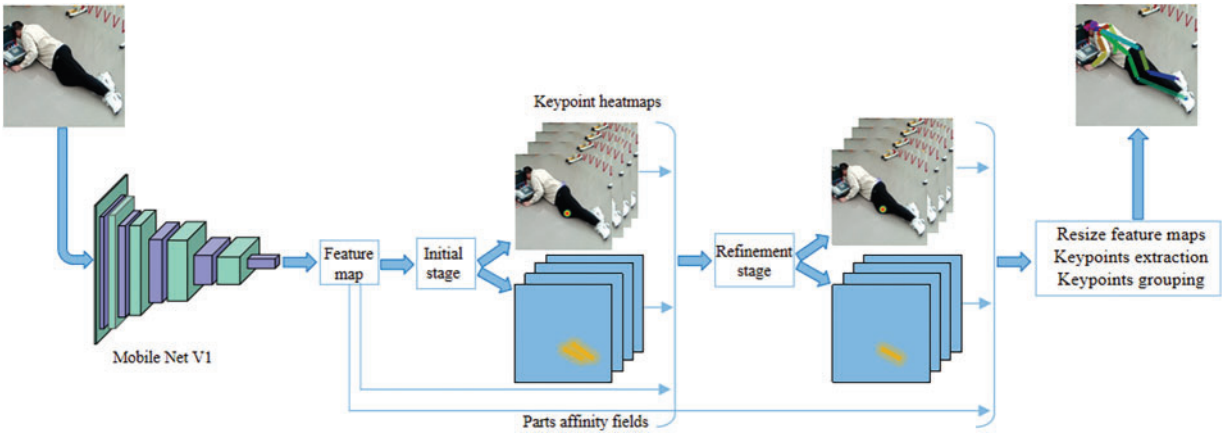

Figure 2: The architecture of the Lightweight Pose Estimation Network (LPEN). Given the video sequence within the human action as the input, the proposed LPEN can automatically obtain the human pose with the 2D skeleton positions

Specifically, color images of size w × h are analyzed by the pre-trained MobileNet V1 network to produce a set of feature maps F. The set of feature maps is input to the Initial Stage where the network outputs a set of Part Affinity Fields (PAFs): $L_1 = \phi_1(F)$, where ϕ_1 is the CNNs used for inference at this stage. In the subsequent Refinement Stage, the initial PAFs predictions L_1 and the original image features F are combined to produce more refined PAFs predictions: $L = \phi_2(F, L_1)$, where ϕ_2 is the CNNs used in this stage. Iteratively, the above process starts from the latest PAFs prediction to repeatedly detect confidence maps: $S_1 = \rho_1(F, L)$ and $S_2 = \rho_2(F, L, S_1)$, where ρ_t is the CNNs used for inference at stage t.

As the confidence maps are based on the latest and most refined PAFs predictions, the difference between stages is minimal. To guide the network to predict PAFs in the first branch and confidence

maps, a loss function is applied at the end of each stage. Specifically, two loss functions f_t^L and f_t^S for the PAFs branch and the confidence map branch at stage t are defined as follows:

$$f_t^L = \sum_{c=1}^{C} \sum_p W(p) \cdot \| L_t^c(p) - L_*^c(p) \|_2^2 \tag{1}$$

$$f_t^S = \sum_{j=1}^{J} \sum_p W(p) \cdot \| S_t^j(p) - S_*^j(p) \|_2^2 \tag{2}$$

where J is the number of confidence maps for different body parts, C is the count of vector fields corresponding to pairs of these parts, $L_*^c(p)$ is the ground truth of PAFs, $S_*^j(p)$ is the ground truth of the confidence maps, and W is a binary mask that is zero when pixel p lacks annotation. The final objective of the whole network is to minimize the sum of loss functions across all layers.

$$f = \sum_{t=1}^{T} f_t^L + f_t^S \tag{3}$$

Overall, images captured by the monitoring system are first processed by a pre-trained MobileNet to extract image features, then rapidly detect key points in the Initial Stage, and further optimized in the Refinement Stage to achieve more accurate pose estimation.

