

Received 16 January 2024, accepted 6 February 2024, date of publication 22 February 2024, date of current version 5 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368869

RESEARCH ARTICLE

Application of a Two-Dimensional Regression Network Algorithm Model Based on Local Constraints in Human Motion Recognition

LIJUN WANG¹, ZIXU WANG², AND LIJUAN ZHOU³

¹School of Computer Application Technology, Changchun University of Technology, Changchun 130012, China

²School of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China

³School of Journalism and Communication, Changchun University of Technology, Changchun 130012, China

Corresponding author: Lijuan Zhou (zhoulijuan_cc@126.com)

This work was supported by the Education Department of Jilin Province under Grant JJKH20230772CY.


ABSTRACT As the behavior analysis of human body is more and more used in the fields of intelligent monitoring and motion analysis, it is of great significance to conduct research. The current two-dimensional regression network algorithm models in human motion recognition and estimation do not consider the interrelationships between human joint points, resulting in missing connections between joint points and low accuracy of feature maps. Therefore, this study proposes an improved two-dimensional regression network algorithm model based on local constraints and relational networks, and verifies its effectiveness. The experimental results show that, considering only local constraints, the proportion of the head in the correct key points of the improved algorithm in the wrist joint score is 84.72%, while the comparison algorithm is 84.55%, an increase of 1.17%. The maximum value is 88.7% when the number of regression network modules is 8. In practical applications, the actual label results of indoor and outdoor environments are basically consistent with those of the detected image, but there are errors under indoor occlusion conditions. Considering both local constraints and relational networks, the improved algorithm has variant standard scores of 98.8%, 95.3%, 93.3%, 89.4%, 95.1%, 96.2%, and 94.2% for the correct percentage of 7 joint points, respectively, which are higher than the comparison algorithm. Overall, the proposed two-dimensional regression network algorithm based on local constraints and relationship networks has practicality and effectiveness, which can be effectively applied in practical human motion recognition.

INDEX TERMS Enter two-dimensional regression network algorithm, human body movements, relationship network, PCKh, joint point.

I. INTRODUCTION

Human pose estimation refers to locating the position of human joint points in a given image or video, and the actual output is the coordinates of human joint points. In human motion recognition and three-dimensional pose estimation, two-dimensional human pose estimation is an important foundation [1]. The popularization and development of mobile devices have created a great market for the application of human posture estimation, such as motion clas-

sification, rehabilitation medicine, autonomous driving, etc. In human-computer interaction, human posture estimation is deeply embedded in people's daily lives through various products and software. In the field of intelligent monitoring, the use of human posture estimation can effectively monitor personnel, improve efficiency, and enhance the ability to process massive data. In the field of entertainment and sports, using human posture estimation can make people's entertainment life more blocked, and at the same time, athlete movements can be standardized by recording their movements. In addition, human pose estimation can improve the understanding ability of computer vision systems for human

The associate editor coordinating the review of this manuscript and approving it for publication was Davide Patti .

behavior in the field of motion recognition. It can provide more accurate and rich input features for action recognition [2], [3]. The mainstream two-dimensional human body pose estimation algorithm mostly based on figure structure model, although the figure structure model was successful in some attitude estimation task, but based on the figure structure model of human body pose estimation algorithm need to detect various parts in the image, and in the real environment, due to the background noise, human appearance changes, detect individual parts of the human body is very difficult. Such as using hybrid local deformable model said the target human body part to provide more abundant body, using hidden support vector machine training model, and put forward a based on cascade detection method, with enhanced regression tree as local human posture regression model, from the root of the dependency node (detect the center of the border), along the dependent path estimate the local posture, and eventually combined into human body posture. However, the two-dimensional human pose estimation algorithm based on the regression method has a large motion range due to the diversification of human activities, which leads to the inaccurate detection of joint points. At the same time, because the human joint points do not exist alone, they have strong structured information, this information is not considered in the current regression-based method. In general, the current two-dimensional human pose estimation algorithm, human posture and human movement are very free, and there are inevitably occlusion problems, and the movement range of each joint point of the human body is large, resulting in poor accurate estimation. In order to fully focus on the relationship between joint points in human pose estimation, improve the disadvantage of low accuracy of human pose estimation due to large freedom of joint points, and solve the problem of human pose estimation in occlusion state. Therefore, this study proposes a two-dimensional regression network algorithm using local constraints, and introduces a relational network to optimize it while considering occlusion interference. We also propose a two-dimensional regression network algorithm for human motion recognition integrating local constraint and relational networks. The aim is to solve the current shortcomings of human pose estimation application in motion recognition and strengthen the connection between network structures during information transmission, thus improving the accuracy of joint point estimation. The research innovation points are as follows: first, according to the decomposition of the human body posture from the whole to the local, the connection between the joint point and the joint point is determined to form a local component. The local component relationship of the human body is added to the objective function to form the penalty term, which together constitutes the objective function of the network. This shape constraint enables the network to implicitly learn the spatial distribution between the joint points. Secondly, based on the data analysis of two-dimensional human pose, it is combined with the two-dimensional human pose estimation algorithm based on local pose constraint. The proposed algorithm solves

the problem of joint occlusion in the regression algorithm, and can realize the accurate positioning of occluded joints. It provides an effective method for human pose estimation in the fields of motion classification, rehabilitation medical treatment, abnormal behavior detection and automatic driving, and promotes the application of human posture estimation in life.

This study is divided into four parts in total. The first part is a summary and discussion of the current application of human posture estimation in human motion recognition. The second part is the analysis of a two-dimensional regression network algorithm model for facial recognition of human movements. The third part is to verify the performance of the improved algorithm. The fourth part is a summary of the entire article.

II. RELATED WORK

Human pose estimation is crucial for human motion recognition and pedestrian recognition. With the rapid development of deep learning, the effectiveness of human pose estimation in human motion recognition is constantly improving both domestically and internationally, and it has begun to be widely applied in the related fields of computer vision. Gosztołai A et al. has improved the 2D pose estimation method by introducing 3D pose enhancement and deep learning to address the current issues in 3D pose estimation and motion recognition. This method improves the performance of human 3D pose estimation and motion recognition under occlusion conditions [4]. Bhavanasi G et al. improved the two-dimensional pose estimation method by utilizing graph machine neural networks and supervised machine learning classification methods to address the related issues in indoor human motion recognition. This method not only improves detection accuracy, but also effectively enhances the robustness of mobile devices for action recognition applications [5]. Li S et al. improved the two-dimensional pose estimation algorithm by utilizing machine learning and deep learning classifiers to address the issues related to pose estimation and action recognition in the daily lives of elderly people. They proposed a fusion ensemble classification algorithm, which effectively improves the accuracy of human pose estimation and action recognition [6]. Chen et al. proposed a related structure for full convolutional propagation with long hop connections by introducing deep learning algorithms to address the current issues of 3D pose estimation and action recognition. This model effectively improves the accuracy of human 3D motion recognition while reducing deep blur during motion recognition [7].

In addition, Dubey et al. comprehensively discussed the deep learning models used for pose estimation and action recognition in the field of human pose estimation in static image understanding. While overcoming the problem of human motion recognition in complex situations, this study also provides a new path for the application of human posture estimation in practical life [8]. Vishwakarma D K et al. provided an improved manual model for human motion recognition by utilizing deep learning models and 3D

sequence skeleton data. This model improves both detection efficiency and recognition accuracy [9]. Liu S et al. proposed a lightweight pose estimation network by utilizing polarization self attention mechanism to address the relevant issues in the practical application of human pose estimation in human motion recognition. It effectively enhances the performance of human motion recognition by improving the accuracy of key joint point regression [10]. Arab H et al. improved the current pose estimation method by utilizing a four noise Weibo amplifier and multiple two-dimensional convolutional layers to address related issues in human motion monitoring and recognition. This method improves the accuracy of human motion recognition on the basis of improving the accuracy of human motion classification [11].

The above numerous studies indicate that there are problems with incomplete detection and low accuracy in the current application of human posture estimation in human motion recognition. However, the current methods for estimating two-dimensional human posture do not take into account the structured information of the human body. At the same time, relevant algorithms only utilize local network analysis and contextual information to solve occlusion problems, only considering the detection of individual joint points, resulting in low detection accuracy. Based on this, this study utilizes local constraints to enhance the utilization of structured information, and integrates relational networks to enhance the ability to handle occlusion problems. The improved two-dimensional regression network algorithm is innovative.

