Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Computational study of multimodal action and identity recognition systems for bandwidth-constrained scenarios

Ziyang Guo[1,✉]

[1] *Department of Information Science and Technology, Shanghai Ocean University, Shanghai, 201316, China*

ABSTRACT

With the rapid development of video surveillance and multimedia applications, video data is requiring higher bandwidth demands for its transmission, storage, and retrieval. This paper presents a novel approach to video processing based on skeletal information and the recognition of identities. The skeletal data enables the extraction of skeletal data features from video frames and integrates this with the recognition of identities in such a way that the video data gets segmented into skeletal data, identity information, and other relevant data. A multimodal approach like this one spans a broad range in data transmission volume, optimizes bandwidth use, and significantly improves storage efficiency and increases retrieval speed. Experimental results have verified that the proposed method is able to transmit information with efficacy even in complex scenarios and further enable significant improvement in the accuracy and speed of performing storage and retrieval tasks. Such improvements turn into an effective solution for real-time monitoring, behavior analysis, and identity recognition applications featuring strong robustness and adaptability.

Keywords: Multi-Modal Fusion, Action Recognition, Identity Recognition, Skeletal Data, Frame Interval Extraction, Data Compression, Dual-Stream Architecture, text processing

1. Introduction

Human action and identity recognition, i.e., recognizing and understanding human actions, is essential for many real-world applications and is a core enabler for a wide range of applications such as smart homes, security monitoring, and human-computer interaction [1-2]. In smart homes, through human action and identification, the system can automatically sense the behavior of the residents to provide personalized smart services [3-4]. For example, recognizing a user's gestures can trigger the operation of specific devices, enabling residents to control lighting, home appliances, and other devices more conveniently, enhancing the comfort and convenience of life [5]. In security monitoring, human movement and identification can be used to detect abnormal behavior or potential threats and improve the

✉ Corresponding author.
E-mail address: a1002655181@163.com (Z. Guo).
Received 05 February 2024; Revised 12 April 2024; Accepted 05 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-408

intelligence of the monitoring system [6-7]. By analyzing human movements, the system can detect abnormalities and alerts in a timely manner, which can help prevent crime and protect public safety [8]. In the field of human-computer interaction, human movement and identification enable users to interact with computing systems through natural movements without touching or using other input devices [9-11]. This intuitive interaction not only enhances the user experience, but also expands the application scenarios, making the interaction design in the fields of virtual reality and augmented reality more flexible and natural.

In recent years, the development of computer vision has led to high-precision human movement recognition using pictures as well as videos, but vision-based recognition methods have their shortcomings, and the accuracy can be seriously degraded in dark or occluded situations [12-14]. On the other hand, although smart devices based on Wifi and wireless communication are booming in the field of perception, factors such as bandwidth limitation under scene conditions lead to a reduction in the accuracy of their recognition of perceived actions and identities [15-16]. There is an urgent need to implement a robust and lightweight model that is capable of multimodal data integration to more accurately understand user actions and respond accordingly, as well as algorithms and optimization strategies to reduce the system's demand on hardware resources and ensure that the model runs smoothly on different devices [17-18]. This has certain significance for the wider application of smart home.

This paper proposes an action and identity recognition system based on multi-modal fusion (shown in Figure 1), which uses the Kinect v2 sensor to simultaneously acquire RGB video and skeletal data, and performs local preprocessing and fusion through a dual-stream architecture. Compared with traditional methods that only rely on RGB or single modal data, this system streamlines and extracts features of the data on the monitoring end. This strategy not only significantly reduces the amount of raw data that needs to be uploaded to the back-end server or cloud for processing, and also retains key identification information, so that efficient and accurate action and identity recognition can still be achieved in bandwidth-limited scenarios.

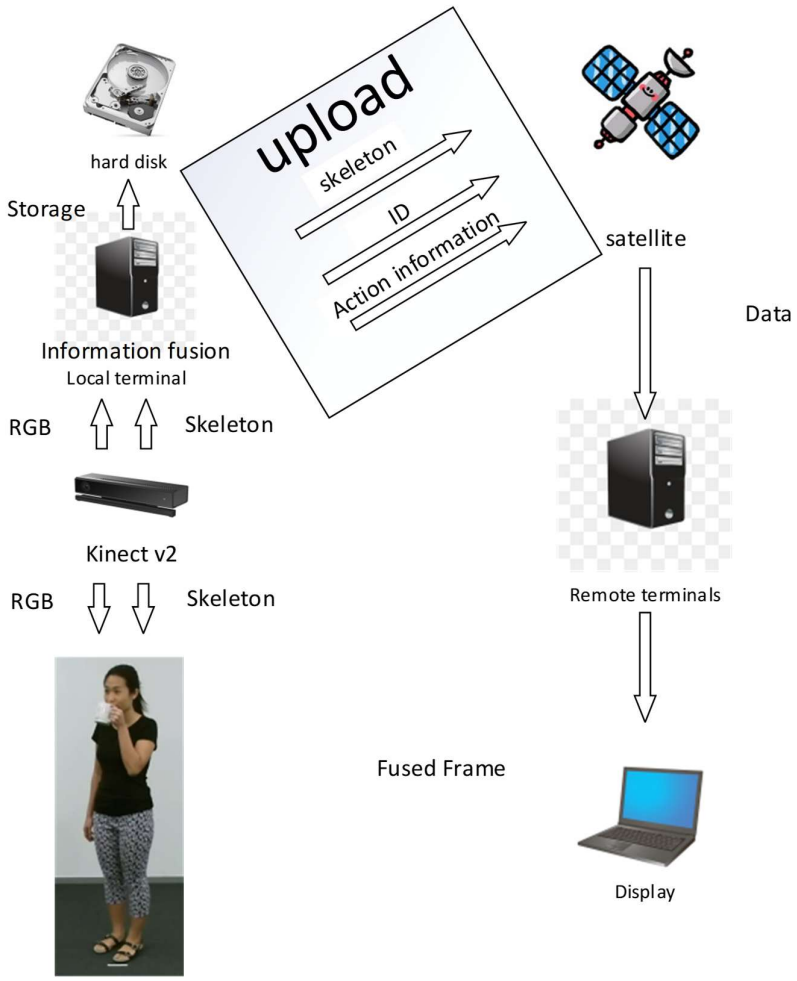

Fig. 1. Data processing flow framework in monitoring applications

The innovation of this system is that it adopts a multi-modal optimization strategy of frame interval extraction and text processing of skeletal data: while ensuring the integrity and discriminability of action and identity features, the scale of data to be transmitted is greatly reduced. Compared with the traditional method of using only RGB or a single modality, this front-end feature extraction and data reduction strategy not only improves the real-time performance and computing efficiency of the system, but also can effectively adapt to remote Bandwidth limitation in surveillance scenarios provides a more practical solution for action and identity recognition in complex environments.

