IJLAI Transactions on Science and Engineering (ISSN: 3006-3418) 2024, vol. 2, no. 3, pp: 36–46

A Review of Multi-modal Human Motion Recognition Based on Deep Learning

Ye Li[1], Yifan Pan[1], and Xinhui Wu[1]

Software College, Shenyang Normal University
Shenyang 110034 China
liye@163.com
Received June. 28, 2024; Revised and Accepted July. 25, 2024

Abstract. Human motion recognition is a research hotspot in the field of computer vision, which has a wide range of applications, including biometrics, intelligent surveillance and human-computer interaction. In vision-based human motion recognition, the main input modes are RGB, depth image and bone data. Each mode can capture some kind of information, which is likely to be complementary to other modes, for example, some modes capture global information while others capture local details of an action. Intuitively speaking, the fusion of multiple modal data can improve the recognition accuracy. In addition, how to correctly model and utilize spatiotemporal information is one of the challenges facing human motion recognition. Aiming at the feature extraction methods involved in human action recognition tasks in video, this paper summarizes the traditional manual feature extraction methods from the aspects of global feature extraction and local feature extraction, and introduces the commonly used feature learning models of feature extraction methods based on deep learning in detail. This paper summarizes the opportunities and challenges in the field of motion recognition and looks forward to the possible research directions in the future.

Keywords: Human motion recognition, Computer vision, Multi-modal, Deep learning.

1. Introduction

In recent years, with the popularity of video shooting equipment such as digital cameras and smart phones and the extensive promotion of video application software, network video business has shown an exponential growth trend, and video carriers have become an important medium for the dissemination of information in People's Daily production and life. There is huge information hidden in video. The huge number of users and the rapid growth of the network video market have brought great challenges to the management, storage and identification of network video, so the network video business has been increasingly valued by all parties [1,2]. In the research field of human-focused Computer Vision (CV) Such as Hand Gesture Recognition [3], Human Pose Estimation [4], Gait Recognition and other tasks, Human Action Recognition (HAR) [5] task is widely used in many fields such as human-computer interaction, smart home, automatic driving, virtual reality, etc. It is becoming an important research topic in the field of computer vision. The main task of human action recognition in video is to help the computer identify the human action in the video, and deduce the human movement pattern by analyzing the video content, so as to establish the mapping relationship between the video information and the categories of human action. The accurate identification of human movements in videos is conducive to the unified classification and management of massive related video data by Internet platforms, and helps to create a harmonious network environment. In addition, the development of HAR technology also promotes the maturity of video anomaly monitoring business, which can assist social security administrators to quickly predict crisis events in public places, and timely monitor users' abnormal behaviors (such as fainting, wrestling, etc.) in family life in order to seek medical treatment in time [6,7]. Therefore, it is of great academic significance and application value to study the task of human motion recognition in video.

The realization process of action recognition task can be divided into two steps: action representation and action classification. Action representation, also known as feature extraction, is considered to be the most important task of action recognition. In this paper, feature extraction algorithms related to human action recognition are mainly divided into traditional manual feature extraction methods and deep learning-based methods to extract manually designed features and trainable features from videos [8]. Traditional feature extraction methods rely on expertise in related fields and often need to design specific features according to different tasks. The performance of recognition algorithms relies heavily on the database itself, which increases the complexity of processing on different data sets, resulting in poor generalization ability and generality. Moreover, in the era of information explosion, the explosive growth of video data undoubtedly brings great challenges to the production of manual

features, so people are more inclined to use non-artificial methods to extract more general feature representations to meet the needs of realistic tasks.

Deep Learning (DP) breakthroughs in speech, image recognition and other fields have encouraged its application in the field of computer vision [9,10]. With the explosion of massive data and the rapid development of GPU and other hardware devices, deep learning is more in line with the characteristics of The Times, which improves the possibility of quickly mining useful information from large-scale video data, and gradually becomes an indispensable research method in HAR task. The method based on deep learning constructs a hierarchical learning and training mode, establishes a progressive learning mechanism between input and output data with the help of models and labels, and independently obtains the action representation of the original video data, thus overcoming the defects of manual feature design, and is a more efficient feature extraction method with better generalization performance.

This paper classifies and summarizes feature extraction methods in the field of human action recognition from two aspects: traditional manual feature extraction methods and deep learning-based feature extraction methods. Finally, it summarizes the difficulties and challenges in the field of action recognition, and summarizes the possible research directions in the future.

2. Moving Object Detection Technology

2.1. Detection of Moving Objects

Object detection has many applications, such as human motion recognition, face recognition and pedestrian detection. The detection of moving objects is the first step in the video analysis process, which can be done in each frame or when the object first appears in the video. It handles the movement of objects of interest by eliminating stationary background objects. In the field of human motion, a good target detection algorithm is very important and often has a huge impact on the experimental results. At present, common target detection algorithms are mainly divided into three types [11,12].

By calculating the difference between two or more frames, the interframe difference method can further distinguish the moving object. The method assumes that the adjacent frames in the video sequence have relatively strong correlation, and the calculation is simple and easy to implement. For a variety of dynamic environments, this method has strong adaptability, and can effectively detect pixels that do not change significantly in adjacent frames, especially if the scene changes due to motion. However, the final result of non-stationary target detection is not accurate because it is difficult to obtain the full contour of the moving target18. And for the target movement too fast and too slow are not suitable, too fast will produce ”ghost”, too slow will produce ”holes”.

