Hindawi
Complexity
Volume 2024, Article ID 9845724, 1 page
https://doi.org/10.1155/2024/9845724

Retraction

Retracted: Dance Movement Recognition Based on Feature Expression and Attribute Mining

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] X. Zhai, "Dance Movement Recognition Based on Feature Expression and Attribute Mining," *Complexity*, vol. 2021, Article ID 9935900, 12 pages, 2021.

Hindawi
Complexity
Volume 2021, Article ID 9935900, 12 pages
https://doi.org/10.1155/2021/9935900

Research Article

Dance Movement Recognition Based on Feature Expression and Attribute Mining

Xianfeng Zhai

College of Music and Dance, Guangxi Science & Technology Normal University, Guangxi, Laibin 546100, China

Correspondence should be addressed to Xianfeng Zhai; zhaixianfeng@gxstnu.edu.cn

Received 24 March 2021; Revised 13 April 2021; Accepted 25 April 2021; Published 3 May 2021

Academic Editor: Zhihan Lv

There are complex posture changes in dance movements, which lead to the low accuracy of dance movement recognition. And none of the current motion recognition uses the dancer's attributes. The attribute feature of dancer is the important high-level semantic information in the action recognition. Therefore, a dance movement recognition algorithm based on feature expression and attribute mining is designed to learn the complicated and changeable dancer movements. Firstly, the original image information is compressed by the time-domain fusion module, and the information of action and attitude can be expressed completely. Then, a two-way feature extraction network is designed, which extracts the details of the actions along the way and takes the sequence image as the input of the network. Then, in order to enhance the expression ability of attribute features, a multibranch spatial channel attention integration module (MBSC) based on an attention mechanism is designed to extract the features of each attribute. Finally, using the semantic inference and information transfer function of the graph convolution network, the relationship between attribute features and dancer features can be mined and deduced, and more expressive action features can be obtained; thus, high-performance dance motion recognition is realized. The test and analysis results on the data set show that the algorithm can recognize the dance movement and improve the accuracy of the dance movement recognition effectively, thus realizing the movement correction function of the dancer.

1. Introduction

Human motion recognition is the recognition of human body posture by extracting features from images [1, 2]. This technology can be used in intelligent dance training. By extracting the characteristics of dancers' images, the posture skeleton diagram of dancers can be obtained, which can be used to identify the movements of dancers, evaluate the posture of dancers, and make correction [3, 4].

Early human motion recognition focused on human body contour features or component models. Jalal et al. [5] designed a human motion recognition algorithm based on component detection by using boosting classifier to extract edge force field features. Heng [6] proposed a human body pose estimation method based on the combination of histogram and color features. However, due to the complexity of human body attitude changes, traditional methods are difficult to achieve effective attitude estimation. Therefore, the method based on deep learning is gradually used in human pose estimation. Mohammadimanesh et al. [7] designed an hourglass-shaped neural network to extract multiscale features for human motion recognition. Yang et al. [8] proposed a partial affinity domain approach to obtain the human skeleton. In addition, a number of deep learning-based human motion recognition algorithms have been proposed [9, 10]. These algorithms can be used for dance movement recognition and assist dancers in training. Dancers' movements change rapidly, their postures are changeable, and the scale of human skeleton is various on the 2D map, which brings the challenge to the intelligentization of dancers' auxiliary training.

But none of the current motion recognition features feature dancers. The attribute feature of dancer is the important high-level semantic information in the action recognition. Attribute recognition can help the model to find more precise feature expression, thus improving the

performance of motion recognition. Therefore, a dance movement recognition algorithm based on feature expression and attribute mining is designed to learn the complicated and changeable dance movements. Firstly, the original image information is compressed by the time-domain fusion module, and the information of action and attitude can be expressed completely. Then, a two-way feature extraction network is designed, which extracts the details of the actions along the way and takes the sequence image as the input of the network. Then, in order to enhance the expression ability of attribute features, a multibranch spatial channel attention integration module (MBSC) based on an attention mechanism is designed to extract the features of each attribute. Finally, using the semantic inference and information transfer function of the graph convolution network, the relationship between attribute features and dancer features can be mined and deduced, and more expressive action features can be obtained; thus, high-performance dance motion recognition is realized.

The rest of our paper is organized as follows. Related work is introduced in Section 2. Section 3 describes the algorithm proposed in this paper. Experimental results and analysis are discussed in detail in Section 4. Finally, Section 5 concludes the whole paper.

2. Related Work

The reason why human motion recognition has always received continuous attention from universities, research institutions, and even related companies is that motion recognition technology has matured and applied in multiple fields. It can not only help people's lives become more colorful but also make it more convenient for human beings. For example, motion recognition technology has been fully applied in intelligent monitoring, human-computer interaction, motion analysis, content-based video retrieval, and information preservation in intangible cultural heritage [11, 12].

Adopt appropriate feature extraction methods to characterize the key information of human actions. Obtaining features that contain rich information is the basis and guarantee of the follow-up action recognition link. Feature quality is related to the final recognition accuracy. In an ideal situation, the extracted features should be robust to changes in the appearance information, background environment, perspective, and execution of the action. At the same time, the feature description must contain sufficient information and be discriminative to facilitate effective action classification. The following introduces the global features and local features, respectively.

2.1. Global Feature Representation Method. The global feature representation method is to extract motion information from the overall motion video and is not sensitive to partial occlusion. Its main advantage is that it can well capture the global motion structure. In addition, since the global feature representation method usually takes the entire action video as a frame input and pools or filters the pixel features of the video to form a feature with a relatively simple structure, it requires less computational cost. In the global feature, the contour contains the boundary information of the human body motion area, which can better represent the overall information of the human body target.

Malbog et al. [13] used Canny edge detection to obtain the shape of the action and made it an action template, extracting key frames from the video and matching the saved action template and then distinguishing the type of human action. Alp and Keles [14] combined "motion energy image" and "motion history image" into a time-domain template and calculated the horses between the motion to be recognized and the template. The distance is given as the basis for the action category. He et al. [15] extended the two-dimensional HOG feature to the three-dimensional space and used the integral video method to calculate the HOG gradient value in the three-dimensional space to form the 3D-HOG feature. Hamzah et al. [16] used the RANSAC method to optimize the optical flow field and performed a two-dimensional Fourier transform on it, taking the maximum value of the obtained multiple Fourier coefficients as the extracted feature vector. Ullah et al. [17] extracted features based on the optical flow field, and each pixel of each frame of video image corresponds to a 12-dimensional feature vector one to one. Then solve the covariance matrix for the eigenvectors corresponding to all the pixels in the motion state in the video segment. Then, recognition is performed under the framework of sparse linear representation.

