frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 11 October 2022
DOI 10.3389/fnbot.2022.918434

Check for updates

OPEN ACCESS

EDITED BY
Carmelo Pino,
University of Catania, Italy

REVIEWED BY
Peter Sutor,
University of Maryland, United States
Yunchao Tang,
Zhongkai University of Agriculture and Engineering, China

*CORRESPONDENCE
Xiaodong Liu
liuxiaodongxht@qq.com

RECEIVED 12 April 2022
ACCEPTED 10 August 2022
PUBLISHED 11 October 2022

CITATION
Liu X, Xu H and Wang M (2022) Learning joints relation graphs for video action recognition.
Front. Neurorobot. 16:918434.
doi: 10.3389/fnbot.2022.918434

COPYRIGHT
© 2022 Liu, Xu and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Learning joints relation graphs for video action recognition

Xiaodong Liu*, Huating Xu and Miao Wang

School of Software, Henan Institute of Engineering, Zhengzhou, China

Previous video action recognition mainly focuses on extracting spatial and temporal features from videos or capturing physical dependencies among joints. The relation between joints is often ignored. Modeling the relation between joints is important for action recognition. Aiming at learning discriminative relation between joints, this paper proposes a joint spatial-temporal reasoning (JSTR) framework to recognize action from videos. For the spatial representation, a joints spatial relation graph is built to capture position relations between joints. For the temporal representation, temporal information of body joints is modeled by the intra-joint temporal relation graph. The spatial reasoning feature and the temporal reasoning feature are fused to recognize action from videos. The effectiveness of our method is demonstrated in three real-world video action recognition datasets. The experiment results display good performance across all of these datasets.

KEYWORDS

action recognition, deep learning, convolutional neural network, GCN, joints relation

Introduction

With the rapid development of internet and mobile intelligent devices, multimedia applications, including urban security, medical treatment, education, communication, industrial production, and cultural film, are more and more widely used. They generate huge amounts of videos every hour and now account for about 80% of the data. YouTube has one billion users, with 300 h of video uploaded and 50 million h of video watched per minute. In addition to Internet video, surveillance video is also an important source of video data. There are about 30 million surveillance cameras in China, and 60 EB of video data are generated per month. Thus, how to effectively analyze and understand these data and create greater economic and social benefits has been a hot topic.

Video action recognition is one of the most important tasks for comprehensively understanding video content. It has many potential applications, such as human–robot interaction (de Carvalho et al., 2022), fruit picking (Wu et al., 2022), agricultural (Tang et al., 2020), and warning decision (Fang et al., 2021). There are various attempts at video action recognition based on spatial and temporal features. Two-stream convolutional neural network (CNN) (Simonyan and Zisserman, 2014) consists of spatial and temporal sub-networks. The spatial streaming sub-network extracts a static feature of the video frame, and the temporal streaming sub-network extracts optical flow features containing motion features of the video. Then, the spatial and temporal features are aggregated, and the action feature of the video is produced. Although the two-stream CNN can utilize the spatial and temporal features of the video, there are some limitations. First, spatial

CNN only extracts a single video frame and the information of the rest video frames is not utilized. Second, optical flow extraction is very time-consuming. To avoid the waste of video frames, Wang et al. (2017) fuse spatial features of all video frames and optical flow features to improve the accuracy of action recognition, and this method also needs to consume too much time. 3D Convolutional network (C3D) (Tran et al., 2015) expands the two-dimensional convolution neural network to three-dimensional, which can learn both spatial and temporal features. Its performance is better than two-stream CNN and does not need to extract optical flow. However, C3D adds many weight parameters and needs more training data. P-CNN (Chéron et al., 2015) clips the body joints in the video frame based on human pose estimation and extracts the spatial and temporal features of each body joint by CNN. This method trains an independent CNN network for each joint and fuses the temporal and spatial features extracted from different joints. Although this method successfully captures local information, the internal dependencies between joints are not modeled. To capture the internal dependencies between joints, recent methods (Li M. et al., 2019) construct a skeleton graph and apply graph convolution networks (GCN) (Defferrard et al., 2016) to extract correlated features. However, the internal relationship between the evolution of high-level semantics has not been established, and lacks the middle layer for modeling the semantics of the target joints, so the potential of using joints of the semantics in video action recognition based on the deep learning method cannot be fully exploited. Although the research on video action recognition has got a lot of achievements, it still faces some challenges.

First, the action of the video is mainly depending on the movement of body joints. As an example, let us try to estimate the action of these videos in Figure 1. In Figure 1A, we mainly judge the action of the video by the movement of the hand and elbow. Other information can as an effective supplement to video action recognition. Similarly, detailed estimations can be made in Figures 1B–D. Therefore, how to make full use of the movement of the body joints is a significant challenge for video action recognition.

Second, the movement of body joints is the low-level feature. It contains less information than high-level semantics. If we can extract the movement of high-level semantic of body joints, the action recognition accuracy can be further improved. Therefore, how to model the movement of high-level semantics of body joints is another challenge for video action recognition.

