# Cross-Dimensional Feature Fusion MLP Model for Human Behavior Recognition

Jianfeng Zhang, Tianwei Shi, Wenhua Cui, Ye Tao, Huan Zhang

*Abstract*—**Aiming at the problem that human behavior is difficult to identify, a bottom-up and horizontal connections Cross-dimensional Feature Fusion MLP model (CFF-MLP) was proposed in this paper. To improve the detection performance of human behavior and realize the effective capture of continuous behavior features, a human key point Vector tracking (VT) algorithm based on optical flow was designed in the CFF-MLP model. This model can obtain the behavioral features and local dependencies of different layers hierarchically. Furthermore, it enhances the generalization ability of the model. The validity of the proposed model is verified and analyzed using KTH dataset and compared with other research methods. The experimental results show that the CFF-MLP model can effectively improve the detection ability of human micro-motion, and its average accuracy rate reaches 94.23%. It improves the detection efficiency of human behavior.**

*Index Terms*—**CFF-MLP model; Human action detection; KTH database; Vector tracking**

## I. Introduction

With the continuous development and innovation of computers and technology, human behavior detection has gained increasing attention. It is widely used in various fields such as human-computer interaction and intelligent sports, etc. Human movements generally include walking, running, swinging, squatting, sitting, jumping and so on. Human action recognition still faces great challenges due to the continuity, complexity, and ambiguity of the spatial background when the human body moves, which affects the final detection result. Traditionally, there are two difficulties in human motion detection: space complexity and time difference [1]. Spatial complexity, namely human movement, is usually restricted by the illumination intensity of the camera, the different angles of the image, and the complexity of background. Simultaneously, the partial occlusion of human body and the background of multi-person recognition also affect the recognition accuracy algorithm. Time difference is reflected by the fact that the starting time of human behavior cannot be predetermined, and the time interval is also variable. All these parameters have a huge impact on the recognition efficiency.

Since the deep learning method is efficient at staged learning through the input layer, huge data is widely available and has high computer power. In recent years, scholars have conducted a lot of experiments on human action recognition on Convolutional Neural Networks (CNN) [2]-[5]. Wang et al. proposed a Temporal Segmentation Network (TSN) that introduces the latest ResNet and Inception V3 deep model architectures to further improve the action recognition method [6]. Carreira et al. extended the filters and the pooling kernels of the ultra-deep image classification ConvNet to 3D images, forming a new convolutional model Two-Stream Inflated 3D ConvNet and realizing the focus of human spatiotemporal feature capture [7]. It used the Recurrent Neural Networks (RNN) to process characteristic learning of human movements [8]-[10]. Donahue et al. proposed a Long-term Recurrent Convolutional Network (LR-CN). This method combines convolutional layers and long-range temporal recursion to realize behavior judgment [11]. Jaouedi et al. used a human behavior recognition method, based on the fusion of sequential visual features and motion paths, to complete behavior detection [12]. Latha et al. established a three-layer neural network and performed detection on the KTH dataset by separating the human body from the background [13]. Salahuddin et al. extracted three behaviors of walking, running, and waving on the KTH dataset. They used the K-means clustering to reduce the data dimension and deployed a classifier to identify these three behaviors [14]. Yi et al. trimmed and calculated the redundant trajectory of human video in KTH dataset and combined appearance and motion saliency to obtain complementary information for the classification of the different behaviors [15]. Kai et al. used the space and the time channels of the VGGNET16 network structure to detect human behavior [16]. Liu et al. constructed a CNN human behavior recognition algorithm model to collect image data from KTH dataset and train it, but the accuracy of the used algorithm was not ideal [17]. Hongwei and Haibo combined the Faster R-CNN algorithm with the batch normalization algorithm and the online hard case mining algorithm to realize human behavior detection [18]. Emanuel et al. built a human action recognition system

Manuscript received June 22, 2022; revised October 14, 2022.

This work was supported by the Natural Science Foundation project of Liaoning Province (2021-KF-12-06), Department of Education of Liaoning Province (2020FWDF01), and Project of Liaoning BaiQianWan Talents Program.

Jianfeng Zhang is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: jianfengzhang177@163.com).

Tianwei Shi is an associate Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author, phone: 139-9805-3962; e-mail: tianweiabbcc @163.com).

Wenhua Cui is a Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: cwh@systemteq. net).

Ye Tao is a lecturer of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: taibeijack @163.com).

Huan Zhang is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: hzzhbest@163.com).

based on a single image or video capture snapshot. The models' input layer were 50 points from $x$ and $y$ coordinate of 25 keypoints from OpenPose, and the output layer was the numerical representation of 11 human action labels. Then used as inference engines to recognize human action from images and real-time video [19]. Although these methods identified the human behavior, they use the most discriminative small segment in an action category and failed to comprehensively utilize multiple small segments. In addition, the recognition accuracies were not ideal, the network lacked generalization ability and the amount of calculation was large.

To deal with the above problems, this paper proposes a novel detection model: the Cross-dimensional Feature Fusion MLP model (CFF-MLP). The CFF-MLP includes the Vector tracking (VT), Axial Shifted MLP (ASMLP), and CFF-based stages. This model uses the OpenPose network to extract the key points of human skeleton and develops a VT algorithm to calculate the matrix information output by the OpenPose network. Based on the Optical Flow (OF) method, VT algorithm performs feature extraction with time series according to the angle value and distance between the coordinates of the key points of the bones. The ASMLP stage is used to place the spatially shifted features on the horizontal and vertical directions in the same module and combine these features using the channel hybrid MLP. Among these methods, each module is horizontally connected to the Up-Sampling Layer of the CFF-based stage. This latter connects the model network layers horizontally to better map behavioral features at different scales. The KTH dataset is used for verification experiments and the average accuracy rate reaches 94.23%.

