electronics

MDPI

Article

Human Action Recognition for Dynamic Scenes of Emergency Rescue Based on Spatial-Temporal Fusion Network

Yongmei Zhang [1,*], Qian Guo [1], Zhirong Du [1,2] and Aiyan Wu [1]

[1] School of Computer Science and Technology, North China University of Technology, Beijing 100144, China
[2] School of Electrical and Control Engineering, North China University of Technology, Beijing 100144, China
* Correspondence: zhangym@ncut.edu.cn

Abstract: Targeting the problems of the insufficient utilization of temporal and spatial information in videos and a lower accuracy rate, this paper proposes a human action recognition method for dynamic videos of emergency rescue based on a spatial-temporal fusion network. A time domain segmentation strategy based on random sampling maintains the overall time domain structure of the video. Considering the spatial-temporal asynchronous relationship, multiple asynchronous motion sequences are increased as input of the temporal convolutional network. spatial-temporal features are fused in convolutional layers to reduce feature loss. Because time series information is crucial for human action recognition, the acquired mid-layer spatial-temporal fusion features are sent into Bidirectional Long Short-Term Memory (Bi-LSTM) to obtain the human movement features in the whole video temporal dimension. Experiment results show the proposed method fully fuses spatial and temporal dimension information and improves the accuracy of human action recognition in dynamic scenes. It is also faster than traditional methods.

Keywords: spatial-temporal fusion; human action recognition; two-stream convolutional neural network; emergency rescue; spatial-temporal asynchronous information

Citation: Zhang, Y.; Guo, Q.; Du, Z.; Wu, A. Human Action Recognition for Dynamic Scenes of Emergency Rescue Based on Spatial-Temporal Fusion Network. *Electronics* **2022**, *12*, 538. https://doi.org/10.3390/electronics12030538

Academic Editor: Fernando De la Prieta

Received: 30 October 2022
Revised: 15 December 2022
Accepted: 18 December 2022
Published: 20 January 2023

1. Introduction

Human action recognition has always been one of the most challenging problems in the field of computer vision [1]. Video is a kind of data over time with a strong temporal correlation. Each pixel in a video has great similarity and strong spatial correlation.

Most of the subjects in videos are people, so the human action recognition technology has piqued considerable research interest as a novel application. The development of artificial intelligence provides a broad space for developing human action recognition technology in the form of virtual reality, intelligent monitoring, motion analysis, human-computer interaction, etc. [2]

In recent years, various natural and man-made disasters have had a great impact on people's lives. In the face of emergencies, identifying specific situations that need emergency responders is critical [3]. Applying action recognition technology to rescue scenarios, such as major traffic accidents, major terrorist attacks, and earthquakes can effectively improve emergency response by medical rescue team members and is conducive to providing auxiliary decisions for decision-making at the disaster site.

Identifying the actions and current state of both first-responders and victims is crucial in such situations. It would be helpful to have a more comprehensive grasp of the on-site rescue work to achieve efficient guidance and accurately and quickly implement the rescue.

In this work, we propose a two-stream asynchronous fusion network based on Temporal Segment Networks (TSN) and Bi-LSTM for human action recognition in emergency rescue classification of entire video sequences. The main contributions include the following:

(1) This paper further refines the currently available dataset in the literature [4]. The dataset was constructed with reference to the AVA dataset production method. To improve

the annotation efficiency, Faster R-CNN was used to detect human positions, including actions and human annotations using Via software. To prevent the overfitting of the network and increase the sample diversity of data, action sequence video data are collected as the emergency first-responder action dataset for annotation, and a data enhancement method is used for data augmentation so that it can better reflect the emergency rescue scene.

(2) A spatial-temporal asynchronous fusion network is proposed. TSN was used to randomly sample video fragments. An RGB image corresponding to each segment and its two-stream field before and after a specified period of time is then input into the spatial and temporal flow networks, respectively, to extract spatial and temporal features. The fusion of spatial-temporal asynchronous information is also realized in the convolutional layer. The fused features were input into Bi-LSTM to extract temporal features and finally implemented for human behavior recognition using Softmax. By modeling the asynchronous relationship between the moving image and the motion sequence (optical flow), the long-term motion can be modeled.

(3) Experiments are conducted on the improved emergency rescue dataset provided in the literature [4] and the publicly available dataset UCF101. We conduct experimental analysis to compare our spatial-temporal fusion network model with other methods, which verifies that the presented method improves the accuracy of action recognition.

The remainder of this paper is structured as follows: Section 2 shows the related work. The proposed spatial-temporal fusion network model is shown in Section 3. The proposed method is presented in Section 4. Experiments and results are described in Section 5, while Section 6 concludes the work of this paper.

2. Related work

2.1. Traditional Human Action Recognition Methods

Human action recognition methods include feature extraction and action recognition. Traditional feature extraction methods mainly include global and local feature extraction. Global feature representation is a method of comprehensively describing the overall structure and shape features of moving objects, such as silhouette-based features and optical-flow-based features. A. Mahjoub et al. [5] computed a depth motion map for each sequence to represent the action motion features from the depth data.

The global feature method is greatly affected by occlusion, visual angle change, and noise, which cannot effectively capture the changes in viewpoint and occlusion. The method based on local features does not rely on the global features of the image and only extracts local features, so it is less affected by noise and messy background, has good robustness, and it is more widely used.

The identification method using local feature representation requires the extraction of the rich interest points from the video and the use of local descriptors to express the points of interest, finally gathering them together. This approach obtains local features directly from the point of interest on the image, thus eliminating the pre-processing step [6]. Methods based on the point-of-interest description are often suitable for simpler scenarios and can lead to decreased recognition performance if the video background is more complex.

The extraction and description of the movement trajectory based on tracking points are also the focus of scholars. The Dense Trajectory (DT) method proposed in the literature [7] densely sampled the points of interest in the framed image and used the dense optical flow method to track these points of interest and connect them into trajectories. Along the motion trajectory of the feature points, the features are extracted as motion descriptors. Bag of features (BOF) is used to encode the feature groups to obtain the features, thereby describing the video behavior.

Abdelbaky A. et al. [8] used a 2D convolutional network, PCANet, as an unsupervised feature extractor instead of 3D convolutional networks to learn the spatial-temporal features of video actions. The learned spatial and temporal features are combined with BOF and a vector local aggregation descriptor encoding scheme. Chen T. et al. [9] improved the

contour features and realized action recognition based on the improved features and multiple features.

Because features are highly susceptible to extrinsic factors, the recognition of feature-based video behavior is different in different scenarios. Select representative feature descriptions based on a specific task, and the selected features will have a great influence on the recognition accuracy [10]. Some current methods require high computational costs while achieving high recognition accuracy, such as dense trajectory algorithms. Such algorithms require a large number of trajectory operations, so the recognition process is relatively complex and has great limitations. As a result, traditional methods are gradually being replaced by deep-learning-based methods [11].

2.2. Human Action Recognition Methods Based on Deep Learning

According to the different neural network structures used for recognition, action recognition methods based on deep learning can be divided into action recognition methods based on two-stream convolutional neural networks (CNN), action recognition methods based on 3D CNN, and action recognition methods based on long short-term memory (LSTM). A comparison of the three methods is shown in Table 1.

Table 1. A comparison of the three deep learning action recognition methods.

Deep Learning Methods	Advantages	Disadvantages
Two-stream CNN methods	Two-stream obtains shape and motion information, respectively, and the recognition effect is better than single-flow networks.	Use optical flow information with large computation.
3D CNN methods	Directly capture spatial-temporal features from the original video sequences.	Considering partial continuous frames, features cannot be learned at the entire video level.
LSTM methods	Combined with CNN, combining both strengths, to better capture temporal information.	Association information before and after the action is not considered.

(1) Action recognition methods based on two-stream CNN:

Inspired by the ventral and dorsal human visual processes, the spatial-temporal two-stream network was proposed by Simonyan et al. [12]. However, the proposed two-stream network is difficult to handle for long timing and complex motions. To address this problem, Wang et al. [13] presented TSN, which introduced a sparse sampling strategy.

Thereafter, researchers proposed many improved models based on the two-stream network structures. For example, Zhuang et al. [14] proposed a novel three-flow spatial-temporal attention-enhancing feature fusion network for action recognition. A two-stream 3D ConvNet fusion framework was proposed by Wang et al. [15].

(2) Action recognition methods based on 3D CNN:

Ji et al. [16] first proposed 3D CNN and applied it to action recognition. Based on this work, Tran et al. [17] presented convolutional 3D (C3D). The network utilizes 3D convolution and 3D pooling to process the input video frames. However, there are still some limitations. 3D convolution produces a large number of parameters, which greatly increases the computation cost. Moreover, 3D convolution models both temporal and spatial information, and easily leads to overfitting problems.