3.2 Pose-Aware Action Analysis Module

The Pose-Aware Action Analysis Module (PAAM) is used to obtain the positional coordinates of human skeletal points in the image. Through these obtained coordinates, we could calculate some critical parameters such as the speed of the body's center of gravity descent v_{fall}, the angle deviation of the body trunk from the vertical axis θ, the ratio of the body's boundary height to width ratio, as well as the positions of the left and right ankle (x_{ankles}, y_{ankles}). These parameters contribute to providing a basis for the judgment of the Abnormal State Inference Module (AAIM) in the subsequent analysis.

3.3 Abnormal Action Inference Module

The Abnormal Action Inference Module (AAIM) consists of two modules: The Fall-down Inference Module and the Trespass Inference Module. The former is designed to detect abnormal actions related to fall-down, while the latter is responsible for detecting trespass.

3.3.1 Fall-Down Inference Module

The Fall-down Inference Module aims to determine whether a person has fallen. Fall-down is a momentary action accompanied by rapid changes in body posture. We utilize the descent speed of the body's center of gravity, the angle of deviation between the body trunk and the vertical axis of the frame, and the height-to-width ratio of the body boundary as key indicators to assess the occurrence of a falling. Specifically, we represent the body's center of gravity (x_{cg}, y_{cg}) using the midpoint between the left hip (x_{lHip}, y_{lHip}) and right hip (x_{rHip}, y_{rHip}). The formula for calculating the coordinates of this midpoint is as follows:

$$x_{cg} = \frac{x_{lHip} + x_{rHip}}{2}, \; y_{cg} = \frac{y_{lHip} + y_{rHip}}{2} \tag{4}$$

The variation in the position of this midpoint is utilized to estimate the descent speed. This speed can be instantaneous or the average speed within a specific time window. In this system, we choose the

latter one to estimate the descent speed. The descent speed v_{fall} can be calculated using the following formula:

$$v_{fall} \approx \frac{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}{\Delta t} \tag{5}$$

where (x_1, y_1) is the position of body center in the first frame and (x_2, y_2) is the position of body center after t seconds (after 5 frames in our work).

The descent speed of the center of gravity is a crucial indicator for determining a fall, as fall-down action typically involves a rapid increase in the center of gravity's speed. In addition, we also analyze the tilt angle of the human body trunk relative to the vertical axis θ. In this work, the line connecting the neck and the center of gravity represents the human body trunk. Thus θ can be formulated as:

$$\theta = arctan\left(\frac{|y_{cg} - y_{Neck}|}{|x_{cg} - x_{Neck}|}\right) \tag{6}$$

where (x_{Neck}, y_{Neck}) is the position of the neck.

Typically, when the human body trunk deviates from the vertical axis by 15–30 degrees, it may indicate a risk of falling. Therefore, it is also an important auxiliary condition for determining whether a fall has occurred. To reduce the possibility of misjudgment in special cases such as bending or squatting, we have added a third criterion: The height-to-width ratio of the body bounding box:

$$ratio = \frac{h}{w} \tag{7}$$

where h and w are the height and width of the body bounding box, respectively. When a person is upright, this ratio tends to be relatively high. Conversely, when a person falls, the ratio decreases due to the more horizontal orientation of the body. Therefore, it can also serve as a measure to determine a fall. To assess a fall event more accurately, we choose to comprehensively consider three core parameters. For example, even if someone's descent speed is not particularly fast, the system can still identify a fall if there are significant changes in trunk tilt angle or height-to-width ratio. To achieve this, we assign weights to each indicator based on experimental data and prior experience, ensuring the influence of each condition in the final decision. We first standardize each parameter:

$$\boldsymbol{v}_{norm} = \frac{\boldsymbol{v}_{fall}}{\boldsymbol{v}_{max}},\ \boldsymbol{\theta}_{norm} = \frac{\boldsymbol{\theta}}{\boldsymbol{\theta}_{max}},\ \boldsymbol{ratio}_{norm} = \frac{\boldsymbol{ratio}}{\boldsymbol{ratio}_{max}} \tag{8}$$

where $\boldsymbol{v}_{max}, \boldsymbol{\theta}_{max}, \boldsymbol{ratio}_{max}$ are the maximum value of each parameter. Thus we would obtain a score to estimate falling:

$$S = w_v \times v_{norm} + w_\theta \times \theta_{norm} + w_r \times ratio_{norm} \tag{9}$$

Finally, the score obtained above is compared with the predefined threshold S_f. If the score surpasses this threshold, the system determines that a falling has occurred, recording the timestamp of the falling. The formula is as follows:

$$IsFall = \begin{cases} 1, & if \quad S \geq S_f \\ 0, & otherwise \end{cases} \tag{10}$$

The analytical approach provides a more holistic and flexible framework for assessing the risk of fall-down, significantly enhancing its practical value in safety scenarios within the context of power training.

