Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 5, pp. 3394–3407, DOI 10.12694/scpe.v25i5.3190

FEATURE EXTRACTION OF GYMNASTICS IMAGES BASED ON MULTI-SCALE FEATURE FUSION ALGORITHM

KUN TIAN* AND QIONGHUA XIA†

Abstract. The feature extraction and analysis of gymnastics images is an essential foundation for estimating human posture. The primary step is to obtain various joint points of athletes based on basic information such as human texture and contour in the sports images. The reconstruction and analysis of the human skeleton is completed based on the feature data of the joint points. In this process, traditional algorithm models often have certain shortcomings in the accuracy of feature extraction for motion images. This paper combines multi-scale feature fusion algorithms to construct a gymnastics motion image feature extraction model, which can achieve more accurate and efficient analysis and research for the feature extraction process of motion images, further improving the detection accuracy of joint points in motion images; this lays an essential foundation for feature extraction of gymnastics images. At the same time, it also provides more methods for skeleton reconstruction based on image feature information during the motion process, improving the efficiency and accuracy of reconstruction.

Key words: Multi scale feature fusion; gymnastics; feature extraction; loss function

1. Introduction. With the continuous development and improvement of scientific and technological informatization, modern technology can support humans to obtain a large amount of image and video information from various channels. A common image processing technology is a technology that uses computers to process and analyse collected images and video information to meet human needs, covering various aspects of human clothing, food, housing, and transportation. The extraction of human skeleton features in images has always been an important research field in image processing and computer vision, and research in this field continues to drive the continuous updates and progress of modern audio and video technology [1-3]. The research and analysis of skeletal information during human motion lays an important foundation for further processing of images and videos, and can also help people analyse the behaviour and key actions of the target human body. For gymnasts, analysing and researching motion images is an extremely important process to improve and enhance their key movements. Therefore, using modern computer technology to extract and analyse the features of gymnastics images provides more means for this process. The human skeleton extraction algorithm divides the human skeleton into multiple joint points, such as the head, buttocks, shoulders, wrists, etc. Then, by analysing the position, direction, and motion of each joint point, the human skeleton information is obtained [4-5]. Further analysis of human posture and behaviour is carried out through the drawn human skeleton, in order to obtain the activity and motion information of the human body in the image.

The extraction of human skeletons in gymnastics images can be divided into two dimensions: two-dimensional human skeleton extraction and three-dimensional human skeleton extraction. 3D human skeleton extraction analyzes images obtained from 3D cameras such as Kinect to obtain the 3D shape or coordinates of human skeleton points. 2D human skeleton extraction mainly analyses 2D images obtained by ordinary image acquisition devices and obtains the 2D coordinates of human skeleton points. Compared to 2D skeleton extraction, 3D skeleton points are denser, modeling is more complex, and 3D cameras are more expensive [6-7]. In most cases, two-dimensional images are obtained for the analysis of gymnastics images. Therefore, this article mainly focuses on the feature extraction of two-dimensional gymnastics images.

The applications related to human pose estimation are all based on obtaining clear and accurate human skeletons in motion images. If the accuracy of feature extraction in motion images is insufficient, it will lead to

*School of Sports and Art, Guangzhou Institute of Physical Education, Guangzhou 510500, China
†Department of physical education, Guangdong University of Foreign Studies, Guangzhou 510420, China (Corresponding author, Qionghua_Xia23@outlook.com)

significant deviations in the analysis of human behaviour and movements. Therefore, improving the accuracy of feature detection and extraction in gymnastics images is of great significance. In recent years, the rapid development of the hardware field has led to the continuous enhancement of computer computing power, and more high-performance algorithms for extracting gymnastics motion images are emerging. The accuracy of human skeleton extraction is constantly improving, and the extraction technology, as the foundation of human posture recognition, will play an increasingly important role in more fields [8].

Early human pose estimation methods mainly relied on manually designed features and template matching, but this method was limited by the complexity and robustness of feature design. With the rise of deep learning, methods based on Convolutional Neural Networks (CNN) have gradually taken a dominant position. These methods train networks with a large amount of data to automatically learn the mapping from the original image to pose information, significantly improving the accuracy and robustness of estimation.

In recent years, researchers have begun to focus on more complex scenes and poses, such as multi-person pose estimation, 3D pose estimation, etc. These issues are more challenging as they require handling occlusion, changes in perspective, and changes in scale. To address these issues, researchers have proposed methods such as multi-scale feature fusion, multi-perspective fusion, and spatiotemporal modeling, further improving attitude estimation performance.

Feature extraction is a crucial step in computer vision tasks, aiming to extract helpful information from the original image for subsequent tasks. Selecting feature extraction techniques is crucial for the final performance in human pose estimation. Traditional feature extraction methods mainly rely on manually designed feature descriptors, such as SIFT, HOG, etc. Although these methods perform well in some simple scenarios, they are challenging to cope with complex and ever-changing human postures and environments. With the development of deep learning, CNN-based feature extraction methods have gradually become mainstream. CNN can automatically learn the mapping from the original image to high-level semantic features, effectively extracting helpful information for pose estimation. In addition, researchers have proposed methods such as multi-scale feature fusion and attention mechanisms to improve the accuracy and robustness of feature extraction.

Gymnastics image analysis faces many challenges. Firstly, gymnastics movements are complex and varied, requiring accurate capture and recognition of various subtle movements. Secondly, gymnastics competition scenes often have occlusion and changes in perspective, which puts higher demands on the robustness of image analysis algorithms. In addition, the differences in body shape and movement habits among different athletes also pose challenges to image analysis. In response to these challenges, researchers have begun to explore methods based on deep learning and multi-scale feature fusion. The mapping relationship from the original image to the gymnastics movements is automatically learned by training a deep neural network model. Meanwhile, multi-scale feature fusion technology can extract feature information at different scales to better cope with complex and ever-changing gymnastics movements.