To address these issues, Carreira et al. [18] proposed I3D networks to obtain greater spatial-temporal resolution by expanding the 2D convolution operations of InceptionV1 networks into 3D convolution. Diba et al. [19] extended the 3D convolution to DenseNet and presented the Temporal 3D Convnets (T3D). T3D adds a temporal transition layer (TTL) to obtain rich temporal information at different scales and capture short, medium, and long

temporal information. Qiu et al. [20] proposed Pseudo-3D Residual Networks (P3D) for the problem of a large number of C3D parameters. The 3D convolution is decomposed into 2D space and 1D time convolution by convolution factors. This method increases the network depth and diversity by separating and flexibly combining the time and space domains and improves the recognition accuracy.

Most of the previous networks have extended the 2D CNN from the time dimension. However, this is not necessarily the best choice. The X3D network was proposed by Feichtenhofer [21]. The X3D extended at multiple scales, including dimensions, such as input frame number, input frame size, sampling frame rate, number of convolutional kernels, depth feature graph width, network depth, and other dimensions. The computational amount and parameters required for X3D are exponentially reduced with state-of-the-art performance.

(3) Action recognition methods based on LSTM:

LSTM-based action recognition methods usually combine CNN and LSTM to build networks. This method takes full advantage of CNN and LSTM, CNN extracts the features of spatial dimension, and LSTM extracts information of temporal dimension. LSTM can solve the gradient disappearance problem, thus handling long video data well. Combining CNN and LSTM is mainly able to better capture spatial-temporal information for fusion.

Donahue et al. [22] proposed a Long-Term Recurrent Convolutional Network (LRCN) that combined the traditional CNN and LSTM to extract the spatial-temporal information of the video. The input of the LRCN network can be either a single-frame image or a video with temporal information. In the action recognition neural network structure proposed by Ou et al. [23], CNN is used to separately extract local spatial and local motion information, and LSTM is adopted to extract feature information in video sequences and obtain the context relation of the local spatial-temporal information. Ge et al. [24] improved the Faster R-CNN framework by introducing LSTM, obtaining the spatial features of the action by Faster R-CNN and the temporal features of the action by LSTM. A more accurate recognition effect is obtained by combining the clues of the auxiliary regions.

3. A spatial-Temporal Fusion Network Model

Currently, 3D CNN and two-stream CNN are mainly used for spatial and temporal fusion. 3D CNN captures spatial-temporal features directly from the original video sequences with universal applicability. However, it does not consider the relationship between the spatial and temporal features, and the densely sampled video frames of the 3D CNN will also produce a large number of parameters, which greatly increases the computation. Two-stream CNN, respectively, extracts the spatial and temporal information through two parallel spatial and temporal networks, which is more conducive to processing and fusing the two, and its training complexity is small. The two-stream network structure is shown in Figure 1.

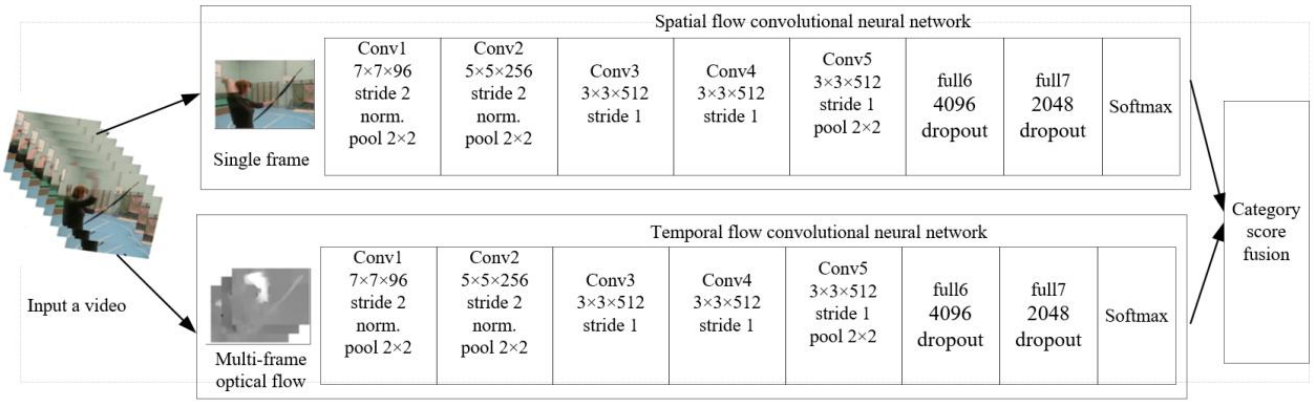

Figure 1. Two-stream network structure.

As can be seen in Figure 1, the spatial flow network processes the static data through multiple convolutional layers and fully connected layers to provide information about the scenes and objects of the network. Using the same convolutional network structure as the spatial flow network, the temporal flow network takes the stack of adjacent *L*-frame image optical flow as input, processes the dynamic information, and represents the motion and temporal information through multiple frames of optical streams. After Softmax, the two-stream network undergoes independent training for fusion operations.

Action recognition research has been greatly influenced by the two-stream network, yet there is still room for improvement. The fusion method is only the direct fusion of the prediction results for the two-stream network classifier, which is relatively simple to implement, but it fails to fully integrate the temporal and spatial information. The spatial stream operates on a single frame. The dense optical flow of the temporal flow can only learn the running information between adjacent frames, and cannot conduct feature learning for the whole video. Therefore, it cannot learn complex and long-term videos, and the network recognition effect is limited.

For this reason, subsequent researchers proposed various methods to improve the two-stream network. A common solution is to use a more intensive image frame sampling method to obtain the long-term information of videos, but it contains a lot of redundant data and increases the cost. TSN offers a time domain segmentation structure with a sparse sampling frame, and this structure can remove some redundant information and extract action information from the whole video.

The TSN framework uses sparse sampling methods to extract short segments of the entire video, randomly samples segments in short sequences, and fuses the category scores of different segments by the segment consensus function. The two-stream CNN utilized in the TSN framework merely extracts the motion and spatial information independently, and finally combines them without taking into account the correlation between the spatial and temporal information. The framework fused after the fully connected layer destroys the space-time properties to some extent. In addition, there are synchronous and asynchronous relationships between the motion information of the video behavior and the scene information in the process of action, but TSN does not consider the relationships.

In view of the defects of the traditional fusion spatial-temporal network and the spatial-temporal asynchronous relationship of actions, this paper adopts the time domain segmentation strategy for TSN to sample segments randomly. The sampled RGB video frames and the two-stream fields before and after a period of time are, respectively, sent into the spatial and temporal flow networks to learn the spatial-temporal features of the video. Both of the spatial-temporal asynchronous information in the convolutional layer will be fused. The fused feature information contains the matching relationship between the motion sequence information and the single-frame image. It is now possible to effectively combine the asynchronous spatial-temporal information of the video by performing long-term motion modeling and extracting temporal features using Bi-LSTM. The final classification is achieved by the Softmax classifier to complete the behavior recognition, and the proposed structure of the spatial-temporal fusion network model is shown in Figure 2.

The leftmost nine images in Figure 2 represent the input images in groups of three, and each group of images is the same, and the three images of the same group represent the three channels of the image. The following three boxes show the early-stage spatial-temporal fusion, the mid-term spatial-temporal fusion, and the late spatial-temporal fusion. The upper side image of the box indicates the information about the three channels for the frame image, and the lower side of the gray image shows the information about the two directions of the overlapping optical flow field for the single frame image corresponding to the time.