The first step of background subtraction method is background modeling to get the reference model. The presence of moving objects is determined by the pixel difference between the current video frame and the previous frame, which is utilized to achieve the detection of moving objects. There are two common methods of background subtraction: recursive and non-recursive. Recursive algorithms do not involve buffering of background estimates, but recursively update a single background model for each given input frame. Therefore, the current background may be affected by input frames from the distant past, and the advantage of this algorithm is that it requires less storage space, but it will last longer in the case of background errors. For example, adaptive background, Gaussian mixing and approximate median methods are recursive methods. The background estimation involved in non-recursive algorithms uses a sliding window method, which stores a buffer of the first few video frames and uses the change of each pixel within the buffer to estimate the background image [13-15]. The technique does not rely on the history of the frame stored in the buffer, so it has a strong adaptability, and the method can detect the complete moving area well, but in the case of large buffers that need to deal with slow traffic, the storage requirements will be significant, thus limiting its application scope. In addition, it is also sensitive to noise and local motion.

The optical flow field of the image is calculated by optical flow method, and cluster processing is carried out according to the optical flow distribution characteristics of the image. It uses the vector characteristics of each pixel in each frame image in the video sequence to detect the moving region in the video. Once the target moves, the optical flow vector formed by the moving target will change, thus realizing the detection of the moving target. This method can obtain complete target detection and motion information from the background, but the calculation is complicated and the anti-noise performance is poor, which is difficult to meet the requirements of real-time applications.

2.2. Action Segmentation

Actions in real life occur continuously in time, without any beginning or end mark, and may be a combination of consecutive actions in space and time, with some corresponding transitional actions in the middle. The function

of action segmentation is to separate these basic actions from the continuous video sequence to obtain data containing sufficient motion information and smaller capacity. At present, most of the literature on human movement recognition is based on the recognition of the previous timely segmentation of the movement. Therefore, motion segmentation is also a very important work, and its goal is to find a way to detect and determine the moment of human motion in the video sequence. Previous studies on human motion segmentation can be divided into three categories:

1. Supervised methods.
   There are real tags available, and the model is built on a video dataset of these tags. In order to avoid the high cost of manual annotation for training, Zellers et al. [16] used movie script as a means of weak supervision, but the script only provided implicit, noisy and inaccurate information about the type and location of operations in the video, so they used kernel-based discriminant clustering algorithm to solve this problem. The algorithm locates the operation in the weakly labeled training data, and then uses the obtained action sample to train the time action detector. Cao et al. [17] trained a single HMM according to the STIP characteristics of each dynamic posture, and then used the cascading HMM to automatically segment and mark the entire Sun Salutation yoga movement sequence.
2. Unsupervised methods.
   Video sequences are modeled without using real labels and have clear training phases. By extracting spatio-temporal interest points, Niebles et al. [18] represented video sequences as a collection of spatio-temporal words. The algorithm was implemented by using latent topic models, such as probabilistic latent semantic analysis models and latent Dirichlet assignments to automatically learn the probability distribution of space-time words and intermediate topics corresponding to human action classes [19].
3. Unsupervised segmentation technique.
   Split different actions in videos of human activity without training. The first set of techniques is to cluster from all videos composed of frames to extract meaningful action primitives. Steenbeek et al. [20] first described continuous actions as a series of dynamic systems that significantly enhanced the expressive power of the model while retaining many of the computational advantages of using dynamic models. Second, methods were derived to incorporate view and rate invariance into these models in order to cluster similar actions together. Finally, algorithms were proposed to learn model parameters from video streams and demonstrate how to cluster a single video sequence into different clusters, where each cluster represented an action. The second set of techniques, such as Aligned Cluster Analysis (ACA) and Hierarchical Aligned Cluster Analysis, which addressed variability in the time scale of human behavior. The oligomerization analysis proposed by Suslick [21] et al. was a robust method to temporarily segment the motion capture data stream into actions. The ACA extended standard kernel K-means clustering in two ways: clustering means to include a variable number of features, and the Dynamic Time Warping (DTW) kernel was used to achieve time invariance. The HACA was an unsupervised hierarchical bottom-up framework called hierarchical pair oligomerization analysis [22]. HACA found m disjoint segments for the partition of a given multidimensional time series, such that each segment belonged to one of k clusters, and combined kernel K-means with generalized dynamic time alignment kernel into the cluster time series data. These unsupervised segmentation techniques divided video data into m segments of variable length and cluster these segments into one of k action clusters.

To sum up, motion segmentation is a very challenging work, and a fast, accurate and robust motion segmentation algorithm for a motion speed and time scale is a very important link in the field of intelligent recognition.

3. Motion Recognition Based on RGB Video

RGB is one of the important channels of RGB-D data. Compared to depth images and bone data modes, the main features of RGB data are shape, color, and texture, which bring the advantages of extracting points of interest and optical flow. This section mainly summarizes some current human motion recognition methods for RGB modes from two aspects: manual features and deep learning.