III. ANALYSIS OF TWO-DIMENSIONAL REGRESSION NETWORK ALGORITHM MODEL FOR HUMAN MOTION RECOGNITION

The current regression network algorithm models related to two-dimensional pose estimation for human motion recognition have many shortcomings. Therefore, this section mainly proposes a two-dimensional regression network algorithm model using local constraints to address the issue of feature map accuracy, as well as an improved algorithm that introduces relational networks for optimization based on it.

A. TWO-DIMENSIONAL HUMAN MOTION RECOGNITION AND ESTIMATION BASED ON LOCAL CONSTRAINTS

The current two-dimensional regression network algorithm models used for human motion recognition and estimation do not consider the issues of missing connections between human joint points and low accuracy of feature maps due to the mutual connections between joint points. Therefore, this study proposes a two-dimensional human motion estimation algorithm based on local motion constraints, while considering the factors of local joint occlusion. Based on this algorithm, a relational network is introduced to optimize it. Each joint in human motion has characteristics such as greater degrees of freedom, higher flexibility, and constantly changing joint states in different movements. Based on this, most of the current research methods use cascaded neural

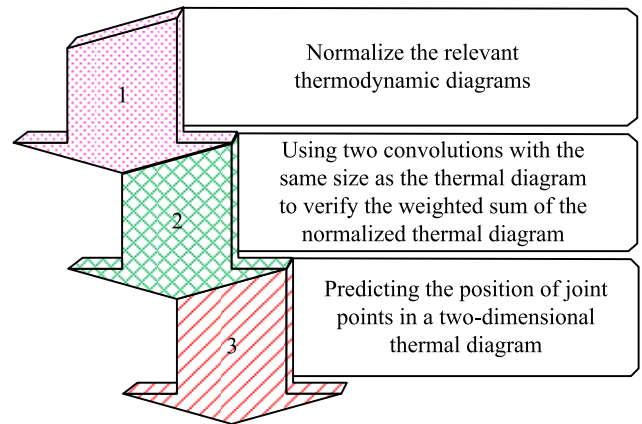


FIGURE 1. The specific process of end-to-end training methods based on regression networks.

networks to extract multi-scale features. It repeatedly uses a bottom-up or top-down approach to calculate the position of human joint points, thereby capturing their spatial position information and ultimately obtaining the coordinates of human joint points. Nevertheless, these methods did not consider the correlation between human joint points [12], [13], [14].

Therefore, the algorithm proposed on the basis of local action constraints mainly combines the advantages of regression networks being able to train end-to-end through coordinates and the human body graph structure being able to model the spatial constraints of human joint points. A loss function for local action constraints is designed, and this structure is introduced into the loss function by simulating the structure of human joint points. The basic idea of the algorithm is to consider that each joint point in the human body is not completely independent and has strong spatial constraints. In response to this issue, this study intends to adopt the idea of local pose constraints, and construct an objective function that utilizes pose constraints at the cost of the similarity between the pose datasets of each joint point and the actual pose datasets in the plane. By constructing a loss function with local pose constraints, implicit learning of the position distribution between connected nodes is hoped to be achieved.

Among them, the current methods for estimating human motion mainly rely on thermal maps and regression. The thermal maps method usually detects each joint point separately, and then gathers them together in the post-processing stage to form an action prediction. The regression method utilizes a function to directly correspond the input image to the joint positions of the body. Compared to the former, this method has the advantage of directly providing motion prediction of the actual coordinates of joint points, without the need for additional steps and post-processing. Therefore, this study chose the regression method to construct the algorithm [15], [16], [17]. Figure 1 shows the specific content of using regression methods.

In Figure 1, the process of using the regression method first involves normalizing the relevant thermodynamic maps. The purpose is to keep the weighted sum in the range of 0-1 for the next step. Secondly, two convolutions with the same size as the thermal map are used to verify the weighted sum of the normalized thermal map. It is equivalent to multiplying and then adding the corresponding elements. Finally, predict the position of joint points in the two-dimensional thermal diagram. The human pose regression method for static images based on Softmax function yields an end-to-end trainable method without requiring a manually generated heatmap and a simple model framework [18]. The calculation expression for normalizing the thermodynamic diagram is equation (1).

$$\Phi(g_{k,l}) = \frac{e^{g_{k,l}}}{\sum_{i=1}^M \sum_{j=1}^G e^{g_{i,j}}} \quad (1)$$

In equation (1), $\Phi(g_{k,l})$ represents the normalized heat map. g represents the thermal diagram, while $g_{k,l}$ represents the actual size of the relevant position (k, l) in the thermal diagram. $M \times G$ represents the actual size of the thermodynamic diagram. e represents a natural constant. The corresponding expression in the weighted sum of the thermodynamic diagram is equation (2).

$$\xi_a(g) = \sum_{i=1}^M \sum_{j=1}^G K_{k,l,a} \Phi(g_{k,l}) \quad (2)$$

In equation (2), $\xi_a(g)$ represents the weighted sum of the thermodynamic diagram. a represents the given component p or q . K represents the $M \times G \times 2$ weight matrix corresponding to coordinate (p, q) . It can be represented by two component matrices, as shown in (3).

$$K_{k,l,p} \frac{k}{K}, K_{k,l,q} \frac{l}{K} \quad (3)$$

Finally, the expression of the joint point prediction formula is shown in equation (4).

$$q = (\xi_p(g), \xi_q(g))^T \quad (4)$$

In order to integrate the density function layer (Soft argmax) into the deep network, it is necessary to take the derivative of $g_{k,l}$, as expressed in equation (5).

$$\frac{\partial \xi_a(g_{k,l})}{\partial g_{k,l}} = K_{k,l,a} \frac{e^{g_{k,l}} \left(\sum_{i=1}^M \sum_{j=1}^G e^{g_{i,j}} - e^{g_{k,l}} \right)}{\left(\sum_{i=1}^M \sum_{j=1}^G e^{g_{i,j}} \right)^2} \quad (5)$$

Among them, by combining local detection and contextual information, the heat map representation can be indirectly learned, and the regression network model can be optimized by utilizing end-to-end regression losses. Therefore, the network architecture of the regression network model is shown in Figure 2.

In Figure 2, the actual input of the regression network model is $256 * 256 * 3$. The architecture is divided into three parts in total. The first part is the stem layer, which mainly consists of the initial fourth version module (Inception v4)

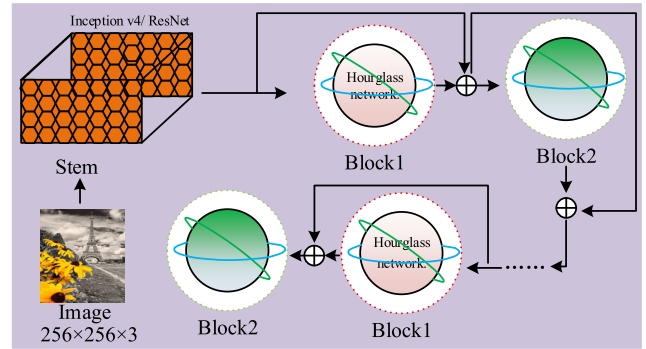
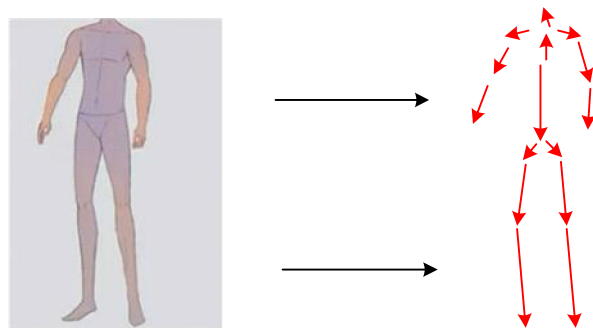


FIGURE 2. Schematic diagram of network architecture for regression network model.

and a residual separable network. Its function is to extract relevant initial features to obtain a $64 * 64 * 512$ feature map. The second part is the Block1 structure, which utilizes the relevant architecture of the Hourglass network. The difference lies in replacing all residual networks with residual separable networks to obtain a $256 * 64 * 64$ feature map. The third stage is the Block 2 structure, which divides the actual input features into m predicted joint points. The actual regression network model includes a total of 8 sets of Block1 layer and Block layer structures. In addition, the graph structure model is a classic object statistical model. The basic idea is that the description of an object consists of a set of deformable structures, and the shape of each part is modeled separately. The deformable target structure is expressed using a spring like connection between two parts.