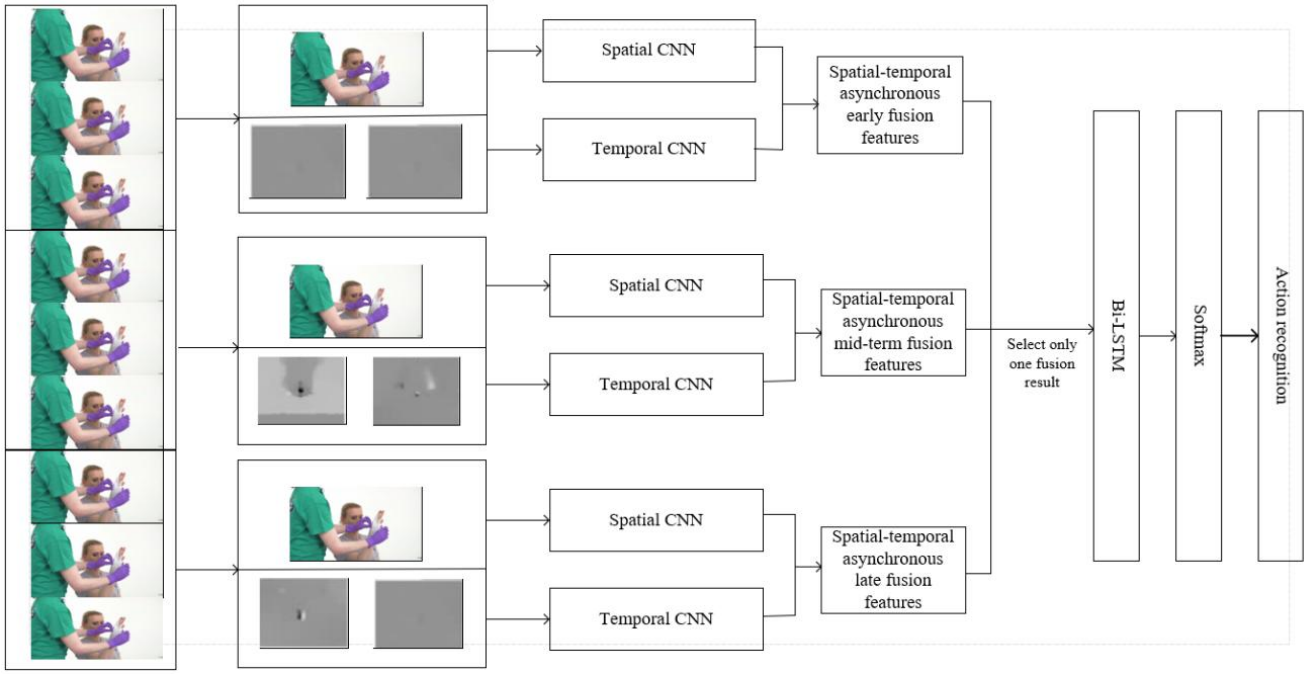

Figure 2. The proposed structure of the spatial-temporal fusion network.

The same video is input into the early-stage fusion, mid-term fusion, and late fusion to extract temporal and spatial features separately. The two features of the different fusions are then fused, respectively, and the feature with the best effect is selected from the fused features as the input of Bi-LSTM to further mine the asynchronous information.

The spatial-temporal fusion network adopts the sparse sampling strategy of the TSN framework, models the long-range time structure based on the segmented sampling, and obtains the video-level prediction results, to effectively learn the action model using the entire action video. Considering the spatial-temporal asynchronous relationship between motion and space, the asynchronous spatial-temporal relationship is also obtained based on the extraction of the spatial-temporal synchronous relationship by inputting asynchronous motion sequence graphs to achieve effective action recognition. Extract the spatial-temporal information by integrating it into the convolutional layer of the spatial-temporal two-stream network, avoid going through the fully connected layer, and effectively reduce the feature loss. Inputting the fusion feature map into Bi-LSTM can model the temporal relationship of video segments, and give the correlation relationship before and after video frames. The network can extract and fuse spatial-temporal relations well to realize action recognition.

4. A Human Action Recognition Method Based on Spatial-Temporal Fusion Network

A sparse sampling method is used to sample the videos. For an input video V, divide it into k segments $\{S_1, S_2, \ldots, S_k\}$ of equal time, and then randomly sample a segment $\{T_1, T_2, \ldots, T_k\}$ from the corresponding segments. The data are preprocessed to extract the RGB image of each segment T_k and its corresponding x and y directional optical flow graphs before and after a period of time, and the optical flow covers the front and back of the image. In this way, not only the motion information on the current action but also its asynchronous motion information is considered.

To obtain each fragment, the paper adopts a sparse sampling method. Utilize the RGB image, and the x direction and y direction optical flow graphs containing time series information, i.e., including not only a small segment of optical flow graph corresponding to the image, but also the optical flow graphs before, during, and after its motion. Assuming that the duration of the motion information is L frames, the input optical flow information involves $2 \times L$, and fully contains the asynchronous information of time and space.

For the tth video frame, its spatial features are extracted by the corresponding one RGB image, the extraction of motion information starts from the tth frame optical flow graph, and then superimposes the horizontal and vertical optical flow fields of N subsequent video frames, that is, input the optical flow fields of $[t, t + N]$ into the temporal convolutional neural network for analysis. The traditional two-stream methods merely extract an RGB image and a set of such optical flow features as input for each sampling segment, but this is only effective for actions with synchronous spatial and temporal information. Some actions are not always synchronous between the spatial and motion information, there are early or late cases, namely asynchrony.

For example, in the carrying behaviors of emergency rescue, when the motion extracted by the optical flow information block is exactly the carrying process of the paramedics, a simple two-stream network such as image and optical flow cannot accurately distinguish the carrying of emergency rescue from ordinary carrying behaviors. However, in the process of emergency rescue, there are frequent behaviors such as simple treatment and moving the injured who need to be carried on stretchers before the carrying action, as well as professional rescue after the carrying action. These actions either advance or lag behind the carrying actions on the image. Thus, in the process of emergency rescue, there is very often strong advance and lagging asynchronous information between the movement of the behavior and image.

Traditional two-stream networks do not consider the asynchronous characteristics between spatial and motion information, so the proposed method considers the asynchronous motion features with spatial information, to obtain more spatial-temporal relationships for effective action recognition. In addition to extracting the N frame optical flow field synchronized with the tth frame image of the traditional methods, the presented method also extracts more optical flow features asynchronously associated with the tth frame image. Take Δ as the time interval, extract a total of $[t - 2\Delta, t + 2\Delta + N]$ frame light-flow fields, and the obtained time domain features are used as the input of the temporal flow network.

Spatial-temporal fusion network enables action classification by extracting and incorporating spatial and temporal features of videos. The traditional two-stream network integrates the spatial-temporal information after passing through the fully connected layer but destroys feature information. To make good use of the extracted space and motion features in the pixel-level correlation to achieve full fusion, this paper compares the different fusion methods, fuses the spatial-temporal features in advance right in the convolutional layer, and the fusion feature sequences contain the synchronous and asynchronous correlations between the motion sequence information and single frame image. The asynchronous relationship between the motion sequences and the space is different for different actions. Therefore, modeling the temporal features is also required for the mid-level spatial-temporal fusion feature graphs following the fusion of motion and spatial information. This paper introduces Bi-LSTM to extract temporal features from the fused feature sequences. It can solve the problem that TSN does not consider the correlation of spatial-temporal information without destroying its spatial-temporal characteristics. At the same time, it can further extract the synchronous and asynchronous relationships between motion and scene information. The flowchart of the proposed spatial-temporal fusion network method is shown in Figure 3.

4.1. Spatial Flow Convolutional Neural Networks

Through the sparse sampling of the videos, a set of single-frame RGB images are obtained as the input of the spatial flow CNN, and the actions are discriminated by the static spatial appearance information. Static RGB images use three channels to store pixel information and represent the shape. For the actions with the obvious correlation between objects and scene information, the actions can be classified through the scenes and the objects in the video frame, such as bandaging.

The actions and certain objects are inseparable in the videos, for example, bandaging the wound requires gauze. The spatial flow CNN extracts the spatial features of actions

by identifying the background and shape information in the RGB frames. Therefore, the spatial flow networks can effectively recognize video behaviors by directly using image classification networks.