The top-down global feature representation needs to obtain the action area of the human body based on the human body detection and positioning technology and adopt the overall coding method for this area to further express the action information. Therefore, this method usually contains very rich global information, and the obtained features are also very discriminative. However, in the entire global feature extraction process, a lot of preprocessing work is required, and the effect of action recognition under complex background and occlusion problems is not optimistic.

2.2. Local Feature Representation Method. The local feature representation method first extracts the feature points of interest from the video image, so as to extract feature information from the surrounding area of the feature points of interest. Then, under the framework of Bag of Visual Words, a mathematical model is constructed for various features to predict the category labels of human actions. The local features calculated directly from the video have become popular in action recognition because of their strong robustness to complex backgrounds and their independence in target detection and tracking.

Wang et al. [18] compared the performance of dense action representation and sparse representation in local action feature evaluation. Zhao et al. [19] applied the trajectory features of the action to the recognition model and proposed a combination of HOG, HOF, and MBH features to form a dense trajectory feature representation method

with rich information and strong robustness. Since the global smoothing constraint is applied to the trajectory, the resulting trajectory is more robust. In order to reduce the computational complexity, Sun et al. [20] combined the idea of saliency detection to only sample relatively significant regions in the video that contain human motion. Lim et al. [21] integrated the sampling method based on the motion boundary into the dense sampling, and the invalid sampling points were significantly reduced. With the development of pixel-based underlying features, previous methods have made the greatest effort to capture the motion information embedded in continuous body movements. At present, the most effective action feature is IDT and its optimized feature representation, combined with Fisher vector coding, can obtain a good recognition effect. The existing action feature representation based on machine learning is also constructed on the basis of the underlying feature representation. For example, Zhao et al. [22] proposed a two-stream structured CNN model, which uses a spatial network to obtain surface information of independent frames such as objects and scenes. The temporal network mainly captures the motion information in the optical flow image.

Similar to the human action recognition method based on traditional methods, the human action recognition method based on deep learning also has a series of works using attributes [23]. Experiments have shown that pedestrian attributes perform well against changes in perspective, posture, and illumination. Rueda and Fink [24] combined pedestrian ID and attribute classification loss and contrast loss to train human action recognition tasks. Zhao et al. [25] proposed a semantic attribute learning model, which trains the attribute semantic description and is suitable for human action recognition. Wang et al. [26] trained the attribute classifier on a separate attribute data set and integrated it into the human action recognition model.

Specific dance action recognition is an important application field of human body action recognition [27]. The dance movement recognition technology can help dancers correct wrong postures and help intelligent dance-assisted training. Raz et al. [28] regard human posture estimation as a detection problem and perform human posture estimation by returning to the heat map of the key points of human posture. Chou et al. [29] proposed a human posture estimation algorithm based on hourglass, which can obtain multiscale features while having a more concise structure. Rogez et al. [30] proposed a real-time 2D pose detection method for multiple people. The main principle of this method is to associate body parts with corresponding individuals through partial affinity domain learning. In order to improve the detection performance of the algorithm for complex key points, Okuno et al. [31] used a global network to detect simple key points and then used RefineNet to detect complex key points for pose estimation. This network structure is called CPN.

3. Dances' Actions Recognition

3.1. Multifeature Fusion Expression. In order to improve the recognition performance of the model for complex dancers' actions, this paper designs a feature expression method based on multifeature fusion. The expression module of multifeature fusion proposed in this paper is mainly composed of a time-domain fusion module and a two-way feature extraction module. The overall structure of the model is shown in Figure 1.

The time-domain fusion module compresses the data amount of the original image through the two-frame fusion algorithm and the three-frame fusion algorithm, reduces the redundant information of the original image, and retains most of the information of the original image [32–36]. The bidirectional feature extraction module designed the bidirectional three-dimensional ConvNets networks Net-1 and Net-2. Net-1 forwards the fusion data to the network for feature extraction, and Net-2 reverses the fusion data to the network for feature extraction and then carries out two-way feature weighted fusion.

3.1.1. Time-Domain Fusion Module. Not only does the signal have high and low-frequency components, but also the image has high- and low-frequency components. For images, low-frequency components show the overall structure, and high-frequency components show detailed features. Obviously, there is redundancy in low-frequency components.

The use of Gaussian filtering to extract low-frequency information in the video is to calculate the transformation of each pixel in the image through the normal distribution, which is defined in the 2-dimensional space as

$$g(x, y, z) = \frac{e^{-\left((x^2+y^2)/2z^2\right)}}{2\pi z^2}. \tag{1}$$

In the formula, the variable z is the scale parameter, and the larger z, the more intense the smoothing. Assuming that the 2-dimensional image is $I(a, b)$, the low-frequency image $L_f(a, b)$ is the convolution of the two; namely,

$$L_f(a, b) = g(x, y, z) * I(a, b). \tag{2}$$

The main purpose of the time-domain fusion module is to reduce the data volume of the original video, reduce the redundant information contained in the video, and compress the data volume of the original video. Behaviour actions are composed of a series of images, containing a lot of information that has nothing to do with behaviour. Denote the image sequence as I_i (a, b), I_{i-1} (a, b), and $I+1$ (a, b) represent the previous frame and the next frame of the video sequence, respectively. First, the current frame I_i (a, b) is used to smoothly obtain the low-frequency information of

the image, and then the low-frequency information and high-frequency information $I_{i+1}(a, b)$ are reconstructed to obtain the fusion image, where the high-frequency component is its phase. The adjacent frame is the original image without Gaussian filtering. Finally, downsampling is performed in the spatial dimension to reduce the size of each frame of image.

The calculation of the two-frame fusion process is

$$\text{Fuse}(a,b) = \alpha * L_f(a,b) + \beta * I_{i+1}(a,b). \tag{3}$$

In the formula, $L_f(a, b)$ is the low-frequency pixel value of the image, $I_{i+1}(a, b)$ is the high-frequency component of the image, Fuse (a, b) is the pixel value of the fused image, and α and β are the weights of image's pixel values. In this paper, the values of α and β are 0.6 and 0.4, respectively.

The calculation of the three-frame fusion process is

$$T_f(a,b) = \alpha * L_{f_{i-1}}(a,b) + \beta * I_i(a,b) + \lambda * L_{f_{i+1}}(a,b). \tag{4}$$

In the formula, $I_i(a, b)$ is the low-frequency pixel value of the image, $L_{f_{i-1}}(a,b)$ and $L_{f_{i+1}}(a,b)$ are the high-frequency components of the image, $T_f(a, b)$ is the fusion image pixel values, and α, β, and λ are the weights of the image. In this paper, the values of α, β, and λ are 0.25, 0.5, and 0.25, respectively.