The current regression network takes the entire image as input, utilizes the features of convolutional neural networks for global extraction, and then predicts the position of each point. This method can cause the loss of internal relationships between joint points, resulting in inaccurate detection of joint points. The human body diagram structure model represents the human body as a collection of human body components, utilizing the joint points between adjacent components to solve the problems existing in existing regression networks. The basic idea is to treat the adjacent joint points of the human body as a component and construct a model based on the constraint information between the components. Therefore, this study incorporates the constraint relationship between components between adjacent joint points into the objective function summary of the regression network model as a penalty term for the actual prediction results. The process, together with the square loss function as the final objective function, solves the defect of only the square loss function as the actual objective function. This study improves the objective function by constructing the relationship between points based on the connections between human joint points. Therefore, the schematic diagram of the human body structure is shown in Figure 3.

In Figure 3, the human body typically consists of 16 joints, including the enlarged head, lower jaw, left shoulder, left wrist, etc. By modeling human joints and connecting the



(a) Physical structure of the human body (b) Decomposition of human body structure

FIGURE 3. Schematic diagram of human body structure.

corresponding joint points based on the graph structure model, 15 components can be formed, including the head (head joint and lower jaw), trunk (neck and abdominal joint points), and so on. On the surface, human movements appear to be a holistic process. But when it comes to a specific part, the corresponding analysis can be carried out using the idea of segmentation. Therefore, this study mainly applies the idea of from local to global to form the overall part of human motion recognition. Based on this, a two-dimensional human motion estimation algorithm is constructed using local motion constraints. In the improved objective function under this algorithm, the training network model uses the sum of L1 and L2 loss functions to calculate the actual loss function. The regression network calculates the loss function based on the final obtained relevant coordinates. According to formula (5), the Soft argmax function can be integrated into a trainable architecture based on the principles of backpropagation and chain differentiation. Its expression is shown in equation (6) [19].

$$\mathfrak{S}_y = \frac{1}{\mathfrak{N}_J} \sum_{n=1}^{\mathfrak{N}_J} \left\{ \|y_n - \hat{y}_n\|_1 + \|y_n - \hat{y}_n\|_2^2 \right\} \quad (6)$$

In equation (6), \mathfrak{S}_y represents the final loss function value. \mathfrak{N}_J represents the number of joint points. y_n represents the true value of the joint point position. \hat{y}_n represents the predicted value of the actual output of the model. Estimate the relevant positions of joint points in actual training, using them as penalty terms for the objective function. This study used the cross entropy loss function, which is expressed in equation (7) [20].

$$\mathfrak{S}_\eta = \frac{1}{\mathfrak{N}_J} \sum_{n=1}^{\mathfrak{N}_J} \left\{ (\eta_n - 1) \log(1 - \hat{\eta}_n) - \eta_n \log \hat{\eta}_n \right\} \quad (7)$$

In equation (7), \mathfrak{S}_η represents the loss function of the probability of joint position. η_n represents the probability of the predicted joint position related to the actual output of the model. $\hat{\eta}_n$ represents the true value of the actual probability of the corresponding joint point. To fully utilize the structural information between joint points and enable them to

utilize more joint constraint information during the matching process, it is proposed to describe it as the link relationship between a set of joint points and their adjacent joint points. And the association relationship between each set of joint points and their corresponding edge is independent of each other. The graph structure model represents the human body as an undirected graph model $H = (B, F)$. Vertex B corresponds to multiple joint points of the human body. For each pair of interconnected human joint points b_i and b_j , there exists edge $(b_i, b_j) \in F$. The loss function expression of the joint point edge structure is shown in equation (8).

$$\mathfrak{S}_Z = \frac{1}{\mathfrak{N}} \sum_{r=1}^{\mathfrak{N}} \|\phi_{\zeta_{ij}} - H_{\zeta_{ij}}\|_2 \quad (8)$$

In equation (8), \mathfrak{S}_Z represents the loss function value of the joint point edge structure. \mathfrak{N} represents the actual number of joint edge structures in the human body. $\phi_{\zeta_{ij}}$ represents the predicted value of the human joint point edge structure output from the training model. $H_{\zeta_{ij}}$ represents the true value of the edge structure between the corresponding joint points. ζ_{ij} represents the difference between the human joint points b_i and b_j . Finally, the penalty loss function in the objective function is constrained using hyperparameters to obtain the final improved objective function, as expressed in equation (9).

$$\mathfrak{S} = \mathfrak{S}_y + \alpha \mathfrak{S}_\eta + \beta \mathfrak{S}_Z \quad (9)$$

In equation (9), \mathfrak{S} represents the improved objective function value. α and β represent hyper-parameters.

B. A 2D REGRESSION NETWORK ALGORITHM CONSIDERING LOCAL JOINT OCCLUSION

For visible joint points, local action constraints can better constrain adjacent joint points by learning structural information. However, under the influence of occlusion interference between human joint points, it is difficult to determine the position of other joint points through local joint points. Therefore, this study introduces a relational network to improve the two-dimensional human motion estimation algorithm proposed using local constraints. Relationship networks achieve object recognition by extracting

relationships between objects. It interacts by introducing the appearance and geometric relationships between different objects, thereby achieving modeling of the relationships between objects. Due to the limited number of relationship parameters, it can be used to improve the performance of object detection and achieve a complete end-to-end object detector. The structure of the target relationship network is shown in Figure 4.

In Figure 4, the relational network obtains the relationship between objects to identify them. By introducing the appearance and geometric relationship between different objects, the modeling of the relationship between objects is realized. There are few relationship parameters, which can be used to improve the performance of target detection and achieve a complete end-to-end target detector. The actual input of the target relationship network includes the appearance features and set features of the target. When multiple targets are given, the calculation expression of the geometric features related to the m -th target and all targets is shown in equation (10).

$$U_R(v) = \sum_{\vartheta} \varpi_{\vartheta v} \cdot (W_s \cdot U_A^{\vartheta}) \quad (10)$$

In equation (10), $U_R(v)$ represents the output of the relationship module. ϑ represents the serial number of the target. $\varpi_{\vartheta v}$ represents the weight coefficient matrix. W_s represents a linear transformation. U_A^{ϑ} represents the appearance feature of the m -th target. The output of the relationship module is determined by multiplying the appearance features of all objects by a dimension of W_s and the weight coefficient matrix actually learned in the network to determine the actual impact of other targets on the current target as the output of the module. Finally, the actual obtained relationship features and original features are fused using the Concat function, as shown in equation (11).

$$U_A^{\vartheta} = U_A^{\vartheta} + \text{Concat} \left[U_R^1(n), \dots, U_R^N(n) \right] \quad (11)$$

In Equation (11), $U_R^1(n), \dots, U_R^N(n)$ represents the geometric correlation features of the target 1 to N , respectively. On this basis, this study embeds this relational network into the detection architecture of the regression network in Figure 2 to improve performance, as shown in Figure 5.

From Figure 5, the overall architecture of the network is mainly a framework for human motion estimation based on regression networks. The network input is a $256 * 256 * 3$ size image. The feature extraction network formed by the Stem network has a convolutional kernel size of $3 * 3$, resulting in a feature map with a size of $64 * 64$ and a number of 576. The feature extraction network is based on the structure of the Inception v4 network, consisting of convolutional layer, pooling layer, and batch normalization layer. By modifying the extracted feature maps, the feature maps of each joint point can be obtained, resulting in a feature map of $n * 32 * 32$. The result of Block 2 layer is a feature map of 16 joint points. Use the feature maps of 16 joint points to input into the relationship network construction module in the

regression network, so that different relationship features can be combined. At the same time, merge it with the original feature information of the joint point to form the final feature of the joint point. The actual input is a $32 * 32 * 16$ feature map, which serves as an input for the subsequent network. Then, the feature values of each joint point are input into the association network and weighted for fusion to obtain the final feature values. At the same time, multiple modifications are made to the network to obtain the coordinates of the connecting nodes. The improved relational network is mainly used to merge the feature modification layer through the relationship module after the feature modification layer in the regression network. The input of the relationship module is the appearance feature of the n -th joint point, that is, the object size, color and other characteristics. And the appearance features of the n -th joint points, representing the position of the object, there are multiple relational modules, compared to the neural network, we have many different channels on each layer to learn the features between different joint points, and finally the output of the module is the fusion of the current target and the appearance features of each target. Each relational module is input with the two features of all the joint points, obtaining the combination of different relational features, and fuse with the original feature information of the joint points as the final feature of the joint points.