The multi-scale feature fusion algorithm can comprehensively describe the characteristics and patterns of gymnastic movements by fusing feature information from different scales. Multi-scale feature fusion algorithms can extract feature information from different scales, such as athlete joint positions, movement trajectories, muscle morphology, etc., and combine temporal information to analyze actions continuously. In addition, multi-scale feature fusion algorithms can be combined with other advanced technologies, such as attention mechanisms, spatiotemporal modeling, etc., to improve the accuracy and robustness of gymnastics image analysis. For example, by introducing attention mechanisms, it is possible to automatically focus on feature regions that are more critical for recognizing gymnastics movements. Through spatiotemporal modeling techniques, it is possible to capture better and analyze the temporal changes and spatial relationships of gymnastics movements [9-10].

This article combines the current problems and related shortcomings in this field. It combines the unique characteristics of gymnastics images to construct a multi-scale fusion algorithm-based gymnastics image feature extraction model. In response to some joint points that are difficult to detect due to slight occlusion caused by environmental disturbances, multi-person interference, and human posture, this article gradually improves the performance of the constructed network model by enriching feature representation, enhancing the utilization of relevant features, effectively balancing, and fusing features of different scales. At the same time, combined with difficulty mining mechanisms, it improves the detection accuracy of challenging joint points in the feature

extraction process of moving images.

This study aims to develop a gymnastics image feature extraction model based on a multi-scale feature fusion algorithm to address the current accuracy and efficiency challenges in gymnastics image analysis. This model is expected to achieve more precise capture and comprehensive feature expression of gymnast movement details, providing strong technical support for subsequent gymnastics movement recognition, evaluation, and teaching. Athletes' movements are complex and varied in gymnastics, often containing rich spatial and temporal information. Therefore, for feature extraction of gymnastics images, it is necessary to consider feature information at different scales to capture the details and overall structure of athlete movements fully. The multi-scale feature fusion algorithm can effectively combine feature information from different scales, improving the accuracy and robustness of feature extraction. This study will first collect much gymnastics exercise image data and carry out preprocessing and annotation work. Then, design a multi-scale feature fusion algorithm that extracts and fuses image features of different scales to construct a comprehensive and effective feature representation. On this basis, further optimization of the algorithm parameters and structure is needed to improve the accuracy and efficiency of feature extraction.

2. Basic Theory of Multiscale Feature Fusion Algorithms.

2.1. Image Multi-scale Characteristics and Smoothing Models.

2.1.1. Image Multiscale Characteristics. For different scene images, observation can quickly and accurately identify objects of interest from the image, which is a visual characteristic that benefits from the observer's ability to adaptively adjust the distance between the human eye and objects according to different scenes to analyse image content. When the human eye observes an image at close range, even subtle changes in colour in local areas of the image are clearly visible, which is beneficial for the observer to obtain detailed information such as image edges and textures. However, image textures can worsen the consistency of colour distribution in the image area, and form a pseudo edge effect inside the area, thereby affecting the observer's understanding of the main contour of the image object [11-13]. As the distance between the human eye and the image continues to increase, observers can more easily capture the overall overview of the area where the semantic object is located and the contours between different objects. At the same time, the attention to subtle changes caused by the internal texture of the object is gradually decreasing.

In order to overcome the influence of image texture on feature extraction and combine the variation of image visual effect with observation distance, it is more important to establish a multi-scale edge-preserving smoothing model for images, which is used to obtain the smoothing components of the original image at different smoothing scales. As the smoothing scale continues to increase while protecting the edges of the image, the texture details inside the image area are gradually smoothed. By combining multi-scale edges and color distribution features of the image, accurate extraction of foreground objects in natural images is achieved [14]. To learn the global features of an image from multiple cascaded features of different scales, this section proposes an encoded Transformer network, which uses the Transformer to learn the segmented features of images at different scales, thereby obtaining the full-scale feature representation of the image. As shown in Figure 2.1, a schematic diagram of the network structure of the scale encoding Transformer model is provided.

2.1.2. Image Edge Preserving Smoothing Model. For gymnastics images, it can be seen as an organic combination of the object's main structure and texture details. The object's main structure is usually represented as a single or multiple areas with consistent brightness/colour in the image, while the image texture is usually represented as a periodic pixel fluctuation on the object's surface [15]. Therefore, image edge preserving smoothing can be understood as preserving important geometric attributes such as the main contour of the solid object, removing texture details on the surface of the solid object. For an input image $u^0(x, y)$, formula (2.1) is given:

$$u(x, y) = u^0(x, y) - t(x, y) \tag{2.1}$$

Among them, x represents the spatial position of image pixels in the horizontal direction; y represents the spatial position of image pixels in the vertical direction; $T(x, y)$ represents the texture component of image $u^0(x, y)$; The smoothing component $u(x, y)$ is a simplified approximation of the original image, mainly used to extract the overall overview features of each object in the image.

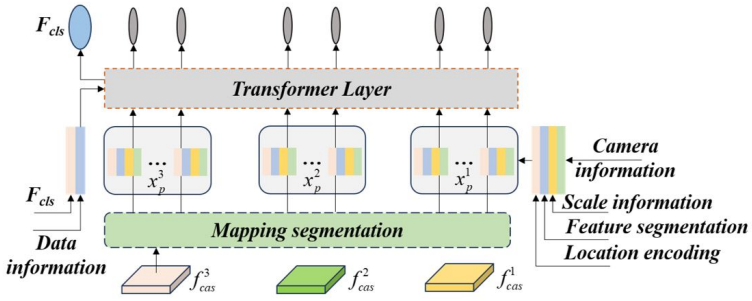


Fig. 2.1: Scale Encoding Transformer Network Architecture

Due to the fact that the smoothing component *u(x, y)* discards the texture information in the original image, its pixel brightness or colour only changes at the edge of the image, while the degree of change inside the image area is relatively weak. Its mathematical representation is similar to a piecewise constant function. In mathematics, gradients are often used to describe the sudden changes in brightness or colour of image pixels [16]. Therefore, in order to measure the approximation degree between the smooth component and the piecewise constant function, a gradient function is used to calculate the overall change in the colour or brightness of *u(x, y)* pixels in the smooth component, as shown in formula (2.2):

$$\Delta L = \int_{\Omega} f(|\nabla u|) \mathrm{d}\Omega \tag{2.2}$$

Among them, Ω represents the domain of image theory; Represents a gradient function, $f(|\nabla u|)$ is also known as a diffusion function in the field of image processing; ∇u represents the gradient information of the smooth component *u(x, y)*, which satisfies the description given in formula (2.3):

$$\begin{cases} \nabla u = [u_x, u_y] \\ |\nabla u| = \sqrt{u_x^2 + u_y^2} \end{cases} \tag{2.3}$$