After presenting the main contributions of this work, this paper will be divided as follow: in Section II, the methodology of work are provided. Section III presents the simulation and the experiments and finally, and Section IV concludes this work and proposes future ideas.

## II. Methodology

### A. Architecture

The overall architecture of the CFF-MLP model is shown in Fig. 1. Video-based human action detection usually requires complex processing. When processing with long videos, action sequences are usually chosen. Action sequences will affect the accuracy of recognition due to their huge computational load and great influence by background factors. The VT of human key points is adopted to shorten the calculation time of the model and to eliminate the influence of complex background on the recognition accuracy. The OpenPose network is used in the frame extraction processing of the video. The human body pose is estimated through the RGB image and the feature matrix information output by OpenPose is inserted into the designed CFF-MLP. The CFF-MLP model can process the multi-scale changes of the human body and strengthen the detection of small movements. In this model, the extracted coordinates of the human skeleton points enter first the VT stage for operation, and the OF vector is solved by calculating the angle and the displacement of these key points. A time-series is introduced into the algorithm to realize the action extraction between consecutive frames of the video. A patch of size is first divided into multiple patch tokens in the ASMLP stage, the number of channels for each of which is determined by patch embedding. After ASMLP stage cross receptive field (horizontal and vertical) sampling and axial transfer to extract features, each cascaded stage is horizontally connected to the Up-Sampling Layer (UPL) of CFF-based stage to extract features of different sizes from different layers of the model. Since the extracted features have different dimensions, a Smooth Layer (SL) is designed in the CFF-based stage for processing. The CFF-based stage adapts to the final feature addition, to obtain the final detection result.

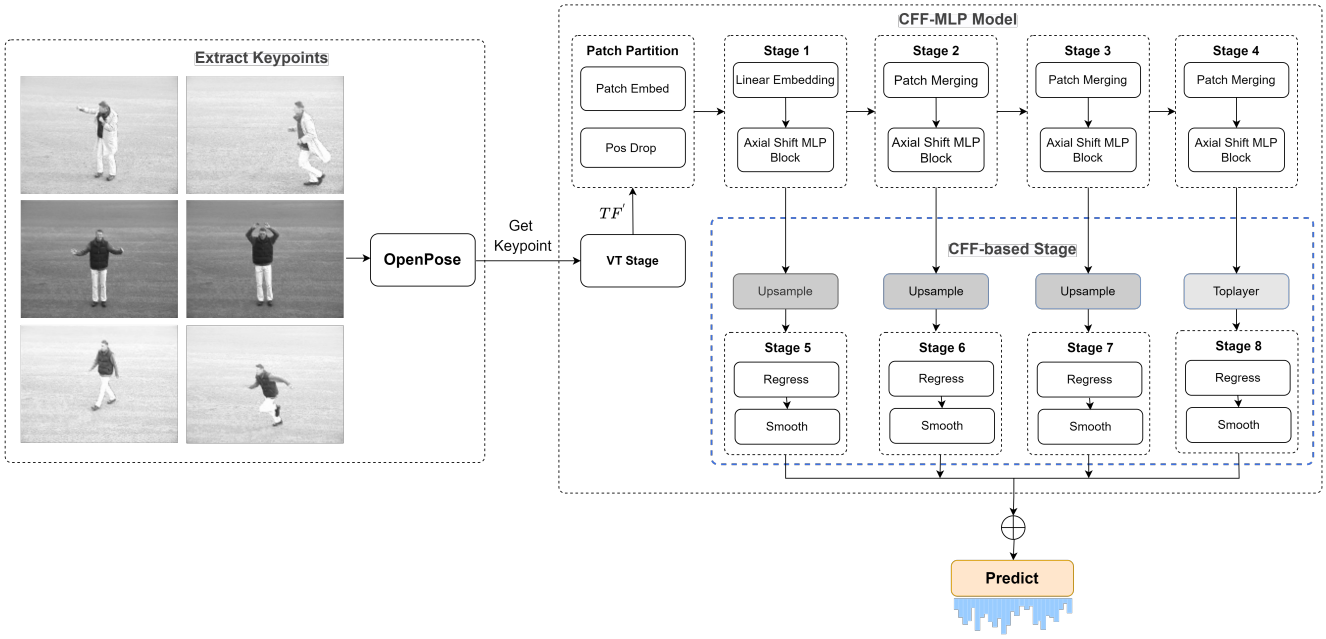


Fig. 1. Overall architecture of the CFF-MLP model

*B. VT Stage*

The angular displacement algorithm is usually used to solve the aperture problem in OF-based algorithms. OF maps the human behavior process on a two-dimensional plane with the corresponding image pixel motion direction and speed. Applying OF to human behavior detection can reflect, more significantly, the correspondence between the current frame and the previous frame. During the same object moves between different frames, it is assumed that the OF brightness does not change and the adjacent pixels have the same motion. Defining I as the pixel value of the pixel point $(x,y)$ at time *t*, the pixel point moves by $\Delta x$ , $\Delta y$ , $\Delta t$ between the two frames. The pixel value $I(x,y,t)$ will be described as:

$$I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t) \tag{1}$$

Assuming that the movement of the pixel point is very small, Eq. (1) adopts Taylor series expansion and eliminates the same terms. Then the Eq. (2) and Eq. (3) are obtained respectively.

$$u = \frac{dx}{dt}, v = \frac{dy}{dt} \tag{2}$$

$$I_x = \frac{\partial I}{\partial x}, I_y = \frac{\partial I}{\partial y}, I_t = \frac{\partial I}{\partial t} \tag{3}$$

According to the derivation process of Eq. (1), Eq. (2) and Eq. (3), the Eq. (1) can be redefined as that:

$$I_x u + I_y v + I_t = 0 \tag{4}$$

where $(u,v)$ is the desired OF vector. Since Eq. (4) is a constraint equation and has two unknowns, it cannot be solved precisely. VT algorithm can better judge the action state of the human body through conditional constraints and then solve the aperture problem. This algorithm uses OpenPose to frame and preprocess the human behavior video in KTH dataset, then the $25\times 3$ human skeleton key point matrix can be obtained.