3.1.2. Two-Way Feature Extraction Module. The two-way feature extraction network structure used in this paper is shown in Figure 2. It can be seen from Figure 2 that both Net-1 and Net-2 networks consist of 6 convolutional layers, 4 pooling layers, 6 activation functions, and 3 batch normalization layers. The size of the convolution kernel of each Conv3D layer is $3 \times 3 \times 3$, and the number of filters of the 6 Conv3D layers is 32, 64, 128, 128, 256, and 256 in order. Except for Conv3D, there is a Relu layer and a BN layer behind each Conv3D layer. The kernel size of the first pooling layer is $2 \times 2 \times 1$, and the step size is $2 \times 2 \times 1$. Only spatial aggregation is performed on the first layer Conv3D, and the amount of time dimension information is retained. The kernel size of other pooling layers is $2 \times 2 \times 2$, and the step size is $2 \times 2 \times 2$, which reduces the space size and time length to a ratio of 4 and 2. At the same time, random inactivation is added after each level of pooling layer mitigation model overfitting. Using the BN layer through normalization methods, the input value of each layer of the neural network conforms to the standard normal distribution, which avoids the disappearance of the gradient during the backpropagation process, greatly improves the training efficiency, improves the classification accuracy, and reduces the superdifficulty of parameter adjustment.

The difference between Net-1 and Net-2 networks is the video input layer and the pooling layer. When the Net-1 network receives data, it is input in the order of original data, two-frame fusion data, and three-frame fusion data, and maximum pooling is used. The main purpose is to retain texture features. When the Net-2 network receives data, it is input in the reverse order of the Net-1 network, and the average pooling layer is used to preserve the overall data characteristics.

3.2. Attribute Mining. Pedestrian attributes express the high-dimensional semantic information of pedestrians and can provide detailed information for pedestrian reidentification tasks, such as hair length and sleeve length. Character attributes are highly robust to lighting, posture changes, and camera changes. For this reason, this paper introduces the attribute information into the action recognition task simply and directly.

In terms of the mode of action, the attention model selected in this paper is the soft attention mode. Using the soft attention method can use the existing task backpropagation for learning without introducing additional tasks. In terms of scope, this paper uses a combination of spatial attention and channel attention to implement the attribute attention model. In order to further improve the performance of the attention model, this paper uses a multibranch integration method to combine the results of multiple attention models.

Model ensemble is an important method in the field of machine learning. Model integration refers to a type of method that uses a certain integration strategy to fuse multiple models or the results of multiple models together as the final model or final result. By using appropriate integration methods, the performance of the model can be effectively improved.

Spatial attention can use existing task backpropagation for learning without introducing additional tasks. The advantage is that it is self-directed and does not need to introduce additional tasks. Since the spatial attention model is not an attention label because of its guidance information, it belongs to semisupervised learning. Considering that the attention model learned by semisupervised learning may be biased, this paper uses the method of model integration to avoid it. Therefore, channel attention is introduced, multiple attention models are trained at the same time, and the outputs of multiple models are finally integrated as the final output result of the model.

The advantage of soft attention is that it is self-directed and does not need to introduce additional tasks. At the same time, the soft attention model belongs to semisupervised learning because its guidance information is not an attention label. Considering that the attention model learned by semisupervised learning may be biased, this paper uses the method of model integration to avoid it. Train multiple attention models at the same time, and finally integrate the outputs of multiple models together as the final output result of the model, which is called the multibranch spatial channel attention integration module (MBSC). As shown in Figure 3, the results of multiple attention branches trained separately are first integrated to obtain an attention map. There are multiple options for specific integration strategies:

$$\text{attention}_{\text{all}} = \frac{1}{m}\sum \text{attention}_i, \tag{5}$$

$$\text{attention}_{\text{all}} = \text{vote}\left(\text{attention}_1, \ldots, \text{attention}_m\right), \tag{6}$$

$$\text{attention}_{\text{all}} = \max\left(\text{attention}_1, \ldots, \text{attention}_m\right). \tag{7}$$

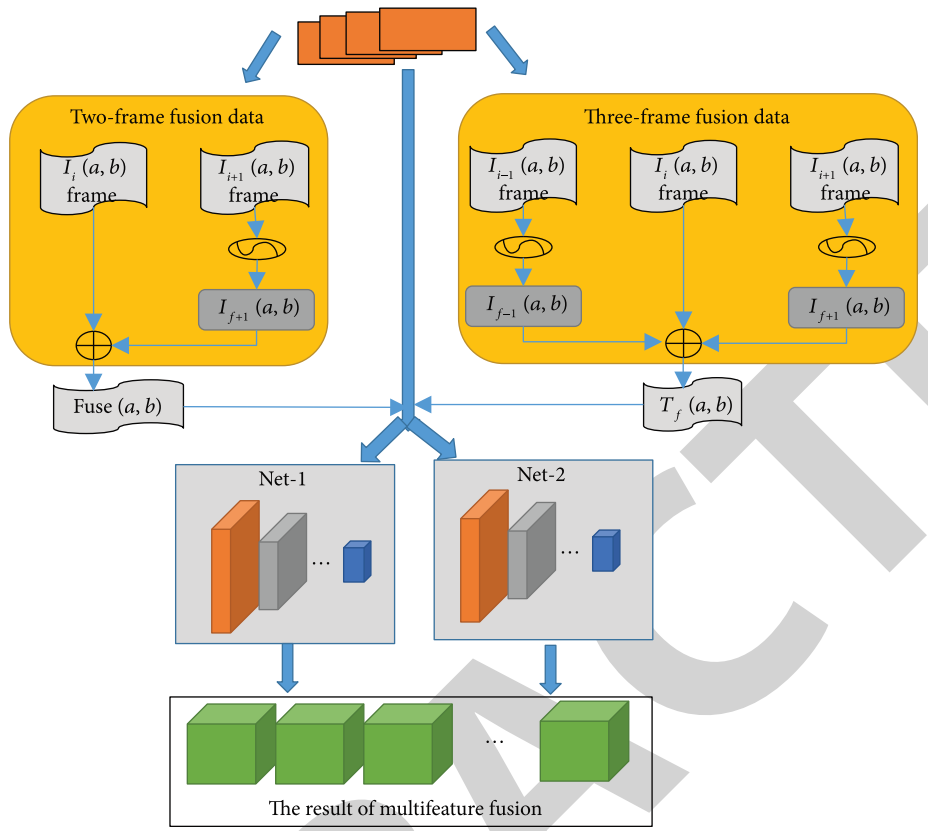


Figure 1: Multifeature fusion expression module.