The texture components *t(x, y)* describe slight pixel perturbation changes in the image area. Due to the limited range of values for image pixel brightness or color, the overall variation of texture components *t(x, y)* satisfies the inequality (2.4):

$$0 \leq \frac{1}{|\Omega|} \int \left[u^0(x,y) - u(x,y)\right]^2 \mathrm{d}\Omega = a_0^2 \leq C \tag{2.4}$$

Due to the equality constraints in the above model, the Lagrange multiplier method is first used to transform it into the following unconstrained functional optimization problem, as shown in formula (2.5):

$$M(u^0, u) = \int_{\Omega} f(|\nabla u|) \mathrm{d}\Omega + \frac{\lambda}{2} \int_{\Omega} (u - u^0)^2 \mathrm{d}\Omega \tag{2.5}$$

Among them, the first term is the regularized energy term, which suppresses image texture details by imposing constraints on the gradient amplitude of the image; The second item is the data fidelity term, which uses the difference measure between the original image u^0 and the smoothing component u to protect important detail features such as foreground contours; λ is a LaGrange multiplier used to balance the smoothness of the region and the strength of edge protection. Since equation (2.5) can be regarded as the independent variable u as the functional of the smooth component ∇u and its gradient image, as shown in equation (2.6):

$$F(u, \nabla u) = f(|\nabla u|) + \frac{\lambda}{2} \left(u - u^0\right)^2 \tag{2.6}$$

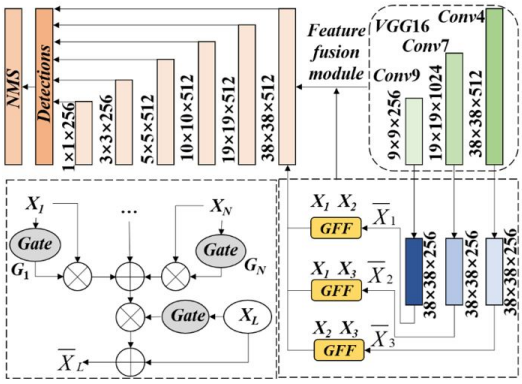


Fig. 2.2: Network Architecture for Multi-scale Feature Fusion

In addition, according to the Euler Lagrange equation, when there is an optimal solution to the above equation, formula (2.7) is satisfied:

$$F_u - \frac{\delta}{\delta_x}\left[F_{u_x}\right] - \frac{\delta}{\delta_y}\left[F_{u_y}\right] = 0 \tag{2.7}$$

2.2. Multi-scale Feature Fusion Network Architecture. The main challenge faced by the extraction and analysis of gymnastics image features is to analyze the ever-changing and complex situation of the movement process, which increases the difficulty of accurately detecting targets and is prone to missed detection. In addition, the contradiction between detecting and analyzing details of different body parts will become more prominent, resulting in low detection accuracy [17-18]. Therefore, fundamental research topics further enhance the model's ability to process details in gymnastics images, accurately extract features in complex environments, and accurately judge their status.

Traditional standard algorithms only perform independent output predictions on feature layers of different scales, with no connection between layers. Shallow feature maps are beneficial for target localization but lack sufficient semantic information. As convolution neural networks deepen, feature maps can represent more semantic information, which is beneficial for target recognition but not conducive to target localization. Therefore, traditional networks need help solving the contradiction between target recognition and localization, and they cannot obtain helpful multi-layer information in images [19-21]. Based on the analysis of traditional network models, this paper combines multi-scale fusion algorithms to construct a network model with higher accuracy in feature extraction in gymnastics motion images. This model combines detailed features with global semantic features through a multi-feature layer fusion strategy and a lightweight and efficient feature fusion module. This can effectively alleviate the contradiction between target localization and recognition in traditional network target detectors. As shown in Figure 2.2, a network architecture for multi-scale feature fusion is presented.

This architecture adopts the idea of multi-scale feature fusion to strengthen the connections between various feature layers, combines the advantages of convolutional neural networks in high and low layers, and combines the useful feature information of high and low layer feature maps [22]. By increasing the semantic information of shallow feature maps and the positioning information of deep feature maps, the detection accuracy of targets is improved, and missed detections are reduced to achieve better performance.

3. A Feature Extraction Model for Gymnastics Images in Multiscale Feature Fusion Algorithms.

3.1. A Moving Image Enhancement Network Based on Multiscale Features.

3.1.1. Scale Selection and Loss Function. In the multi-scale smoothing process of gymnastics motion images, the texture area of the image gradually tends to become smooth as the smoothing scale increases,

which enhances the consistency of colour distribution in the image area. However, when the smoothing scale is too large, the image may have a false overlap effect between the front and background contours due to being too smooth, leading to poor foreground contour positioning accuracy. In addition, the execution time of the algorithm increases with the iteration process, and the larger the smoothing scale, the more time it consumes [23-24]. However, due to the unknown optimal scale for image edge smoothing, this article designs iteration termination conditions for the algorithm based on the feature extraction results of gymnastics motion images at different scales to ensure an appropriate scale for feature extraction to stop. Firstly, the Jaccard Distance is used to measure the similarity of the feature extraction results of the selected images at adjacent scales, and a similarity index is defined as shown in formula (3.1):

$$Js(i) = \frac{Card(T_F^i \cap T_F^{i-1})}{Card(T_F^i \cup T_F^{i-1})} \tag{3.1}$$