3.3.2 Trespass Inference Module

The Trespass Inference Module (TIM) is utilized to assess whether personnel have entered prohibited areas. The system continuously tracks the ankle positions of individuals in the video and compares them with predefined boundaries of restricted zones. A restricted zone is defined as a polygon or rectangular area P with boundaries determined by a series of coordinate points $\{(x_i, y_i) \in \text{boundary}(P)\}$. If the ankle coordinates are within the restricted zone, the system deems it an abnormal trespass and issues a warning. To reduce false alarms, we consider the temporal variation in ankle positions $\Delta_{t_{pass}}$. A time threshold Δt_{in} is presented to determine whether the ankles have remained within the restricted zone for a duration exceeding the threshold. For instance, brief contact with the edge of the restricted zone would not trigger an immediate alert, differentiating between actual trespass and incidental or unintentional touching the crossings. Meanwhile, to mitigate the impact on the accuracy of the pose estimation algorithm, we have appropriately expanded the edges of the restricted zone to eliminate potential jitter issues. Let the expanded region as P′, the width of expansion as d, and the new restricted zone P′ is computed as follows:

$$P' = \left\{ (x_i', y_i') | \sqrt{\left(x_i' - x_i\right)^2 + \left(y_i' - y_i\right)^2} \leq d, \forall (x_i, y_i) \in P \right\} \tag{11}$$

The final judgment formula for the Trespass Inference Module is:

$$IsPass = \begin{cases} 1, & if \quad (x_{ankles}, y_{ankles}) \in P' \; and \;\; \Delta_{t_{pass}} \geq \Delta t_{in} \\ 0, & otherwise \end{cases} \tag{12}$$

where (x_{ankles}, y_{ankles}) is the poison of the ankles acquired in the LPEN.

4 Experiments

To well validate the performance of the proposed system, we conduct validation experiments on the collected data.

4.1 Dataset

We collect 300 original videos from the surveillance in the electric power training scenario, encompassing a broad spectrum of situational occurrences. First, each video is customized and sampled as one video clip containing the fall-down, non-fall-down, trespass, and non-trespass motions. The videos were recorded indoors by eight personnel, varying in age, gender, and body type, to ensure effective motion recognition across diverse individuals. These video clips were categorized into left and right viewing angles, with lengths ranging from 19 to 57 s, and a uniform resolution of 1280 × 720. Second, we employ 10 students to annotate the video clip based on the semantic motion by choosing one label, namely fall-down, non-fall-down, trespass, or non-trespass. Finally, we obtained 85 video clips with the label fall-down, 65 video clips with the label of non-fall-down, 80 video clips with the label of trespass, and 70 video clips with the label of non-trespass. The training and testing set are divided as 2/3 for training, and 1/3 for testing, which are detailed in Table 1.

4.2 Result and Analysis

To illustrate the superior performance of the proposed method, we compare the proposed system and the OpenPose-based system. Here, for fair comparison, the OpenPosed-based system is instead of the proposed LPEN and embedded into the proposed system. The recognition accuracy obtained by

these two methods on the collected dataset is listed in Table 2. We can see that the proposed method achieves the highest accuracy on both the recognition tasks of fall-down and trespass. It is noted that the accuracy of the proposed system in terms of trespass recognition is higher than that in terms of fall-down recognition. This is because the motion of trespass is more obvious than the motion of fall-down in the visual space.

Table 1: The statistics of the dataset

Types	Category			
	Fall-down	Non-fall-down	Trespass	Non-trespass
Training set	57	43	53	47
Testing set	28	22	27	23
Dataset	85	65	80	70

Table 2: The comparison of recognition accuracy between the proposed method and the related method

Method	Recognition accuracy (%)		
	Fall-down	Trespass	Overall
OpenPose-based system	95.2	97.5	96.35
The proposed system	97.6	98.7	98.15

LPEN aims to estimate the positions of all human skeletons. To validate the estimation accuracy of the proposed LPEN in terms of the pose estimation, we adopt the PCK (Percentage of Correct Key points) evaluation to test the performance of LPEN by comparing it with OpenPose. Table 3 lists the comparison between LPEN and OpenPose. It can be seen that the proposed LPEN achieves the best estimation accuracy, namely gaining an improvement of 3.724% compared with OpenPose.