2. Related Work

Recently, much attention has been paid to action recognition with multi-modal fusion. Single-modal methods, such as RGB video-based approaches, require abundant visual information for effective capturing of the scene and background details. However, these methods obviously depend on environmental conditions, and their accuracy drops drastically in some instances where the background is complicated. To complement these limitations, skeletal data was introduced by researchers, and the stability of skeletal data while representing the joint position has empowered high recognition accuracy in changing environmental conditions. For example, the ST-GCN model constructs spatiotemporal graphs of human skeletons and uses GCNs for effective feature extractions in space and time, which empowers a high performance bound of action recognition with skeletal data effectively.

While the skeleton data has indeed behaved well in some cases, single-modal data still lacks the potential in capturing subtle actions with rich contextual information. To get rid of the limitations and

raise performance, multi-modal fusion methods have come into being, aiming at fusing the RGB video with skeletal data to make up for the shortcomings of single-modal ones. Recently, identity recognition has added important monitoring capabilities to multi-modality action recognition. For instance, the integration of action recognition and identity recognition models enables the system to provide accurate identity identification with simultaneous behavioral analysis. Multi-modal fusion has proved to bring not only high accuracy in action recognition but also improvement in robustness within a complicated environment. For instance, the combination of the SlowFast network and ST-GCN brought about impressive advantages of multi-modality in such dynamic and complex scenarios.

In addition to improving recognition precision, how to effectively transmit data is another key problem that needs to be solved in real-world applications of multi-modal action recognition systems. High-quality video streams combined with skeletal data easily lead to a high volume of data, which increases bandwidth demand and transmission latency. Previous studies have proposed strategies toward optimization for data transmission, including the use of data compression and streamlining. Text processing of skeletal data can reduce considerably the amount of data to be transmitted, by reduction to only key skeletal coordinates as well as action categories.

Besides, techniques regarding frame interval extraction are widely applied to perform optimization in video transmission, where some frames are skipped, which reduces the number of transmitted frames. Hence, both the bandwidth required and the computational resource consumed go down. On this basis, our system goes one step further: it not only uses frame interval extraction technology on video data, but also adopts text processing strategies on skeletal data. By compressing skeletal information into a compact text form, we reduce data transfer to a minimum while preserving key motion and identity features. This comprehensive solution of multi-modal fusion and data optimization enables the system to maintain high recognition accuracy and efficient data processing capabilities under limited bandwidth conditions. Experimental results show that compared with traditional methods, our integrated solution not only significantly improves recognition performance in complex environments, but also significantly improves transmission efficiency.

3. Main Methods

This study proposes an action and identity recognition system that is multi-modal fusion-based with the intention of complex action recognition and identity recognition in videos obtained from surveillance. The system realizes a dual-stream architecture integrating RGB video streams with skeletal data for efficient action-identity recognition. Each dual-stream architecture has two separate branches, one processing the RGB video modality and another designed to handle the skeletal modality. A dual-stream approach allows the system to take full advantage of the complementary information provided by different modalities to enhance the robustness and accuracy of action recognition.

The system contains four main modules: RGB modality processing, skeletal modality processing, identity recognition, and action recognition with multi-modal fusion. Final recognition results are exported in JSON format for easy subsequent transmission and use as shown in Figure 2.

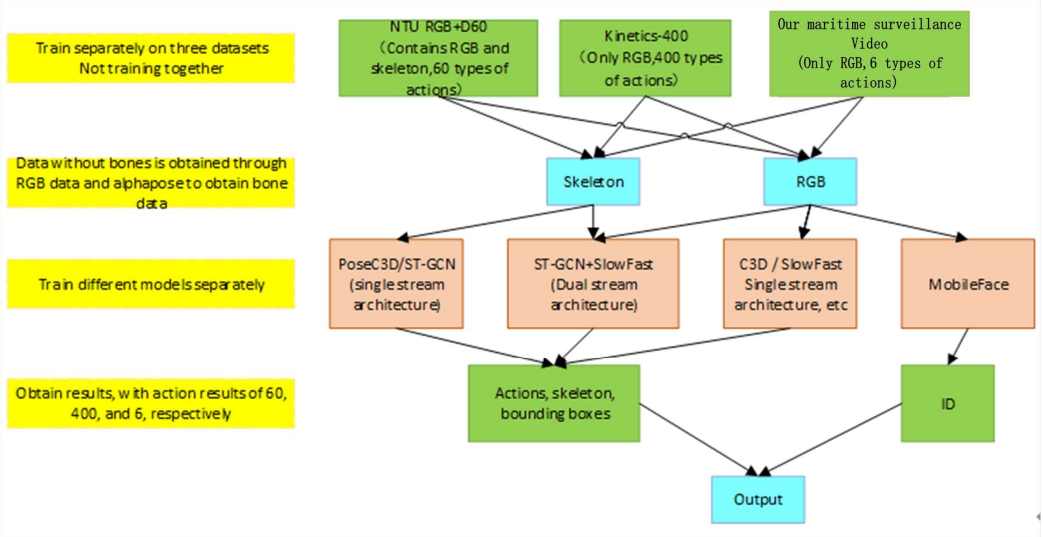

Fig. 2. Proposed System and Benchmark Methods: Datasets, Architectures, and Outputs

3.1. RGB Modality Processing Module

Instead of using the typical C3D, we employed the SlowFast network to extract spatiotemporal features from RGB videos. The SlowFast network is a neural network designed specifically for video understanding. Its main feature is to capture features of different time scales through a dual-path architecture, which is different from C3D which can only capture information of a single spatiotemporal scale. The SlowFast network is a two-path network where one pathway operates on sparse inputs and aims to capture slowly varying spatial features, while the other pathway operates on dense inputs and focuses on capturing rapidly changing detail features. By fusing these two complementary pathways, the SlowFast network can encode both action variations and detailed information across frames of the video more effectively.

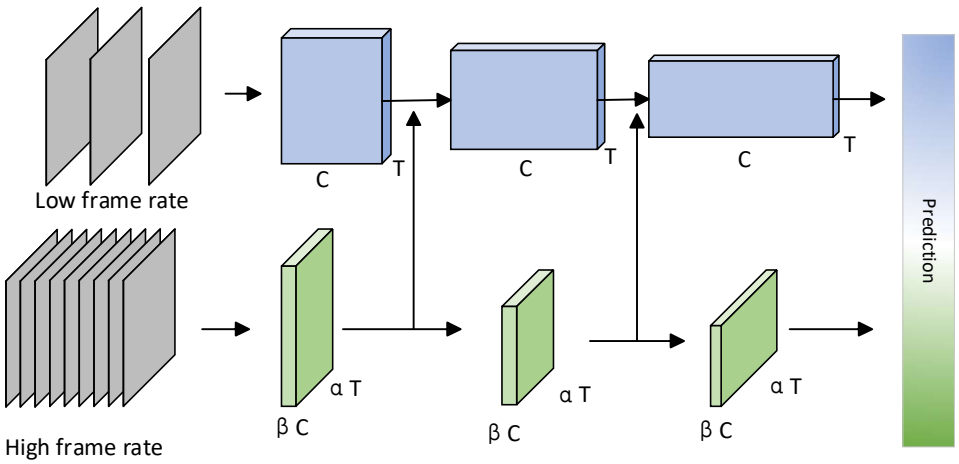

Fig. 3. SlowFast Network

As shown in Figure 3, the result from the fast pathway is fed into the slow pathway through lateral connections, allowing the slow pathway to include information processed in the fast pathway. It is evident that the shapes of the data differ between the two pathways; the fast pathway is represented as {αT, S2, βC} and the slow pathway as {T, S2, αβC}, necessitating a transformation of the fast pathway's output before integration into the slow pathway.