The similarity index *Js(i)* uses the Intersection over the Union standard to quantify the similarity of adjacent scale foreground extraction results. The larger the value, the higher the similarity of adjacent scale foreground extraction results. The original gymnastics motion image generally contains many textures and colors. When the smoothing scale is low, the residual textures in the smoothing components can lead to poor parameter estimation accuracy. As the smoothing scale increases, the texture of moving images is gradually removed, the accuracy of parameter estimation continues to improve, and the accuracy of foreground extraction and similarity index *Js(i)* continues to improve [25-27]. When the smoothing scale is too large, the image is excessively smoothed, resulting in a pseudo overlap between the image's front and background contour boundaries. The segmentation curve crosses the foreground boundary, decreasing the foreground extraction accuracy and similarity index *Js(i)*. Therefore, based on the similarity index *Js(i)* and the trend of algorithm operation time with smoothing scale, the iteration termination condition of the algorithm in this paper is defined as shown in formula (3.2):

$$\xi\left[Js(i+1)\right]0 \tag{3.2}$$

Among them, the meaning of $\xi[\square\cdot]$ represents the operator of backward differentiation, as shown in formula (3.3):

$$\xi\left[Js(i+1)\right] = Js(i+1) - Js(i) \tag{3.3}$$

The multi-scale pyramid network based on attention mechanism uses thermal graph regression to conduct back error propagation, predict the Gaussian heat map of each joint point in the gymnastics motion image, and finally obtain the coordinate position of the joint point by finding the peak. Assuming we give a training set {T, J}, where T is the training set image set and J is the annotated joint point set, where the coordinates of the k-th joint point are (j_k^1, j_k^2) and the size of the thermal map is H×H. Bi (i=l,..,5) represents the i-th branch, and the thermal annotation process for each joint point is shown in formula (3.4):

$$g(j_k, x, y) = \frac{1}{\sqrt{2\pi\sigma_{Bi}^2}} e^{\frac{-(x-j_k^1)^2+(y-j_k^2)^2}{2\sigma_{Bi}^2}} \begin{cases} x = 1, 2, ..., H \\ y = 1, 2, ..., H \end{cases} \tag{3.4}$$

The multi-scale pyramid network based on attention mechanism will undergo error backpropagation in both stages 1 and 2. In stage 1, each branch will calculate a loss function, and then calculate their average value as the loss function for stage 1, as shown in equation (3.5):

$$L_1 = \frac{1}{k \times b} \sum_{i=1}^{b} \sum_{m=1}^{k} \left[y_{Bi,m} - g(j_m)\right]^2 \tag{3.5}$$

Calculate the loss function of all joint points, then classify the k joint points with more significant loss functions as the more challenging to detect joint points. At the same time, to prevent network oscillations caused by a significant difference between the maximum and minimum values of the loss function in a batch of

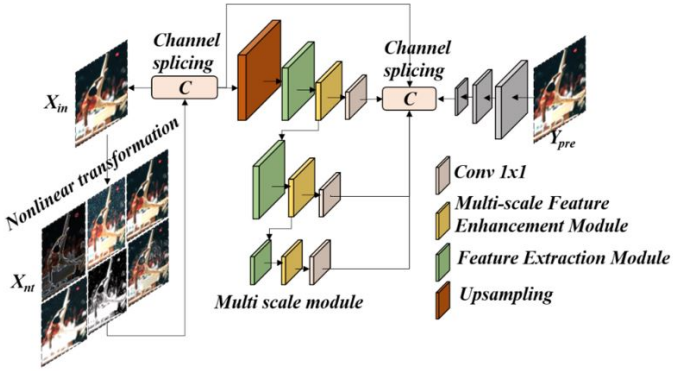


Fig. 3.1: A Model Framework for Multi-scale Feature Fusion Image Enhancement Network

gymnastics training images, the loss function in Stage 2 adopts an equalization strategy on a batch-by-batch basis. In summary, the definition of the loss function for stage two is shown in formula (3.6):

$$L_2 = \frac{1}{k \times n} \sum_{t=1}^{n} \sum_{m=1}^{k} \left[\hat{y}_{t,m} - g(j_m) \right]^2 \tag{3.6}$$

3.1.2. Multi Scale Feature Enhancement Module. Based on the above analysis and discussion, combined with the idea of end-to-end multi-scale feature fusion network, this paper constructs a multi-scale feature fusion algorithm to alleviate the problems of artifacts and noise in enhanced images. Firstly, non-linear transformation is performed on the gymnastics motion image to expand the low grayscale portion of the image and compress the high grayscale portion to enhance dark details. Then, channel fusion is performed with the original image before entering the network to enrich the original features; Subsequently, the fused features of the two images are fed into a multi-scale feature enhancement module, which includes an FEM, MFEM, and up sampling block. As the depth of the network layer increases, shallow features will decrease. Therefore, the network adopts a multi-scale feature fusion method to fuse low-level information with high-level information, reducing the amount of information lost due to functional loss [28]. The encoder in the enhancement module consists of convolutional kernels of size 3×3, while the decoder consists of transposed convolutions and activation layers of the same size. Then fuse each extracted scale feature with the input image to output the final enhanced image. As shown in Figure 3, a model framework for multi-scale feature fusion image enhancement network is presented.

In the gymnastics motion image feature extraction model based on a multi-scale fusion algorithm, the feature extraction module consists of a convolution layer and a ReLU activation function. The convolution kernel size of the convolution layer is 3×3, with a step size of 1. After each layer of convolution, there is a ReLU activation function to enhance the nonlinear representation of the network. After passing through the feature extraction module FEM at each scale, it will serve as the input of MFEM, and the output of the previous layer's enhancement module will serve as the input of the next layer's feature extraction module. In the multi-scale module, three feature extraction modules with different scales can extract image features of different scales.

The enhancement module in this article adopts an encoder decoder structure, where the encoding part consists of a convolutional kernel of size 5×5 and an activation function layer. The encoder, as a whole, presents a structure where the scale of the feature map gradually decreases, continuously reducing the resolution of the feature map to capture contextual semantic information. The decoder section consists of a deconvolution layer and an activation function layer, which also includes four stages. At each stage, after up sampling the input feature map, in order to reduce feature loss, it is concatenated with the corresponding proportion of the feature map in the encoder [29]. Unlike U-net, the decoder in this article does not use pooling layers, which can reduce data loss during the feature extraction process of gymnastics images.