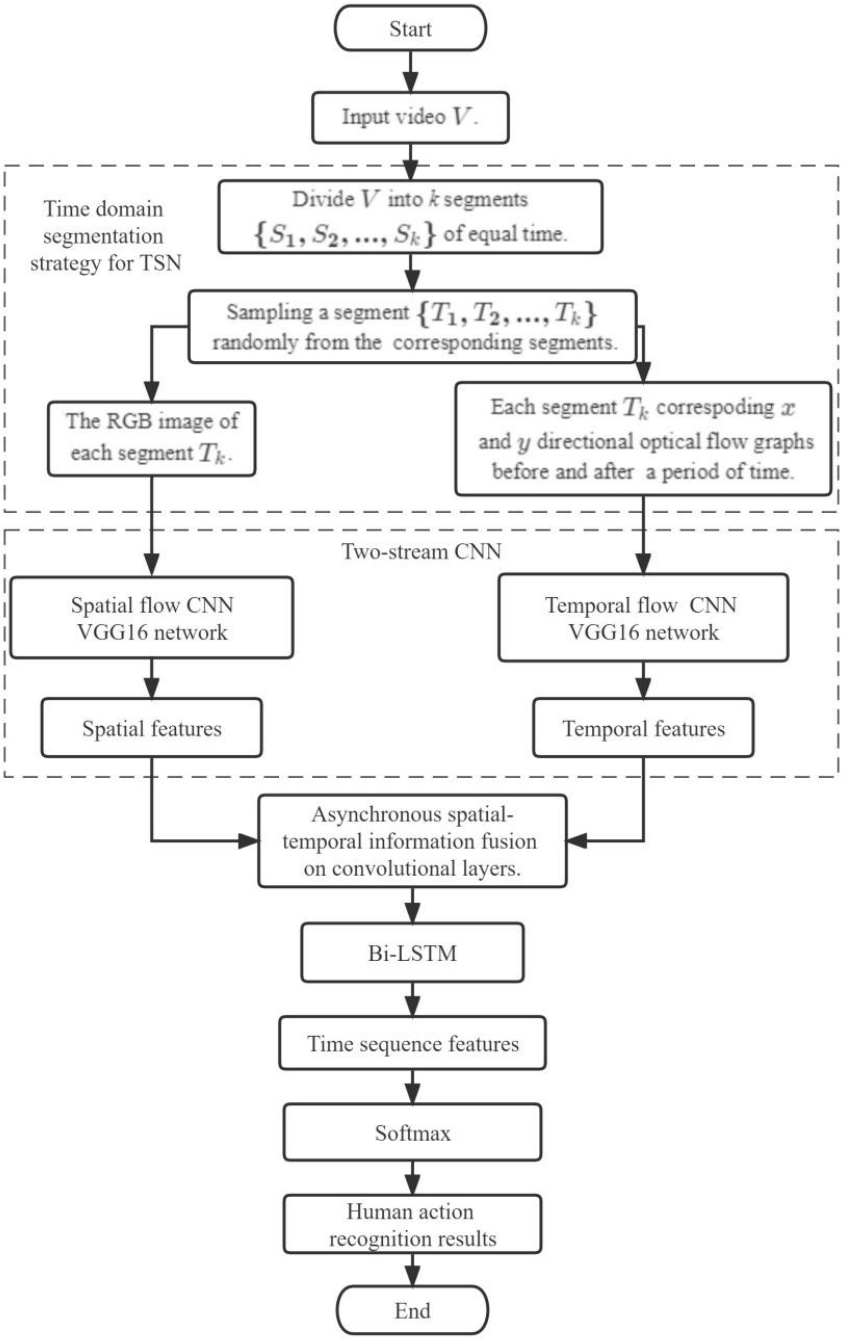

Figure 3. The flowchart of the proposed spatial-temporal fusion network method.

Both the spatial and temporal streams are detected by CNN in the two-stream network architecture. The spatial flow CNN focuses on extracting the features of the RGB image sequences, while the temporal CNN processes the optical flow information between the adjacent frames [25]. Spatial and temporal convolutional neural networks adopt the same network structure. The CNN used in this paper is the VGG16 network [23], and the network structure is shown in Figure 4.

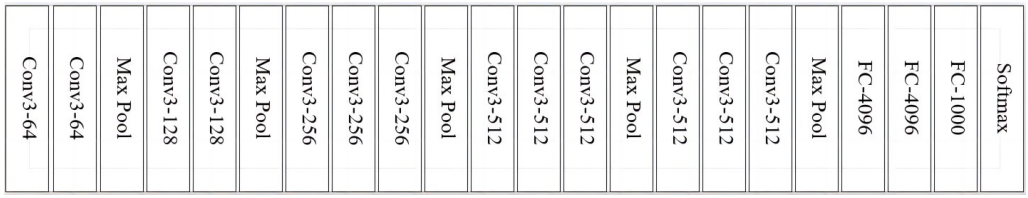

Figure 4. The VGG16 Network Structure.

4.2. Temporal Flow Convolutional Neural Networks

The optical flow is utilized to measure the motion information about the video behaviors in the temporal network of the two-stream network [26]. A video is a combination of video frame images composed of continuous pixels. The optical flow method can detect the speed and direction of the target movement, which is judged by viewing the intensity changes of the pixels between continuous images. It expresses the changes in the images and contains the movement information about the targets, so the temporal features of the moving targets can be obtained [27].

In a temporal network, the input is $2L$ image frames consisting of stacked x directional horizontal optical graphs d_x of L continuous video frames and y directional vertical optical graphs d_y. Assuming that the video frame size is $w \times h$, for any input video frame τ, the input optical flow field block $I_\tau \in R^{w \times h \times 2L}$ of the temporal stream can be calculated by Equation (1).

$$\begin{aligned} I_\tau(u, v, 2k-1) &= d^x_\tau(u, v), \\ I_\tau(u, v, 2k) &= d^y_{\tau+k-1}(u, v), \\ u = [1; w], v &= [1; h], k = [1; L] \end{aligned} \tag{1}$$

where I_τ represents the superposition of optical flow field blocks, namely the input of the temporal network. d^x_τ and d^y_τ, respectively, indicate the horizontal and vertical optical flow fields at time τ. L is the number of video frames, and (u, v) is the offset.

In addition to inputting the superimposed optical flow of the $2 \times L$ ($L = 10$) continuous frames at the time of the motion, the proposed model also extracts the optical flow information of $2k$ before and after as the input of the temporal network. This is because the optical flow from the motion start frame (t_0) to the motion end frame ($t_0 + L$) is the motion information synchronized with the spatial information. In addition to the information, there is also motion information that is asynchronous with spatial information, which is critical to action recognition. Therefore, the model also inputs k frames before the start of motion ($t_0 - k$) and k frames after the end of motion ($t_0 + L + k$) as the input of the temporal network.

Spatial-temporal information is asynchronous. For the two actions of infusion and injection, the motion of touching the human body with a needle tube is very similar. If only the optical flow information at this moment determines the action category, similar motion sequences will easily lead to misjudgment. Therefore, in addition to the frame of the motion moment, it is necessary to input the optical flow information before and after to assist the judgment. For the action at time t, in addition to inputting the RGB image and optical flow at time t, the optical flow information at time t, $t - k$, $t + L/2$, $t + L$, and $t + L + k$ should be input. The presented two-stream network structure is shown in Figure 5.

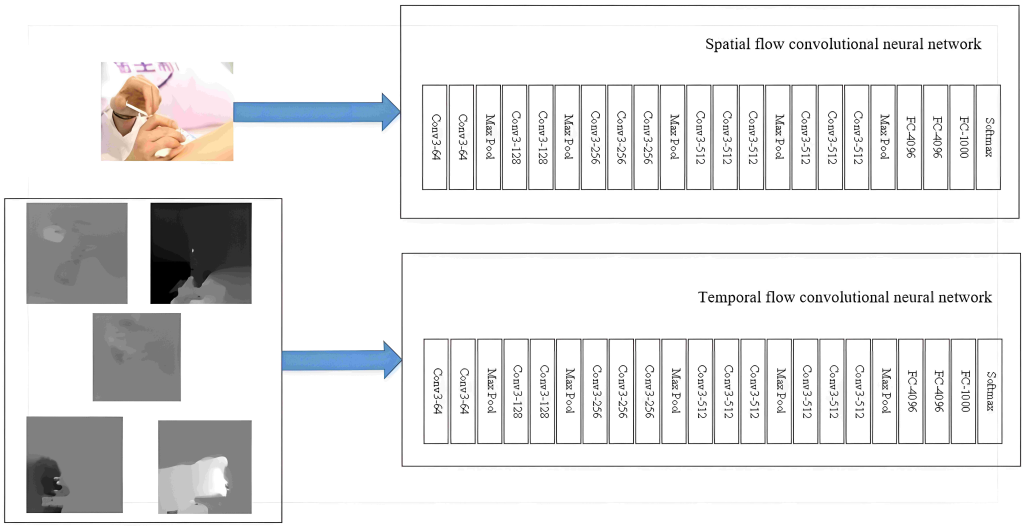

Figure 5. The presented two-stream network structure.

4.3. Spatial-Temporal Feature Learning and Fusion

The spatial and temporal flow networks in two-stream neural networks obtain the corresponding classification results before the fusion to realize the final identification. The fusion of spatial-temporal networks mainly makes full use of the spatial and motion features of the videos and combines the correlation between the spatial and motion features, to judge the different behavior types. For example, in the bandaging behavior, the spatial flow network can identify the shape information of the hands and triangle towel, and the temporal flow network extracts the periodic action of the hands in a specific spatial position, so combining both of them can identify the bandaging action. However, the fusion of category scores after the fully connected layer cannot achieve the true sense of correlation fusion. To fully exploit the connection between spatial and temporal properties, the fusion of spatial and temporal streams needs to be thoroughly studied. To integrate the spatial and temporal network streams, this paper investigates three potential spatial and temporal fusion techniques.

(1) Early-stage fusion:

Early-stage fusion is performed before the input network by fusing the sparsely sampled single-frame RGB image and *L*frame superposition optical flow fields, i.e., three-channel information of the frame images and two directions of light flow field information are fused to form 3 + 2*L* channels, then input to the network to extract spatial-temporal features and achieve action classification. The early-stage fusion process is shown in Figure 6.

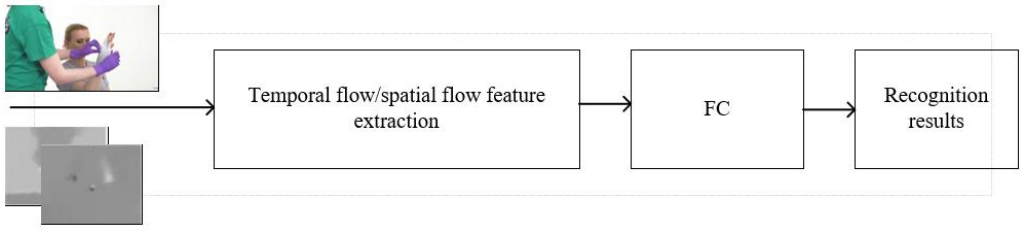

Figure 6. Early-stage fusion process.