Table 3: The accuracy (%) comparison between LPEN and OpenPose

Method	Body					Mean
	Neck	Left hip	Right hip	Left ankle	Right ankle	
OpenPose	89.021	86.634	89.021	77.326	78.758	84.152
LPEN	94.033	88.544	89.976	83.532	83.293	87.876

As mentioned before, the proposed abnormal action recognition system is not only effective for recognizing abnormal actions but also is efficient in terms of computation. This belongs to the proposed lightweight LPEN, as the key module in the proposed abnormal action recognition system. Thus, we conduct the computation comparison between the proposed LPEN and the OpenPose model in terms of the GFLOPs, Parameters, Fps, Start-up time, and Memory Consumption. The comparison results are listed in Table 4. We can see that the proposed LPEN has lower GFLOPs, a smaller number of parameters, a shorter start-up time, and a lower Memory Consumption, compared with those of

OpenPose. Meanwhile, for processing the same videos, the proposed LPEN can reach 23.59 Fps, which is higher than the 20.76 Fps of OpenPose. Overall, through the computation comparison, the proposed method shows better efficiency than the current representation model, namely OpenPose. In summary, it is evident that the utilization of our Lightweight Pose Estimation Network (LPEN) is more suitable for deployment in practical electric power training scenarios.

Table 4: The comparison of computation between the proposed LPEN and the related model. The notion ↓ denotes the smaller value is better, while the notion ↑ denotes the larger value is better

Type	OpenPose	LPEN
GFLOPs↓	136.1	9.0
Parameters (total)↓	25.94 M	4.1 M
Fps (average) ↑	20.76	23.59
Start-up time (average)↓	0.858 s	0.092 s
Memory consumption (average) ↓	232.5 MB	171.4 MB

Finally, we also conduct a qualitative analysis to illustrate the effectiveness of the proposed system. We report some results of abnormal action recognition obtained by the proposed system. Fig. 3 shows the interface of the proposed system. Fig. 4 shows the state of fall-down recognition obtained by the proposed system. Fig. 5 shows the state of trespass recognition obtained by the proposed system. We can see that the proposed system can accurately recognize the state of fall-down action and the state of trespass action.

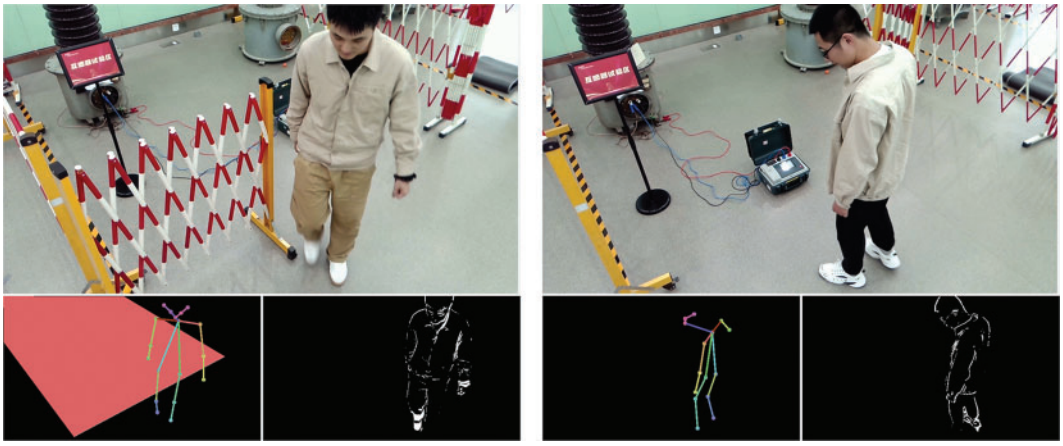

Figure 3: The interface of the proposed abnormal action recognition system

In addition to the aforementioned successful results, we also report some false results in terms of fall-down action and trespass action obtained by the proposed system, as shown in Figs. 6 and 7. The findings indicate that the system exhibits a higher accuracy rate in recognizing actions when the majority of the monitored individual's body is within the surveillance area. However, in some cases where only a minimal portion of the body is exposed to the surveillance field of view, the system may produce erroneous judgments due to the limitations of the pose estimation algorithm. Although such occurrences are relatively infrequent, they can be mitigated by deploying multiple cameras throughout the entire project area, thereby reducing the likelihood of such misjudgments.