To solve the problem of modeling the semantics of body joints and their movement evolution, this paper proposes a joint spatial-temporal reasoning (JSTR) framework for video action recognition. The JSTR consists of two sub-network. One is the pose estimation sub-network, which is used to generate the position of body joints. The other is the action recognition sub-network which takes as input the results of the pose estimation sub-network, and features of body joints are generated. Then a joints spatial reasoning (JSR) graph and intra-joint temporal relation (IJTR) graph are built, respectively, which is used to capture position and temporal relations between joints. Finally, the spatial reasoning feature and the temporal reasoning feature are fused to recognize the action from videos.

In summary, the contributions of this paper include:

- A geometric constraints-based human pose estimation method is proposed to improve the accuracy of human pose estimation.
- We design a joints spatial reasoning network for capturing position relation between body joints.
- We design an intra-joint temporal relation network for capturing the temporal evolution of body joints.

To verify the effectiveness of our proposed method, we compare the performance with the state-of-the-art method on three real-world video action recognition datasets: UCF101, HMDB51, and Kinetics. In the experiments of UCF101 and HMDB51, we obtain an excellent accuracy of 96.3 and 72.4%, respectively. In particular, on the large-scale kinetics dataset, we obtain an excellent top-1 accuracy of 75.8% and the top-5 accuracy of 92.6% on the validation set.

The rest of this paper is organized as follows: Section Related work presents the related work. Section The proposed methods gives a detailed description of the proposed method. Section Experiments shows the experimental results. Finally, Section Conclusion summarizes our conclusions and gives future work.

Related work

There are various attempts at video action recognition based on spatial and temporal feature fusion. Feichtenhofer et al. (2016a,b) fused CNNs both spatially and temporally and then combined them with ResNets (He et al., 2016) to extract better spatio-temporal features. Sun et al. (2015) propose factorized spatio-temporal convolutional networks that factorize the original 3D convolution kernel learning as a sequential process of learning 2D spatial kernels. The computation cost of 2D CNN is low, but it cannot model the time information, so the accuracy is relatively low. 3D CNN can model time information and can obtain better accuracy, but the computation cost is too high. TSM (Lin et al., 2019) can facilitate information exchanged among neighboring frames by shifting part of the channels along the temporal dimension. It can be inserted into 2D CNNs to achieve temporal modeling, which balances accuracy and computation cost. STM (Jiang et al., 2019) contains a channel-wise spatial-temporal module to present the spatial-temporal features and a channel-wise motion module to efficiently encode motion features. TEA (Li et al., 2020) was designed to capture both short- and long-range temporal evolution by a temporal excitation and aggregation block. However, different features