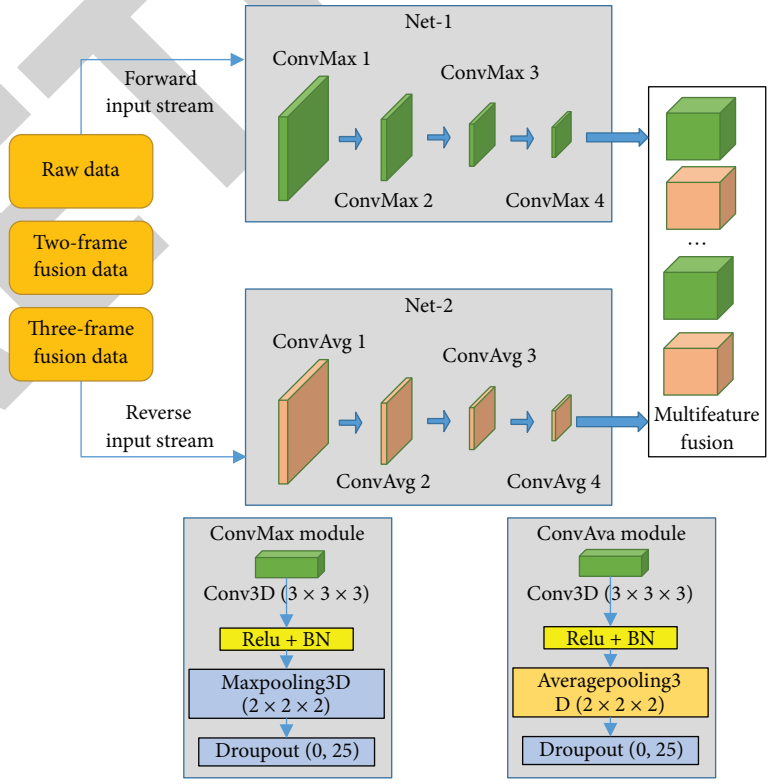


Figure 2: Two-way feature extraction module.

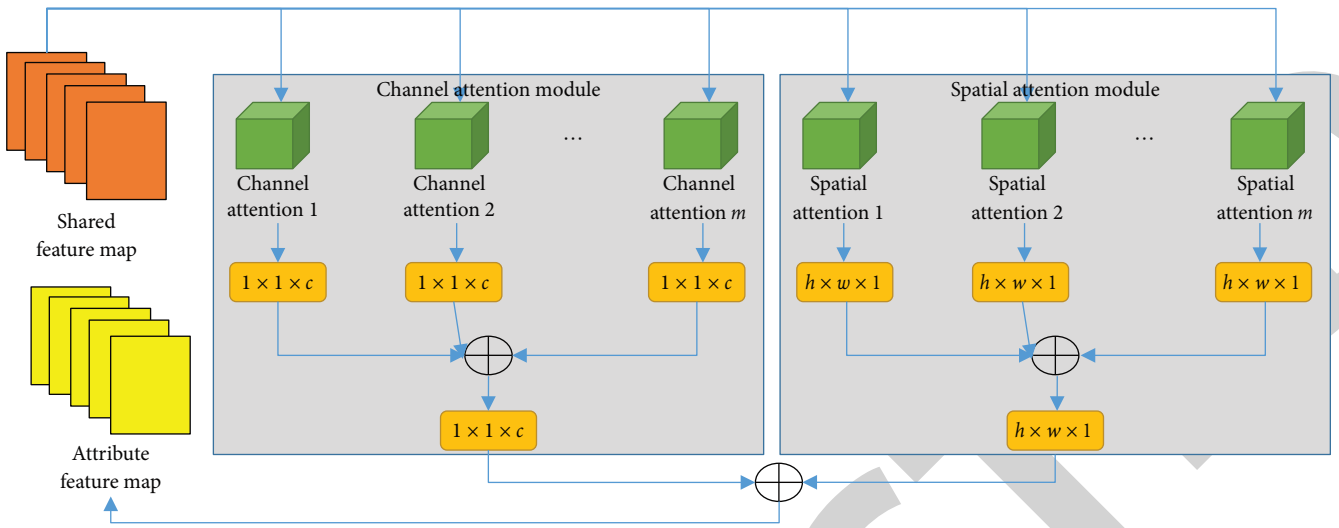


Figure 3: Multibranch spatial channel attention integration.

Among them, the variable attentional represents the attention result after integration, attention$_1$,..., attention$_m$ represent the attention result sequence before integration, and the variable m represents the number of branches. The symbol max represents the maximum value operation. The symbol vote means to vote. Formulae (5)–(7), respectively, represent three different integration strategies: averaging, maximizing, and voting. In this paper, formula (5) is selected as the integration strategy for finding the mean value. Finally, this paper combines the spatial attention map and the channel attention map to form a spatial channel attention map through the matrix dot product method.

$$F_{\text{all}} = \text{spatial\_attention} * \text{channel\_attention}. \quad (8)$$

Among them, the variable F_{all} represents the spatial channel attention map, and the dimension is $h \times w \times c$. The spatial channel attention map is multiplied by the shared feature map to obtain the attribute feature map for specific attributes. The attribute feature map is first reduced by global average pooling, and then the convolution kernel is 1×1 for feature extraction and dimensionality reduction to obtain $1 \times 1 \times c'$ attribute features. Finally, the obtained features are input into the attribute classifier for attribute classification. As shown in Figure 4, it is the overall structure diagram corresponding to the action recognition algorithm based on feature expression and attribute mining.

4. Results and Discussion

This paper is based on feature expression and attribute mining. In the experimental results, we analyse the performance from the perspectives of multifeature fusion, attribute mining, and the combination of the two.

4.1. Data Preprocessing and Parameter Setting. The dance data sets are used in the experiment of this paper, namely, DanceDB data set and FolkDance data set. In the DanceDB data set, emotion markers are used for each dance category. The FolkDance dance data set is divided into four groups of dances in total, each group contains a number of subdivided dance moves, the action categories are relatively rich, and each set of dance moves is relatively complex and challenging.

During the experiment, a small batch of data is used for training, and the batch is processed once every 16 batches, and the number of iterations' epoch is set to 100. The accuracy of the model is improved by automatically adjusting the learning rate. The adjustment process adopts a linear attenuation method, and the initial learning rate is set to 0.01.

4.2. Multifeature Fusion Performance. This part shows the performance of the fusion of different features. This part not only shows different features but also shows the performance of single feature and multiple feature fusion.