$$KeyPoint_t = \begin{bmatrix} KeyPoint_{0,x,t} & , & KeyPoint_{0,y,t} \\ KeyPoint_{1,x,t} & , & KeyPoint_{1,y,t} \\ \vdots & & \vdots \\ KeyPoint_{i,x,t} & , & KeyPoint_{i,y,t} \end{bmatrix} \tag{5}$$

where *i* represents the selected *i*-th skeleton key point ( $0 \le i \le 24$ ) and *t* is the current time. If the picture being processed is from successive video frames, it uses the key points to calculate *u* and *v* representing the direction of motion.

$$u = \frac{dx}{dt} = \frac{\Delta x}{\Delta t} = \frac{KeyPoint_{i,x,t} - KeyPoint_{i,x,t-1}}{\Delta t} \tag{6}$$

where $\Delta x$ represents the relative displacement between the current frame and the previous frame. The value of $\Delta t$ is 1. The matrix $u'$ is obtained by the operation on the continuous video stream.

$$u' = \begin{bmatrix} [u_{0,0}, u_{1,0}, \cdots, u_{i,0}, \cdots, u_{24,0}] \\ [u_{0,1}, u_{1,0}, \cdots, u_{i,1}, \cdots, u_{24,1}] \\ [u_{0,t}, u_{1,t}, \cdots, u_{i,t}, \cdots, u_{24,t}] \\ [u_{0,n}, u_{1,n}, \cdots, u_{i,n}, \cdots, u_{24,n}] \end{bmatrix} \tag{7}$$

Similarly, the matrix $v'$ can be obtained according to Eq. (6) and Eq. (7).

$$v' = \begin{bmatrix} [v_{0,0}, v_{1,0}, \cdots, v_{i,0}, \cdots, v_{24,0}] \\ [v_{0,1}, v_{1,0}, \cdots, v_{i,1}, \cdots, v_{24,1}] \\ [v_{0,t}, v_{1,t}, \cdots, v_{i,t}, \cdots, v_{24,t}] \\ [v_{0,n}, v_{1,n}, \cdots, v_{i,n}, \cdots, v_{24,n}] \end{bmatrix} \tag{8}$$

Performing vector transformation on the obtained $u'$ and $v'$ to obtain the distance between the two coordinate points is initiated.

$$D_{i,t} = \sqrt{u_{i,t}^2 + v_{i,t}^2} \tag{9}$$

Using the OF vector to construct the feature vector, the feature matrix *F* at time *t* is given by:

$$F = \begin{bmatrix} 0 \\ \sqrt{u_{0,t}^2 + v_{0,t}^2} \\ \sqrt{u_{1,t}^2 + v_{1,t}^2} \\ \vdots \\ \sqrt{u_{24,t}^2 + v_{24,t}^2} \end{bmatrix} \tag{10}$$

The feature vector *F* represents the displacement of a single tracked skeleton key points. In human behavior recognition, the angle change between skeleton key points can better reflect the change characteristics than the movement direction of a single key point. In this paper, the angles of adjacent feature points are represented as directional changes.

$$\theta_{m,n,t} = \arctan\left(\frac{KeyPoint_{m,y,t} - KeyPoint_{n,y,t-1}}{KeyPoint_{m,x,t} - KeyPoint_{m,x,t}}\right) \tag{11}$$

where $\theta$ is the angle between two key points of the human body, *m* is the first point in the coordinates of the selected human key points, *n* is the second point, *t* and *t-1* represent the current time and the previous moment, *x* and *y* are the coordinates of the current point respectively. Combining the *s* pairs of feature point angles $\theta_{m,n,t}$ at the same time *t* into a matrix $\theta'$ :

$$\theta' = \begin{bmatrix} \theta_{m_1,n_1,t} \\ \theta_{m_2,n_2,t} \\ \vdots \\ \theta_{m_s,n_s,t} \end{bmatrix} \tag{12}$$

where $(m_1,n_1)$ and $(m_2,n_2)$ are the selected first and second pair of key points. *s* is the number of selected key point pairs. Each pair of the selected key points corresponds to two displacements. $\theta'$ is combined with the displacement to construct the matrix $F'$.

$$F' = \begin{bmatrix} \begin{bmatrix} \theta_{m_1,n_1,t} \\ \vdots \\ \theta_{m_{12},n_{12},t} \\ \vdots \\ \theta_{m_s,n_s,t} \end{bmatrix} & \begin{bmatrix} \sqrt{u_{m_1,t}^2 + u_{m_1,t}^2} \\ \sqrt{u_{m_2,t}^2 + u_{m_2,t}^2} \\ \vdots \\ \sqrt{u_{m_s,t}^2 + u_{m_s,t}^2} \end{bmatrix} & \begin{bmatrix} \sqrt{u_{n_1,t}^2 + u_{n_1,t}^2} \\ \sqrt{u_{n_2,t}^2 + u_{n_2,t}^2} \\ \vdots \\ \sqrt{u_{n_s,t}^2 + u_{n_s,t}^2} \end{bmatrix} \end{bmatrix} \tag{13}$$

To capture the continuous changes of human behavior in video frames, the time series is introduced into the matrix $F'$. The following equation can be obtained by extracting the continuous *k*-frame behaviors in the video.