3.2. Skeletal Modality

In the skeletal modality processing module, skeletal data are captured by Kinect v2. Kinect v2 is a multimodal sensing device that integrates an RGB camera and a depth camera. Its core technology is based on the Time-of-Flight (ToF) principle. Using the depth camera, Kinect v2 emits active infrared light and receives its reflection to precisely measure the depth information of each pixel in the scene. Additionally, the RGB camera captures high-resolution color images. These multimodal data streams are fused through its built-in SDK, enabling the automatic tracking and extraction of human skeletal data.

The skeleton tracking functionality of Kinect v2 leverages its SDK's built-in algorithms to extract the three-dimensional coordinates (x, y, z) of 25 key points on the human body in real time, encompassing the head, torso, limbs, and joint positions. These key point coordinates are calculated and output in real time by the Kinect v2 sensor and algorithms during the recording process. To ensure compatibility with standard two-dimensional skeleton models and reduce computational complexity, the three-dimensional key points extracted by Kinect v2 are projected onto a two-dimensional image plane in this study.

The conversion formula for the relevant two-dimensional coordinates (u, v) is as follows:

$$\mathrm{u} = f_x \cdot \frac{X}{Z} + c_x \tag{1}$$

$$\mathrm{v} = f_y \cdot \frac{Y}{Z} + c_y \tag{2}$$

Where f_x and f_y are the focal lengths of the camera, c_x and c_y are the coordinates of the optical center of the image, and Z is the depth value.

Compared to traditional action recognition methods that rely solely on RGB videos, the use of skeleton data effectively avoids external interferences such as complex backgrounds or occlusion, providing more stable and accurate features for action recognition in complex scenes.

In this study, although Kinect v2 provides three-dimensional coordinates (x, y, z) for 25 key points of the human body, we selected only 17 key points for experiments to improve computational efficiency and compatibility with other datasets, as shown in Figure 4. These 17 key points primarily include critical regions of the head, torso, upper limbs, and lower limbs. These points sufficiently capture the main movement information of the human body. The removal of the 8 additional key points has minimal impact on action recognition accuracy while significantly reducing computational overhead.

Additionally, since Kinect v2 was officially discontinued in 2017 and cannot serve as a long-term, widely applicable solution, we propose a more general key point extraction method in this study. For RGB videos without skeletal information, we use AlphaPose, a deep learning-based pose estimation model, to extract 17 key points from video frames. AlphaPose can accurately extract key points that conform to the COCO standard from RGB videos and output two-dimensional coordinates (x, y), providing reliable input for action recognition models. In comparative evaluations, AlphaPose outperforms other mainstream pose estimation methods in multi-person scenarios, particularly in cases involving overlapping poses or complex actions. For the action recognition task in this study, choosing AlphaPose ensures the high quality and robustness of skeletal key points. By utilizing this approach, the system can perform skeletal data extraction and multimodal fusion even in the absence of a depth sensor.

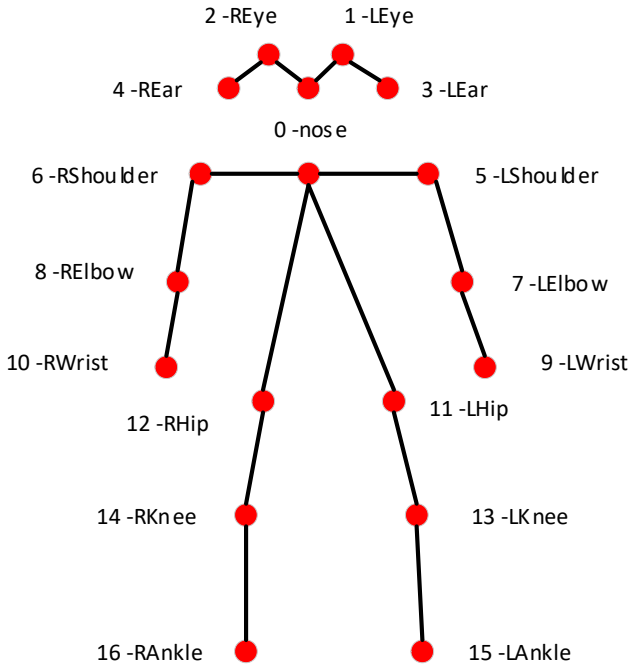

Fig. 4. Output Joint Coordinates

In this system, we employed a Graph Convolutional Networks (GCN)-based skeletal feature extractor to capture dynamic information from skeletal data. We adopt ST-GCN to process skeletal data. GCN is a deep learning model capable of processing graph structured data. Convolutional Neural Network (CNN) is used to process grid structured data, while GCN is specifically designed to process graph data, which has a structure of nodes and edges. GCN can capture the relationships and dependencies between nodes in a graph by performing information transfer and convolution operations between them. Especially in human pose estimation, the joints and skeletal structure of the human body can be represented as a graph, and the lines connecting the joints represent the edges of the graph. Through GCN, the spatial relationships between various joint points in the human body can be effectively extracted, thus enabling better posture analysis.

Specifically, we used the Spatiotemporal Graph Convolutional Network (ST-GCN) to process skeletal data (shown in Figure 5), which combines graph convolutional networks with deep learning models based on time series modeling, aiming to handle spatiotemporal features in video data. ST-GCN models the human skeleton data as a graph structure, capturing the relationships between the joints of the skeleton in the spatial domain and the dynamic changes of the skeleton at different times in the temporal domain, in order to extract the spatiotemporal relationships between the joints of the human body. The advantage of ST-GCN lies in its ability to directly convolve on the graph structure, adapting to the natural graph structure of skeletal data, thus enabling more effective learning of the dependency relationships between human bones.