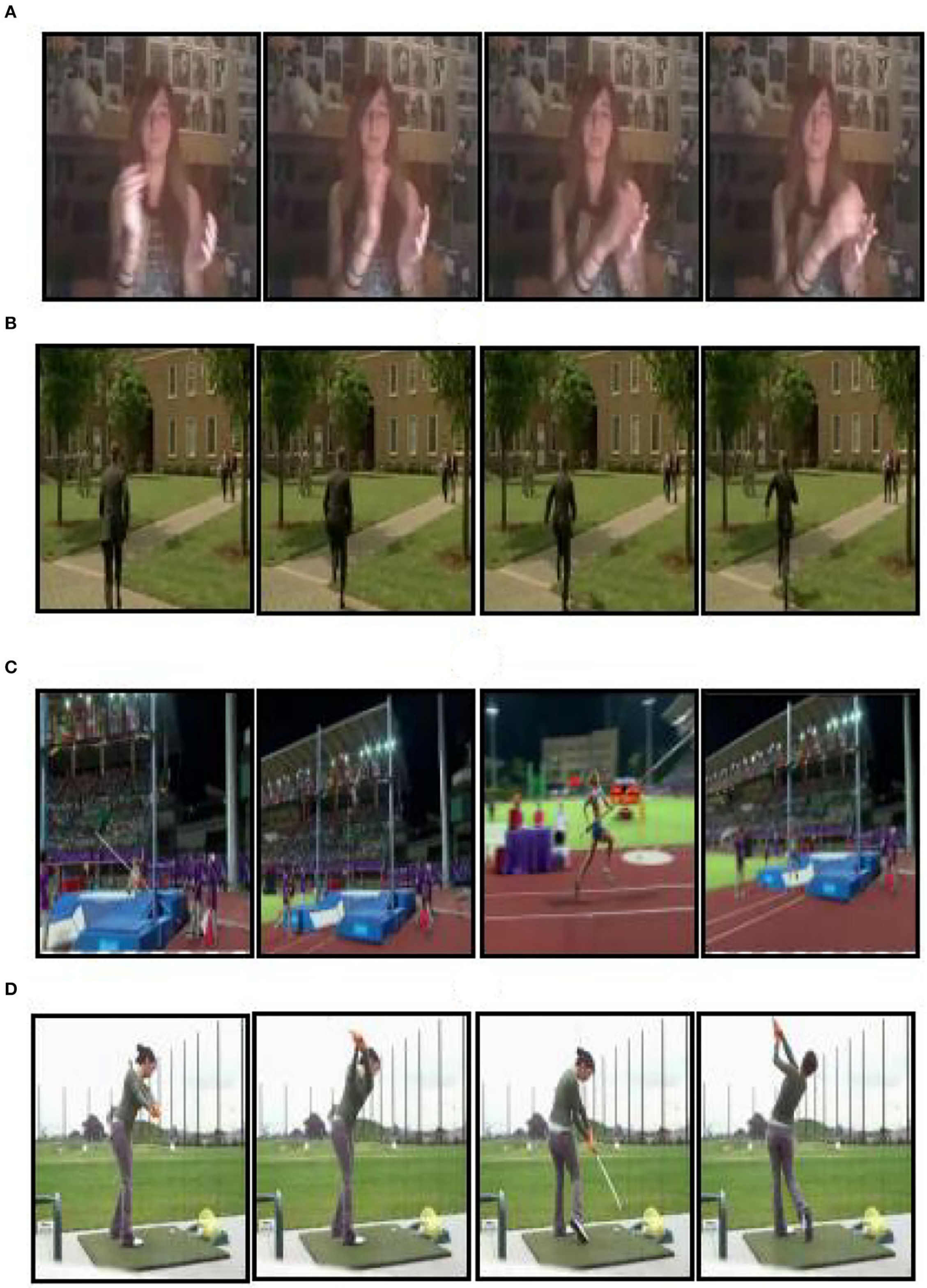

FIGURE 1
(A–D) The motivation of JSTR.

have different contributions to video action recognition. The difference in the contribution of different features on video action recognition cannot be fully utilized by simply combining the different features.

Attention mechanism can make full use of the difference in the contribution of different features. Simonyan and Zisserman (2014) uses both RGB and stacked optical flow as appearance and motion information, respectively. The accuracy is significantly boosted by simply fusing probability scores. VideoLSTM (Li et al., 2018) is an end-to-end sequence learning model, in which hard-wires convolution in the soft-attention LSTM. Xiang et al. (2018) propose a local feature integration framework based on a purely attention-based local feature integration, and carefully analyze the effect of different attention mechanisms. The attention mechanism can focus on important features and meanwhile ignore irrelevant signals and noise, so that the recognition accuracy can be improved.

The internal dependencies between human body joints are often ignored by the traditional action recognition methods, which will result in missing abundant information. To capture dependencies between joints, recent methods apply graph convolution networks (GCN) to extract correlated features. ST-GCN (Yan et al., 2018) learn spatial-temporal features simultaneously. ST-GCN mainly tries to aggregate wider-range features, but node features during long diffusion might be weakened. AS-GCN (Li M. et al., 2019) attempts to capture richer dependencies among body joints and useful non-local information. MUSLE (Li D. et al., 2019) builds space-time graphs and clusters the graphs into compact sub-graphs on each scale. In this paper, the JSTR model position and temporal relations between joints.

The proposed methods

Overview

The overall network framework of the joints spatial-temporal reasoning framework (JSTR) is illustrated in Figure 2. It consists of the pose estimation sub-network and the action recognition sub-network. First, we uniformly sample a set of T frames from the video and extract their feature vectors by a backbone network, such as Resnet (He et al., 2016). Second, the last convolution layer feeds into a pose estimation sub-network and generates the position of body joints, which is used as the input of the action recognition sub-network. Third, the features of each joint are extracted by RoIAlign (He et al., 2017) according to the positions of body joints. After that, a non-linear transformation is performed to get an L dimensional feature vector for each body joint. Fourth, the features of body joints will be split into two branches: one branch is a joint spatial reasoning (JSR) sub-network, which is used to build a relation graph to capture position relations between joints. The other branch is an intra-joint temporal relation (IJTR) sub-network, which models the temporal information between body joints. Then, the output features of JSR and IJTR are fused to form joints spatial-temporal reasoning feature for recognizing action from videos.

Pose estimation

The accuracy of pose estimation is important for our JSTR network, and it is a key module for feature extraction of body joints. To further improve the accuracy of pose estimation, we extend the structured feature learning framework (Xiao et al., 2016) and give a geometric constraint-based human pose estimation method (GCHP), which is illustrated in Figure 3.

We first apply the pose estimation CNN network to generate the positions of body joints and the distance between body joints can be obtained by the positions of body joints. Thus, the loss function of joints' distance can be obtained by calculating the real distance and the generated joint position. Let d_{ij} denotes the relative distance between the i-th joint and the j-th joint, and the loss function of joint distance can be expressed as follows.

$$L_d = \sum_{n=1}^{N} \sum_{i=1}^{P} d_{ij} \tag{1}$$

where L_d is the loss function, N is the number of training images, and P is the number of body joints.

Let $\mathrm{L_p}$ is the loss function of joints' position, and it can be calculated by

$$L_p = \sum_x \sum_y m(x,y) \sum_k t_x(x,y) \log \left(\frac{e^{z_k(x,y)}}{\sum_{k'} e^{z_{k'}(x,y)}} \right) \tag{2}$$

where (x,y) is the position of the body joint, k is the index of the joints, $\mathrm{t_k}$(x,y) is the real label of the position (x,y), and $\mathrm{z_k}$(x,y) is the predicted value.

Geometric constraint function L is a combination of the position loss function and distance loss function, and it can be calculated by

$$L = L_p + \lambda L_d \tag{3}$$

where λ is used to balance the weight of position loss function and distance loss function.

Note that our geometric constraints-based human pose estimation method can be directly inserted into the existing pose estimation pipeline. They can effectively strengthen feature representations and do not increase too many parameters.