(2) Mid-term fusion:

Mid-term fusion is the fusion in the network. The single-frame image and the overlapping optical flow field of its corresponding time are sent as input to the spatial flow and the temporal flow networks, respectively, and the spatial features and motion features of the video are extracted from the multi-layer convolutional layers. The extracted spatial-

temporal features are fused in the convolutional layers to generate the feature graphs and spatial-temporal feature vectors, and then the classifier is used to classify the actions.

Figure 7 shows the mid-term fusion process. Mid-term fusion mainly includes summation fusion, maximum fusion, and mean fusion.

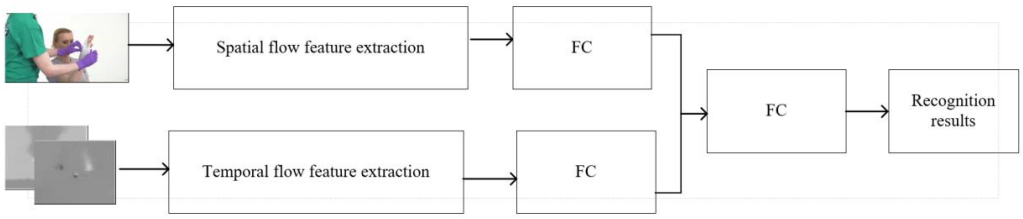

Figure 7. Mid-term fusion process.

(3) Late fusion:

Most of the fusion methods adopted by the traditional two-stream networks are late fusion. After the video information is input to the spatial and temporal flow networks, the corresponding category scores are obtained through feature extraction, and the two scores are directly fused to obtain the final recognition results, as shown in Figure 8.

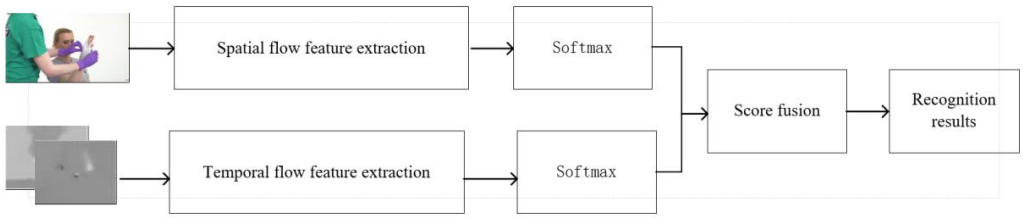

Figure 8. Late fusion process.

Assuming that f_{st} and f_{tp} are eigenvectors, respectively, extracted from the spatial and temporal flow CNN. Calculating the score by the Softmax classifier is shown in Equations (2) and (3).

$$p(j|f_{st}) = S_{st}^{j} = \frac{exp(\theta_{st}^{j} \cdot f_{st})}{\sum_{j'=1}^{n} exp(\theta_{st}^{j} \cdot f_{st})} \tag{2}$$

$$p(j|f_{tp}) = S_{tp}^{j} = \frac{exp(\theta_{tp}^{j} \cdot f_{tp})}{\sum_{j'=1}^{n} exp(\theta_{tp}^{j} \cdot f_{tp})} \tag{3}$$

where θ_{st}^{j} and θ_{tp}^{j} represent the Softmax classifier parameters in the spatial and temporal flow CNN, respectively. $p(j|f_{st})$ and $p(j|f_{tp})$ denote the posterior probabilities that f_{st} and f_{tp} belong to the jth category [28].

The obtained spatial-temporal high-level features are fused in the convolutional layer to form a spatial-temporal fusion feature graph. In this way, the pixel-level fusion can be directly realized without passing through the fully connected layer, and the spatial-temporal correlation information can be extracted to achieve full fusion without affecting any features.

The proposed model in this paper changes the traditional fusion approaches by fusion in the convolutional layer and considers the synchronization and asynchronism of spatial-temporal information to fully fuse spatial-temporal features without destroying the spatial-temporal features. The input of the original two-stream network is a moving image and a motion sequence, and the spatial-temporal synchronization relationship is considered. The presented method adds another motion sequence to the input of the temporal flow to extract the asynchronous motion information with the moving image to model the long-term motion.

4.4. Bi-LSTM Time-Series Feature Learning Network

The Bi-LSTM network in Figure 3 extracts the fused temporal features and contains the matching relationship between motion information and multi-frame images. In the matching process, the asynchronous relationships between motions and images are different for different categories of behaviors, and thus further deep learning of the fused temporal features is required after the fusion of motion and image features. To further mine the synchronous and asynchronous information of the spatial-temporal networks, this paper introduces the Bi-LSTM network to construct the long-term motion model of the fused sequences. Bi-LSTM is a good method to model temporal data. In Bi-LSTM, the input at a certain moment will depend on the video frame information before and after it, which can well satisfy the asynchronous relationship of video actions and fully consider the temporal information. Compared with LSTM, Bi-LSTM can obtain stronger temporal information and realize the effective integration of video asynchrony information in the case of learning the front and rear video information (Algorithm 1). The pseudo-code of the Bi-LSTM temporal feature learning network algorithm is as follows. The meaning of each symbol in the pseudo-code is shown in Table 2.

Algorithm 1: Bi-LSTM temporal feature learning network algorithm.

function $y_t = f(x_t)$
  Current input x_t is the same as the past output $h^j_{t\text{-}1}$
  $f(t)^j = \sigma(W^j_f \cdot h^j_{t\text{-}1} + U^j_f \cdot x_t + b^j_f)$
  $i(t)^j = \sigma(W^j_i \cdot h^j_{t\text{-}1} + U^j_i \cdot x_t + b^j_i)$
  $a(t)^j = tanh(W^j_a \cdot h^j_{t\text{-}1} + U^j_a \cdot x_t + b^j_a)$
  $o(t)^j = tanh(W^j_o \cdot h^j_{t\text{-}1} + U^j_o \cdot x_t + b^j_o)$
  $c(t)^j = c(t\text{-}1)^j \odot f(t)^j + i(t)^j \odot a(t)^j$
  $h(t)^j = o(t)^j \odot tanh(c(t)^j)$
  $y_t = softmax((h^1_t + h^2_t)/2)$
end function

Table 2. The meaning of each symbol in the pseudocode.

Symbol	Meaning
y_t	The temporal feature sequence obtained from the Bi-LSTM temporal feature learning network
x_t	The fusion feature sequence passing through the two-stream fusion network at time t
j	The directions of the input sequence, where $j = 1$ represents x_t to be the forward input and $j = 2$ represents x_t to be the reverse input
$h^j_{t\text{-}1}$	The input value of the LSTM network at time $t - 1$ with the jth type of input
f	Forget gate
i	Input gate
a	Feature extraction operations
o	Output gate
W	The weight of the output value at time $t - 1$
U	The weight of the input value at time t
b	The bias

5. Experiment Results

5.1. Datasets

This paper studies the action recognition for the whole emergency rescue video sequences, which is the classification of emergency rescue actions for the whole video data, so the annotation file is in the form of $\{video, action\_id\}$ and the action ($action\_id$) is specified action for each video ($video$). Experiments mainly adopt an improved self-built emergency rescue dataset and the publicly available dataset UCF101.

5.1.1. An Improved Emergency Rescue Dataset

Based on referring to the abundant information, there are few action datasets in emergency rescue scenes. The existing datasets for spatial-temporal action recognition usually provide sparse annotations for composite actions in brief video clips. The emergency rescue dataset used in experiments is an improved research result of the authors [4].

The video dataset of spatiotemporally localized Atomic Visual Actions (AVA) densely annotates 80 atomic visual actions in 430 15-min video clips, where actions are localized in space and time, resulting in 1.58M action labels with multiple labels per person occurring frequently. The AVA dataset defines atomic visual actions using movies to gather a varied set of action representations. This departs from existing datasets for spatial-temporal action recognition, which typically provides sparse annotations for composite actions in short video clips. AVA, with its realistic scene and action complexity, exposes the intrinsic difficulty of action recognition. Since there are many people and multiple actions in the identification scenes of dynamic emergency rescue, the self-built dataset of literature [4] is built with reference to AVA.

The data divided the actions in the dynamic scenes of emergency rescue into daily actions and medical rescue actions including carrying, cardio-pulmonary resuscitation (CPR), bandage, infusion, injection, oxygen supply, standing, walking, running, lying, sitting, and crouching/kneeling.