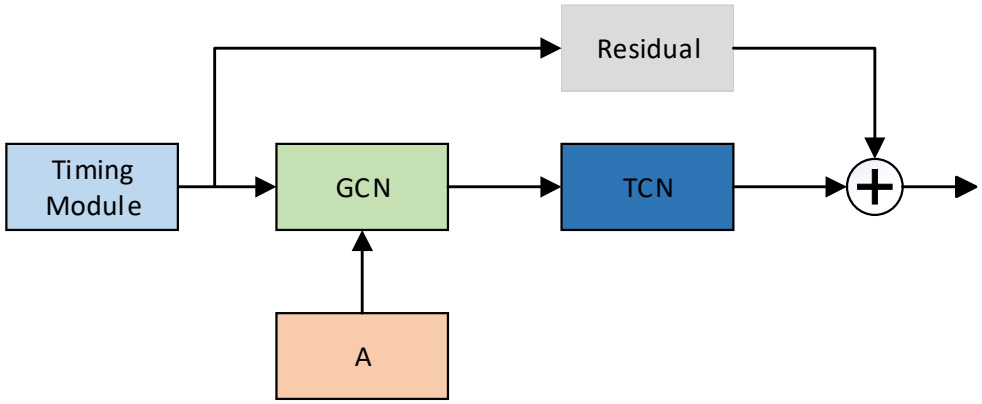

Fig. 5. Simple ST-GCN Module

The basic operation of ST-GCN can be divided into two parts: Spatial Graph Convolution and Temporal Convolution.

Spatial Graph Convolution:

$$X' = \sum_{k=1}^{K} A_k X W_k \tag{3}$$

A_k is a different adjacency matrix. X is the input feature, and the dimension is usually $C_{in} x C_{out}$, where C_{out} is the number of channels for the output feature. The weight matrix corresponding to W_k. X' is the feature obtained through spatial convolution.

Temporal Convolution:

$$Y = \sigma(W_t * X') \tag{4}$$

W_t is the weight of the time convolution. $*$ is convolution. σ is the activation function. Y is the final output.

The overall representation is:

$$Y = \sigma\left(W_t * \left(\sum_{k=1}^{K} A_k X W_k\right)\right) \tag{5}$$

3.3. Identity Recognition Module

Identity recognition serves as an important link in multi-modal video processing for accurate person identification. This paper focuses on the task of identity recognition, which involves distinguishing and identifying individual identities from multiple subjects in a closed-set scenario. We adopt the compact MobileFaceNet model with low computational complexity, making it well-suited for real-time deployment in resource-limited environments. MobileFaceNet allows not only high accuracy in identity recognition but also significantly reduces computational resource consumption, enabling efficient real-time monitoring.

Identity recognition in multi-person video scenes is often challenging due to the need to accurately distinguish between different individuals. This study adopts the regional multi-person pose estimation (RMPE) method provided by AlphaPose. First, each individual in the video is detected and cropped. Subsequently, the system performs skeleton extraction, action recognition, and face recognition on each cropped single-person image, ensuring that the identities of different individuals can be accurately distinguished within the same frame. This approach enables the system to effectively extract and identify each person's identity information, even in scenarios involving a large number of individuals.

First, MobileFaceNet normalizes the input image, detects the face position, and extracts feature vectors (shown in Figure 6). These feature vectors represent an individual's identity features, and the system measures similarity by calculating the Euclidean distance between the input image and known feature vectors in the database for efficient identity recognition. If the similarity exceeds a preset threshold, the identity is confirmed, and the information is recorded; otherwise, it is labeled as "unknown identity detected."

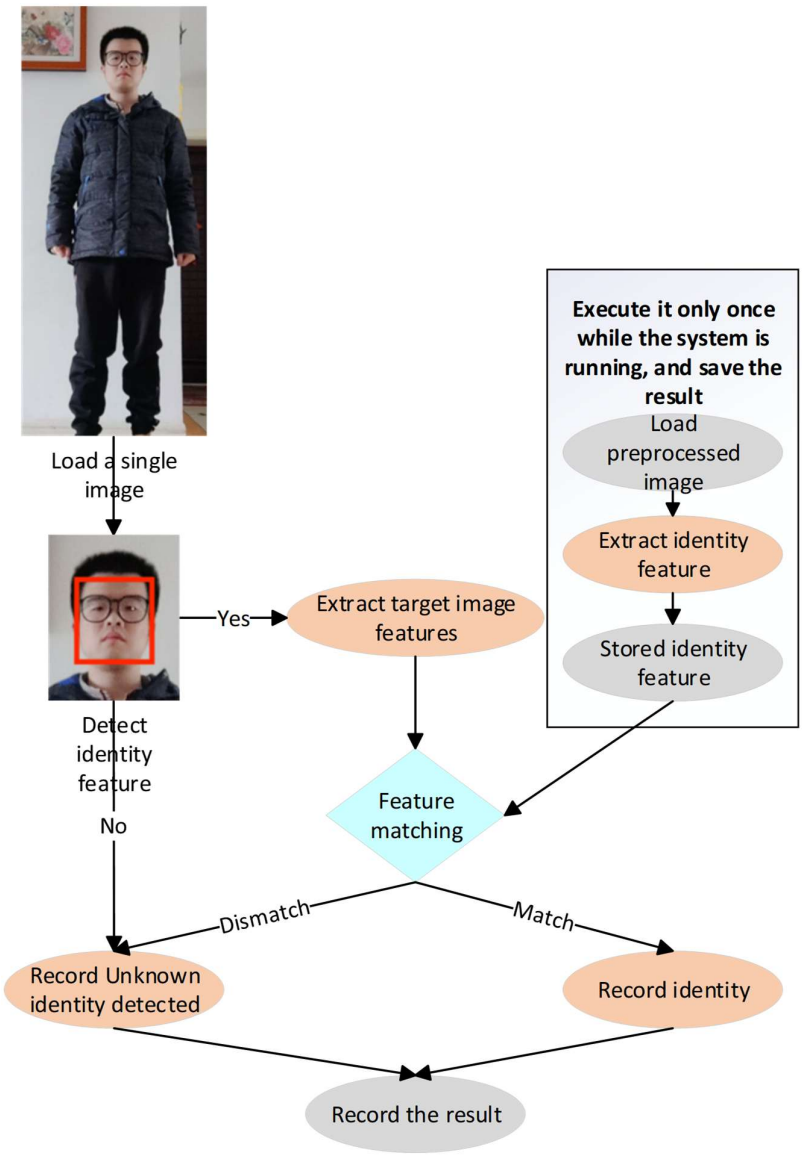

Fig. 6. Identity Recognition Detection

3.4. Action Recognition and Multi-Modal Fusion Module

The key to multimodal fusion lies in effectively integrating the complementary information of different modalities to enhance the robustness and accuracy of action recognition. RGB features capture visual context, providing static background and object interaction information, while skeleton features focus on human posture and dynamic action details, such as limb movement trajectories and amplitudes. By combining the strengths of these two modalities, the model significantly improves recognition robustness across various scenarios, particularly in environments with complex backgrounds or multiple individuals (as illustrated in Figure 7).