From the experimental results in Figure 5, it can be seen that the recognition rate of dance movements in each group is still relatively low for each single feature. The HOG feature is used to characterize the local appearance and shape of the action. When the similarity between the actions in the dance combination is too high, it will increase the difficulty of recognition and affect the recognition accuracy. In addition, the results of the first two groups also show that when the action similarity in the dance combination is very low, the performance of the HOG feature in this paper is better than the HOF feature for complex dance action recognition. In this paper, the recognition rate of the audio signature feature in the four groups is not much different, which also shows that the audio feature has maintained a good recognition rate and is not affected by factors such as complex actions. Compared with a single feature, the method in this paper has a relatively large improvement in the recognition rate of each group of dance movements. Although facing the influence of similar dance moves, the recognition rate of the method based on multifeature fusion in this paper is higher than the recognition rate of all single features in the two

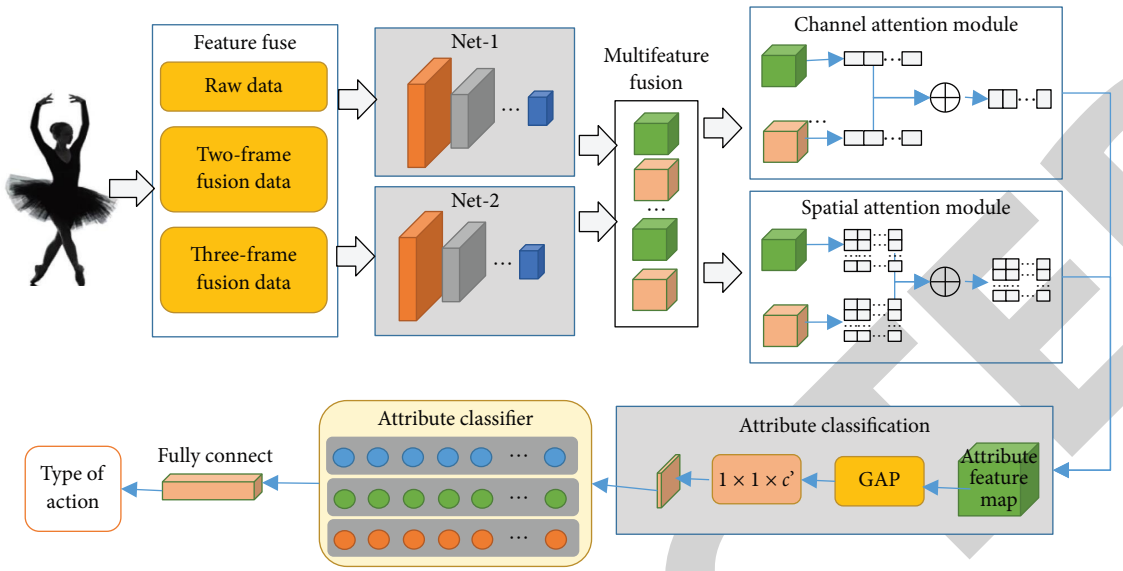


Figure 4: Structure diagram of action recognition algorithm based on feature expression and attribute mining.

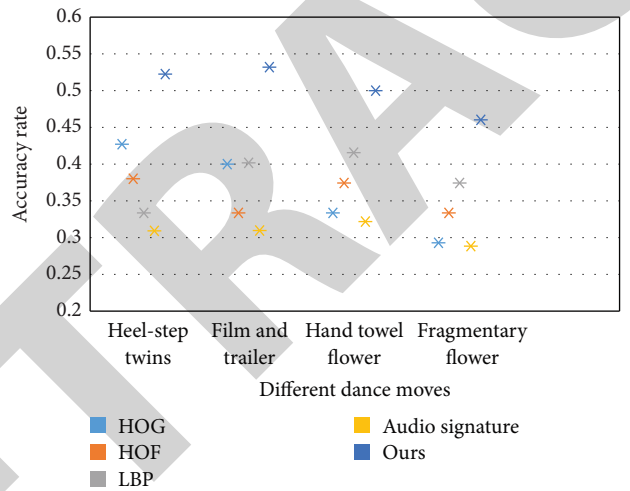


Figure 5: Comparison of experimental results between the algorithm in this paper and other algorithms on different dance combinations.

combinations, respectively. Comparing the first two groups at the same time, although there is a decline, the method in this paper is more robust than the performance of a single feature in the same group. This shows that, in dance movement recognition, by fusing features and then learning the corresponding weights in the training process, the influence of complex movements and similar factors can be reduced to maintain a certain accuracy.

The accuracy of various human action recognition methods (including HOG, LBP, HOG, and SIFT fusion, LBP and SIFT fusion, and the algorithm proposed in the paper) is compared as follows.

From Figure 6, we can see that the recognition accuracy of the fusion algorithm in the paper in the DanceDB data set and FolkDance data set is higher than that of other algorithms. However, the recognition rate of HOG feature and LBP feature fusion method is higher than that of any single feature, which shows that feature fusion can effectively improve the recognition accuracy. Among several multi-feature fusion algorithms, the multifeature fusion proposed in this paper has a slightly higher recognition accuracy in both databases than other methods.

The dance movement recognition algorithm based on multifeature fusion will be tested on the database, and the test results are shown in Figure 7. In Figure 7, (a) represents the original image; (b) represents the key points of the real human skeleton in the figure; (c) represents the key points of the human skeleton predicted by the hourglass method; (d) represents the key points of the human skeleton predicted by the FPN-based algorithm; (e) represents the key points of the human skeleton predicted by the algorithm in this paper. It can be seen from the results in Figure 7 that, in the first column of simple dance moves, the three methods can basically predict the key points of the dancer. But for the

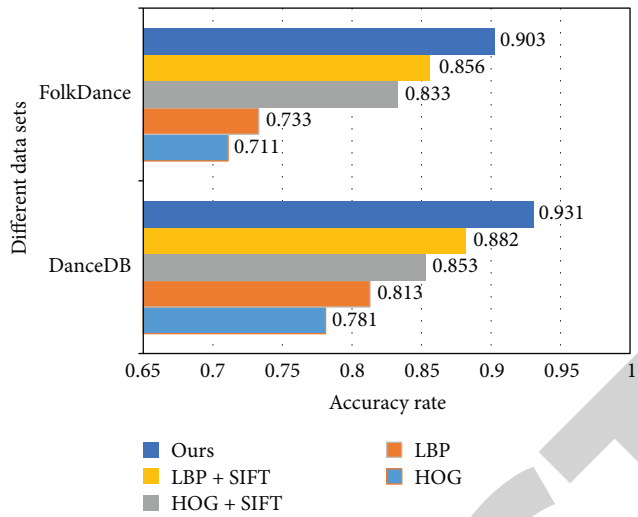


Figure 6: Comparison of experimental results using different feature extraction algorithms.

second, third, and fourth columns of dance moves in Figure 7, the algorithm in this paper has a richer semantic feature after multifeature fusion and successfully predicted the key points.