Action recognition framework

In the video action recognition sub-network, a joint spatial-temporal reasoning graph is built to perform relational

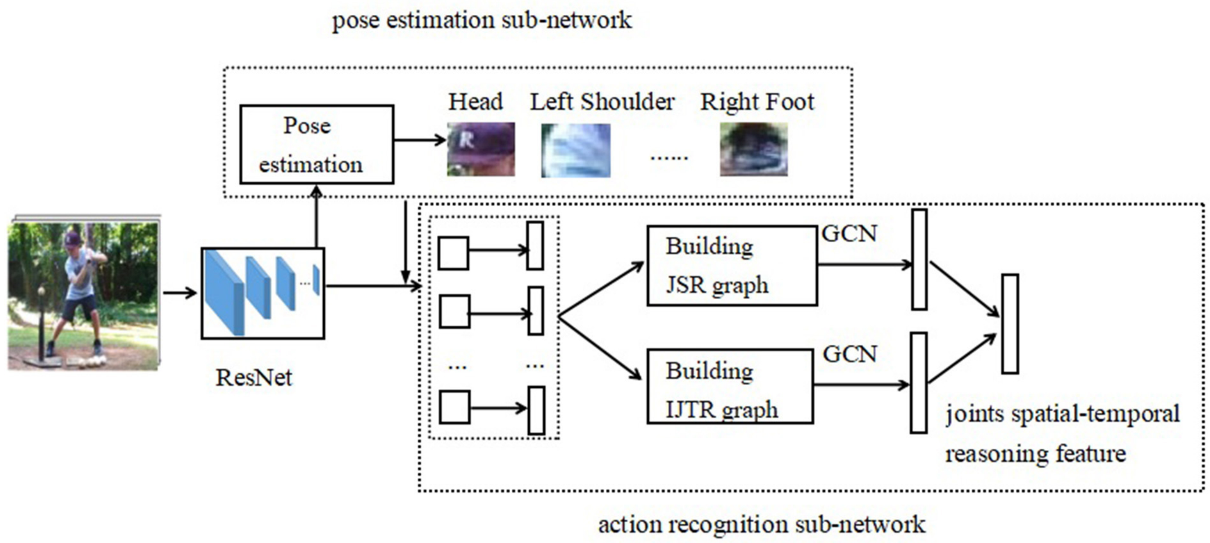
pose estimation sub-network
Pose estimation
Head
Left Shoulder
Right Foot
......
ResNet
Building JSR graph
Building IJTR graph
GCN
GCN
joints spatial-temporal reasoning feature
action recognition sub-network

FIGURE 2
The framework of JSTR. The video frames are fed into a backbone network, such as Resnet. The last convolution layer feeds into a pose estimation sub-network and generates the position of body joints. The features of each joint are extracted by RoIAlign according to the positions of body joints. Upon these original features and positions of joints, a spatial-temporal reasoning graph is built. Then, the spatial features and temporal features are fused to recognize the action from videos.

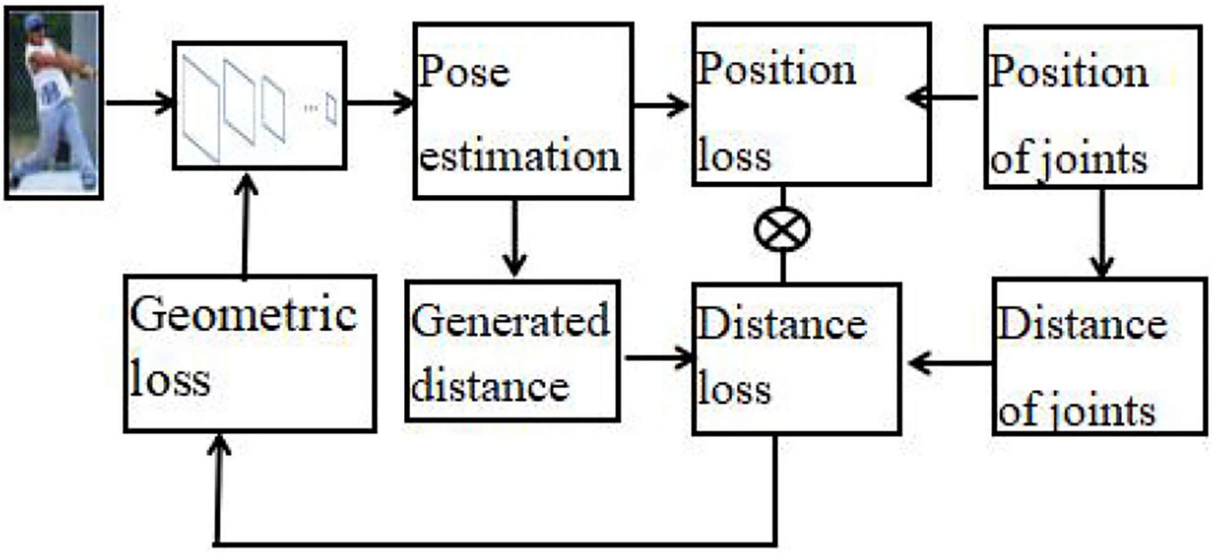

FIGURE 3
Geometric constraint-based human pose estimation.

reasoning on joints for action recognition. The positions of body joints are obtained by the pose estimation sub-network, which is the input of the last convolution layer of the video action recognition sub-network. Features of each joint are obtained by RoIAlign (He et al., 2017). Afterward, upon these original features and positions of joints, we build a spatial-temporal reasoning graph, where each node denotes a joint. For the spatial graph, each edge is a scalar weight, which is computed according to the two joints' features and their relative position. For the temporal graph, each edge represents the temporal information of the same joint. The GCN is applied to conduct relational reasoning based on a spatial-temporal reasoning graph. After graph convolution, the spatial-temporal relational representation for joints is generated. Then the spatial features and temporal features are fused to recognize the action from videos.