$$TF' = \begin{bmatrix} \begin{bmatrix} \theta_{m_1,n_1,t} & \theta_{m_1,n_1,t+1} & \cdots & \theta_{m_1,n_1,t+k} \\ \theta_{m_2,n_2,t} & \theta_{m_2,n_2,t+1} & \cdots & \theta_{m_2,n_2,t+k} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{m_s,n_s,t} & \theta_{m_s,n_s,t+1} & \cdots & \theta_{m_s,n_s,t+k} \end{bmatrix} \\ \begin{bmatrix} \sqrt{u_{m_1,t}^2+u_{m_1,t}^2} & \sqrt{u_{m_1,t+1}^2+u_{m_1,t+1}^2} & \cdots & \sqrt{u_{m_1,t+k}^2+u_{m_1,t+k}^2} \\ \sqrt{u_{m_2,t}^2+u_{m_2,t}^2} & \sqrt{u_{m_2,t+1}^2+u_{m_2,t+1}^2} & \cdots & \sqrt{u_{m_2,t+k}^2+u_{m_2,t+k}^2} \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{u_{m_s,t}^2+u_{m_s,t}^2} & \sqrt{u_{m_s,t+1}^2+u_{m_s,t+1}^2} & \cdots & \sqrt{u_{m_s,t+k}^2+u_{m_s,t+k}^2} \end{bmatrix} \\ \begin{bmatrix} \sqrt{u_{n_1,t}^2+u_{n_1,t}^2} & \sqrt{u_{n_1,t+1}^2+u_{n_1,t+1}^2} & \cdots & \sqrt{u_{n_1,t+k}^2+u_{n_1,t+k}^2} \\ \sqrt{u_{n_2,t}^2+u_{n_2,t}^2} & \sqrt{u_{n_2,t+1}^2+u_{n_2,t+1}^2} & \cdots & \sqrt{u_{n_2,t+k}^2+u_{n_2,t+k}^2} \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{u_{n_s,t}^2+u_{n_s,t}^2} & \sqrt{u_{n_s,t+1}^2+u_{n_s,t+1}^2} & \cdots & \sqrt{u_{n_s,t+k}^2+u_{n_s,t+k}^2} \end{bmatrix} \end{bmatrix} \quad (14)$$

### C. ASMLP Stage

ASMLP stage uses axial displacement (vertical and horizontal displacements) and channel projection for feature capture. It is mainly divided into three parts: Patch Partition, Patch Merging and AS-MLP Block.

--The Patch Partition is used to divide the original image into multiple Patch token areas. The size of all areas combined is $48\times H\times W$. $H$ and $W$ represent the width and height of the original image. To prevent overfitting, a Pos Drop Layer is added to this module. By traversing the nodes of each layer, the node retention probability ranges between 0 - 1 and it is set to reduce the node weight;

--The Patch Merging is used to merge the output features processed by the Patch Partition and use a Linear Layer to transform the size of the features;

--The AS-MLP Block is the core part of ASMLP stage. It is mainly connected by Norm Layer, MLP, Axial Shift and Residual. Axial Shift represents the axial displacement operation. The network structure diagram of AS-MLP Block is shown in Fig. 2. In the Conv Layer, the $1\times1$ convolution is selected to greatly improve the nonlinear characteristics and the expressive ability of the model while keeping the scale of the feature map unchanged. However, with the model expression capability enhancement, the gradient will disappear in the process of back propagation. To reduce detection error rate, the output and input of each layer are linearly superimposed by nonlinear transformation by using residual connection.

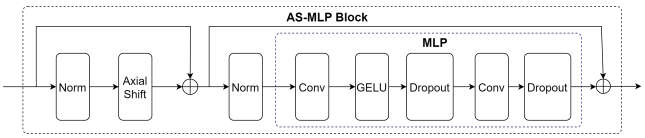


Fig. 2. AS-MLP Block network structure diagram

In Axial Shift, channel projection is used to extract features that are translated along the vertical and horizontal spatial directions as shown in Fig. 3. The input dimensions of the horizontal and the vertical shifts are both $C\times h\times w$. Assumed that $C=3$, $w=5$, the number represents the feature index and $P$ is Zero padding. The input features are divided according to the value of $C$. Then, a zero-padding operation is performed, and the selected features are extracted for the projection of the next channel. The features of different execution directions are recombined to realize the interaction of spatial information of different positions.

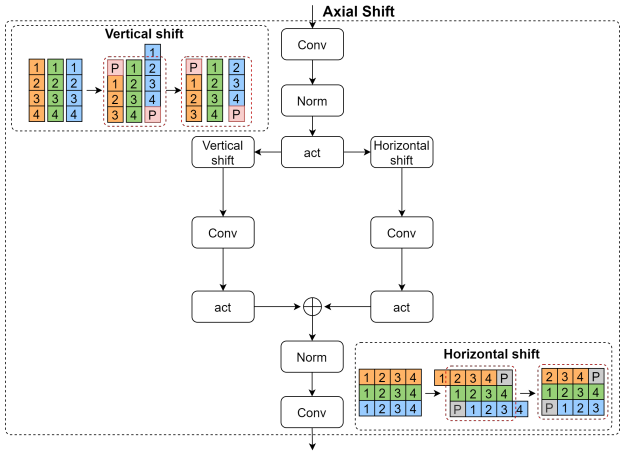


Fig. 3. Axial Shift

### D. CFF-based Stage

To solve the multi-scale change problem in human behavior detection, the CFF-based stage uses a bottom-up hierarchical structure with lateral connections to construct the features of each scale. this method can effectively avoid the scale robustness of the model to the original input features. Based on the Up-Sampling Layer, the module adds the Regress and the SL. They are connected and added horizontally with the up sampling to improve the detection ability of the changes even for small behavior amplitudes of human body. The $1\times1$ convolution is used in the Regress Layer whereas the Gaussian Error Linear Unit (GELU) is utilized to increase the robustness of the algorithm. The SL is applied to smooth the sequence to avoid interference during image acquisition and transmission. The network structure diagram of CFF-based stage is shown in Fig. 4.