4.3. Performance Analysis of Attribute Mining. This part shows the performance of attribute mining. This part also shows the performance analysis of different attention mechanisms.

Figure 8 lists the experimental results under different module settings. MBSC is the multibranch spatial channel attention model mentioned above. Similarly, MBS and MBC refer to the multibranch spatial attention ensemble model and the multibranch channel attention ensemble model, respectively. According to the data in Figure 8, the following conclusions can be drawn:

(1) Both spatial attention and channel attention mechanisms can improve the performance of action recognition to a certain extent. Relatively speaking, the channel attention mechanism improves more. This shows that the idea of using the attention mechanism for attribute mining to improve the performance of action recognition is correct.

(2) Combining spatial attention and channel attention can further improve the results of the experiment, which shows that spatial attention and channel attention are complementary, and combining the two is beneficial to the study of action recognition.

In summary, the experiment verifies the effectiveness of the multibranch spatial channel attention integration module.

In the multibranch spatial channel attention integration module, the number of branches m as an important superparameter will have an important impact on the experimental results. Specifically, if the number of branches used is insufficient, invalid information is more likely to affect the effect of the integration. On the contrary, if the number of branches is too large, part of the weaker but still valid information will be ignored, and the training process will be too complicated and time-consuming. During the experiment, $m \in [1, 8]$ is an integer, where 1 represents the case of non-multi-branch. In order to make it easier to observe the changes in the effect when the number of branches m changes, the line chart used in this paper represents the case of different branch numbers. The experimental results on the two data sets are shown in Figure 9. Observing the curve changes in Figure 9 in the process, it can be found that as the number of branches increases, the effect improves more obviously at the beginning, and the effect begins to deteriorate after reaching the apex, but the deterioration process is very slow. This shows that the increase in the number of branches can help the model focus on more important areas and features, and unimportant and less important information is gradually eliminated. According to the experimental results, $m = 4$ is selected as the superparameter set in the experiment.

4.4. Comparison of the Method in this Paper with the Current Typical Mainstream Models. This part combines multifeature expression and attribute mining and compresses with the current mainstream algorithms.

In order to evaluate the performance of the algorithm in this paper, compared with the traditional algorithm, dual-stream structure, 3D convolutional neural network, LSTM structure, and CPN structure, the results are shown in Figure 10. The experiment tested the average recognition accuracy of each algorithm under 5-fold crossover on two data sets. On the DanceDB data set, the method in this paper is superior to other algorithms. Compared with the classic dual-stream network and CPN, the accuracy has been improved. On the FolkDance data set, the algorithm in this

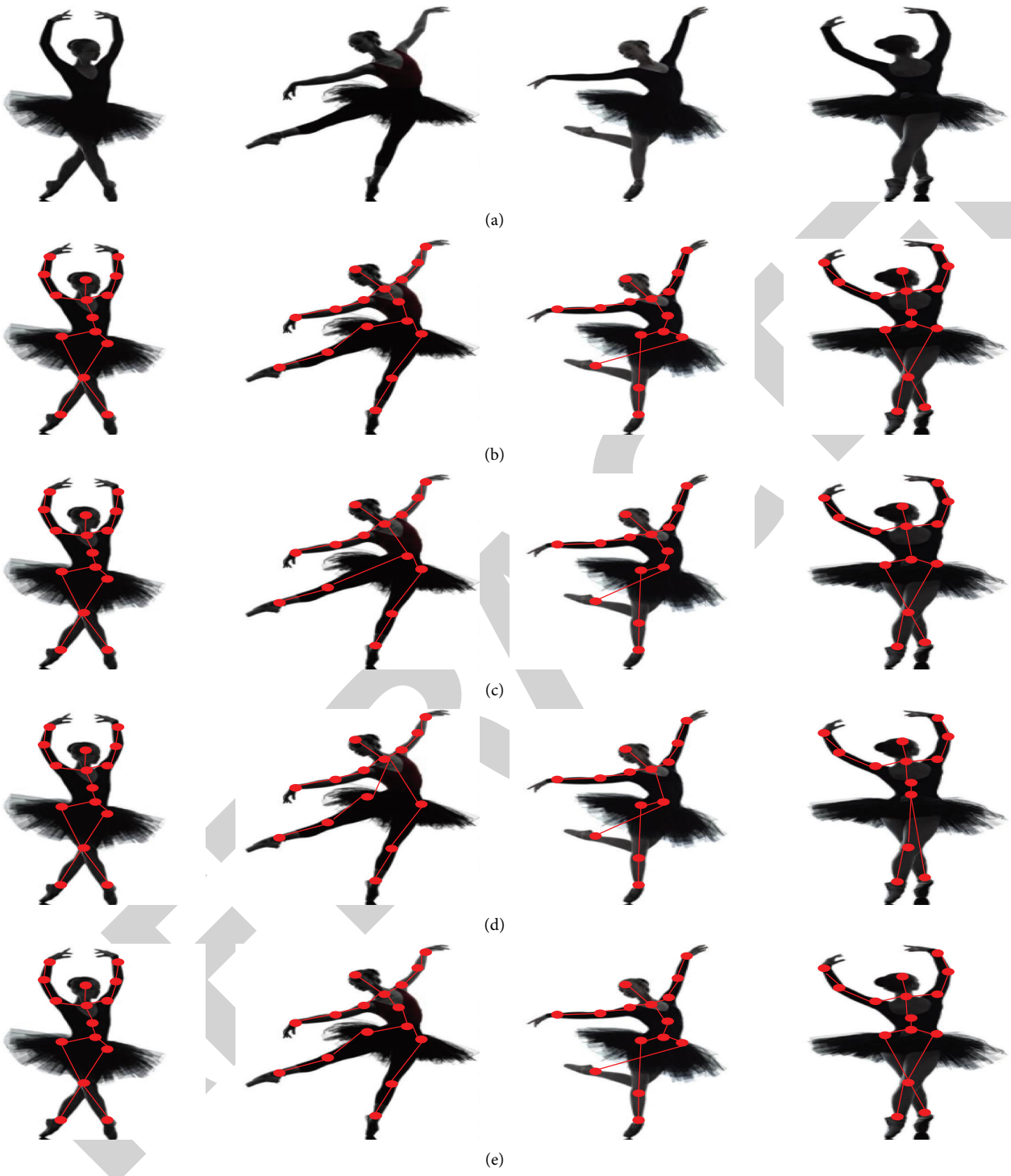


Figure 7: Skeleton key point prediction based on multifeature fusion. (a) Original image; (b) key of real human skeleton; (c) hourglass predictions; (d) FPN prediction; (e) our prediction.