3.2. Construction of a Feature Extraction Model for Gymnastics Images Based on Multiscale Feature Fusion Algorithm. . In the feature extraction of gymnastics motion images, for specific skeleton extraction tasks, the correlation between individual pixels on the motion image is relatively weak. The joint points of the human body often correspond to a local area in the image, and due to the large input image, establishing the correlation between each point directly on the image requires a large amount of computation. Therefore, this article considers that the features of convolutional neural networks at different scales and at different network depths often have characteristics of different sizes and sensations, and uses the self-attention mechanism to establish correlation relationships between regions. The basic description is shown in formula (3.7):

$$y_i = \frac{1}{C} \sum_{\forall region\, j} S(region_i, region_j) \times g(region_j) \tag{3.7}$$

Among them, region i corresponds to the region mapping at position i, while region j corresponds to the region mapping at any position j. S (·) establishes the similarity relationship between regions, and g (·) calculates the feature representation at region j. Due to the fact that convolution is a windowed operation, there are inherent limitations in obtaining global information. Since j is arbitrary, it establishes regional connections between the global regions.

Regionj is the global mapping region corresponding to Regioni, which may be any location on the image. The detection of human joint points is not isolated. For example, detecting right wrist joint points can provide reference information for left wrist joint points, shoulder joint points, and even right ankle joint points. Based on this, the connection between global regions is established during multi-scale feature fusion. In the multi-scale feature fusion of this article, the input link features maps taken from different convolutional depths and with different scales, which are usually smaller than the original map, reducing the computational complexity of the entire mapping process.

Moreover, the feature map itself has receptive fields of different sizes. Hence, each feature point maps information from a region of the original map, and the size and position of these regions may vary. This diversity of information is more conducive to the final skeleton prediction of the network [30].

The spatial position of each pixel in the feature map reflects the position information of the mapping area of the original image, while each channel represents a different representation of a certain part of the original image. Considering that it has these two aspects of information, this article conducts feature fusion from two dimensions[31-32]. Assuming that the feature maps of the two input scale branches are F_i^h and F_j^l, and F_j^l represents the feature representation at position i of the low scale feature map, and F_i^h represents the feature representation at position i of the high scale feature map, the spatial feature fusion process can be represented by formula (3.8):

$$y_i^s = \frac{1}{C(F)} \sum_{\forall F_j^l} f_s(F_i^h, F_j^l) \times g_s(F_j^l) \tag{3.8}$$

The weighted object is a high scale feature, as shown in equation (3.9). Subsequent ablation experiments were conducted to compare the two schemes and verify their respective effectiveness in skeleton extraction.

$$y_i^c = \frac{1}{C(F)} \sum_{\forall F_j^l} f_c(F_i^h, F_j^l) F_i^h \tag{3.9}$$

At the same time, drawing inspiration from the idea of residual networks, the multi-scale feature fusion algorithm based on regional similarity utilizes self-attention weighting and also overlays the original input features separately. The entire feature fusion process can be represented by formula (3.10):

$$F_i = F_i^h + upsample(F_j^l) + \alpha y_i^s + \beta y_i^c \tag{3.10}$$

Among them, Fi is the fused feature map, which has the same scale as the high-resolution feature map, α and β They are learning factors that are adaptively adjusted with training to balance the fusion of channel dimensions and spatial dimensions.

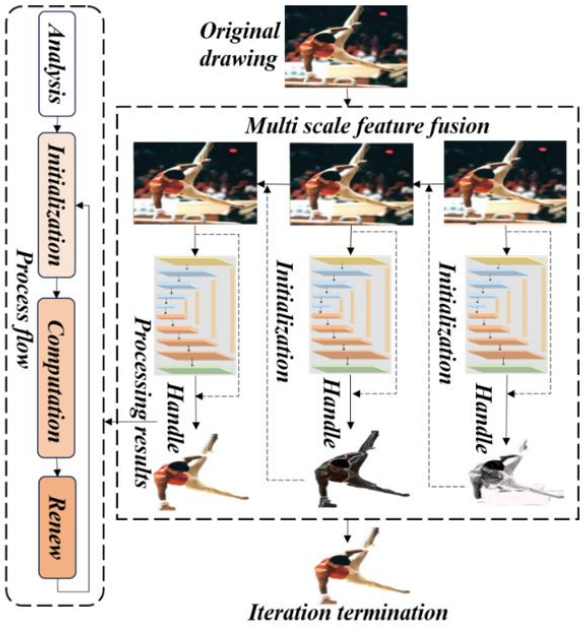


Fig. 3.2: Flow Diagram of Proposed Model

Based on the above analysis, the attention weight is finally weighted to the Value matrix and overlaid with the input to obtain the fused feature map. This process can be represented by formula (3.11):

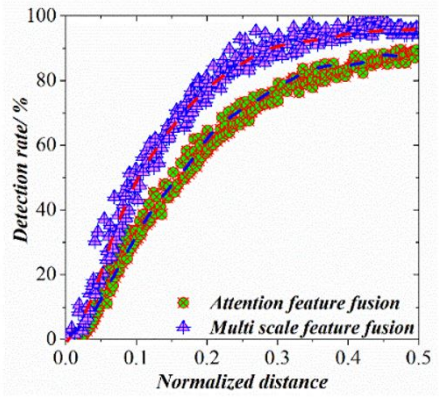
$$y^s = \alpha \times \mathrm{At}^s \times Value + F_h \tag{3.11}$$

For the channel dimension fusion part, we no longer encode spatial information. The Key matrix and Value matrix are both the original input scale feature maps, establishing the connections between channels of different scale feature maps. The original input low scale feature map Fs is up sampled and dimensionally adjusted to obtain a key matrix with a size of $B{\times}N2{\times}C$. The original input high scale feature map is dimensionally adjusted and transposed to obtain a Value matrix with a size of $B{\times}C{\times}N2$. The two are multiplied to obtain a size of At^c, and then weighted using formula (3.12) to obtain the fused feature map. The entire process can be expressed as:

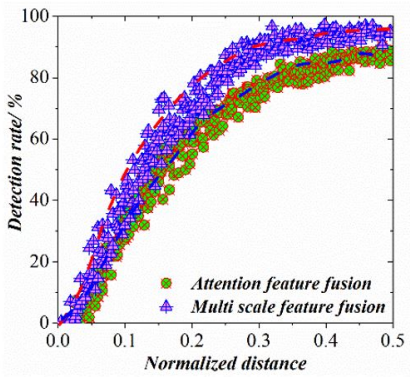
$$y^c = \beta \times \mathrm{At}^c \times Value + F_l \tag{3.12}$$

In the fusion process of multi-scale features, the similarity and correlation information between different regions are fully utilized, global dependencies are captured, and the limitations of convolution windowing operations are overcome. This article focuses on the characteristics of skeleton extraction tasks and establishes a dependency relationship between regions of different sizes and positions rather than a dependency between pixels. This model utilizes non-local operations for feature fusion. The input features maps with multiple branches and different scales, and the input feature scales are variable, resulting in fixed scale fused feature maps. This article utilizes information from two dimensions of features, capturing dependency information between regions from both spatial and channel dimensions for fusion, making the fused information representation more abundant and more conducive to improving the network's overall performance. Based on the establishment of the above model, as shown in Figure 3.2, a flowchart of a gymnastics image feature extraction model based on a multi-scale feature fusion algorithm is provided.