3.1. Manual Feature Method Based on RGB Video

In the field of motion recognition, silhouette features, optical flow features, temporal and spatial features are widely used. External factors such as clothing color, lighting conditions, and background clutter are not helpful in the recognition task, and silhouette features are less affected by these irrelevant features of the image, so they are widely adopted. Hebbar et al. [23] proposed a spatio-temporal contour representation called Silhouette Energy Image (SEI), as well as a variety of variable action models to characterize motion and shape attributes to automatically identify human behavior in daily life. They built a variable (or adaptive) model based on the SEI and

suggested parameters, combining the spatio-temporal properties of the energy template (SEI and variable action model) to represent a global motion descriptor as a representation of shape and motion in human motion recognition. Lin et al. [24] showed that occlusions in the interaction space of dynamic objects could be detected, and that their 3D shape was fully restored as a result of silhouette reconstruction shape, and provided a Bayesian sensor fusion formula to process all occlusions cues that occurred in multi-view sequences, which could represent human movements. However, the problem that the silhouette was shielded by the human body had not been well solved. Optical flow feature is another important feature in motion recognition. Optical flow is defined as the motion of a single pixel on the image plane. Optical flow is often a good approximation of the real physical motion projected onto the image plane, and most methods for calculating optical flow assume that the color/intensity of a pixel is constant as it shifts from one video frame to the next. Wang et al. [25] used optical flow to realize remote motion recognition, and introduced a novel motion descriptor based on optical flow measurement into the spatio-temporal volume of each stable human body image, and regarded optical flow as a spatial mode of noise measurement. Instead of precise pixel displacement, these modes were carefully smoothed and aggregated to form spatiotemporal motion descriptors, which were finally identified using the relative similarity measure of the nearest neighbor frame. Liu et al. [26] proposed a new method of motion representation based on optical flow analysis and Random Sample Consensus (RANSAC) method. RANSAC was an iterative method of estimating the parameters of a mathematical model from a set of observations containing interior points and outliers, and could be used to filter out any unwanted points of interest around a scene and retain only those that were relevant to a particular point. In this way, the area of the human body within the frame was estimated, and this rectangular area was divided into multiple smaller areas or blocks, and then the percentage of the frame-by-frame change of the point of interest in each block was recorded, and the matrix constructed by the strategy was used as the feature vector for a particular action.

3.2. Deep Learning Method Based on RGB Video

In recent years, deep learning methods have also been used to achieve motion recognition in RGB video. The methods using deep learning are mainly divided into two categories: CNN and RNN.

For CNN-based methods, there are currently four main methods for encoding spatiotemporal structure information. The first approach uses CNNS to extract features from individual frames and then fuses time-domain information. Zhang et al. [27] studied four time fusion methods and proposed the concept of slow fusion, that is, higher levels can obtain more global information in space and time dimensions. This realization extends the connectivity of all convolutional layers in time and realizes time convolution. The second method is to extend the convolution operation to the time domain. Chen et al. [28] developed a new 3D-CNN action recognition model. The model extracted features from space and time dimensions by performing 3D convolution, thereby capturing motion information encoded in multiple adjacent frames. The developed model generated multiple channels of information from the input frame, and the final feature representation combined information from all channels. However, this method broke down the video sequence into short segments and aggregated the video level information by late fraction fusion.

4. Traditional Manual Feature Extraction Method

Most traditional motion recognition algorithms rely on manual feature design methods, whose purpose is to analyze the motion pattern of human body from the original video input and extract the corresponding underlying features, and transform the video data information into feature vectors that can be understood by the classification model, so as to map the original video data into the corresponding action category labels. Video data not only contains static scene information, but also contains rich dynamic changes. Therefore, for video classification, robust video feature representation not only meets the two basic characteristics of differentiation and effectiveness, but also needs to contain a lot of time information and spatial information, which increases the difficulty of manual feature design. Traditional manual features are mainly divided into global features and local features.

4.1. Global Feature Extraction

The global feature representation of the motion is based on the overall description of the moving target, usually the human body in the video needs to be segmented by background subtraction or target tracking, and then the global feature extraction is carried out. Common global features include human contour-based features, skeleton-based features and optical flow-based features.

1. Features based on human body contours. Most of the early action recognition studies relied on human body contour features, and reflected the arrangement and combination of human action sequences in the time domain through specific algorithm design, which usually required the establishment of sample templates for each action category in advance. In the classification process, the action template to be measured is compared with the established standard template, and then the template with the greatest similarity is selected as its final classification result. Common human body contour features include Motion Energy Images (MEI) [29] and Motion History Images (MHI) [30]. Human movement information is preserved by observing coarse-grained image movement associated with a given action from a specific direction. Among them, MEI describes the position movement of human body in space and the spatial distribution of its energy, reflecting the contours of movement and the intensity of movement. By observing the brightness change of the human body in the same position in a certain period of time in the video frame, MHI reflects the time when motion occurs and its change in the time domain. Reference [31] extracted the enhanced Gabor tensor features of moving fragments on the basis of capturing their MEI, and finally made subspace projection to obtain effective motion descriptors. Based on the MHI features of image sequences, reference [32] used different geometric moments to encode the features, which improved the computational efficiency without loss of information. Feature extraction methods based on human contours are widely used because of their low computational cost and strong robustness [33].