Building JSR graphs

Graph model has been found to be effective at representing the spatial and temporal relations in visual content. Therefore, the graph structure is utilized to model body joints relation for action recognition.

Graph definition

We consider a joints spatial relation (JSR) graph. The nodes $V = \{(f_i, p_i) | i = 1, 2, ..., P\}$ are a set of P body joints, where f_i is the d-dimension appearance feature of the i-th joint, and $p_i = (x_i, y_i)$ is coordinates of the i-th body joint. Let $A \in R^{P \times P}$ be the adjacent matrix of the JSR graph, which represents pair-wise relation among joints. The relation value A_{ij} is the affective coefficient, which is used to measure the intensity of the features of nodes.

To obtain sufficient representational power to capture the underlying relation between two joints, both distance information and angel information need to be considered. Moreover, the distance relation and angel relation have different position attributes. Therefore, the relation value can be defined as a composite function below:

$$A_{ij} = h(f_d(p_i, p_j), f_a(p_i, p_j)) \tag{4}$$

where $f_d(p_i, p_j)$ denotes the distance relationship between two joints, and the angel relation is calculated by $f_a(p_i, p_j)$. The function f() is a fusion function that fuses distance and angel relation.

Next, we will give the definition of distance relation, angel relation, and the fusion function.

Let $p_i = (x_i, y_i)(i = 1, 2, ..., P)$ be the coordinates of the i-th joints, the distance relationship between the i-th joint and the j-th joint can be expressed as:

$$f_d(p_i, p_j) = \left\| l_j - l_i \right\|_2 \tag{5}$$

Let $f_a(p_i, p_j)$ be the angel relation between the i-th and j-th body joints, and it can be expressed as:

$$f_a(p_i, p_j) = \arctan \frac{y_j - y_i}{x_j - x_i} \tag{6}$$

The function h fuses distance and angel relation. In our experiments, we adopt the following function to compute the value of A_{ij} (Jianchao and Limin, 2019).

$$A_{ij} = \frac{f_d(p_i, p_j) \exp(f_a(p_i, p_j))}{\sum_{j=1}^{P} f_d(p_i, p_j) \exp(f_a(p_i, p_j))} \tag{7}$$

where we perform normalization on each joint using the Softmax function so that the sum of all the related values of one actor node i will be 1.

Building IJTR graphs

Temporal information of body joints is a crucial cue for activity recognition. We model the temporal information of body joints by the intra-joint temporal relation (IJTR) Graph. The IJTR graph is represented by an adjacent matrix $A^T \in R^{P \times P}$. During training, we randomly sample a set of $K = 2$ frames from the entire video and build IJTR Graph. Therefore, for the adjacent matrix A^T of the IJTR Graph, we directly set $A^T_{ij} = 1$, if the i-th joint and the j-th joint are in two sample frames, and $A^T_{ij} = 0$, otherwise. At testing time, a sliding window approach is used, and the features from all windows are mean-pooled to form a global activity feature.

Reasoning and training on graphs

The convolutional neural network is originally used to extract visual features of images or videos based on 2D or 3D filters. In contrast, a graph convolutional neural network (GCN) can perform message propagation from nodes to its neighbor nodes. Thus, it is usually used to perform relational reasoning. In this paper, GCN is applied to reasoning relations of human body joints.

Given a graph with N nodes, the operation of one graph convolution layer can be expressed as follows.

$$X_{(l+1)} = \sigma((D)^{-\frac{1}{2}} A (D)^{-\frac{1}{2}} X_{(l)} W^{(l)}) \tag{8}$$

where $D \in \Re^{P \times P}$ is the degree matrix of A, $X_{(l)}$ is the output of the $(l-1)$-th layer, W1 is the learned parameters, and $\sigma(.)$ is a non-linear activation function like ReLU.

In particular, in our JSTR framework, the adjacent matrix of the JSR is defined in Section Building IJTR graphs and the adjacent matrix of the IJTR is defined in Section Reasoning and training on graphs. $X^{(0)} = [f(x_1), f(x_2), ..., f(x_P)]$ is the initial feature matrix, where $f(x_i)$ is the feature vector of the i-th joint. The final output of the GCN is updated features of joints, $X_{(l)}$, in the graphs, which can be aggregated into a frame-level vector. In this paper, the GCN output of JSR and IJTR is aggregated into spatial relation features and temporal relation features, respectively, and then these two features are aggregated into spatial-temporal relation features for recognizing action from videos.

Experiments

Dataset

UCF101

UCF101 (Soomro et al., 2012) has 13,320 web video clips and 101 action classes. The videos range from daily life activities to unusual sports and each video. The average accuracy of three training/testing splits is adopted in our experiments.