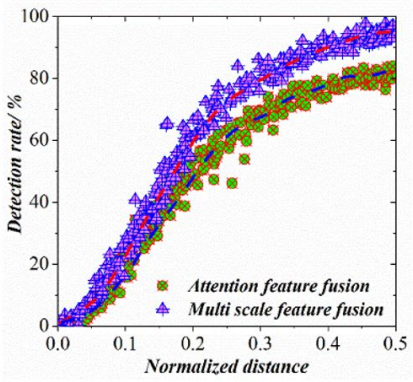
4. Analysis of Model Results. Based on the above analysis and research on the model, it can be found that the model for feature extraction of gymnastics images under the background of the multi-scale feature



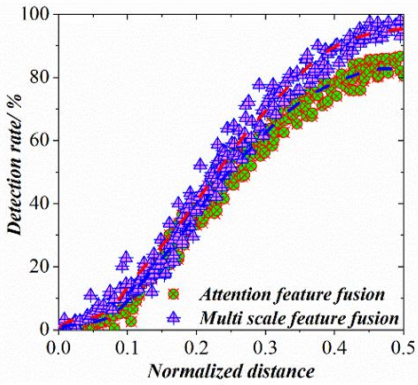
(a) Knee



(b) Ankle



(c) Hip



(d) Wrist

Fig. 4.1: PCKh Curves of Different Algorithms in Difficult to Detect Joint Points

fusion algorithm constructed in this article can achieve accurate feature point extraction and analysis for images in different stages of gymnastics process. Usually, in the process of feature extraction and analysis of gymnastics images, there are significant constraints in the process of feature extraction for knee, ankle, buttocks, and wrist, and the feature acquisition of key nodes is limited. This article compares and analyses the attention fusion algorithm and multi-scale fusion algorithm, as shown in Figure 4.1, and provides a comparison of two different algorithm models. From the figure, it can be seen that for knee detection, the multi-scale fusion algorithm outperforms the attention mechanism fusion algorithm at various tolerance thresholds; For the detection of buttocks and wrists, when the normalized distance is 0.1-0.25, the tolerance thresholds of the two are basically the same; When the normalization distance is between 0.25 and 0.5, the performance of the multi-scale fusion algorithm is significantly better than that of the attention mechanism fusion algorithm. Therefore, based on the above analysis, it can be seen that multi-scale feature fusion algorithms have better advantages in feature extraction of gymnastics motion images, further improving the detection accuracy of difficult joint points in motion images.

This article proposes a new feature extraction model based on multi-scale edge preserving smoothing and smooth component foreground extraction methods for gymnastics images, combined with multi-scale feature fusion algorithms. Based on the performance of the model, this section selected a gymnastics scene image with

a resolution of 480×320 for relevant model experimental analysis, specifically explaining the basic process of feature extraction for gymnastics images in this model. As shown in Figure 4.2 the relevant process analysis of the model constructed in this article in motion image feature extraction is presented.

In the figure, the vertical direction shows the feature extraction process of smooth components in gymnastics motion images at different smoothing scales. The original image can be seen as a smooth component on a rough scale. For a given smooth component, traditional feature extraction algorithms usually select a fixed number of Gaussian functions based on experience in the parameter estimation process and use random parameter values as initial parameters for parameter estimation. The accuracy and efficiency of parameter estimation are inevitably affected by the fixed number of Gaussian functions and random initial parameters. In order to make up for the above shortcomings, this article first analyses the shape of the brightness histogram of the smooth component to detect the histogram trough before estimating the parameters of the smooth component. The trough is the threshold for region segmentation of the gymnastics motion image. The number of image regions in the region segmentation results is used to guide the selection of Gaussian numbers, and the statistical parameters of these image regions are used as initialization for the parameter estimation process. Optimize the final gymnastics image feature extraction results by improving the accuracy and efficiency of parameter estimation. Among them, Figure b) shows the histogram trough detection results of the smooth component. Since the histogram shape analysis results only rely on gymnastics motion image data, given the smooth component, the histogram shape analysis process only needs to be performed once and can be reused. Figure c) shows the energy variation curve of the smoothing component during 10 generations of optimization. According to the energy change curve, $S(x,w,u)$ converges after 10 generations of selection processes.

Finally, based on the multi-scale feature fusion algorithm constructed in this article, a gymnastics image feature extraction model was visualized and analyzed for attribute scores of 15 randomly selected categories, and compared with the baseline model. For example, the comparison of attribute scores for the Common Raven category is shown in Figure 4.3. Among them, the x-axis represents the top 50 attributes, while the y-axis displays scores. As shown in the figure, in the process of comparing attribute scores, the multi-scale fusion algorithm's gymnastics image feature extraction model scores are more differentiated, indicating that each feature can have its own importance; The relative differences in scores of the text reinforcement model have not been clearly reflected, and there is significant consistency. Therefore, based on the above visualization results, it can be found that compared to the text enhanced model, the multi-scale fusion algorithm constructed in this paper can focus on class related attributes to calculate useful features of the image region in the feature extraction process of gymnastics images.