We collected various videos about emergency rescue scenes from a variety of video websites such as YouTube, Tencent Video, and Bilibili, and intercepted the videos to obtain the segments related to emergency rescue operations using the video editing software FFmpeg, and include a total of 700 video segments. In addition to the videos collected in the literature [4], the daily actions also use some segments of the KTH public dataset [29]. To increase the recognition accuracy of small targets for large ranges, some small target data are also added to the dataset. Some examples of the dataset are shown in Figure 9.

Figure 9. Some examples of the dataset.

A bounding box is used to locate a person and his or her actions. For each piece of video data, keyframes are extracted, and human-centered annotations are performed. In each keyframe, each person is marked with the preset action vocabularies of the paper that may have multiple actions.

To improve the annotation efficiency, Faster R-CNN is used to detect the position of the person, and Via software is utilized to annotate actions and people. In the stage of action annotation, this paper deletes all incorrect bounding boxes and adds missing bounding boxes to ensure high accuracy. During the labeling stage, each video clip is annotated by three independent annotators to ensure the accuracy of the dataset as much as possible.

Marking all actions of all people in all keyframes, most person-bounding boxes have multiple labels, which naturally leads to a type imbalance between action categories. Compared to daily actions, there are fewer medical actions. This paper refers to the features of the AVA dataset and runs the identification model on actions without adopting the manually constructed and balanced datasets. For the actions annotated by the self-built dataset, the frequency distribution of the various action categories is counted in Figure 10.

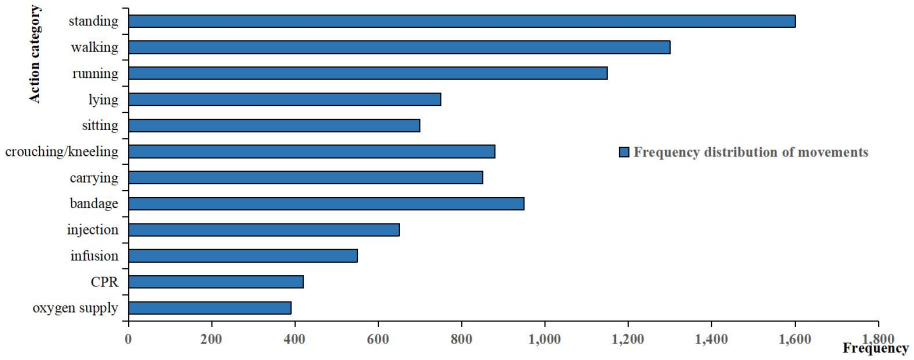

Figure 10. Action category frequency distribution in the self-built dataset.

The number of manual annotation samples is smaller. To prevent the overfitting phenomenon in the network and increase sample diversity, this paper collects action sequence video data as the emergency rescue human action data for annotation and then expands the data through data augmentation methods.

Data augmentation can address the issue of sample class imbalance and prevent overfitting in neural networks. The preprocessing includes short-edge resizing as well as normal operations. During the model training process, image augmentation methods of cropping sampling, translation transformation, and random flipping are used for the images.

The dataset has a great influence on the experiment results. To ensure the accuracy of the dataset, the paper trains the model on both the self-built and UCF101 datasets, and chooses the model with the best result to further test and adjust the dataset.

5.1.2. UCF101 Dataset

The mainstream action recognition dataset UCF101 has 13,320 videos from 101 action categories. The action categories include human–object interaction, human–human interaction, playing musical instruments, body-motion only, and sports. Since most of the available action recognition datasets are unrealistic and performed by participants in stages, UCF101 aims to encourage further research on action recognition by learning and exploring new realistic action categories. The database consists of realistic user-uploaded videos containing camera motion and cluttered backgrounds. UCF101 is currently the most challenging dataset of actions.

5.2. Metrics

This paper uses accuracy to evaluate the recognition results and visualizes the recognition results by adopting a confusion matrix. A confusion matrix is a performance mea-

surement for machine learning classification and is mainly used to count the number of predicted values in the wrong and right categories, respectively [30].

In the confusion matrix, rows represent actual values, columns represent predicted values, and the number of columns is equal to the total number of rows. True Positive (TP) shows the predicted positive and it is true. True Negative (TN) represents the predicted negative and it is true. False Positive (FP) denotes the predicted positive and it is false. False Negative (FN) indicates the predicted negative and it is false. The four types of samples have no intersection, and the sum of TP, FP, TN, and FN is the total number of samples.

The larger the diagonal values of the confusion matrix (TP, TN), the higher the correct classification probability of the model and the better the model performance. For an ordinary binary task, the confusion matrix is shown in Table 3. It is a table with four different combinations of predicted and actual values.

Table 3. Binary confusion matrix.

Confusion Matrix		**True Value**	
		Positive Classes	**Negative Classes**
Predicted value	Positive classes	TP	FP
	Negative classes	FN	TN

5.3. Analysis of the Experiment Results

To reflect the effectiveness of the spatial-temporal asynchronous fusion network alone, the VGG_16 network structure is adopted in the fusion network.

(1) Analysis of experiment results of spatial-temporal asynchronous information

The model inputs a total of $2(L+2k)(L=10)$ superimposed optical flows of successive frames, including synchronous and asynchronous information on spatial and motion sequences. The effect on the model is studied by taking different values of k, and the results are shown in Table 4.

Table 4. Effects of taking different values of k on the recognition results

k	**Accuracy Rate**
5	85.6%
10	86.5%
15	86.1%
20	85.9%

The result is best when k is taken as 10, that is, a superimposed optical flow field from $(t_0 - 20)$ to $(t_0 + 20)$ for consecutive 40 frames. When k is, respectively, taken as 15 and 20, there is too much confusing information, which can easily decrease the recognition accuracy.

When training the spatial-temporal fusion network, the input of the image recognition network is a static image frame at time t_0. The input of the optical flow network is superimposed optical flow fields from $(t_0 - 20)$ to $(t_0 + 20)$ centered on the time t_0, forming a $2 \times 40 = 80$ channel optical flow block, which is cut into 20 channel superimposed optical flow blocks by a sliding window with a step size of 5. Due to the limitation of hardware memory, the batch size used for network training is 8, which is equivalent to randomly sampling 8 frames of static images during each training and the optical flow fields of 20 frames before and after each frame.

Due to the strong spatial-temporal asynchrony of actions, this paper fully utilizes spatial and temporal asynchronous actions to improve accuracy by integrating the asynchronous information of spatial and temporal features. The effects of the spatial-temporal information on the recognition accuracy are shown in Table 5.

Table 5. Effects of spatial-temporal information on accuracy.

Input Information	Accuracy Rate
Spatial-temporal synchronous information	85.7%
Spatial-temporal asynchronous information	86.5%

(2) Analysis of the experiment results of the fusion methods

In this paper, spatial-temporal features are fused in the convolutional layer inside the two-stream fusion VGG16 model. During the experiment, the spatial-temporal feature fusion is performed on the convolutional layer Conv3 of the two-stream structure. The paper compares the separate spatial flow, temporal flow networks, and the two-stream networks, respectively. The action recognition accuracy of the different fusion methods is shown in Table 6.

Table 6. Comparison of action recognition accuracy for different fusion methods.

Actions	Separate Spatial Flow	Separate Temporal Flow	Two-Stream Network	Conv3
Carrying	82	77.6	84.3	85.4
CPR	81.5	76.8	85.2	88.2
Bandage	85.6	80.2	87.2	89.6
Infusion	80.2	76.5	84.9	86.5
Injection	80.8	76.3	84.7	87.3
Oxygen supply	83.4	77.2	83.6	85.4
Standing	81.2	76.8	87.4	88.3
Walking	82.1	79.3	87.3	89.1
Running	80.5	78.2	87.2	89.7
Lying	85.2	75.2	86.1	87.5
Sitting	83.7	77.4	86.5	88.3
Crouching/kneeling	82.3	78.1	85.6	86.9
Average accuracy rate	82.4	77.5	85.83	87.68

As can be seen from Table 6, the fusion of spatial-temporal asynchronous information in the convolutional layer Conv3 has the best effect. Different from the loss of information in the fully connected layer fusion, the fusion in the convolutional layer can not only retain better middle-level information of time and space but also obtain higher accuracy.

(3) Experiment comparison analysis of the action recognition for the spatial-temporal fusion CNN

To verify the effectiveness of the spatial-temporal fusion CNN in action recognition, the proposed method is compared with TSN, two-stream network, and two-stream network + Bi-LSTM methods, as shown in Table 7. It can be seen that the proposed method improves recognition accuracy.

Table 7. Comparison of recognition accuracy for different methods.

Methods	Accuracy Rate
The proposed method	90.1%
TSN	87.8%
Two-stream network	86.2%
Two-stream network + Bi-LSTM	86.8%

To give the model performance more intuitively, a confusion matrix is used to present the degree of confusion between the predicted and actual categories of the model, as shown in Figure 11.