After feature extraction by the dual-stream architecture, we designed a multimodal fusion module to adaptively integrate RGB and skeleton features. Specifically, we introduced a multi-head self-attention mechanism along with dynamic fusion weights to effectively capture the relationships and complementary information between the two modalities. The multi-head self-attention mechanism captures the correlations between RGB and skeleton features and identifies complementary information across different modalities. Dynamic fusion weights further assign the importance of each modality based on scene characteristics. For instance, when an action is closely tied to the scene background, RGB features are given greater importance, whereas in scenarios involving complex interactions between individuals, skeleton features contribute more significantly. This allows the model to dynamically adjust the importance of each modality according to the scene, ensuring effective feature fusion.

With this design, the model leverages both the contextual information provided by RGB and the dynamic details captured by skeleton features, significantly enhancing action recognition performance. Experimental results demonstrate that in complex multi-person scenarios, the fused features effectively reduce misjudgments caused by background interference or partial occlusions, thereby improving the model's robustness and accuracy.

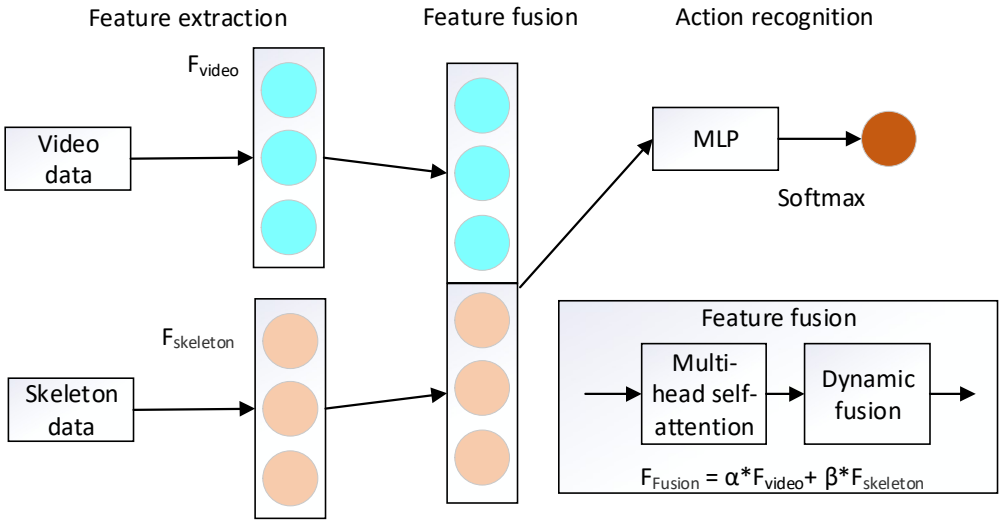

Fig. 7. Multi-Modal Action Recognition for Dual-Flow Architecture

Specifically, we first map the skeletal features and RGB features into one unified feature space using the feature transformation matrices for consistency in the feature dimensions. Then we use the multi-head attention mechanism to fuse the skeletal features Fs and RGB features Fr. The formulation of the fused features Fattention is:

$$Q = W_{\mathrm{q}}\left[F_{\mathrm{s}}, F_{\mathrm{r}}\right] \tag{6}$$

$$K = W_{\mathrm{k}}\left[F_{\mathrm{s}}, F_{\mathrm{r}}\right] \tag{7}$$

$$V = W_{\mathrm{v}}\left[F_{\mathrm{s}}, F_{\mathrm{r}}\right] \tag{8}$$

$$F_{attention} = \mathrm{soft\,max}\left(\frac{QK^{T}}{\sqrt{\mathrm{d_k}}}\right)V \tag{9}$$

$$\mathrm{soft\,max}\left(\mathrm{z}_i\right) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{10}$$

Where $F_s \in \square^{Cs}$, $F_r \in \square^{Cr}$, W_q, W_k, and W_v are the weight matrices for query, key, and value, respectively, obtained by linear transformations of the skeletal and RGB features. $W_i \in \square^{d\times(Cs+Cr)}$. QK^T refers to the calculation of similarity; d is used to scale the dot product results. F is the attention weight distribution normalized by softmax for similarity. Through this process, the fusion module can dynamically adjust the importance weights of each modality according to the characteristics of the input, thereby enhancing the recognition of complex actions.

For enhancing its adaptability, we introduced a dynamic fusion weight strategy. During this process, α and β will automatically change based on the characteristics of the input data in the model. Specifically, as α and β are probability distributions obtained through the Softmax function, their values range from 0 to 1. When the scene is more dependent on motion structures, the tendency of α increases; When the scene is more sensitive to texture and color information, β will increase:

$$[\alpha,\beta] = soft\max\left(F_{attention}\right) \tag{11}$$

Due to the fact that α and β have already been refined into a pair of dynamically adaptive weights through attention mechanism and Softmax function in the previous step, reasonable fusion can be achieved without additional complex operations. Weighted summation can smoothly integrate two modal features and organically integrate their discriminative abilities:

$$F_{\text{fusion}} = \alpha\times F_s + \beta\times F_r \tag{12}$$

The generation of dynamic weights is based on the characteristics of the input features. By learning adaptive weights, the model can automatically adjust the importance of each modality according to the specific scenario. For instance, in a scene with a complex background, skeleton features may be assigned a higher weight, whereas in other situations, RGB features might play a more dominant role.

Once we obtain and process the fused features, the model maps them to specific action classes. For classification, we utilized a simple yet effective classifier composed of a fully connected layer and a Softmax layer: the fused feature vector F_{fusion} undergoes dimensionality reduction and nonlinear transformation through a fully connected layer to produce an intermediate representation F_{fc}:

$$F_{\text{fc}} = ReLU\left(W_{fc}\times F_{fusion} + b_{fc}\right) \tag{13}$$

where W_{fc} and b_{fc} are the weight matrix and bias term of the fully connected layer, respectively. The $ReLU$ function is introduced to add nonlinearity.

Finally, the intermediate representation is mapped to a probability distribution using a Softmax layer to output the probability of each action class:

$$p_i = soft\max\left(F_{fc}\right) \tag{14}$$

In this study, we utilized the Kinetics-400 dataset and a custom dataset. Unlike the independent RGB and skeleton data directly captured by the Kinect V2 sensor, the skeleton information in these two datasets is extracted from RGB videos using an algorithm. Consequently, the skeleton data is strongly correlated with the RGB data since they originate from the same video source.

To address the characteristics of this "homologous" modality, we incorporated several optimization mechanisms into the multimodal fusion method. First, recognizing that skeleton information is derived from RGB data and may contain redundant information, we applied a de-redundancy process during the skeleton feature extraction stage. This process retains only dynamic human posture-related features, minimizing redundant interference between the modalities. Additionally, we employed a multi-head self-attention mechanism to dynamically weight the RGB and skeleton features, capturing

their correlations and complementarities. This ensures that even with homologous modalities, the model can effectively extract meaningful feature interactions.

Furthermore, compared to the data fusion approach used with Kinect V2, the feature coupling between homologous modalities is stronger. Therefore, in the fusion process, greater emphasis should be placed on the complementarity between the two modalities, rather than treating them as completely independent sources. With these optimized designs, our method fully leverages the advantages of homologous modalities while mitigating the redundancy issues that may arise due to data homology.