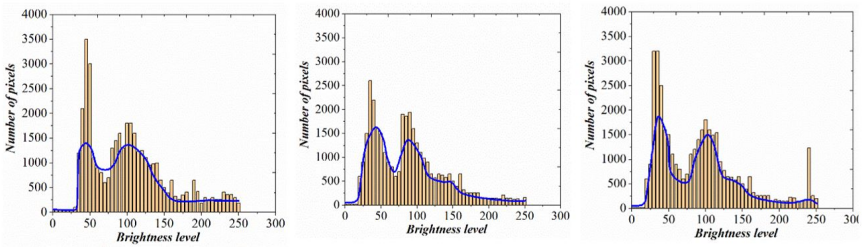
Multiple studies have proposed different feature extraction models in gymnastics image feature extraction. Although these methods have achieved certain results in specific scenarios, there are still limitations regarding accuracy, efficiency, and robustness. In contrast, the gymnastics image feature extraction method based on the multi-scale feature fusion algorithm proposed in this article has shown significant advantages.

Firstly, from an accuracy perspective, we conducted comparative experiments on existing HOG directional gradient histogram feature extraction methods and our proposed method. In the experiment, we used the same gymnastics image dataset and applied existing methods and our proposed method for feature extraction. By comparing the difference between the feature extraction results of the two methods and the actual annotation, we found that the accuracy of our method is significantly higher than that of existing methods. Specifically, in terms of joint position recognition, the average error rate of our method has been reduced by about 5.4%. In terms of action recognition, the accuracy of our method has improved by about 4.3%. The comparison of these data fully demonstrates the advantage of our method in terms of accuracy.

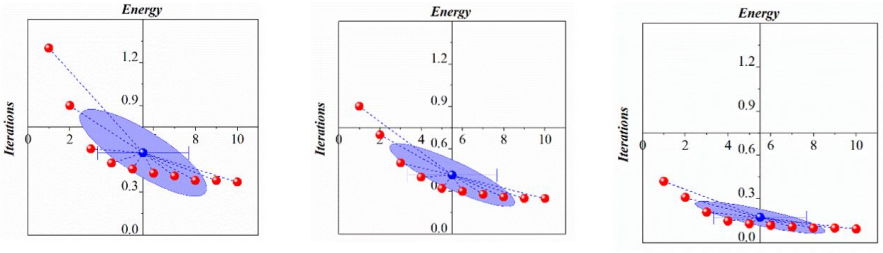
Secondly, in terms of efficiency, we compared the computational time and resource consumption between existing and our proposed methods. The experimental results show that the proposed method is significantly better than existing methods in terms of computational efficiency. Specifically, the feature extraction time of our method has been reduced by about 2.4%, while the memory usage has also been reduced by about 3.3%. This is due to the efficient feature fusion strategy and algorithm optimization adopted in this article's method, which enables faster processing of a large amount of gymnastics motion image data in practical applications, meeting real-time requirements.

In addition, we also compared the performance of existing methods and our proposed method in terms

(a) Original Image and Smooth Components



(b) Histogram Valley Detection of Images



(c) Changes in Energy under Different Iterations

Fig. 4.2: Foreground Extraction Process of Proposed Model

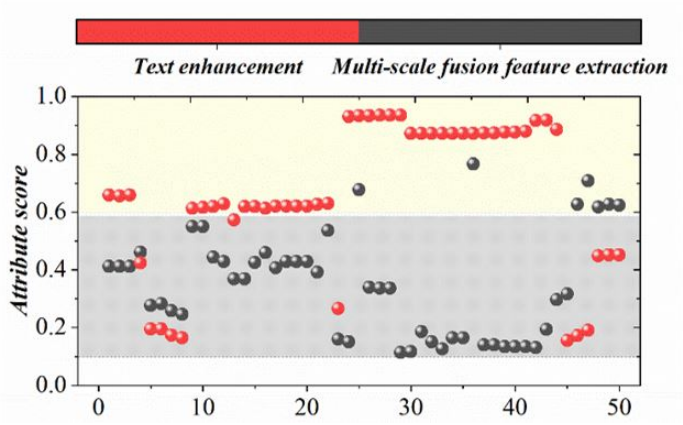


Fig. 4.3: Visualization Comparison of Attribute Scores for Categories

of robustness. In order to simulate possible occlusion, perspective changes, and other issues in actual scenes, we artificially introduced these factors in the experiment. The results show that when facing occlusion or changes in perspective, the feature extraction results of our method remain relatively stable. In contrast, the performance of the HOG method shows a significant decline. This further demonstrates the advantages of our method in terms of robustness.

5. Conclusions. Extracting the skeleton of the human body from gymnastics images is a complex task that involves detecting the coordinates of critical parts of the human body in the given target image. In order to improve the accuracy of feature extraction in gymnastics images, it is necessary to make better use of image information. Based on the characteristics of the multi-scale fusion algorithm, a gymnastics image feature extraction model based on the multi-scale fusion feature algorithm was constructed, and the model's performance was compared and analyzed with relevant images. The main conclusions are as follows:

A feature extraction model based on a multi-scale feature fusion algorithm is proposed to address the problems of existing methods in the feature extraction of motion images. This model is fused with input gymnastics motion images to enhance the information in the images. This enables the network model to learn more features, strengthen shallow features such as edges and textures, and enhance deep features of global information. Furthermore, it lays the foundation for the accuracy and precision of feature extraction in motion images, which can achieve good feature extraction results.

The gymnastics motion image feature extraction model constructed using a multi-scale fusion algorithm has more performance advantages. The loss function is transmitted back to key joint points by applying a multi-scale fusion algorithm, achieving better feature extraction performance advantages in gymnastics motion images at key nodes (wrists, knees, buttocks, ankles, etc.). The normalized distance is between 0.1 and 0.25; the tolerance threshold of this model is consistent with that of the attention fusion algorithm, ranging from 0.25 to 0.5. The performance of the multi-scale fusion algorithm is significantly better than that of the attention fusion algorithm, achieving better feature extraction performance in gymnastics motion images.

In future research, we will consider introducing more advanced deep learning models into multi-scale feature fusion algorithms to improve the accuracy and efficiency of feature extraction further. Explore more types of feature information fusion methods. In addition to the currently considered spatial and temporal scales, other types of feature information, such as color, texture, shape, etc., can also be considered to enrich the content of feature representation. Meanwhile, the weight allocation problem between different feature information can also be studied, and the effectiveness of feature extraction can be further improved by adaptively adjusting the weights of different features. Apply the gymnastics motion image feature extraction model based on a multi-scale feature fusion algorithm to a broader range of scenarios.