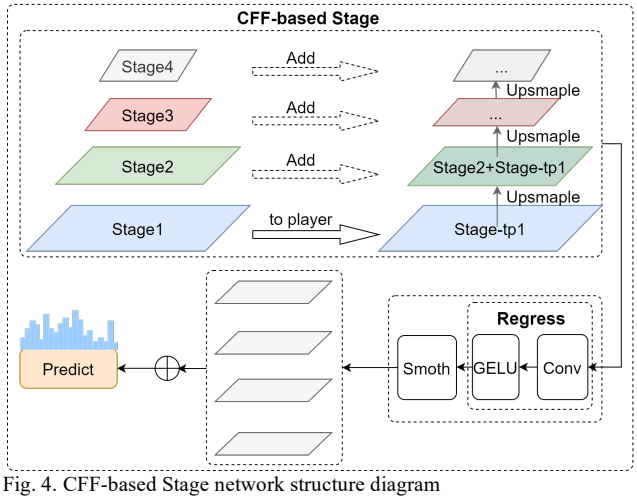


Fig. 4. CFF-based Stage network structure diagram

The Up-Sampling Layer adopts the array sampling method. Since there are differences in the output size of each layer of the CFF-based stage, the Up-Sampling Layer uses bilinear interpolation to operate and fills the outer boundary value with the edge value. When the grayscales of points $(x_0,y_0)$, $(x_0,y_1)$, $(x_1,y_0)$ and $(x_1,y_1)$ are known, the interpolation calculation method is shown that:

$$f(x,y) \approx f(x_0,y_0)(x_1-x)(y_1-y)+f(x_1,y_0)x(y_1-y)+ f(x_0,y)(x_1-x)y+f(x_1,y_1)xy \quad (15)$$

where $f(x,y)$ is the value of the function. The matrix operations are expressed as follow:

$$f(x,y) \approx \begin{bmatrix} x_1-x & x \end{bmatrix} \begin{bmatrix} f(x_0,y_0) & f(x_0,y_1) \\ f(x_1,y_0) & f(x_1,y_1) \end{bmatrix} \begin{bmatrix} y_1-y \\ y \end{bmatrix} \quad (16)$$

The weight coefficient of the original image depends on the distance between the pixels to be interpolated and the original image pixels. The closer the distance is, the greater the weighting coefficient is obtained. Bilinear interpolation reduces some of the visual distortions caused by resizing an image to a non-integer scaling factor. The interpolation principle is shown in Fig. 5.

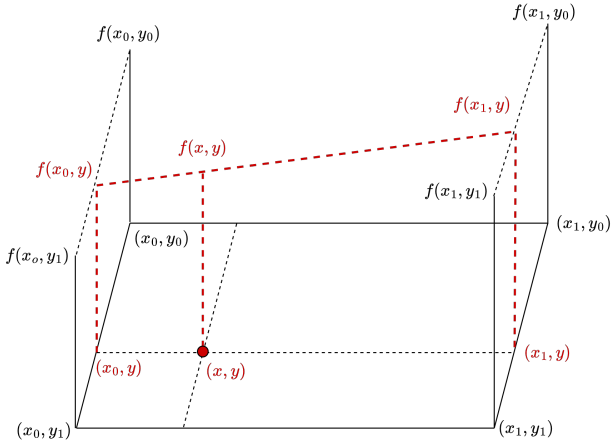


Fig. 5. Principle of Bilinear Interpolation

Geometrically, array sampling input and output pixels are represented as squares rather than points, aligning input and output tensors through the corners of their corner pixels. The schematic diagram of array sampling is shown in Fig. 6.

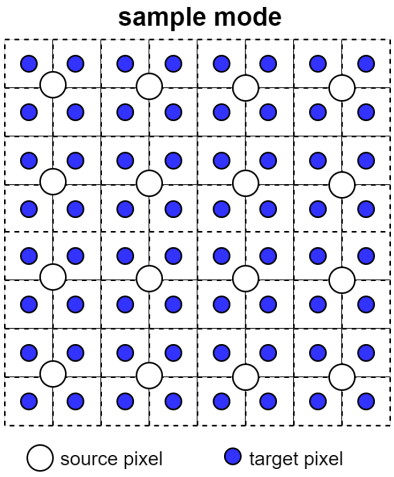


Fig. 6. Sample mode

## III. Experimental

### A. Datasets

The KTH Dataset contains four different experimental scenarios: indoor, indoor scale change, outdoor, and outdoor clothing change. The dataset has a total of 600 video samples with a resolution of 160×120 . 25 subjects completed six different human behaviors of boxing, hand clapping, walking, jogging, running, and handwaving. Each behavior type contains 100 videos. The same subject performed the same behavior from different angles [20].

### B. Experimental Setting

This experiment uses Windows 10 64-bit operating system, Intel(R) Xeon(R) E5-2630 V4 2.20 GHz (2 processors) with 12GB RAM and NVIDIA TITAN X(Pascal). Python version is 3.8 and Pytorch version is 1.9.

### C. Experimental results and analysis

The experimental process is divided into two stages:

1) Evaluating the performance of VT and ASMLP algorithms, and completing ablation experiments on different time series.
2) Adding a CFF-based stage to the model and using the time series of the first stage to verify the effectiveness of the improved method.

In the first stage, the VT and ASMLP algorithms were constructed in the model. Using OpenPose to identify human skeleton key points in successive frames of video and introducing human skeleton key points matrix into the original model were the main tasks to be implemented. Firstly, the VT algorithm was used to calculate the key points matrix of human skeleton, to achieve the key points tracking of different positions of the skeleton and to generate the characteristic vector matrix of displacement and angle change. Secondly, the time series was introduced into the eigenvector matrix for feature capture of continuous actions. The experiment was carried out in five-time dimensions (16, 18, 20, 22, 24 respectively). Training rounds were equal to 3000. The experimental results are shown in Table I.