The improved relational network mainly uses the feature modification layer in the regression network, and the feature map obtained from the feature modification layer is obtained through the relational module to obtain the fused features. The input of the relationship module is the appearance feature of the n -th joint point. In fact, it refers to the external characteristics of an object such as its size, color, shape, etc. The final result is the fusion of the apparent features of existing objects and individual objects. Each association module takes two characteristics of all joint points as input, combines different association characteristics, and then fuses them with the original feature information of the joint points to become the final feature of the joint points. Based on this, the final improved two-dimensional human motion estimation algorithm flow is Figure 6.

From Figure 6, the algorithm first inputs a $32 * 32 * 16$ feature map in the relational network to modify the network layer's output. Secondly, the obtained 16 feature maps are represented as $\{U_A^n, U_D^n\}_{n=1}^{\mathfrak{N}_j}$ (U_A^n and U_D^n represent the image and position features of the n -th joint point, while \mathfrak{N}_j has a value of 16). Next, use other joint points to calculate the relationship features of the current n -th joint point, and then calculate the relationship weights. Finally, the collected 15 relational features are aggregated together to obtain an enhanced feature map. The calculation of relationship features is equation (12).

$$U_R(n) = \sum_m \varpi^{mn} \cdot U_A^m \quad (12)$$

In equation (12), ϖ^{mn} represents the learned relationship weight. The image features of the m -th joint point of U_A^m .

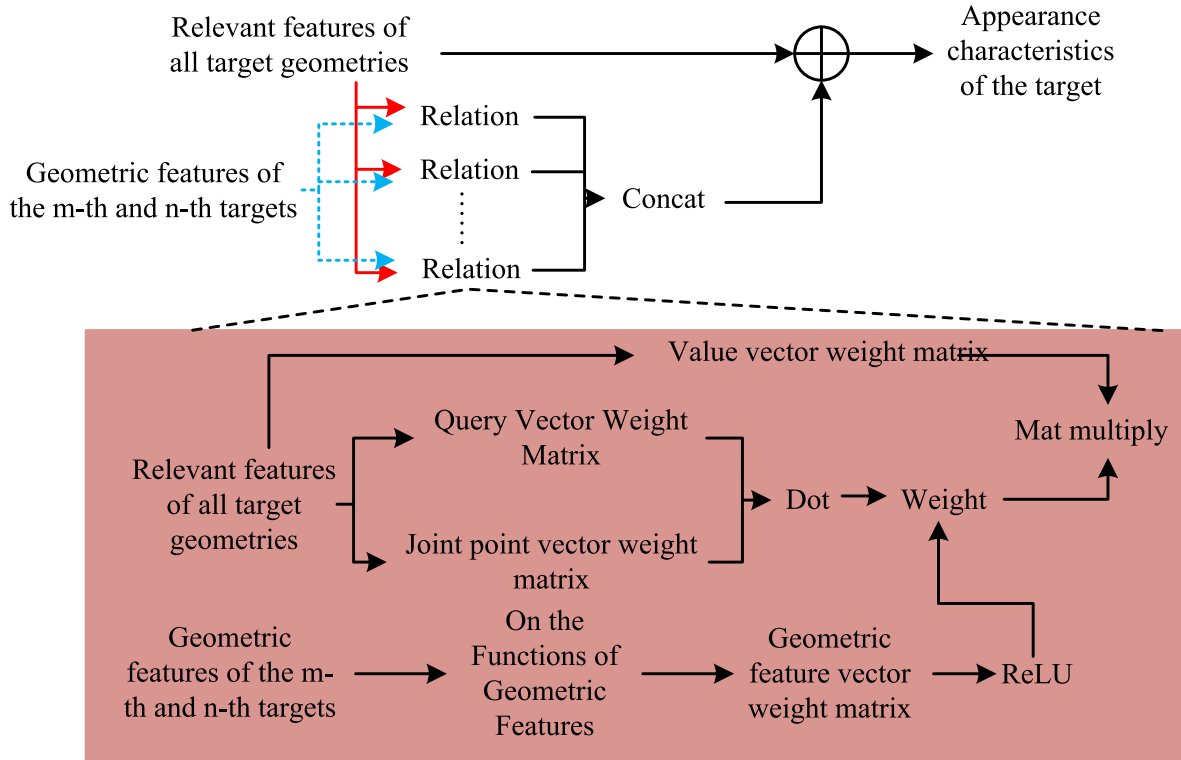


FIGURE 4. Schematic diagram of the structure of the target relationship network.

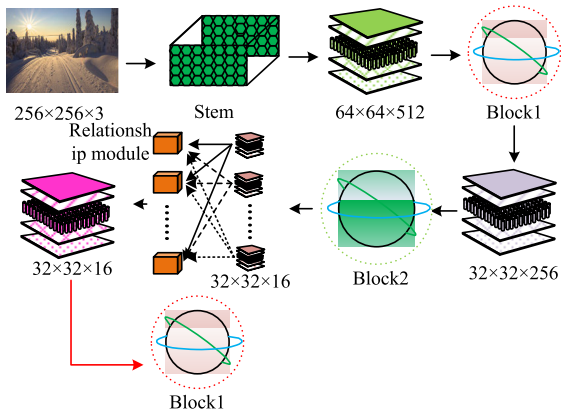


FIGURE 5. Schematic diagram of improved network structure.

Since the weight of the m -th joint point on the n -th joint point is determined by the image feature weight and the position feature weight, the image feature weight produced by the m -th joint point to the n -th joint point is determined by the image features of the m -th joint point and the n -th joint point. Therefore, the corresponding expression of relationship weights is shown in equation (13).

$$\begin{cases} \omega^{mn} = \frac{\omega_D^{mn} \cdot \exp(\omega_A^{mn})}{\sum_o \omega_D^{on} \cdot \exp(\omega_A^{on})} \\ \omega_A^{mn} = \frac{\text{dot}(W_L U_A^m, W_D U_A^n)}{\sqrt{d_k}} \end{cases} \quad (13)$$

In equation (13), ω_A^{mn} represents the weight of image features. ω_D^{mn} represents the weight value of the position feature. o represents the joint number that is not equal to m . W_L and W_D represent the joint vector weight matrix and the query vector weight matrix. d represents the dimension. dot represents a point. The image feature weight generated by the m -th joint point to the n -th joint point in formula (12) is determined by the image features of the m -th and n -th joint points, as shown in equation (14).

$$\omega_D^{mn} = \max(0, W_D \cdot \sigma_D(U_D^m, U_D^n)) \quad (14)$$

In equation (14), σ_D represents a function related to the query vector, which is used to map the positional features between the ϑ -th and ν -th targets into a high-dimensional space. At this point, W_D converts the obtained response features into scalars, which act as the nonlinearity of a Rectified Linear Unit (ReLU). Finally, the calculation expression of the aggregation operation of relational features is shown in equation (15).

$$U_A^n = U_A^n + (U_R^1(n) + \dots + U_R^{15}(n)) \quad (15)$$

IV. PERFORMANCE ANALYSIS OF IMPROVED TWO-DIMENSIONAL REGRESSION NETWORK ALGORITHM

Experimental analysis is needed to address the performance analysis of improved two-dimensional regression network algorithms. Therefore, this section mainly uses theoretical