TABLE 1 Experimental results on the MPII dataset.

Experiment	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
Tang et al. (2018)	95.6	95.9	90.7	86.5	89.9	86.6	82.5	89.8
Xiao et al. (2018)	97.0	95.9	90.3	85.0	89.2	85.3	81.3	89.6
Yang et al. (2020)	97.1	95.9	90.3	86.4	89.1	87.1	83.3	89.9
Zhang et al. (2021)	97.5	96.0	90.7	87.3	89.5	87.7	83.4	90.3
GCHP	98.1	96.2	90.8	87.4	89.9	86.7	83.6	90.6

HMDB51

The HMDB51 dataset (Kuehne et al., 2011) has 6, 766 videos downed from movies and web videos, and each video is labeled with one of the 51 action categories. The average accuracy of three training/testing splits is adopted in our experiments. Videos are subject to different viewpoints, video quality, camera motions, and occlusions.

Kinetics

The kinetics dataset is a large dataset for human action recognition, containing over 240,000 video action clips (Carreira et al., 2017). It uses a training set of 236,763 videos and a testing set of 19,095 videos. There are 400 types of actions.

Implementation details

An input video is uniformly partitioned into 16 segments, and one frame is randomly Sampled from each segment to obtain 16 frames for one video. The pose estimation sub-network generates 16 body joints, and the features of each joint are extracted by RoIAlign and get a 1,024-dimensional feature vector for each joint. Two JSR and IJTR are pre-trained on the training set separately. The network is trained in 150 epochs using the minibatch size of 32. The learning rate starts from 0.001 and multiplies to 0.1 every 30 epochs.

The evaluation of GCHP

The GCHP is evaluated on the MPII (Andriluka et al., 2014) dataset. It contains 25,000 images with 40,000 different human instances and uses a training set of about 28,000 human instances and a testing set of 11,000 human instances, and the total number of body joints used is 16. The PCP (percentage of correct parts) is employed in this experiment. Similar to structured feature learning, the strict PCP, which only when both ends lie within 50% of the ground truth will this prediction is correct, is used. We compare experimental results with three deep learning-based methods (Tang et al., 2018; Xiao et al., 2018; Yang et al., 2020; Zhang et al., 2021). Table 1 shows the experimental results on the MPII dataset. As we can see from Table 1, our GCHP outperforms previous state-of-the-art work on every body part evaluated.

Ablation studies

In this subsection, we perform detailed studies on the action recognition datasets to understand the contributions of our model components to video action recognition. In our experiments, we selected 14 body joints for one frame and sampled 20 frames for one video.

Joints spatial reasoning

We first train the pose estimation sub-network on the MPII dataset (Andriluka et al., 2014), and the trained network is used to predict the position of body joints. In our experiments, we select 14 body joints including head, upper arm, lower arm, upper leg, lower leg, and etc. Then, ROIAlign is used to extract features of these body joints. JSR is built and Table 1 gives the results of the fusion of JSR and baseline method TSN (Wang et al., 2016; Zhang et al., 2021). We can observe that JSR can improve the accuracy by 0.7, 0.6, and 1.1%, respectively, on Kinetics, UCF101, and HMDB51 dataset.

Intra-joint temporal relation

We further evaluate the effect of IJTR. In the training phase, two adjacent video frames are selected to build the IJTR graph. In the test phase, we use a sliding window approach, and the emotion features from all windows are fused in a mean-pooled manner. Table 2 gives the results of the fusion of IJTR and baseline method TSN. As can be shown in Table 3, the accuracy improved by 0.3, 1.1, and 2.2%, respectively, on Kinetics, UCF101, and HMDB51 datasets.

Joints spatial-temporal reasoning

The output of JSR and IJTR is fused and produce the feature of joints spatial-temporal reasoning. Table 4 gives the results of JSTR. Table 4 shows the comparison results of JSTR and baseline method TSN, we can see that the accuracy improved

TABLE 2 Exploration of SA.

Method	Kinetics accuracy (%)	UCF101 accuracy (%)	HMDB51 accuracy (%)
TSN	92.4	94.2	69.4
JSR	93.1	94.8	70.5

TABLE 3 Exploration of PR.

Method	Kinetics accuracy (%)	UCF101 accuracy (%)	HMDB51 accuracy (%)
TSN	92.4	94.2	69.4
IJTR	92.7	95.3	71.6

TABLE 4 Exploration of JSTR.

Method	Kinetics accuracy (%)	UCF101 accuracy (%)	HMDB51 accuracy (%)
TSN	92.4	94.2	69.4
JSTR	93.6	96.3	72.4

TABLE 5 Mean classification accuracy (%) compared with state-of-the-art methods on UCF101 and HMDB51.

Method	UCF101 (%)	HMDB51 (%)
Two Stream (Simonyan and Zisserman, 2014)	88	59.4
TSN (Zhang et al., 2021)	94	68.5
Fusion (Feichtenhofer et al., 2016a)	92.5	65.4
ST-ResNet (Feichtenhofer et al., 2016b)	93.4	66.4
Attention cluster (Xiang et al., 2018)	94.6	69.2
MUSLE (Li D. et al., 2019)	94.8	72.2
Our methods	96.3	72.4

by 1.2, 2.1, and 3.0%, respectively, on kinetics, UCF101, and HMDB51 datasets.