The results show that the model performance reaches the best at the time dimension equal to 22 and the recognition accuracy reaches 93.20%. In the case of 5 groups of different time-dimensions, the recognition accuracy of Jogging and Running has always been at a lower level than other actions. Too short time dimensions may not be able to capture the complete action process of Jogging and Running, and too long-time dimension will be confused with the next complete action. They will have a negative impact on the recognition accuracy. Fig. 7 shows the changes of the training loss value and the test accuracy value in the time-dimension is 22.

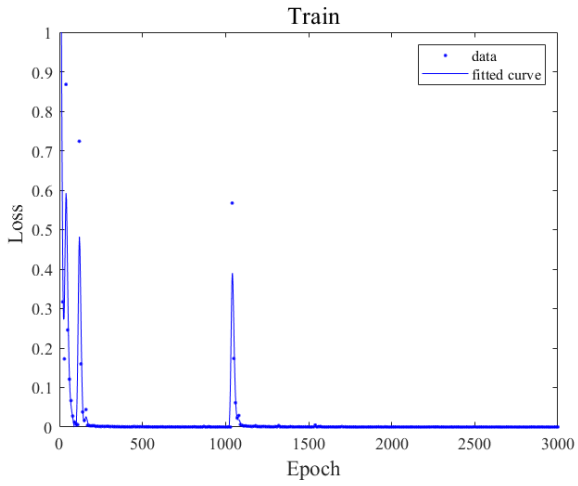


Fig. 7(a). Train loss

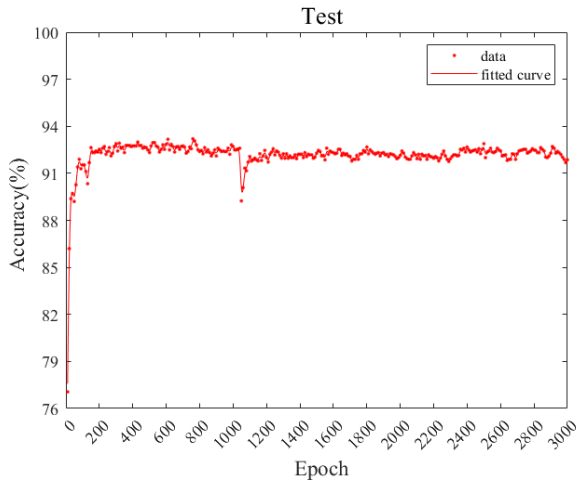


Fig.7(b) Test Accuracy

Fig. 8 shows the confusion matrix of behavior classification in the 22nd time-dimension. The diagonal line represents the number of correctly classified samples. The detection effect of human waving behavior is the most obvious. It can be seen from the confusion matrix that there is a high degree of confusion in the identification of Jogging and Running. The reason behind this behavior is that there is a high similarity between the two behaviors.

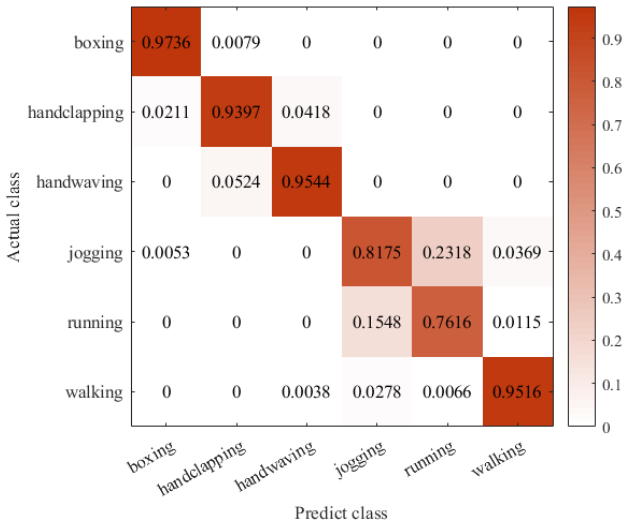


Fig. 8 Confusion matrix for behavior classification with time-dimension 22

In the first stage of the experiment, the research methods of literature [9-10] and the VGG-16 network model were reproduced. Using the same KTH dataset, the spatial channel network and the time channel network were used to extract static and dynamic features, respectively. The training round was based on 5000 samples. The experimental results show that the recognition accuracy of this proposed model is improved by 22.66% and 11.35% compared with the spatial channel network and temporal channel network, respectively. Compared with the VGG-16 network model, the recognition accuracy of this proposed model is improved by 1.12%. The comparative results are shown in Table II. CFF-MLP model has the best recognition accuracy for all six kinds of behaviors and, even, the recognition accuracy of Running and Walking behaviors is significantly improved.

In the second stage, the CFF-based stage is constructed using the CFF-MLP model based on the first stage. The time-dimension is also selected as the previous experiment (e.g., 16 to 24 with an incremental step of 2). The experimental results are shown in Table III.

Added to that, the model performance reaches the best with a recognition accuracy equal to 94.23% in the time-dimension 20. Compared with the first stage, the time-dimension is reduced by 2. Except for the Jogging and Handwaving behaviors, the recognition accuracy of the other four types has been improved. The average recognition accuracy has increased by 1.04%. It is shown that the CFF-MLP model can capture the features of continuous behaviors in a shorter time. Thus, Fig. 9 shows the change of the Training Loss value and the Test Accuracy value in the 20th time-dimension.