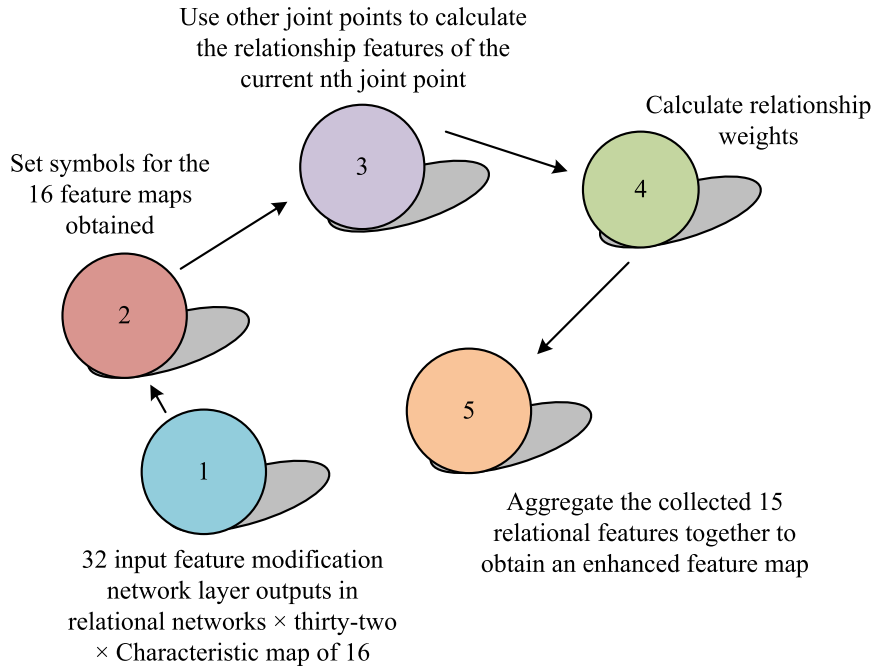


FIGURE 6. Schematic diagram of two-dimensional human pose estimation algorithm flow based on improved regression network and relational network.

experiments and practical application experiments to analyze the algorithm performance.

A. PERFORMANCE ANALYSIS OF TWO-DIMENSIONAL REGRESSION NETWORK HUMAN MOTION RECOGNITION ALGORITHM BASED ON LOCAL CONSTRAINTS

In order to verify the performance of the two-dimensional human motion recognition and estimation algorithm further optimized after the introduction of relational networks, this study first validated the two-dimensional regression network algorithm using local action constraints. Secondly, the improved research algorithm based on it was validated. All experiments were run in keras network, and network optimization was performed on 1080 GPU using Root Mean Square Prop optimizer with a batch size of 12. The first experimental algorithm was set to A, and the second experimental algorithm was set to B. Table 1 shows the initial experimental parameter settings for the two experiments.

In Table 1, the number of iterations for both experiments is set to 120, the batch size is 12, and the initial learning rate is 0.001. In addition, in the first experiment, the input image size was $256 * 256 * 3$, and the experimental environment was set in the deep learning framework (Keras). In the second experiment, both hyper-parameters α and β are set to 0.01. Under this parameter setting, the dataset is selected as the Measurement Point II (MPII) dataset, and the evaluation criteria are selected as the Percentage of Correct Key Points Head (PCKh) to evaluate the accuracy of the model. The threshold is set to 0.5. The action recognition algorithm (C) based on a multitasking framework was introduced in the

TABLE 1. Initial experimental parameter settings for two experiments.

Model parameters of two-dimensional regression network algorithm based on local constraints				
Image size	Experimental environment	Optimizer	General Processing Unit	Batch size
$256 \times 256 \times 3$	Keras	Root Mean Square Prop	1080	12
Iterations	Initial learning rate	Learning rate after 80 iterations	Learning rate after 100 iterations	Learning rate after 120 iterations
120	0.001	20 times decrease from 0.001	400 times decrease from 0.001	8000 times decrease from 0.001
Model parameters of two-dimensional regression network algorithm based on local constraints and relational networks				
α	0.01	β	0.01	
Iterations	120	Batch size	12	
Initial learning rate	0.001	Learning rate after 80 iterations	20 times decrease from 0.001	
Learning rate after 100 iterations	400 times decrease from 0.001	Learning rate after 120 iterations	8000 times decrease from 0.001	

experiment to compare it with the research algorithm A. The joint points were selected as head, shoulder, elbow, wrist, buttocks, knee, and ankle, and the performance indicators of different hyper-parameters were verified. The results are shown in Figure 7.

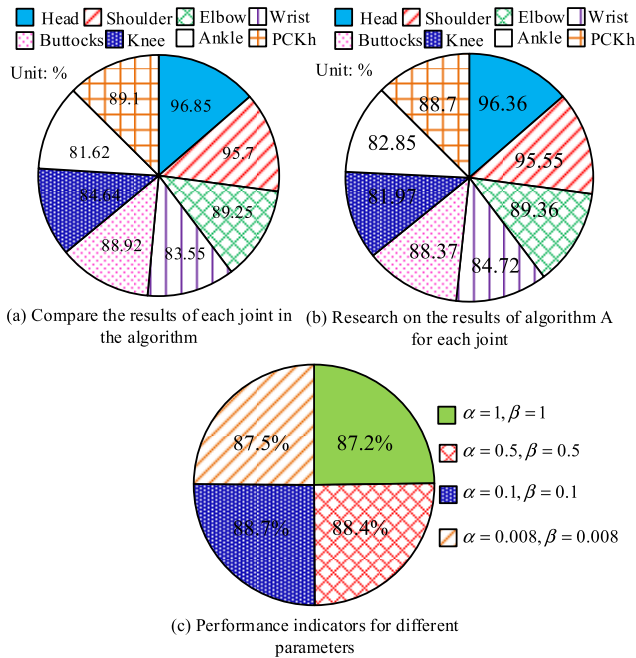


FIGURE 7. Comparison results of different algorithms and performance indicators of different parameters.

Based on Figure 7, the numerical values of the research algorithm in the elbow, wrist, and ankle are 89.36%, 84.72%, and 82.85%, respectively, which are higher than those of the comparative algorithms. However, it is slightly lower than the comparison algorithm in the head, shoulders, buttocks, and knees. In addition, when both α and β values are 1, the PCKh value is 87.2%. When both α and β values are 0.05, the PCKh value is 88.4%. When both α and β values are 0.01, the PCKh value is 88.7%. When both α and β values are 0.008, the PCKh value is 87.5%. Overall, the research algorithm A outperforms the comparison method in only some joints on the MPII dataset, and has a certain improvement effect. Simultaneously studying the algorithm, the PCKh value is the highest and the parameter settings are optimal when both α and β values are 0.01. Based on this, in order to identify gaps more clearly, this study separately extracted PCKh images of the wrist and ankle for analysis. The estimated joint points of different algorithms and the PCKh results of different network module numbers are shown in Figure 8.

Based on Figure 8, there is no significant difference in the numerical values between the two algorithms when the standard distance is between 0 and 0.15 in the comparison of wrist joint points. After exceeding 0.15, the research algorithm A gradually surpasses the comparison algorithm. Overall, the PCKh score of Algorithm A is 84.72%, while the comparison algorithm is 84.55%, an increase of 1.17%. In the comparison of ankle joint points, when the standard distance is between 0 and 0.13, the growth amplitude of the two algorithms remains basically the same. But after exceeding 0.14, the growth rate of algorithm A in the study was greater than that of the comparison algorithm. Overall, the PCKh score value

of Algorithm A studied is 82.85%, which is 1.23% higher than that of the comparative algorithm. In addition, when the number of regression network modules is 8, the PCKh result value is the highest, at 88.7%. Overall, research algorithm A has certain advantages compared to comparative algorithms, and the depth of the regression network will have an impact on the loss function. Its optimal module count is 8. Based on this result, this study applies it to actual indoor and outdoor human motion recognition to verify the performance of the research algorithm in visualizing the running results. The recognition effect of indoor human movements is shown in Figure 9.

Figure 9 shows the physical image in an indoor environment. The research algorithm A is basically consistent with the joint point detection of the real label of the image to be detected, while there is some error in the comparison algorithm. Overall, research algorithms in indoor non-interference environments have high performance in human motion recognition. The recognition effect of outdoor human movements is shown in Figure 10.