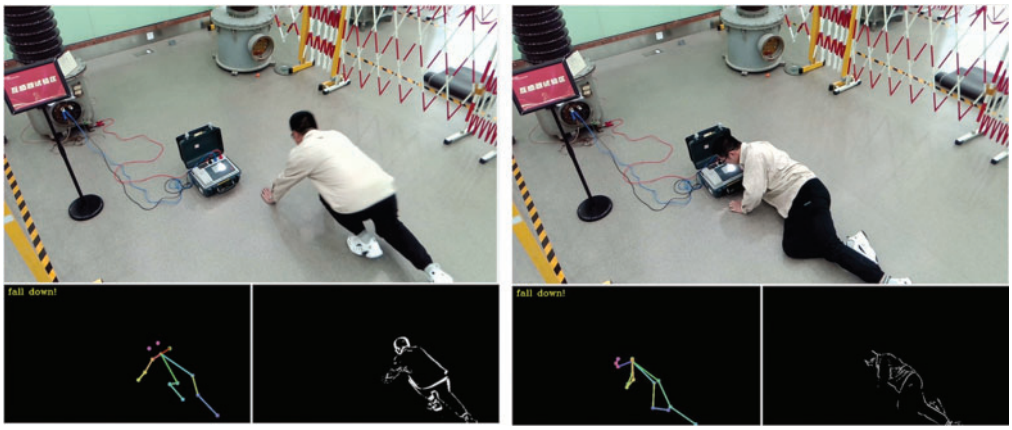

Figure 4: The recognition result of fall-down action obtained by the proposed system

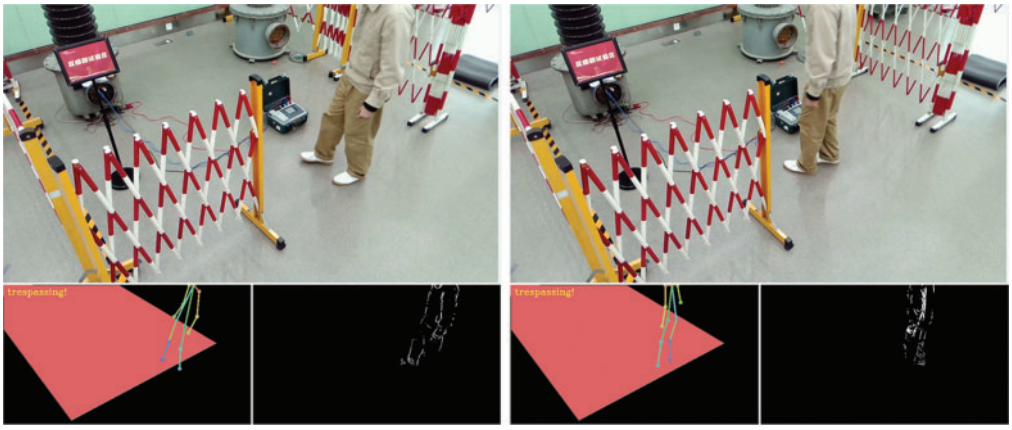

Figure 5: The recognition result of trespass action obtained by the proposed system