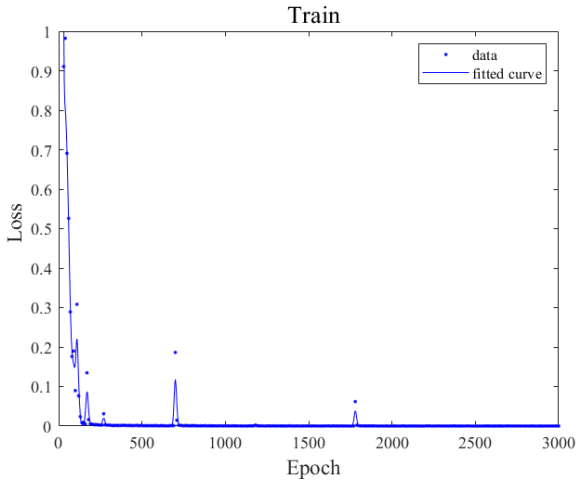


Fig.9(a) Train Loss

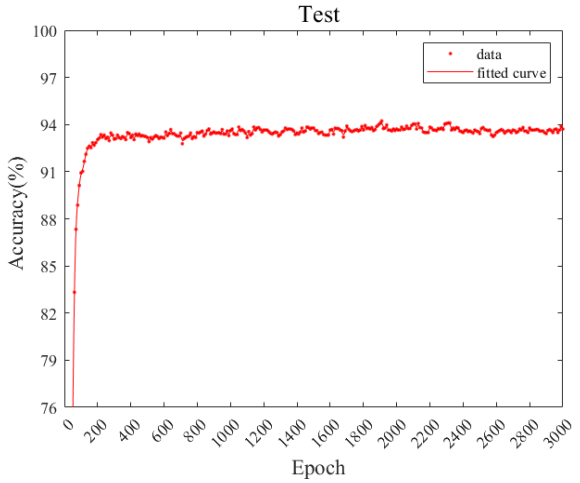


Fig.9(b) Test Accuracy

Fig. 10 shows the confusion matrix of behavior classification in the 20th time-dimension. It can be seen from the confusion matrix that the construction of CFF-based stage in the model can effectively reduce the recognition confusion degree of Boxing, Handclapping, Walking, Running, and other behaviors.

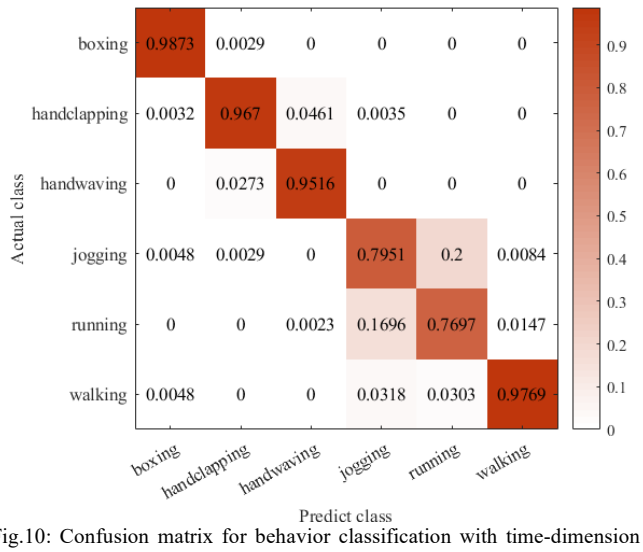


Fig.10: Confusion matrix for behavior classification with time-dimension 20

## IV. Conclusion

This paper proposes a CFF-MLP model consisting of three stages: VT, ASMLP and CFF-based for human behavior detection. It extracts the key points of human skeleton, using OpenPose software, and designs a VT Algorithm based on the OF method. The model architecture is designed in a horizontal connection and bottom-up, and the spatial transfer features in horizontal and vertical directions are extracted by the axial transfer method. The validation experiments are carried out on the KTH dataset and compared with other human behavior detection methods. According to the analysis of the experimental results, the CFF-MLP model has a good generalization ability and has obvious mapping to the behavior characteristics at different scales.

## References


[1] R. F. Li, L. L. Wang, and K. Wang, "A survey of human body action recognition," *Pattern Recognition & Artificial Intelligence*, vol. 27, no. 1, pp. 35-48, 2014.

[2] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream CNN: Learning representations based on human-related regions for action recognition," *Pattern Recognition,* vol. 79, pp. 32-43, 2018.

[3] J. Zhang, C. Wu, and Y. Wang, "Human fall detection based on body posture spatio-temporal evolution," *Sensors (Switzerland),* vol. 20, no. 3, 2020.

[4] L. Liu, "Objects detection toward complicated high remote basketball sports by leveraging deep CNN architecture," *Future Generation Computer Systems,* vol. 119, pp. 31-36, 2021.

[5] X. Li, and S. Sun, "Research on abnormal behavior detection based YOLO network," *Electronic Design Engineering,* vol. 26, no. 20, pp. 154-158, 2018.

[6] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740-2755, 2018.

[7] J. Carreira, and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE*, pp. 6299-6308.

[8] Z. Zhang, Y. Song, and Y. Zhang, "Motion-pose recurrent neural network with instantaneous kinematic descriptor for skeleton based gesture detection and recognition," *Proceedings - 4th Asian Conference on Pattern Recognition, ACPR 2017.* pp. 770-775.

[9] P. Zhang, Z. Su, Z. Dong, and K. Pahlavan, "Complex Motion Detection Based on Channel State Information and LSTM-RNN," *2020 10th Annual Computing and Communication Workshop and Conference, CCWC 2020.* pp. 756-760.

[10] M. Mahedi Hasan, M. Shamimul Islam, and S. Abdullah, "Robust Pose-Based Human Fall Detection Using Recurrent Neural Network," *2019 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2019.* pp. 48-51.

[11] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 4, pp. 677-691, 2017.

[12] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "Deep learning approach for human action recognition using gated recurrent unit neural networks and motion analysis," *Journal of Computer Science,* vol. 15, no. 7, pp. 1040-1049, 2019.

[13] N. S. A. Latha, and R. K. Megalingam, "Exemplar-based learning for recognition annotation of human actions," *Proceedings of the 2020 9th International Conference on System Modeling and Advancement in Research Trends, SMART 2020.* pp. 91-93.