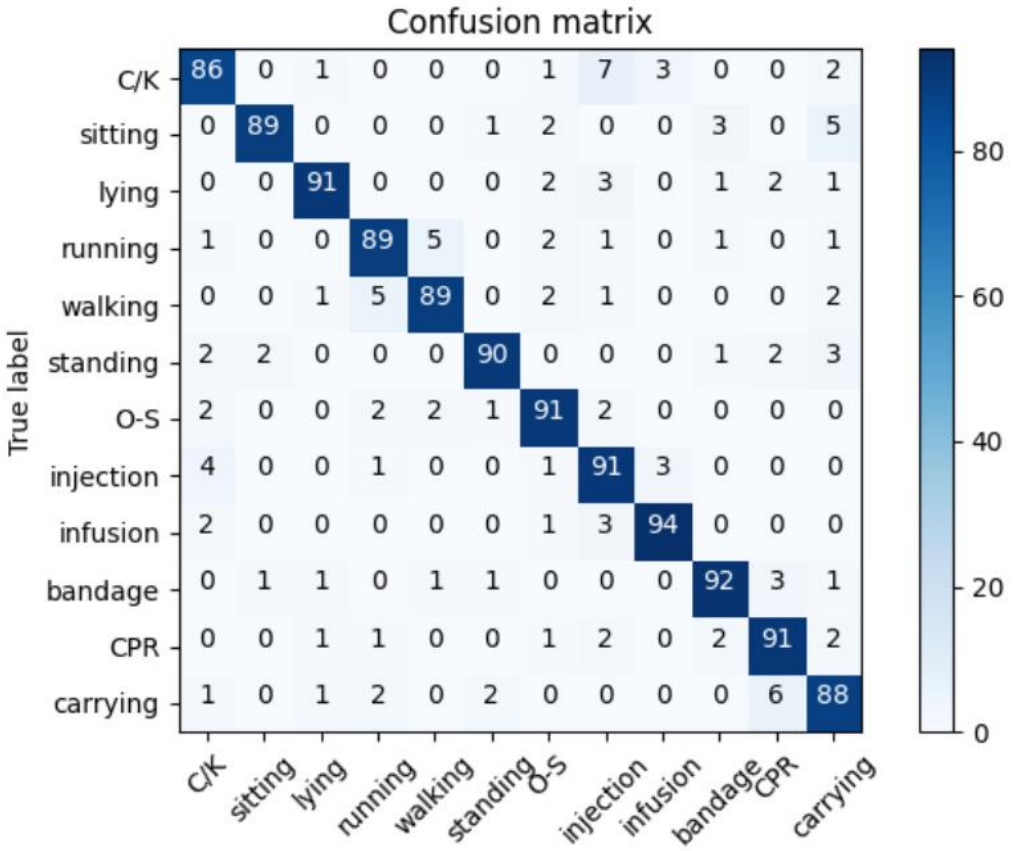

Figure 11. Confusion matrix.

In Figure 11, C/K represents crouching/kneeling, and O-S indicates oxygen supply. As can be seen from the confusion matrix in Figure 11, among the confusing actions, carrying is a 7% probability of being identified as walking because carrying has a certain overlap with walking and running, and other actions are similar. CPR is a 6% probability of identifying as C/K since CPR in emergency rescue situations is mostly in a kneeling position. Injection and infusion can be confused with each other, i.e., 5% of the injections will be predicted to be infusions and vice versa. Other actions are misclassified with a small probability. In daily human actions, the common action of walking is a 3% probability of being recognized as running, 1% probability of being identified as standing, and 4% probability of being recognized as carrying, because it is easy to be misclassified as carrying when several people gather in one place. C/K is a 6% probability of being identified as sitting. The above actions are easily confusing daily actions and medical rescue actions. There are still some misclassifications of easily confusing actions, but the misclassifications have reduced to some extent, and the model has a better ability to distinguish confusing actions.

The recognition results are visualized as shown in Figures 12–14. In Figure 12, the optical flow captures the dynamic action sequence information about the action of dressing a wound, although the injured man is sitting and relatively still, the action sequences of the ambulanceman capture the bandage, so the recognition results are the action of the bandage. Figure 13 shows the recognition results when the persons in the video are carrying uniformly.

Figure 12. Bandage action recognition results in a simpler background.

Figure 13. Carrying recognition results in a simpler background.

Figure 14. CPR recognition results in a more complex background.

For the more complex background situations, when the persons in the video perform different actions, the recognition result is the action with the highest probability of all actions, i.e., performing the most important action. Figure 14 gives the recognition results when the people in the video perform different actions, respectively, standing, C/K, and CPR, the dynamic action of the main person is CPR, and the recognition results are CPR, which means that the main execution action of the video is CPR.

Moreover, to verify the effectiveness of the proposed spatial-temporal fusion model, experiments are also conducted on the mainstream dataset UCF101 to compare the proposed method with the classical and advanced methods.

This paper compares single-flow CNN and various improved methods based on two-flow CNN, including the algorithm based on C3D, the traditional recognition algorithm based on two-stream convolutional networks (Two-stream Convnet), the Long-Term Recurrent Convolutional Networks (LRCN)-based recognition algorithm, two-stream network and LSTM fusion recognition algorithm (Two-stream + LSTM), recognition algorithm fused two-stream network and LSTM in convolutional layer (Two-stream + LSTM + ConvFusion), the improved human action recognition algorithm of Spatial Transformer Networks (STN) and CNN fusion [31], and the two-stream 3D Convnet fusion action recognition algorithm [15]. The comparison results are shown in Table 8.

Table 8. The experiment comparison results on the UCF101.

Methods	Accuracy Rate
C3D(1 nets) [17]	82.3%
C3D(3 nets) [17]	85.2%
Two-stream [12]	88%
LRCN [22]	82.9%
Two-stream + LSTM [32]	88.6%
Two-stream + LSTM + ConvFusion [33]	92.5%
literature [31]	90.5%
literature [15]	92.6%
The proposed method	93.2%

The comparison results show the proposed spatial and temporal fusion method has the best recognition effect, the method can accurately recognize the human action in the videos and verify the effectiveness of the method.

In terms of speed, on the UCF101 public dataset, the time complexity of this method is determined by TSN and Bi-LSTM with a running speed of 197.2 fps. The time complexity of C3D is determined by the convolutional layer with a running speed of 313 fps. The time complexity of the two-stream network is also determined by the convolutional layer with a running speed of 33.3 fps. The LRCN method is simpler than the C3D network structure, has a small number of parameters and is easy to train, and runs faster than the C3D network. The time complexity of literature [32] is determined by Two-stream together with LSTM with a running speed of 29.7 fps. The time complexity of literature [33], literature [31], and literature [15] is determined by the convolutional layer and LSTM. Literature [33] runs at two speeds, 6 fps when the input is an optic flow image and 30 fps when the input is an RGB frame. The time complexity of literature [31] is determined jointly by STN and CNN, and its running speed is 37 fps. Literature [15] runs at two speeds, processed at 186.6 fps when the input is an RGB frame. When the input is an optical flow image, the processing speed is 185.9 fps.

As can be seen from Table 8, the speed of the proposed method is slower than both the C3D and LRCN methods, mainly due to the high time complexity caused by the introduction of Bi-LSTM. Compared with the other methods, the proposed method is fast. Overall, the proposed method has the highest recognition accuracy and relatively fast speed.

In recent years, human action recognition methods have focused on deep learning. At present, the latest methods mainly include the TS-PVAN action recognition model based on attention mechanism [34], skeleton-based ST-GCN for human action recognition with extended skeleton graph and partitioning strategy [35], human action recognition based on 2D CNN and Transformer [36], linear dynamical system approach for human action recognition with two-stream deep features [37], and hybrid handcrafted and learned feature framework for human action recognition [38]. Comparative analysis with the latest methods is as follows.

(1) The TS-PVAN action recognition model based on attention mechanism can adequately extract spatial features and possess certain generalization abilities, but the temporal network of the TS-PVAN cannot efficiently model long-range time structure and extract rich long-term temporal information. This paper introduces Bi-LSTM to model long-term motion and fully mine the long-term temporal information.

(2) The human action recognition method combined Skeleton-based ST-GCN with extended skeleton graph and partitioning strategy can extract the non-adjacent joint relationship information in the human skeleton images, and divide the input graph of Graph Convolutional Network (GCN) into five types of fixed length tensors by the partition strategy, to include the maximum motion dependency. However, this method does not consider the temporal features. The proposed method extracts temporal information using the temporal network of the two-stream network.

(3) 2D CNN is one of the mainstream methods for human action recognition at present. 2D CNN-based framework not only has the advantages of lightweight and fast reasoning ability but also operates on short segments of sparsely sampled whole videos. However, 2D CNN still suffers from the insufficient representation of some action features and a lack of temporal modeling capability. The human action recognition method based on 2D CNN and Transformer adopts 2D CNN architecture of channel-spatial attention mechanism to extract spatial features in frames, utilizes Transformer to extract complex temporal information between different frames, and improves the recognition accuracy. However, Transformer extracts spatial and temporal features in sequential order, and as the number of frames increases, the number of parameters also increases substantially, causing a burden for the calculation. The paper adopts a dual-stream network structure to extract the spatial-temporal information, so the spatial-temporal feature extraction is in parallel, and the TSN sparse sampling strategy is used to avoid the greater computational burden caused by the increase in the number of frames.