   Human profile is a vision-based description method, which is easy to be affected when the observation direction and camera position change, resulting in inaccurate recognition results. In addition, the accuracy of the standard template required in the classification process is high, and the accuracy of the template depends on the sample size of the database to support its calculation.
2. Features based on human models. Since human motion patterns can be abstracted into skeleton movements represented by simple geometric structures, relevant research based on human models has also become an important direction in the field of HAR, which intuitively describes human movements through the changes of human nodes between video frames, and can be divided into two-dimensional representations and three-dimensional representations [34]. The two-dimensional model uses two-dimensional geometric shapes (such as rectangle, oval, irregular shape) to represent each part of the human body, and then estimates the corresponding model parameters and matches the corresponding model through the underlying apparent features extracted from the image, so as to distinguish different body areas such as the head, body, and limbs, and describes the specific movement pattern through the movement and deformation of each geometric figure. In reference [35], irregular two-dimensional silhouette images were used to approximate the human motion contour, and the histogram of graph nodes was extracted from them to obtain the classification feature vector. This method did not require accurate positioning of human joint information and saved calculation expenses. However, the two-dimensional model could not represent the distance information of the human body in the process of movement, so when self-occlusion and collision occur in the process of movement, there could be a large error in the estimated motion mode. In order to alleviate the above problems, reference [36] used depth cameras to estimate the positions of different human skeletal joints, and used the sequence of Angle changes between joints to depict human movements. Reference [37] obtained static human model data through 3D scanning equipment, and then used skin algorithm to bind bone data, so as to reconstruct real-time motion mode. Skeleton features accurately represented the static human posture, but weakened the time evolution of actions. Therefore, reference [39] combined skeleton features with RGB data to construct time images based on RGB data to represent the dynamic changes of actions. Occlusion of human body parts will seriously affect the motion recognition accuracy based on bone data. Depth information contains rich distance information, which alleviates the occlusion problem of bone data. Therefore, reference [40] combined the advantages of the two modes of depth information and bone data to avoid the defects of a single input mode. The three-dimensional model used geometric models such as cylinder or cone to correspond to the human body construction mode, and estimated the relevant data by combining the prior information such as human kinematics and depth information, overcoming the defects of the two-dimensional model in dealing with problems such as self-occlusion and motion collision.

   By applying a unified human model to represent any individual, the mannequin-based approach alleviates the intra-class differences caused by individual changes to a certain extent, but the simple simplification of complex human movements into a rigid geometric model and the simple use of node changes for action recognition will produce large errors. In addition, the depth information required for 3D models needs to be captured by expensive camera equipment, and the construction of models will be more complicated.
3. Characteristics based on optical flow. Optical flow is generally generated by the movement of the foreground target itself, the shift of the camera shooting Angle or the simultaneous occurrence of two phenomena. Its calculation is based on the assumption that the brightness change of the image only comes from the movement of the object, and the brightness change of pixels on adjacent frames is used to reflect the movement of the

human body in the time domain. Reference [41] described long-distance human motion through a descriptor based on computational optical flow, which approximated the smooth trajectory of human motion by tracking each stable human image trajectory and calculating its fuzzy form of optical flow instead of the precise pixel displacement. Reference [42] combined optical flow characteristics with MHI to accurately track the motion state of moving objects in a certain period of time. Optical flow features have been widely used in the field of motion recognition because of their good ability to represent motion in time dimension. However, optical flow features are easily affected by illumination and occlusion, and models using optical flow data as input have large memory requirements and high computational costs.

In general, the representation method based on global features is limited by factors such as camera movement and illumination change, and needs to remove pre-processing operations such as background, foreground extraction, human body positioning and tracking, so it has poor performance in the representation of motion under complex dynamic backgrounds.

4.2. Local Feature Extraction

In order to avoid preprocessing, the local feature representation method focuses on the detection of interest points in the video, and further encodes the local representation of human motion into feature vectors for classification stage, which has good classification effect in specific action recognition tasks. Common local features include features based on spatiotemporal interest points and features based on trajectories.

1. Features based on space-time interest points. The feature extraction method based on spatiotemporal interest points can be divided into two parts: the detection of interest points and the description of feature points. Firstly, the detector detects the spatio-temporal points of interest, the points that suddenly change in the air, and uses the point set composed of spatio-temporal points of interest to represent human actions. Then, descriptors are used to encode the points of interest into feature vectors that the classifier can understand, so as to describe the action information. Because it is easy to capture and insensitive to visual changes, it is highly respected in motion recognition tasks under complex background. The Harris3D feature detector extends the corner detection on the spatial domain to the time domain, and fuses the histogram of gradient (HOG) feature and the optical flow histogram (HOF) feature to obtain local descriptors, and then describes the local motion. On this basis, reference [42] used hash method and sparse coding method to improve the final feature coding, but the descriptors generated by this method were sensitive to noise, scale and Angle changes. In order to overcome the above shortcomings, reference [43] used a Scale-Invariant Feature Transform (SIFT) algorithm to detect the key points in the frame sequence, but this method only considered the appearance information in the spatial dimension, ignoring the evolution of human motion in the time dimension. Therefore, the 3D SIFT operator [44] extends the SIFT operator in time dimension in order to accurately describe the spatiotemporal characteristics of video data and obtain a good local spatiotemporal feature descriptor. However, for fuzzy images and images with smooth edges, fewer feature points are detected, which increases the difficulty of action recognition.