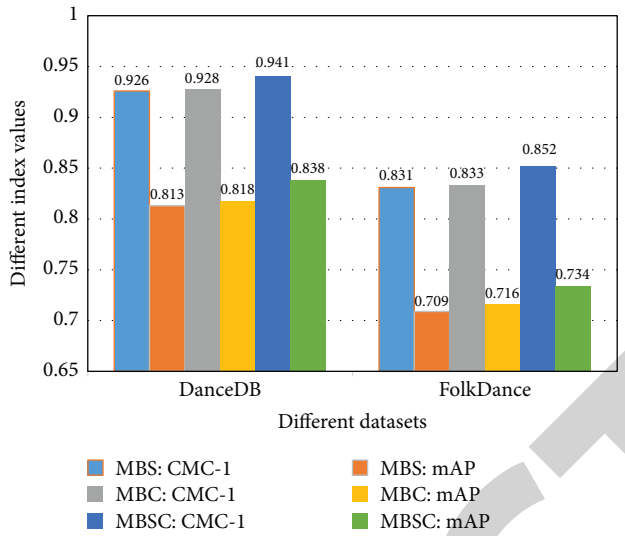
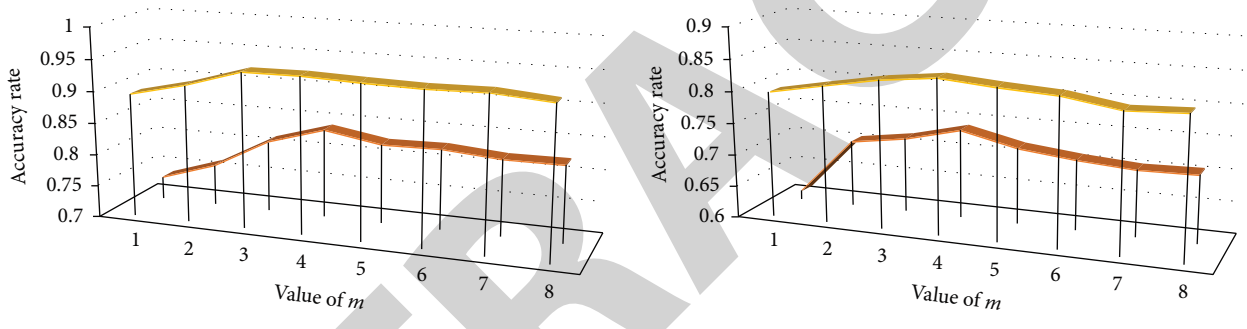


Figure 8: Ablation experiment.



	1	2	3	4	5	6	7	8
Rank-1	0.893	0.914	0.941	0.942	0.939	0.937	0.939	0.932
mAP	0.734	0.763	0.811	0.838	0.823	0.825	0.818	0.819

(a)

	1	2	3	4	5	6	7	8
Rank-1	0.796	0.813	0.829	0.839	0.832	0.827	0.813	0.817
mAP	0.612	0.703	0.716	0.738	0.718	0.708	0.702	0.704

(b)

Figure 9: Branch number curve on the two data sets. (a) Branching curve on the DanceDB data set. (b) Branching curve on the FolkDance data set.

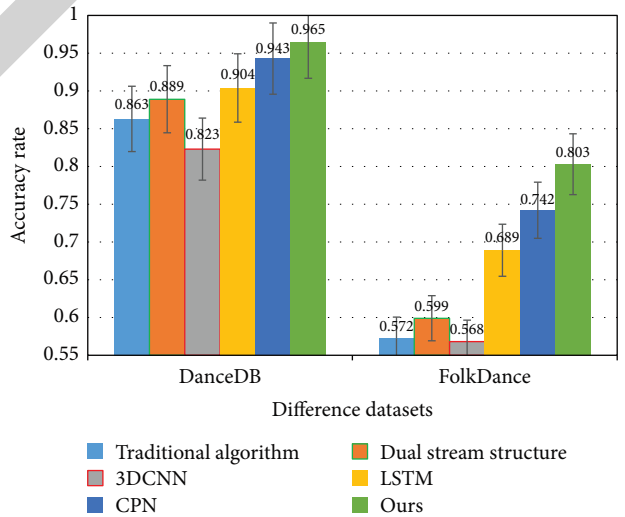


Figure 10: Comparison of the recognition accuracy of the algorithm in this paper on the two data sets.

paper also obtained the best recognition accuracy, and compared with the 3D convolutional neural network, the accuracy rate is also improved.

The method in this paper has excellent recognition accuracy due to the following aspects:

(1) Extract the information of the original video through the fusion of two and three frames in the time domain to improve the ability to analyse the video information, so that it can fully express the behaviour information.

(2) Through the two-way feature extraction network on the network structure, the detailed features and overall features of the behaviour are extracted, and the spatial-temporal features are efficiently extracted.

(3) Further attribute mining is performed on the fusion features, and the attribute features are incorporated into the recognition, thereby improving the recognition performance.

5. Conclusion

This paper designs a dance movement recognition algorithm based on feature expression and attribute mining, which is used to learn complex and changeable dancer movements. First, the time-domain fusion module is introduced to compress the original image information and completely express the information of the action and posture. Then, a two-way feature extraction network is designed to extract the detailed features and overall features of the dancers. Then, the multibranch spatial channel attention integration module of the attention mechanism is used to extract the characteristics of each attribute. Finally, using the semantic inference and information transmission functions of the graph convolutional network to realize the mining and reasoning of the relationship between the attribute characteristics and the dancer characteristics, and obtain the movement characteristics with stronger expressive ability, so as to realize high-performance dance movement recognition. In order to verify the effectiveness of the proposed algorithm, this paper conducts experiments on two commonly used data sets for each module, different parameters, and different variants in the paper to verify whether the three methods proposed in this paper are effective. The results show that the algorithm can recognize dance movements and can effectively improve the accuracy of dance movement recognition, so as to realize the function of correcting dancers' movements.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Kuehne, H. Jhuang, E. Garrote et al., "HMDB: a large video database for human motion recognition," in *Proceedings of the 2011 International Conference on Computer Vision*, pp. 2556–2563, IEEE, Barcelona, Spain, November 2011.

[2] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: a survey," *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018.

[3] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: promising applications for indoor monitoring," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.

[4] P. Wang, H. Liu, L. Wang, and R. X. Gao, "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration," *CIRP Annals*, vol. 67, no. 1, pp. 17–20, 2018.