(4) The human action recognition method combined linear dynamical system approach with two-stream deep features captures the spatial-temporal features of human action using a dual-stream structure. The method operates directly on video sequences. The longer the video sequence is, the more time is consumed. The presented method adopts the time domain segmentation strategy for TSN to randomly sample fragments and speed up the operation.

(5) The human action recognition method based on hybrid handcrafted and learned feature framework uses a two-dimensional wavelet transform to decompose video frames into separable frequency and directional components to extract motion information. The dense trajectory method is used to extract feature points for tracking continuous frames. However, this method can only deal with videos with clear action boundaries, which is also a disadvantage of the proposed method.

6. Conclusions

To address the problem that the spatial-temporal fusion network does not fully fuse the spatial and temporal dimension information which leads to a decrease in human action recognition accuracy, this paper proposes a dynamic scene human action recognition method based on the spatial and temporal fusion network model. Considering the strong asynchrony and time sequence of video action recognition, a spatial-temporal feature asynchrony fusion framework is designed to extract spatial and asynchronous temporal

features for fusion. Utilize Bi-LSTM to fully extract temporal information to capture video-level motion information and fuse spatial-temporal information, and realize human motion recognition by Softmax. The presented method can model long-term motion behaviors by modeling the asynchronous relationship between the moving image and the motion sequence (optical flow). The proposed method is still far away from practical application, so the robustness and real-time performance of the method can be further studied in the future.

Author Contributions: Conceptualization, methodology, data curation, validation, writing—original draft preparation Q.G., Y.Z., Z.D. and A.W.; writing—review and editing, Z.D.; supervision, project administration, Y.Z.; funding acquisition, Y.Z. and A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This paper was funded by National Key Research and Development Program Project (2020YFC0811004), National Natural Science Fund of China (61371143), and R&D Program of Beijing Municipal Education Commission (KM202110009002).

Data Availability Statement: The raw data can be provided on simple request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ye, D.; Qiu, W.G.; Zhang, L.C.; Huang, Y.H. Human action recognition based on 2S-LSGCN. *Comput. Eng. Des.* **2022**, *43*, 510–516.
2. Zhang, H.B.; Fu, D.M.; Zhou, K. Time-sequence-enhanced video action recognition method. *Pattern Recognit. AI* **2020**, *33*, 951–958.
3. Bao, J.Q.; Ye, J.Q.; Zhang, J.Z.; Lu, J.F. The Development and Thinking of China's Social Emergency Force under the New Situation. *China Emerg. Rescue.* **2020**, *6*, 38–42. [CrossRef]
4. Zhang, Y.; Guo, Q. Human Action Recognition Algorithm in Dynamic Scene of Emergency Rescue. In Proceedings of the 2021 IEEE 4th International Conference on Computer and Communication Engineering Technology (CCET 2021), Beijing, China, 13–15 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 12–16.
5. Mahjoub, A.B.; Atri, M. A flexible high-level fusion for an accurate human action recognition system. *J. Circuits Syst. Comput.* **2020**, *29*, 2050190. [CrossRef]
6. Luo, H.L.; Tong, K.; Kong, F.S. Summary of human action recognition in deep learning-based videos. *Electron. J.* **2019**, *47*, 1162–1173.
7. Zhou, H.; Liu, Y.X.; Gong, Y.; Kou, F.Y.; Xu, G.L. Action recognition algorithm based on dense trajectories and optical flow binarization image. *Comput. Eng. Appl.* **2022**, *58*, 174–180.
8. Abdelbaky, A.; Aly, S. Two-stream spatiotemporal feature fusion for human action recognition. *Vis. Comput.* **2021**, *37*, 1821–1835. [CrossRef]
9. Chen, T.T.; Yao, H.; Wei, Y.T.; Zuo, M.Z.; Yang, M.T. Human action recognition based on fusion features. *Comput. Eng. Des.* **2019**, *40*, 1394–1400.
10. Majumder, S.; Kehtarnavaz, N. Vision and inertial sensing fusion for human action recognition: A review. *IEEE Sens. J.* **2021**, *21*, 2454–2467. [CrossRef]
11. Akhtar, M.J.; Mahum, R.; Shafique, F.; Amin, R.; Ahmed, M.-S.; Lee, S.M.L.; Shaikh, S. A Robust Framework for Object Detection in a Traffic Surveillance System. *Electronics* **2022**, *11*, 3425. [CrossRef]
12. Simonyan, K.; Zisserman, A. Two-stream Convolutionalal networks for action recognition in videos. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
13. Wang, L.M.; Xiong, Y.J.; Wang, Z.; Qiao, Y.; Lin, D.H.; Tang, X.O.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 8–10 October 2016; pp. 20–36.
14. Zhuang, D.; Jiang, M.; Kong, J.; Liu, T.S. Spatial-temporal attention enhanced features fusion network for action recognition. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 823–841. [CrossRef]
15. Wang, X.; Gao, L.L.; Wang, P.; Sun, X.S.; Liu, X.L. Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length. *IEEE Trans. Multimed.* **2018**, *20*, 634–644. [CrossRef]
16. Ji, S.W.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef] [PubMed]
17. Tran, D.; Bourdev, B.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatial-temporal features with 3D Convolutional networks. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
18. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733.
19. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Arzani, M.M.; Yousefzadeh, R.; Gool, L.V. 3D convolutional neural networks for human action recognition. *arXiv* **2017**, arXiv:1711.08200.

20. Qiu, Z.; Yao, T. Mei, T. Learning spatio-temporal representation with pseudo-3D residual networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 5534–5542.
21. Feichtenhofer, C. X3D: Expanding architectures for efficient video recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, 13–19 June 2020; pp. 200–210.
22. Donahue, J.; Hendricks, L.A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Darrell, T.; Saenko, T. X3D: Expanding architectures for efficient video recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
23. Ou, H.G.; Sun, J.F. Spatial-temporal information deep fusion network with frame attention mechanism for video action recognition. *J. Electron. Imaging* **2019**, *28*, 023009. [CrossRef]
24. Ge, P.; Zhi, M. Human action recognition based on the inference network. *Comput. Eng. Des.* **2021**, *42*, 853–858.
25. Wang, Y.; Ma, C.H.; Mao, Z.Q. Behavioral identification based on the space-time two-stream fusion network and the Attention model. *Comput. Appl. Softw.* **2020**, *37*, 156–159.
26. Liang, X.; Li, W.X.; Zhang, H.N. A Review of Human Behavior Recognition Methods. *Comput. Appl. Res.* **2022**, *39*, 651–660.
27. Wang, Z.Z.; Zhou, X.Z.; Yan, H. An anomalous behavior detection model based on the dual-stream structure. *Comput. Appl. Softw.* **2022**, *39*, 188–193.
28. Wang, Z.W.; Gao, B.P. Abnormal behavior recognition based on spatial-temporal fused convolutional neural networks. *Comput. Eng. Des.* **2020**, *41*, 2052–2056.
29. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004; pp. 32–36.
30. Yang, X.l. Summary of performance metrics for classification learning algorithms. *Comput. Sci.* **2021**, *48*, 209–219.
31. Yu, H.; Zhi, M. Human action recognition based on the improved CNN framework. *Comput. Eng. Des.* **2019**, *40*, 2071–2075.
32. Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
33. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, NV, USA, 27–30 June 2016.
34. Guo, J.L.; Hu, T.H.; Shi, S.J.; Chen, E.Q. TS-PVAN Action Recognition Model Based on Attention Mechanism. 2022. Available online: http://kns.cnki.net/kcms/detail/21.1106.TP.20221116.1017.008.html (accessed on 17 November 2022).
35. Wang, Q.Y.; Zhang, K.X.; Asghar, M.A. Skeleton-Based ST-GCN for Human Action Recognition With Extended Skeleton Graph and Partitioning Strategy. *IEEE Access* **2022**, *10*, 41403–41410. [CrossRef]
36. Zhu, X.H.; Zhi, M.;Yin, Y.Z. Human action recognition based on 2D CNN and Transformer. *IEEE Access* **2022**, *45*, 123–129.
37. Du, Z.N.; Mukaidani, H. Linear dynamical systems approach for human action recognition with dual-stream deep features. *Appl. Intell.* **2022**, *52*, 452–470. [CrossRef]
38. Zhang, C.; Xu, Y.; Xu, Z.; Huang, Z.; Lu, J. Hybrid handcrafted and learned feature framework for human action recognition. *Appl. Intell.* **2022**, *52*, 12771–12787. [CrossRef]