   Thanks to the development of corner detector, feature detection based on spatiotemporal interest points is easy to extract and widely used. However, it uses the set form of some unrelated points to describe human movement information and is limited to the complexity of human movement in real scenes, so it is difficult to get practical application of this technology.
2. Track-based features. The trajectory of human movement contains abundant motion information, and the d-ifference of the mutation points of the trajectory speed and direction represents different types of motion. The track-based feature extraction method mainly includes three steps: intensive sampling and tracking of the feature points, track-based feature extraction and feature encoding. In order to effectively capture motion information, literature [45] sampled the dense points of local modules of each frame at different scales and tracked them in dense optical flow field to extract the dense trajectory of moving objects. At the same time, starting with the underlying features of the image, the performance was further improved by combining HOG and HOF features on each dense point. Reference [46] defined trajectory motion correlation to determine the corresponding weights of different trajectories in the classification process, so as to weigh the motion trajectories that were more relevant to the target actions. In order to extract high-quality trajectory features, reference [47] improved the dense trajectory features by compensating the camera motion. When tracking the trajectory of the human body in the video, features such as HOG, HOF, MBH and dense trajectory were extracted along the trajectory of the optical flow field, and the features were compiled by using the two methods of characteristic word bag (BoW) or Fisher vector (FV) code to obtain the final video feature representation. Then support vector machine was used to encode the extracted feature representation to a fixed size for the final classification recognition.

Compared with global features, local features do not need to accurately locate the human body, are insensitive to interference such as Angle changes, complex scenes and occlusion, have good stability, strong anti-interference ability, and avoid preprocessing operations. However, manual feature coding requires a large memory overhead. Moreover, local features lack detailed information in appearance, and require additional expertise to design features in specific domains, which has domain limitations and is difficult to generalize.

4.3. Three-dimensional Convolutional Network

The general approach of feature extraction algorithms based on 3D convolutional networks is to take a small number of continuous video frames stacked into a spatiotemporal cube as the model input, and then adaptively learn the spatiotemporal representation of video information through hierarchical training mechanism under the supervision of a given action category label. 3D convolutional networks directly capture distinguishing video features from video data at the same time in both spatiotemporal dimensions without deliberately designing spatiotemporal feature fusion modules, which can effectively deal with the fusion problem of short-term spatiotemporal information and better promote the interaction of spatiotemporal features in the process of recognition and judgment.

1. Model based on standard three-dimensional convolution. Reference [48] extended the two-dimensional convolutional network to three-dimensional space and extracts video features from spatiotemporal dimension. On this basis, a variety of 3DCNN variants were proposed, such as C3D, 13D, Res3D, etc. Thanks to the development of GPU, the method based on 3DCNN has gradually become the mainstream method in the field of video action recognition. Reference [49] used multi-view learning to extract multiple local shape descriptors, and then combined multiple view descriptors with 3DCNN to improve the description ability of classification features. Reference [50] showed that human beings did not pay attention to the whole content when observing the surrounding environment, but focused their attention on the significant areas of the environment. Inspired by this, some scholars introduced an attention mechanism into the design of feature extraction algorithm to help the model allocate more attention resources to the target region in the process of feature learning, thus suppressing redundant information and quickly screening out key information in complex video content. A Convolutional Block is proposed (CBAM) [51] built a hierarchical dual attention mechanism on the basis of the two-dimensional residual network structure, adding channel attention and spatial attention to each residual block in sequence, but the network structure ignored the time information which is crucial to the action recognition task.
   The model based on standard three-dimensional convolution structure has inherent advantages in extracting local spatiotemporal fusion features due to its inherent structure, but it also has many limitations. The number of model parameters required to train for the model based on the standard three-dimensional convolution structure is very large, which increases the computational complexity and storage overhead of the model, and is not conducive to the iterative optimization of the model, resulting in the difficulty for the model to rapidly converge to the optimal solution.
2. Model based on three-dimensional convolution structure deformation. In order to reduce the training parameters of the model, improve the computing speed and reduce the memory consumption, a variety of structural deformations based on standard 3D convolutional networks are proposed. In earlier related studies, researchers approximated a standard three-dimensional convolutional layer with a convolution kernel size of $3 \times 3 \times 3$ to three cascades of convolutional layers with filter sizes of $1 \times 3 \times 1$, $1 \times 1 \times 3$ and $3 \times 1 \times 1$, respectively, to improve the effectiveness of the model [], but this approach is equivalent to deepening the depth of the model three times, resulting in difficult training of the model.
   Computational complexity and memory consumption limit the length of input video data, so the feature extraction model based on 3D convolutional networks can only represent human movements in a short time range. It is difficult to process video data information with a long time span, thus affecting the performance of the model. Therefore, whether the long-term spatio-temporal sequence information can be fully analyzed is the key to improve the accuracy of video action classification.