3.5. Frame Interval Extraction

In practical surveillance systems, video is usually captured at a constant frame rate of 25 per second. While high frame rate processing might improve detection accuracy, the associated computational load increases steeply, especially in resource-constrained situations. Additionally, contiguous frames are often highly redundant, yielding minimal performance gains. This means high-frame-rate processing can impose increased computational burdens, particularly in applications with stringent real-time demands.

To handle this, we proposed a frame interval extraction approach to optimize computational efficiency while effectively capturing motion information (shown in Figure 8). The basic idea is to set a frame interval parameter n, so that only every nth frame is processed, reducing intensive computation. The output frame interval also controls which frames should be displayed or included in the final output. This approach allows the system to save redundant calculation costs without losing much recognition accuracy, making it ideal for long-duration video inference.

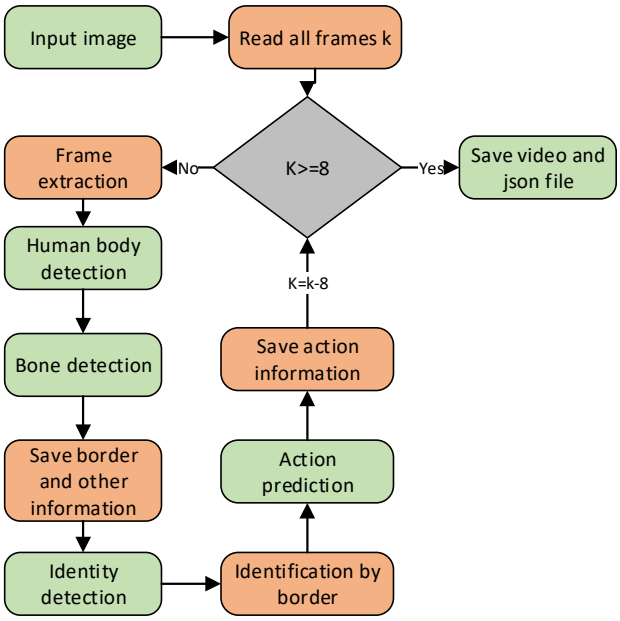

Fig. 8. Frame Interval Extraction Method

3.6. Information Output and Transmission

The final step in the system packages the results of action recognition and skeletal information and identity recognition, formatted in JSON for transmission. These components are encoded within JSON data:

1) Video Frame Information: It includes the timestamp or the sequence number of the video frame.
2) Category of action: This is the classification outcome of the human activity in the video.
3) Skeletal Data: Co-ordinates of a key-point obtained by pose estimation.
4) Identity Information: This is the identified outcome of the identity of the person from the RGB video.

4. Experiment Results and Analysis

4.1. Experiment Setup

These experiments were done on a PC with an NVIDIA RTX 3090 GPU, an Intel Core i7-12700 CPU, and 64GB RAM. PyTorch 2.0.1 was installed on Ubuntu 20.04 in the work. The SlowFast network applied for the RGB modality, while the ST-GCN is taken for the skeletal modality.

The base channel count for the RGB stream was set at 64, while that for the extended channel count was 32. For the ST-GCN, it was set with 17 input channels and utilized three layers of spatiotemporal graphs to extract dual features. An I3D structure was applied for the action recognition head, considering the inputs processed at a dimension of 512, which was outputting 60 action classes with a dropout rate of 0.5(shown in Table 1).

Table 1. Hyperparameter Settings

Hyperparameter	Value
Optimizer	SGD
Initial Learning Rate	0.0075
Momentum	0.9
Weight Decay	0.0001
Learning Rate Scheduler	MultiStepLR
Learning Rate Decay Factor	0.1
Maximum Training Epochs	150

4.2. Datasets

In this work, we adopt two of the most popular benchmark datasets: NTU RGB+D 60 and Kinetics-400 for the action recognition task, both datasets were downloaded from the OpenXLab official website. In addition, a customized maritime surveillance video dataset was used. The three datasets were trained and tested separately to evaluate the model's performance in real-world application scenarios.

NTU RGB+D 60 is a large-scale 3D action recognition dataset with a total of 60 different daily actions. The video database was captured from 40 participants of various ages and genders, so it allows multimodal analysis by RGB videos and skeletal data. Another widely used action recognition dataset is the Kinetics-400 dataset. Each video clip is approximately 10 seconds long and was collected from real-world scenes on YouTube. Additionally, to further evaluate the model's performance in real-world application scenarios, we used a custom marine surveillance video dataset for training and testing. This custom dataset was collected using a standard surveillance camera installed in the cabin. It includes six types of typical surveillance actions, such as standing, sitting, walking, making phone calls, drinking water, and eating, as illustrated in Figure 9. Each action contains approximately 200 samples.

Detailed information about the dataset is presented in Table 2.

Table 2. Statistics of the Dataset

Dataset	Total Samples	Action Classes	Training Samples	Validation Samples
NTU RGB+D 60	56880	60	40320	16560
Kinetics-400	260232	400	240436	19796
Surveillance Video	1200	6	1050	150

Due to data licensing restrictions, the examples presented in this article are simulations of the types of actions we collected in a laboratory environment. However, the overall characteristics of the custom dataset are consistent with those of the data collected in real-world settings, making these examples representative of the action types in the actual dataset.