REFERENCES

[1] Ş'ukr'u, K., G'ung'or, S. & Mehmet, G. A new method based on deep learning and image processing for detection of strabismus with the Hirschberg test. *Photodiagnosis And Photodynamic Therapy*. **44** (2023)
[2] Bahram, R. Efficient and low-cost approximate multipliers for image processing applications. *Integration*. **94** (2024)
[3] Gener, S., Dattilo, P., Gajaria, D. & Others Gpu-based and streaming-enabled implementation of pre-processing flow towards enhancing optical character recognition accuracy and efficiency. *Cluster Computing*. **26** (2023)
[4] Chenglin, W., Huanqiang, H., Kean, L. & Others Attention-guided and fine-grained feature extraction from face images for gaze estimation. *Engineering Applications Of Artificial Intelligence*. **126(PB)** (2023)
[5] Weiyong, R. & Lei, S. Robust latent discriminative adaptive graph preserving learning for image feature extraction. *Knowledge-Based Systems*. **268** (2023)
[6] Weimin, L. Feature Extraction Method of Art Visual Communication Image Based on 5G Intelligent Sensor Network. *Journal Of Sensors*. **2022** (2022)
[7] Lulu, Q., Shijun, C., Junpei, H. & Others Statistical System of Cultural Heritage Tourism Information Based on Image Feature Extraction Technology. *Mathematical Problems In Engineering*. **2022** (2022)
[8] Jing, H., Hongyu, H., Guomin, L. & Others Study on Feature Extraction of Cable Surface Defect Image Based on Morphology and Edge Detection Algorithm. *Journal Of Physics: Conference Series*. **2035** (2021)
[9] Lijie, Z., Haisheng, D. & Donghui, C. An adaptive recognition method for take-off action images of back-style high jump based on feature extraction. *Future Generation Computer Systems*. **2021(prepublish)**
[10] Pengwei, S., Hongyu, S., Hua, Z. & Others Feature Extraction and Target Recognition of Moving Image Sequences. *IEEE AC*. **CESS 8** (2020)

[11] Yang, C. An Image Multi-Scale Feature Recognition Method based on Image Saliency. *International Journal Of Circuits, Systems And Signal Processing*. **15** (2021)
[12] Ying, S., Yaoqing, W., Bowen, L. & Others Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images. *IET Image Processing*. **17** (2022)
[13] Zihang, W., Hui, X., Yuchao, L. & Others Pavement texture depth estimation using image-based multiscale features. *Automation In Construction*. **141** (2022)
[14] Bi, S., Han, X. & Yu Y. An, L. 1image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Transactions On Graphics (TOG)*. **34** (2015)
[15] Sendova, M., Mariana, S. & Matthew, M. Direct surface area measurement from digital images via brightness histogram method. *Measurement Science And Technology*. **31** (2020)
[16] Tong, W., Lin, G., LingXiao, Z. & Others STAR-TM: STructure Aware Reconstruction of Textured Mesh from Single Image. (IEEE transactions on pattern analysis,2023)
[17] Chenyang, B., Shaozhong, C., Weijun, Z. & Others Printing roller image salient object detection based on multi-scale feature fusion. *Journal Of Physics: Conference Series*. **25** (2023)
[18] Kequan, Y., Jide, L., Songmin, D. & Others Multiscale features integration based multiple-in-single-out network for object detection. *Image And Vision Computing*. **135** (2023)
[19] Jinyuan, N., Xinyue, Z. & Jianxun, Z. Multiscale Feature Fusion Attention Lightweight Facial Expression Recognition. International. *Journal Of Aerospace Engineering*. **2022** (2022)
[20] Junwei, L., Lingyi, L., Wenbo, X. & Others Stereo super-resolution images detection based on multi-scale feature extraction and hierarchical feature fusion. *Gene Expression Patterns: GEP*. **45** (2022)
[21] Chao, Y., Sheng, R., Xiang, Y. & Others Crowd density estimation based on multi scale features fusion network with reverse attention mechanism. *Applied Intelligence*. **52** (2022)
[22] Lindeberg, T. Scale-Covariant and Scale-Invariant Gaussian Derivative Networks. *Journal Of Mathematical Imaging And Vision*. **64** (2021)
[23] Zheng, W. & Others Evidence theory based optimal scale selection for multi-scale ordered decision systems. *International Journal Of Machine Learning And Cybernetics*. **13** (2021)
[24] Xie, J., Yang, M., Li, J. & Others Rule acquisition and optimal scale selection in multi-scale formal decision contexts and their applications to smart city. *Future Generation Computer Systems*. **83** (2018)
[25] Shixun, W. & Qiang, C. The Study of Multiple Classes Boosting Classification Method Based on Local Similarity. *Algorithms*. **14** (2021)
[26] Wu, S., Wu, Y., Cao, D. & Others A fast button surface defect detection method based on Siamese network with imbalanced samples. *Multimedia Tools And Applications*. **78** (2019)
[27] Lai, H. & Zhang, P. Few-Shot Object Detection with Local Feature Enhancement and Feature Interrelation. *Electronics*. **12** (2023)
[28] Yunji, Z., Yuhang, Z., Xiaozhuo, X. & Others Fault diagnosis based on feature enhancement and spatial adjacent region dropout strategy. *Journal Of The Brazilian Society Of Mechanical Sciences And Engineering*. **45** (2023)
[29] Shiqi, W., Kankan, W., Tingping, Y. & Others Improved 3D-ResNet sign language recognition algorithm with enhanced hand features. *Scientific Reports*. **12** (2022)
[30] Arghya, P., Jayashree, K., Debashis, N. & Others Feature enhancing image inpainting through adaptive variation of sparse coefficients. *Signal, Image And Video Processing*. **17** (2022)
[31] Ghosh, S., Singh, A., Jhanjhi, N., Masud, M. & Aljahdali, S. SVM and KNN Based CNN Architectures for Plant Classification. *Computers, Materials & Continua*. **71** (2022)
[32] Usman, T., Saheed, Y., Ignace, D. & Nsang, A. Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification. *International Journal Of Cognitive Computing In Engineering*. **4** pp. 78-88 (2023)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Jan 16, 2024
Accepted: Apr 25, 2024