[14] S. Salahuddin, I. Ahmed, M. Rashid, R. Shafi Ur, and N. Minallah, "Automatic Recognition of Human Actions," *2019 13th International Conference on Open Source Systems and Technologies, ICOSST 2019 - Proceedings.* pp. 60-65.

[15] Y. Yi, and Y. Lin, "Human action recognition with salient trajectories," *Signal Processing,* vol. 93, no. 11, pp. 2932-2941, 2013.

[16] K. Zhang, and W. Ling, "Joint Motion Information Extraction and Human Behavior Recognition in Video Based on Deep Learning," *IEEE Sensors Journal,* vol. 20, no. 20, pp. 11919-11926, 2020.

[17] X. Liu, D.-y. Qi, and H.-b. Xiao, "Construction and evaluation of the human behavior recognition model in kinematics under deep learning," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-9, 2020.

[18] M. O. Hongwei, H. Wang, and H. E. University, "Research on human behavior detection based on Faster R-CNN," *CAAI Transactions on Intelligent Systems*, 2018.

[19] Andi W. R. Emanuel, Paulus Mudjihartono, and Joanna A. M. Nugraha, "Snapshot-Based Human Action Recognition using OpenPose and Deep Learning," IAENG International Journal of Computer Science, vol. 48, no.4, pp862-867, 2021

[20] Z. S. David, and A. H. Abbas, "Human action recogntion using interest point detector with kth dataset," *vol,* vol. 10, pp. 333-343.


TABLE I
Experimental Results Of The First Stage

| Time-series | Best_acc | Boxing | Handclapping | Handwaving | Jogging | Running | Walking |
|---|---|---|---|---|---|---|---|
| 16 | 88.48% | 86.6% | 91.98% | 93.02% | 77.59% | 61.14% | 93.69% |
| 18 | 90.88% | 92.56% | 92.48% | 93.39% | 77.46% | 74.60% | 95.51% |
| 20 | 90.92 | 92.56 | 90.95 | 95.62 | 79.51 | 67.88 | 94.97 |
| 22 | 93.20% | 97.36% | 93.97% | 95.44% | 81.75% | 76.16% | 95.16% |
| 24 | 93.08% | 96.18% | 94.63% | 96.26% | 80.60% | 69.85% | 96.21% |

TABLE II
COMPARISON WITH OTHER METHODS IS BASED ON KTH DATASET

| Approaches | Best_acc | Boxing | Handclapping | Handwaving | Jogging | Running | Walking |
|---|---|---|---|---|---|---|---|
| Spatial Channel Networks | 70.54% | 92.77% | 75.45% | 72.21% | 50.32% | 57.63% | 67.98% |
| Time Channel Networks | 78.85% | 91.22% | 83.68% | 79.64% | 59.76% | 50.52% | 88.99% |
| VGG-16 network model | 92.08% | 97.00% | 94.00% | 92.00% | 90.00% | 84.00% | 91.00% |
| Ours:VT +ASMLP Stage | 93.20% | 97.36% | 93.97% | 95.44% | 81.75% | 76.16% | 95.16% |

TABLE III
EXPERIMENTAL RESULTS OF THE SECOND STAGE

| Time-series | Best_acc | Boxing | Handclapping | Handwaving | Jogging | Running | Walking |
|---|---|---|---|---|---|---|---|
| 16 | 89.01% | 85.71% | 94.39% | 93.94% | 74.51% | 69.19% | 92.19% |
| 18 | 91.17% | 93.42% | 96.11% | 91.53% | 76.83% | 72.49% | 95.51% |
| 20 | 94.23% | 98.73% | 96.70% | 95.16% | 79.51% | 76.97% | 97.69% |
| 22 | 93.20% | 97.19% | 95.71% | 95.82% | 82.54% | 68.21% | 94.47% |
| 24 | 93.85% | 95.98% | 96.88% | 96.54% | 88.48% | 65.44% | 96.97% |

**JIANFENG ZHANG** was born in Heilongjiang Province, P. R. China, received the B. Sc degree in Automation from University of Science and Technology Liaoning, Anshan, P. R. China, in 2020.

He is currently pursuing the M. Sc degree in Control Science and Engineering with University of Science and Technology Liaoning, Anshan, P. R. China. He research interest is computer vision.

**TIANWEI SHI** was born in Liaoning Province, P. R. China, received the M. Sc degree in Control Theory and Control Engineering from University of Science and Technology Liaoning, Anshan, P. R. China, in 2010, received the Ph.D. degree in Mechatronic Engineering from Northeastern University, Shenyang, P. R. China, in 2016

He is currently an associate professor in the School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan, P. R. China. The main research directions are brain-computer interface physiological electrical signal analysis and processing, machine vision, etc.

**WENHUA CUI** was born in Liaoning Province, P. R. China, received the M. Sc degree in Electromechanical control and automation from Dalian University of Technology, Dalian, P. R. China, in 1998, received the Ph.D. degree in Control Theory and Control Engineering from Dalian University of Technology, Dalian, P. R. China, in 2014,

She is currently a professor in the School of Computer and Software Engineering, University of Science and Technology Liaoning. She has published more than 20 academic papers and established more than 20 scientific research projects. The main research directions are control theory, sensor measurement and control, intelligent IoT, information security, computer network, machine vision, etc.

**YE TAO** was born in Liaoning Province, P. R, received the B. Sc degree in Computer Science and Technology from Anshan Normal University, received the M. Sc degree in Computer Science and Technology from Northwest Normal University.

He is currently a lecturer in the School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan, P. R. The main research direction is image encryption

**HUAN ZHANG** was born in Liaoning Province, P. R. China, received the B. Sc degree in Software Engineering from University of Science and Technology Liaoning, Anshan, P. R. China, in 2020.

She is currently pursuing the M. Sc degree in Software Engineering with University of Science and Technology Liaoning, Anshan, P. R. China. Her research interest is computer vision.