[5] A. Jalal, M. A. K. Quaid, and M. A. Sidduqi, "A Triaxial acceleration-based human motion detection for ambient smart home system," in *Proceedings of the 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pp. 353–358, IEEE, Islamabad, Pakistan, January 2019.

[6] C. Heng, "New human pose estimation algorithm based on HOG and color features," *Computer Engineering and Applications*, vol. 53, no. 21, pp. 190–194, 2017.

[7] F. Mohammadimanesh, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, "A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 151, pp. 223–236, 2019.

[8] J. Yang, W. S. Zheng, Q. Yang et al., "Spatial-temporal graph convolutional network for video-based person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3289–3299, Seattle, WA, USA, June 2020.

[9] P. Ghosh, J. Song, E. Aksan et al., "Learning human motion models for long-term predictions," in *Proceedings of the 2017 International Conference on 3D Vision (3DV)*, pp. 458–466, IEEE, Qingdao, China, October 2017.

[10] J. Butepage, M. J. Black, D. Kragic et al., "Deep representation learning for human motion prediction and classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6158–6166, Honolulu, Hawaii, July 2017.

[11] E. Barsoum, J. Kender, Z. Liu, and H. P.- GAN, "Probabilistic 3D human motion prediction via GAN," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1418–1427, Salt Lake City, UT, USA, June 2018.

[12] M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, "A robust human activity recognition system using smartphone sensors and deep learning," *Future Generation Computer Systems*, vol. 81, pp. 307–313, 2018.

[13] M. A. F. Malbog, L. L. Lacatan, R. M. Dellosa et al., "Edge detection comparison of hybrid feature extraction for combustible fire segmentation: a Canny vs Sobel performance analysis," in *Proceedings of the 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, pp. 318–322, IEEE, Shah Alam, Malaysia, August 2020.

[14] E. C. Alp and H. Y. Keles, "Action recognition using MHI based Hu moments with HMMs," in *Proceedings of the IEEE EUROCON 2017-17th International Conference on Smart Technologies*, pp. 212–216, IEEE, Ohrid, Macedonia, July 2017.

[15] L. He, S. Wen, L. Wang et al., "Vehicle theft recognition from surveillance video based on spatiotemporal attention," *Applied Intelligence*, vol. 51, pp. 1–16, 2020.

[16] R. A. Hamzah, A. F. Kadmin, S. F. A. Ghani et al., "Disparity refinement process based on RANSAC plane fitting for machine vision applications," *Journal of Fundamental and Applied Sciences*, vol. 9, no. 4S, pp. 226–237, 2017.

[17] A. Ullah, K. Muhammad, J. Del Ser et al., "Activity recognition using temporal optical flow convolutional features and multilayer LSTM," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9692–9702, 2018.

[18] S. Wang, E. Zhu, J. Yin, and F. Porikli, "Video anomaly detection and localization by local motion based joint video representation and OCELM," *Neurocomputing*, vol. 277, pp. 161–175, 2018.

[19] Y. Zhao, Y. Xiong, and D. Lin, "Trajectory convolution for action recognition," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 2208–2219, Montreal, Canada, December 2018.

[20] M. Sun, Z. Zhou, Q. Hu et al., "SG-FCN: a motion and memory-based deep learning model for video saliency detection," *IEEE Transactions on Cybernetics*, vol. 49, no. 8, pp. 2900–2911, 2018.

[21] W. Lim, S. Lee, M. Sunwoo, and K. Jo, "Hierarchical trajectory planning of an autonomous car based on the integration of a sampling and an optimization method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 2, pp. 613–626, 2018.

[22] Y. Zhao, K. L. Man, J. Smith et al., "Improved two-stream model for human action recognition," *EURASIP Journal on Image and Video Processing*, vol. 2020, no. 1, pp. 1–9, 2020.

[23] H. Liang, X. Sun, Y. Sun et al., "Text feature extraction based on deep learning: a review," *EURASIP Journal on Wireless Communications and Networking*, vol. 2017, no. 1, pp. 1–12, 2017.

[24] F. M. Rueda and G. A. Fink, "Learning attribute representation for human activity recognition," in *Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 523–528, IEEE, Beijing, China, August 2018.

[25] X. Zhao, L. Sang, G. Ding et al., "Recurrent attention model for pedestrian attribute recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 9275–9282, Honolulu, HI, USA, February 2019.

[26] J. Wang, X. Zhu, S. Gong et al., "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2275–2284, Salt Lake City, UT, USA, June 2018.

[27] F. Zhang, T. Y. Wu, J. S. Pan et al., "Human motion recognition based on SVM in VR art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, pp. 1–15, 2019.

[28] G. Raz, M. Svanera, N. Singer et al., "Robust inter-subject audiovisual decoding in functional magnetic resonance imaging using high-dimensional regression," *Neuroimage*, vol. 163, pp. 244–263, 2017.

[29] C. J. Chou, J. T. Chien, and H. T. Chen, "Self adversarial training for human pose estimation," in *Proceedigns of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 17–30, IEEE, Honolulu, HI, USA, November 2018.

[30] G. Rogez, P. Weinzaepfel, and C. Schmid, "Lcr-net++: multi-person 2d and 3d pose detection in natural images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 5, pp. 1146–1161, 2019.

[31] A. Okuno, T. Ishikawa, and H. Watanabe, "Rollover detection of infants using posture estimation model," in *Proceedings of the 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pp. 490–493, IEEE, Kobe, Japan, October 2020.

[32] S. Liu, W. Yu, F. T. S. Chan, and B. Niu, "A variable weight-based hybrid approach for multi-attribute group decision making under interval-valued intuitionistic fuzzy sets," *International Journal of Intelligent Systems*, vol. 36, no. 2, pp. 1015–1052, 2021.

[33] X. Zenggang, T. Zhiwen, C. Xiaowen et al., "Research on image retrieval algorithm based on combination of color and shape features," *Journal of Signal Processing Systems*, vol. 93, pp. 1–8, 2019.

[34] J. Yang, C. Wang, B. Jiang et al., "Visual perception enabled industry intelligence: state of the art, challenges and prospects," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2204–2219, 2020.

[35] K. Sim, J. Yang, W. Lu, and X. Gao, "MaD-DLS: mean and deviation of deep and local similarity for image quality assessment," *IEEE Transactions on Multimedia*, p. 1, 2020.

[36] J. Guo, Y. Zhao, Y. Jiang, and H. Song, "Coverage guided differential adversarial testing of deep learning systems," *IEEE Transactions on Network Science and Engineering*, p. 1, 2020.