Figure 10 is a physical image of the outdoor environment without obstruction. The actual detection image of Algorithm A studied in an outdoor environment is not significantly different from its real label, and it remains basically consistent, while the error of the comparison algorithm is significantly greater. Overall, the algorithm has high performance in both indoor and outdoor environments, indicating its effectiveness. However, there were errors in the research algorithm when there was obstruction indoors, so improvements were made to it.

B. PERFORMANCE ANALYSIS OF IMPROVED TWO-DIMENSIONAL HUMAN MOTION RECOGNITION ALGORITHM

In order to verify the performance of Algorithm B, which is improved by introducing a relational network based on Algorithm A, this study conducted experimental analysis on it. The experimental results of different algorithms on the MPII dataset are compared as shown in Figure 11.

D in Figure 11 represents the relational network algorithm. The values of Algorithm B in head, shoulder, elbow, wrist, buttocks, knee, and ankle are 98.75%, 97.52%, 89.88%, 85.12%, 89.21%, 85.12%, and 83.93%, respectively. At the same time, the overall PCKh value is 90.1%. Both comparison values are higher than the comparison algorithm. Overall, the improved algorithm improves the wrist and ankle joint points that are easily occluded in the dataset by 0.4% -1.57% and 1.08% -2.31% compared to the comparative algorithm, indicating the effectiveness and high performance of the research algorithm. The overall changes in PCKh of the elbow, ankle, and knee joints, as well as the PCKh results for different network module numbers, are shown in Figure 12.

Based on Figure 12, in elbow joint recognition, the research algorithm began to significantly outperform the comparison algorithm after the standard distance exceeded 0.05. In wrist joint recognition, the growth rate of the research algorithm

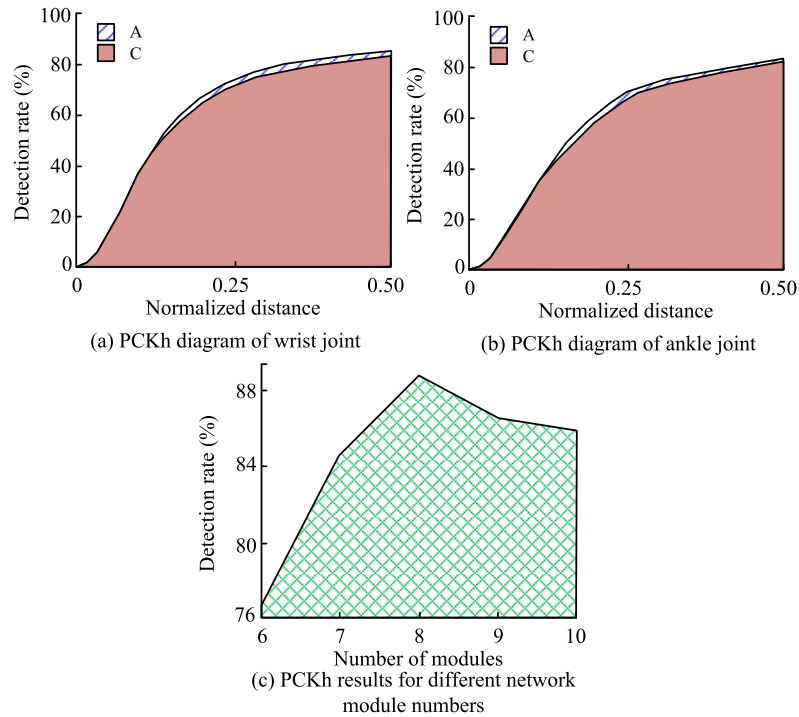


FIGURE 8. Joint points of estimation results of different algorithms and the Percentage of Correct Key Points Head results of different network module numbers.

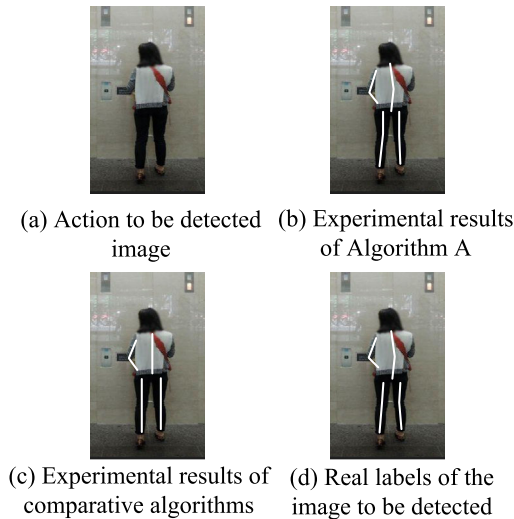


FIGURE 9. Indoor human motion recognition effect.

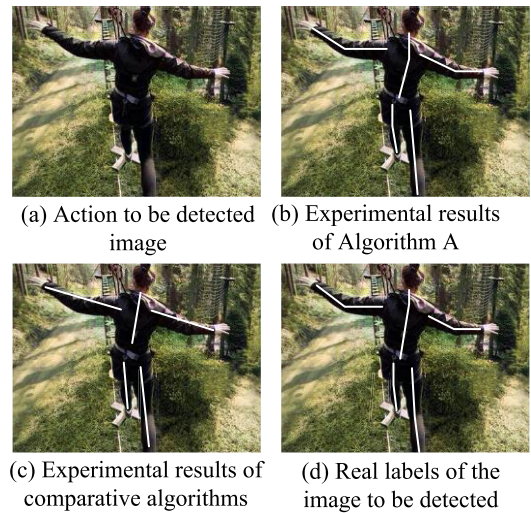


FIGURE 10. Outdoor human motion recognition effect.

starts to be higher than that of the comparison algorithm after the standard distance exceeds 0.15. In knee joint recognition, the growth rate of research algorithms is significantly higher than that of comparative algorithms. In addition, when the number of regression network modules is also 8, the PCKh result value is the highest, at 89.1%. Overall, the improved algorithm has better performance, indicating the effectiveness of the improvement. To further validate the effectiveness of the improved algorithm, this study selected additional indi-

cators for comparison based on the PCKh indicator, namely PCKh@0.5 Score and PCK@0.2 Score. At the same time, convolutional pose machine (E), local heat map regression (F), deep cutting (G), human pose estimation using contextual information (H), and learning feature pyramid algorithm (I) for human pose estimation are introduced and compared with the research algorithm. Table 2 shows the performance of different algorithms on the MPII dataset PCKh@0.5 Result.

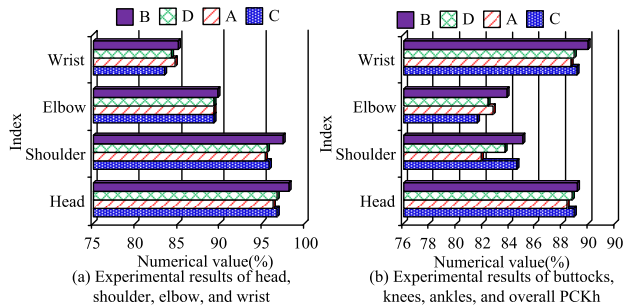


FIGURE 11. Comparison of experimental results of different algorithms on Measurement Point II datasets.

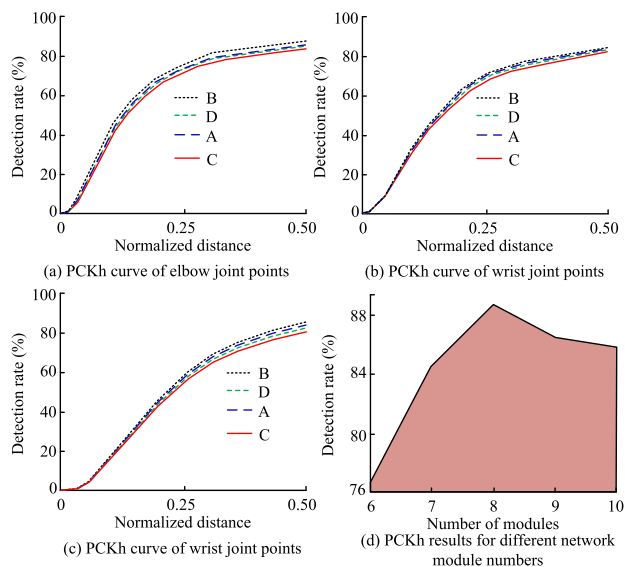


FIGURE 12. Overall changes in Percentage of Correct Key Points Head at elbow, ankle, and knee joint points, as well as Percentage of Correct Key Points Head results for different network module numbers.