5. Conclusion and Future Wortks

The method based on traditional manual feature extraction requires huge memory overhead and computing cost, and relies on the prior knowledge of domain experts, which has strong subjectivity. In many cases, the method based on deep learning has better performance. The feature extraction method based on deep learning benefits from the hierarchical training mode of the neural network, and automatically extracts high-dimensional features from the original video data through the progressive feature extraction mechanism, and fully captures the context semantic information of the video data, thereby increasing the description ability of the model and facilitating

the final recognition and judgment. Feature extraction is directly related to whether the video content can be accurately and fully expressed, and then affect the classification results. However, in the face of the explosive growth of video data, the increasingly complex video content and the real needs of real-time analysis, video feature extraction methods also put forward higher requirements for effectiveness, robustness and timeliness. The challenges in video feature extraction methods and possible future research directions are summarized as follows:

1. Multi-feature fusion. Different types of input are processed by the feature extraction model to obtain different types of features, which describe the human motion pattern in the video from different aspects. The emphasis of each feature is different, so it is easy to use a single feature for subsequent recognition and judgment, which will lead to wrong classification results. Many models are directly based on RGB data for feature extraction. With the application and development of camera equipment, RGB data has the advantages of easy acquisition and rich fine-grained information, and its corresponding features can directly reflect the apparent and detailed texture information of objects. However, due to factors such as camera jitter, environmental illumination and occlusion in the process of video acquisition, RGB data usually carries a large amount of background noise, resulting in the complexity and variability of the space-time dimension of video data, resulting in a large intra-class gap between the same actions of different individuals, and thus affecting the video representation ability of classification features. Combining features of different categories can combine the advantages of each feature to avoid the defects of single feature classification task. At present, some researchers overcome the sensitivity of RGB data to background noise by combining depth information in video data, but the acquisition cost of depth information is high and the recognition accuracy is not ideal. Therefore, it is worth discussing to design more simple and effective additional input modes to generate different types of features, and to characterize human body motion modes through the fusion of multiple features, and to synthesize the advantages of multiple features by utilizing the complementarity between different features.
2. Representation of dynamic information. Dynamic motion information is the content of multi-frame difference in video data, which is used to describe the motion history. How to design a feature extraction mechanism to accurately describe the dynamic evolution of human motion in time dimension is of great significance for the correct differentiation of human motion in video. Some researchers use the optical flow characteristics in the video to characterize the dynamic information of the human body, eliminating the influence of irrelevant background factors while compensating the time information. Although the accuracy is improved, the complexity of optical flow calculation and the memory cost are high, which greatly reduces the effectiveness and practicability of the model. In addition, the optical flow characteristics often need to be calculated in advance, and the generation of optical stream video requires a lot of time and cost, which can not achieve the effect of real-time classification and prediction. Therefore, in order to meet the practical requirements, it is of great practical significance to seek a simple and efficient dynamic representation instead of complex optical flow calculation to reduce memory consumption. In order to meet the real-time requirements, it is also an urgent problem to integrate dynamic feature extraction into the action recognition network for real-time predictive analysis.
3. Feature screening. Video data contains a lot of redundant information, if all features are treated equally, it will lead to a lot of unnecessary features in the process of feature extraction, which will interfere with the recognition results and increase the amount of redundant computation. The attention mechanism can imitate the visual attention mechanism used by humans to observe the world, focusing on the core objects in the space region and the action fragments in the time dimension. In recent years, researchers have devised different mechanisms for space-time attention, They tend to focus on the relevant research of frame-level space-time attention to assist the model to automatically screen important video frames and their corresponding prominent spatial regions. However, the action information contained in adjacent frames is almost the same, so it is difficult to distinguish their importance. Some scholars try to solve the above problems by adding complex regularization, but the calculation amount and complexity of the model also increase Therefore, it is also worth studying to shift the research focus from frame-level attention to clip-level attention and assign different importance scores to different video clips. In addition, different convolutional checks should have different channels to extract different categories of features, so the features corresponding to different channels should also be treated differently. In summary, how to adjust the attention mechanism to assist the model to flexibly select key features is the key to improve the discriminant ability of the final classification features.
4. Multimodal feature mining. At present, most researches on human movement recognition only consider the visual features in the video, and classify human movements based on the intuitively perceived video images. However, the video data in real life not only contains image features, but also contains a lot of voice information and text information. The full use of these data types can assist the model to further explore the deep features and then understand the video content. How to combine the data types of different attributes in the video, display and mine the information contained in all kinds of data, and make use of the complementary

characteristics between the multi-modal features is the key to assist the model to determine the action category and improve the recognition accuracy. Although the introduction of multi-modal data increases the connection between different data types, the mining of multi-modal features requires model training on different data sets, and feature extraction and category prediction for each mode respectively. This also means that the complexity of the model and the cost of model training increase, so designing a model that is easy to train and optimize to generate simple and effective multimodal representation is also a worthy direction of exploration.

In recent years, the application field of human action recognition technology is more and more extensive, covering many fields such as automatic driving, robot and intelligent monitoring, which has important practical significance. This paper gives a comprehensive overview of the feature extraction methods involved in the field of human action recognition in video, summarizes the research status of traditional manual feature extraction methods and feature extraction methods based on deep learning, and analyzes the advantages and disadvantages of various methods. Finally, the existing challenges and possible future research directions in the field of human motion recognition are summarized, in order to help future researchers to understand the relevant research status of feature extraction algorithms in human motion recognition tasks more clearly and explicitly.