Comparison with the state-of-the-art

Now, we compare our model with the state-of-the-art methods. We compare our model with the state of the art methods on UCF101 and HMDB dataset. Our approach obtains robust improvements over the state-of-the-art methods. As can be seen from Table 5, the accuracy of our method on UCF101 and HMDB datasets is 96.3 and 72.4%, respectively. Compared with the state-of-the-art method, it increased by 1.5 and 0.2%, respectively, on the UCF101 and HMDB datasets. We also notice that the methods (Simonyan and Zisserman, 2014; Feichtenhofer et al., 2016a,b; Zhang et al., 2021) fuse spatial and temporal features without attention mechanism and relation reasoning, so their accuracy is relatively low. Attention cluster (Xiang et al., 2018) improves performance by first analyzing the importance of each local feature and then bestowing the global feature. Relation reasoning is adopted in MUSLE (Li D. et al., 2019) and the accuracy can be improved.

TABLE 6 Kinetics top-1 and top-5 accuracy (%).

Method	Top-1 (%)	Top-5 (%)
TSM (Lin et al., 2019)	72.5	90.7
STM (Jiang et al., 2019)	73.7	91.6
TEA (Li et al., 2020)	74.0	91.3
MUSLE (Li D. et al., 2019)	75.1	92
Our methods	75.8	93.6

On Kinetics, Table 6 shows the kinetics top-1 and top-5 accuracy on different methods. TSM (Lin et al., 2019) balances accuracy and computation cost by being inserted into 2D CNNs to achieve temporal modeling. The computation cost is relatively low, and the accuracy is also relatively low. STM (Jiang et al., 2019) uses a unified 2D CNN network that integrates spatial-temporal and motion features. It can extract long-term temporal features without 3D CNN, and the accuracy is higher than TSM. TEA (Li et al., 2020) was designed to capture both short- and long-range temporal evolution, which further improves the accuracy. MUSLE (Li D. et al., 2019) exploits the discriminative sub-graphs across different scales for facilitating spatio-temporal reasoning. The accuracy is higher than those methods without a reasoning relationship. As shown in Table 6, the top-1 and top-5 accuracy of our method are 75.8% and 93.6, respectively. It increases top-5 by 1.6% compared with the state-of-the-art method. There are two reasons that JSTR can achieve excellent performance. First, JSR can capture position relations between joints. Second, the temporal information of body joints is modeled.

Conclusion

To model the relation between body joints, a joint spatial-temporal reasoning (JSTR) framework is proposed. The JSTR consists of two sub-networks. In the pose estimation sub-network, geometric constraints-based human pose estimation is applied to further improve pose estimation accuracy. In the action recognition sub-network, a joints spatial reasoning (JSR) graph and intra-joint temporal relation (IJTR) graph are built, respectively, which is used to capture position and temporal relations between joints. Finally, the spatial reasoning feature and temporal reasoning feature are fused to recognize the action from videos. We validate JSTR in action recognition using three datasets, UCF101, HMDB51, and Kinetics. The experiment results show that JSTR can improve the recognition accuracy compared to the state-of-the-art methods. In terms of future work, we hope to extract spatio-temporal GCN features simultaneously to further improve the accuracy of action recognition and integrate it into end-to-end trained networks.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: University of Central Florida (UCF) Center for Research in Computer Vision, https://www.crcv.ucf.edu/data/UCF101.php.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work was supported by the foundation of the He'nan Educational Committee (21A520006) and the scientific and technological research project of the Henan Provincial Science and Technology Department (182102310919).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andriluka, M., Pishchulin, L., Gehler, P., and Schiele, B. (2014). "Human pose estimation: new benchmark and state of the art analysis," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 3686–3693. doi: 10.1109/CVPR.2014.471

Carreira, J., Zisserman, A., Vadis, Q. (2017). "Action recognition? A new model, and the kinetics dataset," in *IEEE International Conference on Computer, Vision, and Pattern, Recognition* (Honolulu, HI: IEEE), 4724–4733. doi: 10.1109/CVPR.2017.502

Chéron, G., Laptev, I., and Schmid, C. (2015). "P-cnn: pose-based cnn features for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 3218–3226. doi: 10.1109/ICCV.2015.368

de Carvalho, K. B., Basílio, V. T., and Brandão, A. S. (2022). Action recognition for educational proposals applying concepts of Social Assistive Robotics. *Cogn. Syst. Res.* 71, 1–8. doi: 10.1016/j.cogsys.2021.09.002

Defferrard, M., Bresson, X., and Van dergheynst, P. (2016). "Convolutional neural networks on graphs with fast localized spectral fifiltering," in *Advances in Neural Information Processing Systems* (Barcelona: MIT Press), 3844–3852.