TABLE 2. Different algorithms on the MPII dataset Percentage of Correct Key Points@0.5 result.

	Head	Shoulde r	Elbo w	Wrist	Buttock s	Knee	Ankle
E	97.7 %	94.9%	88.6%	83.9 %	88.3%	82.7 %	79.3 %
F	97.8 %	95.0%	89.8%	85.2 %	89.3%	85.6 %	81.6 %
G	96.7 %	95.1%	89.2%	84.3 %	88.3%	83.3 %	77.9 %
H	98.5 %	96.9%	92.7%	88.7 %	91.6%	89.7 %	86.5 %
I	98.4 %	96.6%	92.4%	88.6 %	91.0%	88.5 %	85.9 %
B	99.5 %	97.6%	93.4%	89.1 %	92.7%	90.1 %	88.5 %

In Table 2, the improved algorithm was studied at 7 joint points PCKh@0.5 The scores are 99.5%, 97.6%, 93.4%, 89.1%, 92.7%, 90.1%, and 88.5% respectively, which are higher than the comparison algorithm. Overall, they are 92.5%, significantly higher than the comparison algorithm and only 0.2% lower than the current state-of-the-art method.

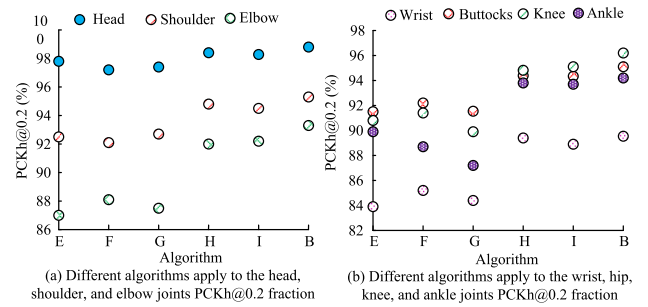


FIGURE 13. Different algorithms on the Measurement Point II dataset Percentage of Correct Key Points@0.2 result.

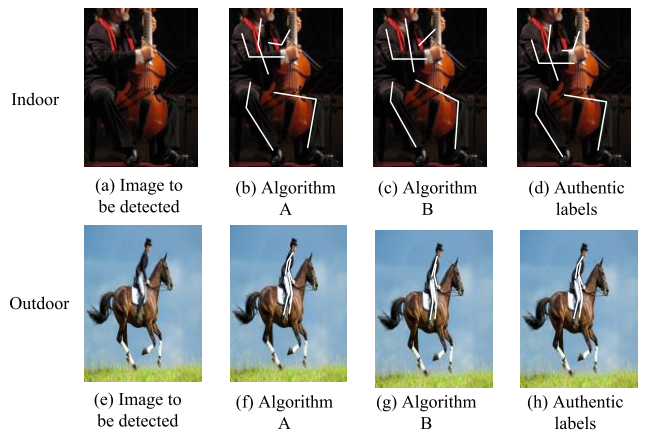


FIGURE 14. Experimental results of improved algorithms in real human indoor and outdoor motion recognition.

Overall, the improved algorithm has high performance in human motion recognition. Different algorithms on the MPII dataset PCKh@0.2 The results are shown in Figure 13.

Based on Figure 13, the improved algorithm is applied at 7 joint points PCK@0.2 The scores were 98.8%, 95.3%, 93.3%, 89.4%, 95.1%, 96.2%, and 94.2% respectively, all higher than the comparison algorithm. Overall, it is 94.3%, significantly higher than the comparison algorithm and only 0.2% lower than the current state-of-the-art method. Overall, the research algorithm has high performance in human motion recognition, surpassing the effectiveness of most algorithm models, fully utilizing the information of local constraints and the connections of relationship networks, thereby assisting in human motion recognition. In order to verify its practical application effect, this study also applies it to indoor and outdoor human motion recognition, and the comparative algorithm is Algorithm A. The results are shown in Figure 14.

In Figure 14, in an obstructed indoor or outdoor environment, the actual detection image of the improved algorithm does not differ significantly from its actual label, while the error of the comparison algorithm is significantly greater. Overall, the algorithm has high performance in both indoor and outdoor environments, indicating its practicality. Overall, this study proposes a two-dimensional regression network algorithm model for human motion recognition using local

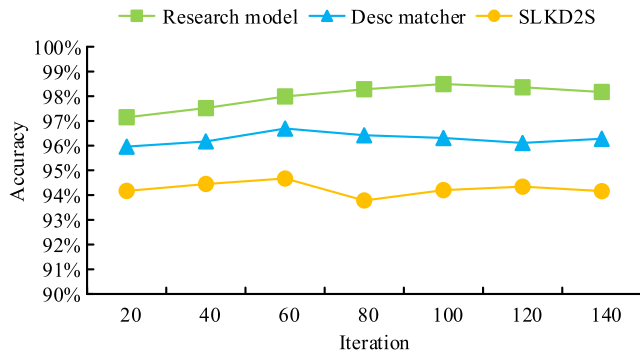


FIGURE 15. The pose estimation accuracy of the research model, Descriptor Matchers and Single branch lightweight knowledge extraction 2S algorithm.

TABLE 3. Summary of the experimental processing.

Joint	Two-dimensional regression network based on local constraints	Improved two-dimensional human motion recognition algorithm
Head	96.36%	98.75%
Shoulder	95.55%	97.52%
Elbow	89.36%	89.88%
Wrist	84.72%	85.12%
Buttocks	88.37%	89.21%
Knee	81.97%	85.12%
Ankle	82.85%	83.93%

constraints and relational networks, which has high performance. The pose estimation accuracy of the research model, Descriptor Matchers and Single branch lightweight knowledge extraction 2S (SLKD2S) algorithm is shown in Figure 15.

Figure 15 shows that the highest accuracy of the research model was 98.49% and the lowest was 97.15%, with an average accuracy of 97.99%. However, the highest accuracy of Desc Matcher and SLKD2S did not exceed 96.7%, with an average accuracy of 96.28% and 94.25%, respectively. The pose estimation accuracy of the visible research model is better than the other models. It can be seen from the test results of the research model that the accuracy of human pose estimation is effectively improved, and the detection of the position of human joints in the image can be used for the subsequent auxiliary understanding of the image content. The test results of the local constraint-based 2-D regression networks and their improved 2-D human motion recognition algorithm are summarized in Table 3.

V. CONCLUSION

The current two-dimensional regression network algorithm model for human motion recognition and estimation does not consider the interrelationships between human joint points, resulting in missing connections between joint points and low accuracy of feature maps. In response to this issue,

this study proposed an improved two-dimensional regression network algorithm based on local constraints, and further improved it by introducing a relational network considering occlusion conditions. At the same time, experimental analysis was conducted on it. The experiment showed that in algorithm A validation, the values of the elbow, wrist, and ankle were 89.36%, 84.72%, and 82.85%, respectively, which were higher than those of the comparative algorithms. Overall, the PCKh score of Algorithm A was 84.72%, while the comparison algorithm was 84.55%, an increase of 1.17%. In practical applications, the actual label results of indoor and outdoor environments were basically consistent with those of the detected image, but there were errors under indoor occlusion conditions. In algorithm B validation, its PCKh scores in head, shoulder, elbow, wrist, buttocks, knee, and ankle were 98.75%, 97.52%, 89.88%, 85.12%, 89.21%, 85.12%, and 83.93%, respectively, which were higher than the comparison algorithms. The PCKh@0.5 scores were 99.5%, 97.6%, 93.4%, 89.1%, 92.7%, 90.1%, and 88.5% respectively, which were also higher than the comparison algorithm. It could effectively recognize human movements under occlusion conditions in practical applications. Overall, the improved two-dimensional regression network algorithm model proposed in this study based on local constraints and relational networks has high effectiveness and practicality. However, when the occlusion area is large, the algorithm struggles to estimate the joint position correctly. Therefore, future research will focus on how to accurately estimate the position of joints when the occlusion area is large, so as to facilitate the development of machine vision field and enrich the human-computer interaction experience of virtual reality, games and others.