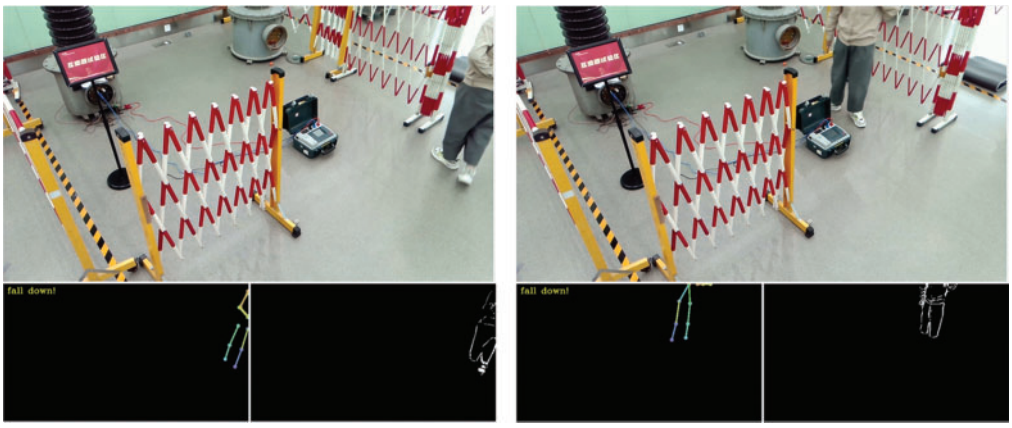

Figure 6: The false result of fall-down action obtained by the proposed system

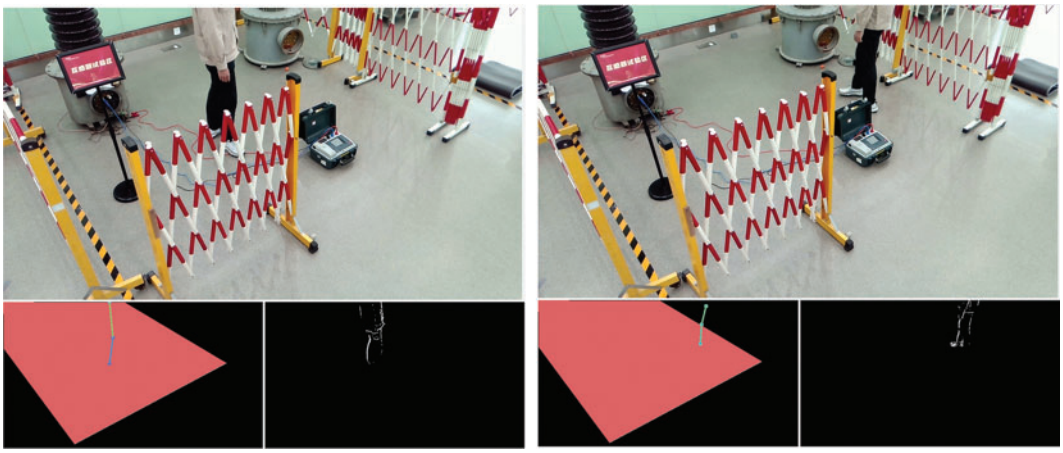

Figure 7: The false result of trespass action obtained by the proposed system

5 Conclusion

Electric power training is an indispensable component to ensure the safety and reliable operation of the system. To address the problem of abnormal action recognition in electric power training scenarios, we present a novel Abnormal Action Recognition (AAR) system embedded with a new Lightweight Pose Estimation Network (LPEN) for effectively and efficiently recognizing the fall-down action and the trespass action. The AAR system consists of four modules, namely Surveillance in Electric Power Training Scenario, Lightweight Pose Estimation Network (LPEN), Pose-aware Action Analysis Module (PAAM), and Abnormal Action Inference Modules (AAIM). LPEN, which includes MobileNet, Initial Stage, and Refinement Stage, is designed to capture the skeleton positions of humans. PAAM aims to acquire the positional coordinates of human skeletal points in the frame. AAIM aims to assess whether there is a current occurrence of abnormal fall-down or unauthorized trespass behavior in the current frame. In the assessment of abnormal fall-down, three aspects are considered as criteria: Falling speed, major angles of the skeletal points, and the bounding box of the human. For the unauthorized trespass, the position of the ankles is the main focus. Extensive experiments are conducted to show the effectiveness and efficiency of the proposed system.

Acknowledgement: Many thanks to Yang Zhu, Bo Jiao and other volunteers during data collection.

Funding Statement: This work has been supportted by Natural Science Foundation of Jiangsu Province (No. BK20230696).

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Yunfeng Cai, Qing Yang; data collection: Long Zhang; analysis and interpretation of results: Jin Tang, Xiaotian Bi; draft manuscript preparation: Yunfeng Cai, Ran Qin, Qing Yang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Due to the nature of this research and the privacy information of volunteers, participants of this study did not agree for their data to be shared publicly.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Alsumaidaee, C. Yaw, and S. Koh, "Review of medium-voltage switchgear fault detection in a condition-based monitoring system by using deep learning," *Energies*, vol. 15, no. 18, pp. 6762, 2022. doi: 10.3390/en15186762.