Fang, S., Liu, F., Xiong, Z., Fang, J., and Li, L. (2021). Seismic performance evaluation of recycled aggregate concrete-filled steel tubular columns with field strain detected via a novel mark-free vision method. *Structures* 37, 426–441. doi: 10.1016/j.istruc.2021.12.055

Feichtenhofer, C., Pinz, A., and Wildes, R. P. (2016b). "Spatiotemporal residual networks for video action recognition," in *Advances in Neural Information Processing Systems* (Honolulu, HI: IEEE), 3468–3476. doi: 10.1109/CVPR.2017.787

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016a). "Convolutional two-stream network fusion for video action recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 1933–1941. doi: 10.1109/CVPR.2016.213

He, K., Gkioxari, G., Dollar, P., and Girshick, R. B. (2017). "Mask R-CNN," in *IEEE International Conference on Computer Vision* (Venice: IEEE), 2980–2988.

He, K., Zhang, X., and Ren, S. J. J., Sun. (2016). "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Jianchao, W., and Limin, W. (2019). "Learning actor relation graphs for group activity recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Seoul: IEEE), 9964–9974. doi: 10.1109/CVPR.2019.01020

Jiang, B., Wang, M., Gan, W., Wu, W., and Yan, J. (2019). "STM: spatiotemporal and motion encoding for action recognition," in *IEEE International Conference on Computer Vision* (Seoul: IEEE), 2000–2009. doi: 10.1109/ICCV.2019.00209

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). "Hmdb: a large video database for human motion recognition," in *IEEE International Conference on Computer Vision* (Barcelona: IEEE), 2556–2563. doi: 10.1109/ICCV.2011.6126543

Li, D., Qiu, Z., Pan, Y., Yao, T., Li, H., and Mei, T. (2019). "Representing videos as discriminative sub-graphs for action recognition," in *International Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 3310–3319.

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. (2019). "Actional-structural graph convolutional networksfor skeleton-based action recognition," in *International Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 3595–3603.

Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., and Wang, L. (2020). "TEA: temporal excitation and aggregation for action recognition," in *International Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 906–915.

Li, Z., Gavves, E., Jain, M., and Snoek, C. G. M. (2018). VideoLSTM convolves, attends and flows for action recognition. *Comput. Vision Image Understand.* 166, 41–50. doi: 10.1016/j.cviu.2017.10.011

Lin, J., Gan, C., and Han, S. (2019). "TSM: temporal shift module for efficient video understanding," in *International Conference on Computer Vision* (Seoul: IEEE), 7082–7092. doi: 10.1109/ICCV.2019.00718

Simonyan, K., and Zisserman, A. (2014). "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems* (Montreal, QC: MIT Press), 568–576.

Soomro, K., Zamir, A. R., and Shah, M. (2012). *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*. Available online at: https://ui.adsabs.harvard.edu/abs/2012arXiv1212.0402S/abstract

Sun, L., Jia, K., Yeung, D. Y., and Shi, B. E. (2015). "Human action recognition using factorized spatio-temporal convolutional networks," in *IEEE International Conference on Computer Vision* (Santiago: IEEE), 4597–4605. doi: 10.1109/ICCV.2015.522

Tang, W., Yu, P., and Wu, Y. (2018). "Deeply learned compositional models for human pose estimation," in *Proceedings of the European Conference on Computer Vision* (Cham: Springer), 197–214. doi: 10.1007/978-3-030-01219-9_12

Tang, Y., Chen, M., Wang, C., Luo, L. J., Lian, G., and Zou, X. (2020). Recognition and localization methods for vision-based fruit picking robots: a review. *Front. Plant Sci.* 11, 510. doi: 10.3389/fpls.2020.00510

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago: IEEE), 4489–4497. doi: 10.1109/ICCV.2015.510

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., et al. (2016). "Temporal segment networks: towards good practices for deep action recognition," in *European Conference on Computer Vision* (Cham: Springer), 20–36. doi: 10.1007/978-3-319-46484-8_2

Wang, Y., Long, M., Wang, J., and Yu, P. S. (2017). "Spatiotemporal pyramid network for video action recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 2097–2106. doi: 10.1109/CVPR.2017.226

Wu, F., Duan, J., Ai, P., Chen, Z., Yang, Z., and Zou, X. (2022). Rachis detection and three-dimensional localization of cut off point for vision-based banana robot. *Comput. Electron. Agric*. 198, 107079. doi: 10.1016/j.compag.2022.107079

Xiang, L., Gan, C., Melo, G. D., Wu, J., and Xiao, L. (2018). "Attention clusters: purely attention based local feature integration for video classification," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 7834–7843.

Xiao, B., Wu, H., and Wei, Y. (2018). "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision* (Cham: Springer), 472–487. doi: 10.1007/978-3-030-01231-1_29

Xiao, C., Ouyang, W., Li, H., and Wang, X. (2016). "Structured feature learning for pose estimation," in *IEEE International Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 4715–4723.

Yan, S., Xiong, Y., and Lin, D. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence* (New Orleans, LA: AAAI), 7444–7452. doi: 10.1609/aaai.v32i1.12328

Yang, C., An, Z., Zhu, H., Hu, X., and Xu, Y. (2020). Gated convolutional networks with hybrid connectivity for image classification. *Proc. AAAI Conf. Artif. Intell.* 34, 12581–12588. doi: 10.1609/aaai.v34i07.6948

Zhang, C., He, N., Sun, Q., Yin, X., and Lu, K. (2021). "Human pose estimation based on attention multi-resolution network," in *International Conference on Multimedia Retrieval* (Taipei: ACM), 682–687. doi: 10.1145/3460426.3463668