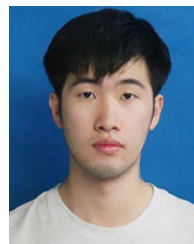
REFERENCES

- [1] S. Oslund, C. Washington, A. So, T. Chen, and H. Ji, "Multiview robust adversarial stickers for arbitrary objects in the physical world," *J. Comput. Cognit. Eng.*, vol. 1, no. 4, pp. 152–158, Sep. 2022, doi: [10.47852/bonviewjcc2202322](https://doi.org/10.47852/bonviewjcc2202322).
- [2] C. Yang, X. Wang, and S. Mao, "RFID-based 3D human pose tracking: A subject generalization approach," *Digit. Commun. Netw.*, vol. 8, no. 3, pp. 278–288, Jun. 2022, doi: [10.1016/j.dcan.2021.09.002](https://doi.org/10.1016/j.dcan.2021.09.002).
- [3] F. A. Dharejo, M. Zawish, Y. Zhou, S. Davy, K. Dev, S. A. Khowaja, Y. Fu, and N. M. F. Qureshi, "FuzzyAct: A fuzzy-based framework for temporal activity recognition in IoT applications using RNN and 3D-DWT," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 11, pp. 4578–4592, Nov. 2022, doi: [10.1109/TFUZZ.2022.3152106](https://doi.org/10.1109/TFUZZ.2022.3152106).
- [4] A. Gosztolai, S. Günel, V. Lobato-Ríos, M. Pietro Abrate, D. Morales, H. Rhodin, P. Fua, and P. Ramdya, "LiftPose3D, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals," *Nature Methods*, vol. 18, no. 8, pp. 975–981, Aug. 2021, doi: [10.1038/s41592-021-01226-z](https://doi.org/10.1038/s41592-021-01226-z).
- [5] G. Bhavanasi, L. Werthen-Brabant, T. Dhaene, and I. Couckuyt, "Patient activity recognition using radar sensors and machine learning," *Neural Comput. Appl.*, vol. 34, no. 18, pp. 16033–16048, Sep. 2022, doi: [10.1007/s00521-022-07229-x](https://doi.org/10.1007/s00521-022-07229-x).
- [6] S. Li, J. Yi, Y. A. Farha, and J. Gall, "Pose refinement graph convolutional network for skeleton-based action recognition," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1028–1035, Apr. 2021, doi: [10.1109/LRA.2021.3056361](https://doi.org/10.1109/LRA.2021.3056361).
- [7] T. Chen, C. Fang, X. Shen, Y. Zhu, Z. Chen, and J. Luo, "Anatomy-aware 3D human pose estimation with bone-based pose decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 198–209, Jan. 2022, doi: [10.1109/TCSVT.2021.3057267](https://doi.org/10.1109/TCSVT.2021.3057267).

- [8] S. Dubey and M. Dixit, "A comprehensive survey on human pose estimation approaches," *Multimedia Syst.*, vol. 29, no. 1, pp. 167–195, Feb. 2023, doi: [10.1007/s00530-022-00980-0](https://doi.org/10.1007/s00530-022-00980-0).
- [9] D. K. Vishwakarma and K. Jain, "Three-dimensional human activity recognition by forming a movement polygon using posture skeletal data from depth sensor," *ETRI J.*, vol. 44, no. 2, pp. 286–299, Jan. 2022, doi: [10.4218/etrij.2020-0101](https://doi.org/10.4218/etrij.2020-0101).
- [10] S. Liu, N. He, C. Wang, H. Yu, and W. Han, "Lightweight human pose estimation algorithm based on polarized self-attention," *Multimedia Syst.*, vol. 29, no. 1, pp. 197–210, Feb. 2023, doi: [10.1007/s00530-022-00981-z](https://doi.org/10.1007/s00530-022-00981-z).
- [11] H. Arab, I. Ghaffari, L. Chioukh, S. O. Tatu, and S. Dufour, "A convolutional neural network for human motion recognition and classification using a millimeter-wave Doppler radar," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4494–4502, Mar. 2022, doi: [10.1109/JSEN.2022.3140787](https://doi.org/10.1109/JSEN.2022.3140787).
- [12] Q. Gao, Y. Chen, Z. Ju, and Y. Liang, "Dynamic hand gesture recognition based on 3D hand pose estimation for human–robot interaction," *IEEE Sensors J.*, vol. 22, no. 18, pp. 17421–17430, Sep. 2022, doi: [10.1109/JSEN.2021.3059685](https://doi.org/10.1109/JSEN.2021.3059685).
- [13] J. Dong, Q. Fang, W. Jiang, Y. Yang, Q. Huang, H. Bao, and X. Zhou, "Fast and robust multi-person 3D pose estimation and tracking from multiple views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6981–6992, Oct. 2022, doi: [10.1109/TPAMI.2021.3098052](https://doi.org/10.1109/TPAMI.2021.3098052).
- [14] H. Liu, T. Liu, Z. Zhang, A. K. Sangaiah, B. Yang, and Y. Li, "ARHPE: Asymmetric relation-aware representation learning for head pose estimation in industrial human–computer interaction," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 7107–7117, Oct. 2022, doi: [10.1109/THI.2022.3143605](https://doi.org/10.1109/THI.2022.3143605).
- [15] T. Bao, S. Q. Xie, P. Yang, P. Zhou, and Z.-Q. Zhang, "Toward robust, adaptive and reliable upper-limb motion estimation using machine learning and deep learning—A survey in myoelectric control," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 8, pp. 3822–3835, Aug. 2022, doi: [10.1109/JBHI.2022.3159792](https://doi.org/10.1109/JBHI.2022.3159792).
- [16] Z. Ni, T. Wu, T. Wang, F. Sun, and Y. Li, "Deep multi-branch two-stage regression network for accurate energy expenditure estimation with ECG and IMU data," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 10, pp. 3224–3233, Oct. 2022, doi: [10.1109/TBME.2022.3163429](https://doi.org/10.1109/TBME.2022.3163429).
- [17] M. Nowak, I. Vujaklija, A. Sturma, C. Castellini, and D. Farina, "Simultaneous and proportional real-time myocontrol of up to three degrees of freedom of the wrist and hand," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 2, pp. 459–469, Feb. 2023, doi: [10.1109/TBME.2022.3194104](https://doi.org/10.1109/TBME.2022.3194104).
- [18] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5137–5146.
- [19] W. Zhang, Y. Chen, W. Yang, G. Wang, J.-H. Xue, and Q. Liao, "Class-variant margin normalized softmax loss for deep face recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4742–4747, Oct. 2021, doi: [10.1109/TNNLS.2020.3017528](https://doi.org/10.1109/TNNLS.2020.3017528).
- [20] Y. Kim, Y. Lee, and M. Jeon, "Imbalanced image classification with complement cross entropy," *Pattern Recognit. Lett.*, vol. 151, pp. 33–40, Nov. 2021, doi: [10.1016/j.patrec.2021.07.017](https://doi.org/10.1016/j.patrec.2021.07.017).



LIJUN WANG was born in Taonan, Jilin, Han Nationality, in June 1970. He received the master's degree in information management in management science and engineering from the Changchun University of Technology, in 2005. He is currently a Professor with the School of Applied Technology, Changchun University of Technology. In 2021, he led and completed the Changchun Science and Technology Plan Project "Public Health Management Platform." His research interests include human motion recognition and data mining.



ZIXU WANG was born in Taonan, Jilin, Han Nationality, in November 1997. He received the bachelor's degree in mathematics and applied mathematics from Northeast Electric Power University, in 2020. He is currently pursuing the master's degree with the School of Computer Science and Engineering, Changchun University of Technology, with a focus on human motion recognition.



LIJUAN ZHOU was born in Baishan, Jilin, Han Nationality, in July 1971. She received the master's degree in computer application technology from Tianjin University, in 2006. She is currently a Professor with the School of Journalism and Communication, Changchun University of Technology. In 2022, she was responsible for the Research Project of Jilin Provincial Department of Education titled "Research on Automatic Import Technology of Finished Parts in Automotive White Body Welding Production Line." Her research interests include machine learning and deep learning.

...