6. Conflict of Interest

The authors declare that there are no conflict of interests, we do not have any possible conflicts of interest.

Acknowledgments. This work was supported by the 2021 Scientific research funding project of Liaoning Provincial Education Department (Research and implementation of university scientific research information platform serving the transformation of achievements).

References

1. Shaw, Kenneth, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos [C] Conference on Robot Learning. PMLR, 2023.
2. Yang, Zhiwei, et al. Video event restoration based on keyframes for video anomaly detection [C] Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
3. Tchantchane, Rayane, et al. A review of hand gesture recognition systems based on noninvasive wearable sensors [J]. Advanced Intelligent Systems, 5.10 (2023): 2300207.
4. Zheng, Ce, et al. Deep learning-based human pose estimation: A survey [J]. ACM Computing Surveys, 56.1 (2023): 1-37.
5. Gedamu, Kumie, et al. Relation-mining self-attention network for skeleton-based human action recognition [J]. Pattern Recognition, 139 (2023): 109455.
6. Yin S, Li H, Sun Y, et al. Data Visualization Analysis Based on Explainable Artificial Intelligence: A Survey[J]. IJLAI Transactions on Science and Engineering, 2024, 2(2): 13-20.
7. Yu J, Lu Z, Yin S, et al. News recommendation model based on encoder graph neural network and bat optimization in online social multimedia art education[J]. Computer Science and Information Systems, 2024. doi: 10.2298/CSIS231225025Y.
8. Dallel, Mejdi, et al. Digital twin of an industrial workstation: A novel method of an auto-labeled data generator using virtual reality for human action recognition in the context of humanCrobot collaboration [J]. Engineering applications of artificial intelligence, 118 (2023): 105655.
9. Sharifani K, Amini M. Machine learning and deep learning: A review of methods and applications[J]. World Information Technology and Engineering Journal, 2023, 10(07): 3897-3904.
10. Yin S, Li H, Laghari A A, et al. An anomaly detection model based on deep auto-encoder and capsule graph convolution via sparrow search algorithm in 6G internet-of-everything[J]. IEEE Internet of Things Journal, 2024. DOI: 10.1109/JIOT.2024.3353337.
11. Liu Y, Liu X, Zhang B. Retinanet-vline: a flexible small target detection algorithm for efficient aggregation of information[J]. Cluster Computing, 2024, 27(3): 2761-2773.
12. Liu Y, Jiang D, Xu C, et al. Deep learning based 3D target detection for indoor scenes[J]. Applied intelligence, 2023, 53(9): 10218-10231.
13. Yang R, Li W, Shang X, et al. KPE-YOLOv5: an improved small target detection algorithm based on YOLOv5[J]. Electronics, 2023, 12(4): 817.
14. Zhang Q, Li Y, Guo C, et al. Marine radar monitoring IoT system and case study of target detection based on PPI images[J]. Expert Systems, 2024, 41(7): e13333.
15. Jiang Y, Yin S. Heterogenous-view occluded expression data recognition based on cycle-consistent adversarial network and K-SVD dictionary learning under intelligent cooperative robot environment[J]. Computer Science and Information Systems, 2023, 20(4): 1869-1883.
16. Zellers R, Lu X, Hessel J, et al. Merlot: Multimodal neural script knowledge models[J]. Advances in neural information processing systems, 2021, 34: 23634-23651.