[2] Y. Zhang, Z. Yan, J. Zhu, S. Li, and C. Mi, "A review of foreign object detection (FOD) for inductive power transfersystems," *eTransportation*, vol. 15, no. 11, pp. 102–116, 2019. doi: 10.1016/j.etran.2019.04.002.

[3] A. Khalaj, T. Scherer, J. Siriwardana, and S. Halgamuge, "Increasing the thermal efficiency of an operational data center using cold aisle containment," presented at the ICIAfS, Colombo, Sri Lanka, Dec. 22–24, 2014.

[4] E. K. Tam, F. Badra, R. J. Marceau, M. A. Marin, and A. S. Malowany, "A Web-based virtual environment for operator training," *IEEE*, vol. 14, no. 3, pp. 802–808, 1999. doi: 10.1109/59.780889.

[5] C. Z. Dong and F. N. Catbas, "A review of computer vision-based structural health monitoring at local and global levels," *Struct. Health Monit.*, vol. 20, no. 2, pp. 692–743, 2021. doi: 10.1177/1475921720935585.

[6] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," presented at the CVPR, Columbus, OH, USA, Jun. 23–28, 2014.

[7] R. M. Haralick, H. Joo, C. Lee, X. Zhuang, V. G. Vaidya and M. B. Kim, "Pose estimation from corresponding point data," *IEEE Trans. Syst.*, vol. 19, no. 6, pp. 1426–1446, 1989. doi: 10.1016/B978-0-12-266719-0.50006-3.

[8] A. Erol, G. Bebis, M. Nicolescu, and R. D. Boyle, "Vision-based hand pose estimation: A review," *Comput. Vis. Image Underst.*, vol. 108, no. 1–2, pp. 52–73, 2007.

[9] J. Wang, X. Long, Y. Gao, E. Ding, and S. Wen, "Graph-PCNN: Two stage human pose estimation with graph pose refinement," presented at the ECCV, Glasgow, Scotland, Aug. 23–28, 2020.

[10] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, 2021. doi: 10.1109/TNNLS.2021.3084827.

[11] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu and J. Shen, "Deep learning-based human pose estimation: A survey," *ACM Comput. Surv.*, vol. 56, no. 1, pp. 1–37, 2023. doi: 10.1145/1122445.1122456.

[12] A. S. Micilotta, E. J. Ong, and R. Bowden, "Real-time upper body detection and 3D pose estimation in monoscopic images," presented at the ECCV, Graz, Austria, May 7–13, 2006.

[13] G. Hidalgo *et al.*, "Single-network whole-body pose estimation," presented at the ICCV, Seoul, South Korea, Oct. 27–Nov. 2, 2019, pp. 6982–6991.

[14] A. Nibali, Z. He, S. Morgan, and L. Prendergast, "3D human pose estimation with 2D marginal heatmaps," presented at the WACV, Wailea, HI, USA, Jan. 7–11, 2019, pp. 1477–1485.

[15] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," presented at the CVPR, Columbus, OH, USA, Jun. 23–28, 2014, pp. 1653–1660.

[16] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," presented at the NIPS, Montreal, Canada, Dec. 1–5, 2014, pp. 1799–1807.

[17] S. E. Wei, V. Ramakrishna, and T. Kanade, "Convolutional pose machines," presented at the CVPR, Las Vegas, NV, USA, Jun. 26–Jul. 1, 2016, pp. 4727–4732.

[18] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," presented at the ECCV, Amsterdam, Netherlands, Oct. 11–14, 2016.

[19] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille and X. Wang, "Multi-context attention for human pose estimation," presented at the CVPR, Honolulu, HI, USA, Jul. 21–26, 2017, pp. 1831–1840.

[20] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan and E. Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," presented at the CVPR, Jun. 19–25, 2021, pp. 13264–13273.

[21] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The devil is in the details: Delving into unbiased data processing for human pose estimation," presented at the CVPR, Seattle, WA, USA, Jun. 14–19, 2020, pp. 5700–5709.

[22] D. Groos, H. Ramampiaro, and E. A. F. Ihlen, "EfficientPose: Scalable single-person pose estimation," *Appl. Intell.*, vol. 51, no. 4, pp. 2518–2533, 2021. doi: 10.1007/s10489-020-01918-7.