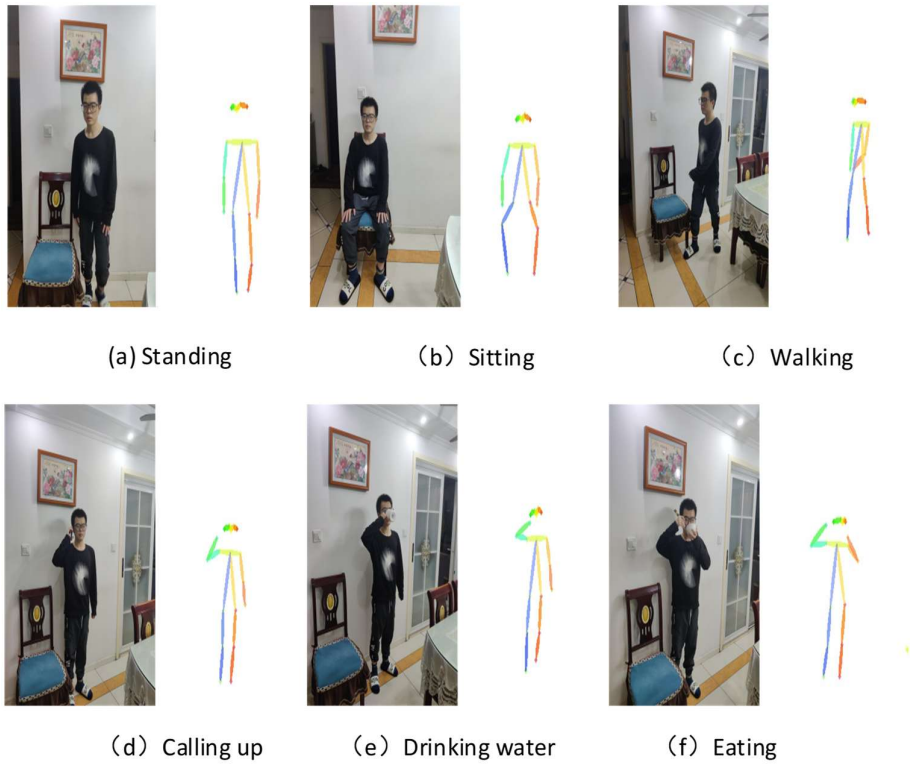

Fig. 9. Examples of Labeled Actions in the Custom Dataset

4.3. Results

The dual-stream architecture model demonstrated strong performance in our experiments on both the NTU RGB+D 60 and Kinetics-400 datasets, as evidenced by improved accuracy and efficiency metrics (shown in Figure 10). On the NTU RGB+D 60 dataset, the top-1 accuracy of the proposed model is 96.52%, meaning it outperformed a single-modal baseline model, SlowFast, by 9.34%. On the Kinetics-400 dataset, the dual-stream architecture model achieved a top-1 accuracy of 85.3%, notably higher in comparison to the existing baseline model, namely SlowFast.

This is primarily because of the effective combination of skeletal and RGB video features in the model, which enables it to capture action information more precisely even in complex situations.

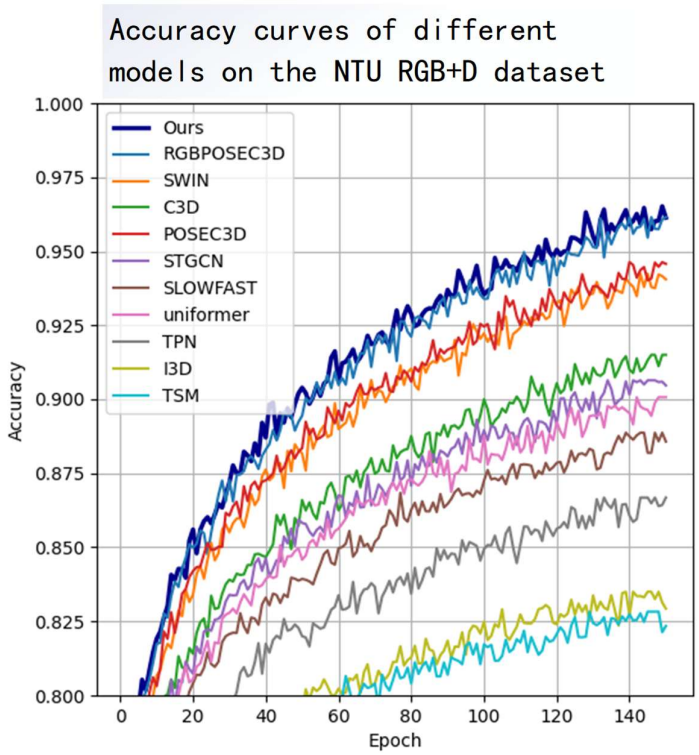

Fig. 10. Accuracy curves for different models

In the field of action recognition, different modalities and methods are widely used to improve recognition performance. This study selected some representative skeleton modality and RGB modality models for comparison, including: spatiotemporal models based on skeleton data such as ST-GCN and POSEC3D, convolutional neural network methods based on RGB video such as C3D, I3D, SLOWFAST, SWIN, TPN, TSM and UNIFORMER, and RGBPOSEC3D that combines RGB and skeleton multimodal fusion. These methods cover a variety of technical routes from single modality to multimodality, providing a comprehensive benchmark for comparing the performance of the model proposed in this study on different datasets (Kinetics400 and NTU RGB+D60). The specific performance of each method is shown in Table 3.

Table 3. Performance comparison of different action recognition methods on different datasets

Method	Modality	kinetics400		NTU RGB+D60	
		top 1	top 5	top 1(1 clip testing)	top 1(10 clip testing)
STGCN	Skeletal	80.6	81.7	90.64	91.69
POSEC3D	Skeletal	81.3	83.8	93.63	94.7
C3D	RGB	83.08	95.93	91.50	91.65
I3D	RGB	74.8	92.07	83.44	83.50
SLOWFAST	RGB	78.65	92.86	87.18	88.87
SWIN	RGB	83.46	95.91	92.29	94.24
TPN	RGB	76.74	92.57	85.49	86.68
TSM	RGB	74.38	91.71	82.57	82.83
uniformer	RGB	80.9	94.6	89.73	90.07
RGBPOSEC3D	RGB+Skeletal	84.7	96.3	96.15	96.23
Ours	RGB+Skeletal	85.3	97.2	96.52	96.91

Figure 11 depicts the recognition accuracy of each action category in the custom-made surveillance video dataset. From this, it would follow that categories "Standing" and "Eating" have the highest value of accuracy, while actions such as "Drinking Water" and "Making a Phone Call" demonstrate a low accuracy because of their big similarity.

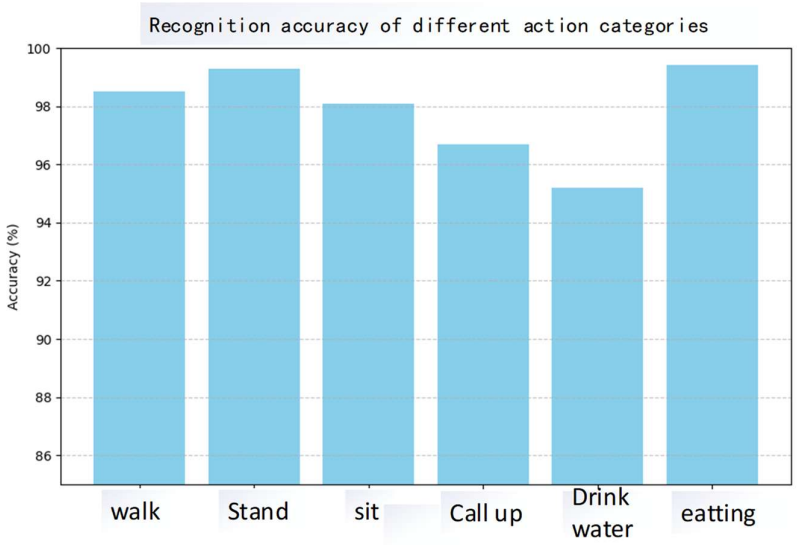

Fig. 11. Accuracy of custom datasets

Before integrating the function of identity recognition, several different but widely used identity recognition models were independently tested against their performances.