17. Cao Y, Xie R, Yan K, et al. Novel dynamic segmentation for human-posture learning system using hidden logistic regression[J]. IEEE Signal Processing Letters, 2022, 29: 1487-1491.
18. Cherian A, Wang J, Hori C, et al. Spatio-temporal ranked-attention networks for video captioning[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020: 1617-1626.
19. Yutong G, Khishe M, Mohammadi M, et al. Evolving deep convolutional neural networks by extreme learning machine and fuzzy slime mould optimizer for real-time sonar image recognition[J]. International Journal of Fuzzy Systems, 2022, 24(3): 1371-1389.
20. Steenbeek H, van Geert P. The emergence of learning-teaching trajectories in education: A complex dynamic systems approach[J]. Nonlinear dynamics, psychology, and life sciences, 2013, 17(2): 233-267.
21. Suslick B A, Alzate-Sanchez D M, Moore J S. Scalable Frontal Oligomerization: Insights from Advanced Mass Analysis[J]. Macromolecules, 2022, 55(18): 8234-8241.
22. Shmilovich K, Mansbach R A, Sidky H, et al. Discovery of self-assembling -conjugated peptides by active learning-directed coarse-grained molecular simulation[J]. The Journal of Physical Chemistry B, 2020, 124(19): 3873-3891.
23. Hebbar S A, Sharma R, Somandepalli K, et al. Vocal tract articulatory contour detection in real-time magnetic resonance images using spatio-temporal context[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7354-7358.
24. Lin W, Zheng C, Yong J H, et al. Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1736-1745.
25. Wang L, Guo Y, Liu L, et al. Deep video super-resolution using HR optical flow estimation[J]. IEEE Transactions on Image Processing, 2020, 29: 4323-4336.
26. Liu Y, Miura J. RDMO-SLAM: Real-time visual SLAM for dynamic environments using semantic label prediction with optical flow[J]. IEEE Access, 2021, 9: 106981-106997.
27. Zhang J, Wang W M, Yang X H, et al. Double-cone ignition scheme for inertial confinement fusion[J]. Philosophical transactions of the Royal Society A, 2020, 378(2184): 20200015.
28. Chen C F R, Panda R, Ramakrishnan K, et al. Deep analysis of cnn-based spatio-temporal representations for action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 6165-6175.
29. Li M, Shen Y, Luo K, et al. Harnessing dislocation motion using an electric field[J]. Nature Materials, 2023, 22(8): 958-963.
30. Knuth F, Shean D, Bhushan S, et al. Historical Structure from Motion (HSfM): Automated processing of historical aerial photographs for long-term topographic change analysis[J]. Remote Sensing of Environment, 2023, 285: 113379.
31. Nahak S, Pathak A, Saha G. Fragment-level classification of ECG arrhythmia using wavelet scattering transform[J]. Expert Systems with Applications, 2023, 224: 120019.
32. Xu G, Wang X, Ding X, et al. Iterative geometry encoding volume for stereo matching[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 21919-21928.
33. Jisi A, Yin S. A new feature fusion network for student behavior recognition in education[J]. Journal of Applied Science and Engineering, 2021, 24(2): 133-140.
34. Rani S, Lakhwani K, Kumar S. Knowledge vector representation of three-dimensional convex polyhedrons and reconstruction of medical images using knowledge vector[J]. Multimedia Tools and Applications, 2023, 82(23): 36449-36477.
35. Li L, Wang J, Yang S, et al. Estimate of three-dimensional Wadell roundness of irregular particles using image processing and topographic analysis[J]. Construction and Building Materials, 2023, 396: 132273.
36. Xing Q, Hong R, Shen Y, et al. Design and validation of depth camera-based static posture assessment system[J]. Iscience, 2023, 26(10).
37. Yang Y, Zhang H, Fernndez A B, et al. Digitalization of Three-Dimensional Human Bodies: A Survey[J]. IEEE Transactions on Consumer Electronics, 2024.
38. Hoang V H, Lee J W, Piran M J, et al. Advances in skeleton-based fall detection in RGB videos: From handcrafted to deep learning approaches[J]. IEEE Access, 2023.
39. Bilakeri S, Kotegar K A. Enhancing person re-identification on RGB-D data with noise free pose-regularized color and skeleton distance features[J]. Engineering Research Express, 2024, 6(1): 015205.
40. Puchaa S, Kasprzak W, Piwowarski P. Human Interaction Classification in Sliding Video Windows Using Skeleton Data Tracking and Feature Extraction[J]. Sensors, 2023, 23(14): 6279.
41. Kumar R, Kumar S. Multi-view multi-modal approach based on 5S-CNN and BiLSTM using skeleton, depth and RGB data for human activity recognition[J]. Wireless Personal Communications, 2023, 130(2): 1141-1159.
42. Zhou H, Liu Q, Wang Y. Learning discriminative representations for skeleton based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 10608-10617.
43. Lovanshi M, Tiwari V. Human skeleton pose and spatio-temporal feature-based activity recognition using ST-GCN[J]. Multimedia Tools and Applications, 2024, 83(5): 12705-12730.
44. Koehl P, Orland H. A Physicists View on Partial 3D Shape Matching[J]. Algorithms, 2023, 16(7): 346.
45. Shaharom M F M, Abd Mukti S N, Raja Maharjan G, et al. Multispectrals three-dimensional model based on SIFT feature extraction[J]. Int J Geoinform, 2023, 19(5): 18.
46. Lin S D, Otoya P E L. LS-SIFT: Enhancing the robustness of SIFT during Pose-invariant Face Recognition by Learning Facial Landmark Specific Mappings[J]. IEEE Access, 2024.
47. Chakraborty D, Chiracharit W, Chamnongthai K. Semantic scene object-camera motion recognition for scene transition detection using dense spatial frame segments and temporal trajectory analysis[J]. IEEE Access, 2024.

48. Huang Y, Yang C, Sun X, et al. Ground-motion simulations using two-dimensional convolution condition adversarial neural network (2D-cGAN)[J]. Soil Dynamics and Earthquake Engineering, 2024, 178: 108444.
49. Ding W, Li W. High speed and accuracy of animation 3D pose recognition based on an improved deep convolution neural network[J]. Applied Sciences, 2023, 13(13): 7566.
50. Sun L, Li N, Zhao G, et al. A three-dimensional human motion pose recognition algorithm based on graph convolutional networks[J]. Image and Vision Computing, 2024, 146: 105009.
51. Luo C, Li B, Liu F. Iterative Back Projection Network Based on Deformable 3D Convolution[J]. IEEE Access, 2023.

Biography

Ye Li is with the Software College, Shenyang Normal University. Research direction is IoT, Information systems, Computer application and AI.
Yifan Pan is with the Software College, Shenyang Normal University. Research direction is IoT, Computer application and AI.
Xinhui Wu is with the Software College, Shenyang Normal University. Research direction is IoT, Computer application and AI.