[23] H. S. Fang, S. Xie, Y. W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," presented at the ICCV, Venice, Italy, Oct. 22–29, 2017, pp. 2334–2343.
[24] Y. Chen, Z. Wang, Y. Peng, and Z. Zhang, "Cascaded pyramid network for multi-person pose estimation," presented at the CVPR, Salt Lake City, UT, USA, Jun. 18–22, 2018, pp. 7103–7112.
[25] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," presented at the ECCV, Munich, Germany, Sep. 8–14, 2018, pp. 466–481.
[26] N. Rodrigues *et al.*, "Top-down human pose estimation with depth images and domain adaptation," presented at the VISIGRAPP, Prague, Czech Republic, Feb. 25–27, 2019, pp. 281–288.
[27] L. Pishchulin, E. Insafutdinov, and S. Tang, "DeepCut: Joint subset partition and labeling for multi-person pose estimation," presented at the CVPR, Las Vegas, NV, USA, Jun. 26–Jul. 1, 2016, pp. 4929–4937.
[28] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," presented at the ECCV, Amsterdam, Netherlands, Oct. 11–14, 2016, pp. 34–50.
[29] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," presented at the NIPS, Long Beach, CA, USA, Dec. 4–9, 2017, pp. 2277–2287.
[30] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," presented at the CVPR, Honolulu, HI, USA, Jul. 21–26, 2017, pp. 7291–7299.
[31] S. Kreiss, L. Bertoni, and A. Alahi, "PIFPAF: Composite fields for human pose estimation," presented at the CVPR, Long Beach, CA, USA, Jun. 16–20, 2019, pp. 11977–11986.
[32] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang and L. Zhang, "HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation," presented at the CVPR, Seattle, WA, USA, Jun. 13–19, 2020, pp. 5386–5395.
[33] Y. Wang, M. Li, H. Cai, W. M. Chen, and S. Han, "LitePose: Efficient architecture design for 2D human pose estimation," presented at the CVPR, New Orleans, LA, USA, Jun. 19–25, 2022, pp. 13126–13136.
[34] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: A survey," *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 663–676, 2019. doi: 10.1145/1122445.1122456.
[35] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," presented at the SIGMOD Conf., Santa Barbara, CA, USA, May 21–24, 2001, pp. 37–46.
[36] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," presented at the SIGMOD Conf., Dallas, TX, USA, May 16–18, 2000, pp. 427–438.
[37] K. Kang, W. Liu, and W. Xing, "Motion pattern study and analysis from video monitoring trajectory," *IEICE Trans. Inf. Syst.*, vol. 97, no. 6, pp. 1574–1582, 2014. doi: 10.1587/transinf.E97.D.1574.
[38] J. G. Lee, J. Han, and K. Y. Whang, "Trajectory clustering: A partition-and-group framework," presented at the SIGMOD Conf., Beijing, China, Jun. 12–14, 2007, pp. 593–604.
[39] Z. Liu, D. Pi, and J. Jiang, "Density-based trajectory outlier detection algorithm," *J. Syst. Eng. Electron.*, vol. 24, no. 2, pp. 335–340, 2013. doi: 10.1109/jsee:2013.00042.
[40] H. Wang and C. Schmid, "Action recognition with improved trajectories," presented at the ICCV, Sydney, Australia, Dec. 1–8, 2013, pp. 3551–3558.
[41] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-temporal data mining: A survey of problems and methods," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–41, 2018. doi: 10.1145/3161602.
[42] M. Ullah, M. M. Yamin, A. Mohammed, S. D. Khan, H. Ullah and F. A. Cheikh, "Attention-based LSTM network for action recognition in sports," *Electron. Imaging*, vol. 33, no. 4, pp. 1–6, 2021. doi: 10.2352/ISSN.2470-1173.2021.6.IRIACV-302.
[43] M. U. Farooq, M. N. M. Saad, and S. D. Khan, "Motion-shape-based deep learning approach for divergence behavior detection in high-density crowd," *Vis. Comput.*, vol. 38, no. 5, pp. 1553–1577, 2022. doi: 10.1007/s00371-021-02088-4.
[44] B. Morris and M. Trivedi, "Learning trajectory patterns by clustering: Experimental studies and comparative evaluation," presented at the CVPR, Miami, FL, USA, Jun. 20–25, 2009, pp. 312–319.
[45] D. Sinha and M. El-Sharkawy, "Thin MobileNet: An enhanced MobileNet architecture," presented at the UEMCON, NY, USA, Oct. 10–12, 2019, pp. 0280–0285.