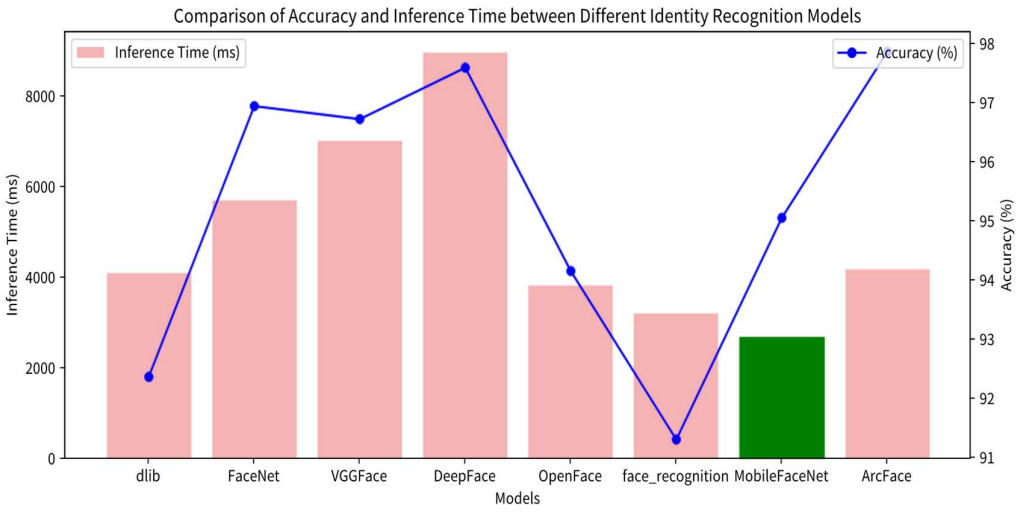

Fig. 12. Comparison of Accuracy and Inference Time between Different Identity Recognition Models

Figure 12 is a depiction of several different models that process 100 facial images showing both accuracy and inference time. As it can be seen, all these models differ significantly in the question of accuracy and speed. MobileFaceNet presents the highest speed with high accuracy, while DeepFace has the highest accuracy but with the longest inference time. Based on this result, high-real-time requirement scenarios will be the best scenarios for both the MobileFaceNet and OpenFace, whereas if the accuracy is of higher importance instead of speed, then DeepFace and ArcFace shall be more applicable.

4.4. Frame Interval Inference Time Analysis

In our experiments, for a 30-second video whose frame rate is 25, there are 750 frames. Without extracting the frame interval, the system handled all 750 frames, which cost 96.06 seconds when performing action recognition. When adding the identity recognition, the time rose to 122.19 seconds, in which 96.06 seconds were used by action recognition and 26.13 seconds were used by identity recognition.

As in the previous experiment, it only processed every n = 8th frame, for a total of 94 frames, significantly reducing the computational load. Using this setting, the action recognition time went down to 11.27 seconds, with the identity being recognized in 3.94 seconds, which results in 15.21 seconds of total inference time. This optimization greatly enhanced the real-time capability, thus making it even more applicable for scenarios that demand high computational efficiency (shown in Table 4).

In order to verify the contribution and efficiency of different modules in the proposed method, we designed the following three model configurations as the basis for ablation experiments:

A: Only the action recognition module is enabled, and the identity recognition module and frame interval reasoning module are not included.

B: Both the action recognition module and the identity recognition module are enabled, but the frame interval reasoning module is not included.

C: The action recognition module, the identity recognition module, and the frame interval reasoning module are enabled, that is, the complete model proposed in this paper.

Table 4. The Efficiency of the Frame Interval Inference Method in This Paper

(Note: The results in this table are only for the model proposed in this paper)

Model	Action Recognition	Identity Recognition	Frame Interval n=8	Inference Time		Total Time (s)
				Action Recognition (s)	Identity Recognition (s)	
A	√	×	×	96.06	-	96.06
B	√	√	×	96.06	26.13	122.19
C	√	√	√	11.27	3.94	15.21

4.5. Transmission Efficiency and Storage Optimization Analysis

The bandwidth and storage resources in most remote surveillance systems or edge computing are very limited. Particularly, the uplink bandwidth of the system is usually only about 1 Mbps in the case of maritime monitoring. Under these conditions, the high-definition video transmission may suffer from delays, and real-time transmission can hardly be performed. Besides, some devices have limited storage capacity incapable of holding massive video data.

In this paper, we present an approach that sends only key skeletal, identity, and action information. The system alleviates, to a large degree, the requirement of full video transmission and hence reduces data volume in bandwidth-limited environments, thereby reducing storage demands as well. Experimental results show that compared with transmitting complete videos, the bandwidth used for transmitting skeletal, identity, and action information is only 42.67%, while original video requires 549.33% bandwidth utilization.

This reduced the transmission time from 164.8 to 12.8 seconds (shown in Table 5).

This multi-modal optimization greatly raises the bar on the transmission efficiency of resource-constrained remote surveillance systems for real-time performance with high efficiency under possible poor bandwidth conditions.

Table 5. Comparison of Bandwidth Utilization Between Raw Video and Skeleton Data

Data Type	Data Volume (MB)	Bandwidth (MBps)	Transmission Time (s)	Bandwidth Utilization (%)	Compression Ratio
Raw Data	20.6	0.125	164.8	549.33	12.88
Skeleton Data	1.6	0.125	12.8	42.67	

This would implicitly mean the loss of some of the visual details and information in the background from the original video, but skeletal information and identity recognition data could grasp the main actions and individual information that meets the functional requirement of remote surveillance or behavior analysis.

4.6. Ablation Study

Then, to assess the contribution of each component in view of the performance of the model, an ablation study was performed. For that purpose, the comparison performances are conducted by removing the multi-attention mechanism and the dynamic weight allocation module that will be described next in Table 6.

Removing the attention mechanism reduced Top-1 accuracy by 1.32% on the NTU RGB+D dataset, showing that multihead attention plays a critical role in complementing different modal features. Without dynamic weight allocation, the accuracy decreased by 0.85%, which may mean that adaptive weighting will change the importance of each modality dynamically for improving the robustness of the model (shown in Table 6).

Table 6. Ablation Experiments: Effects of Different Modules on Model Performance

Model	RGB	Skeletal	Attention	Dynamic Weight Allocation	NTU RGB+D Top-1 (%)	Kinetics-400 Top-1 (%)
Skeletal Only	×	√	√	√	93.63	81.3
RGB Only	√	×	√	√	87.18	78.65
No Dynamic Weight	√	√	×	√	96.1	85
No Attention	√	√	√	×	95.2	84.9
Full Model (Ours)	√	√	√	√	96.52	85.3

4.7. Information Output

We also visually verify the recognition performance for typical actions such as "Standing" and "Sitting" in the information output module presented in Figure 13. The proposed system compresses and stitches the multi-modal information comprising frame data, identity recognition, action recognition, and skeletal information effectively to ensure efficient transmission. In practice, skeletal information and identity ID are then combined with the recognition result of the action, compressing the data before sending. This will strongly reduce the bandwidth used during operations. The system promises high accuracy in recognition while enhancing the efficiency in transmission, hence suitable for maritime surroundings with limited bandwidth.

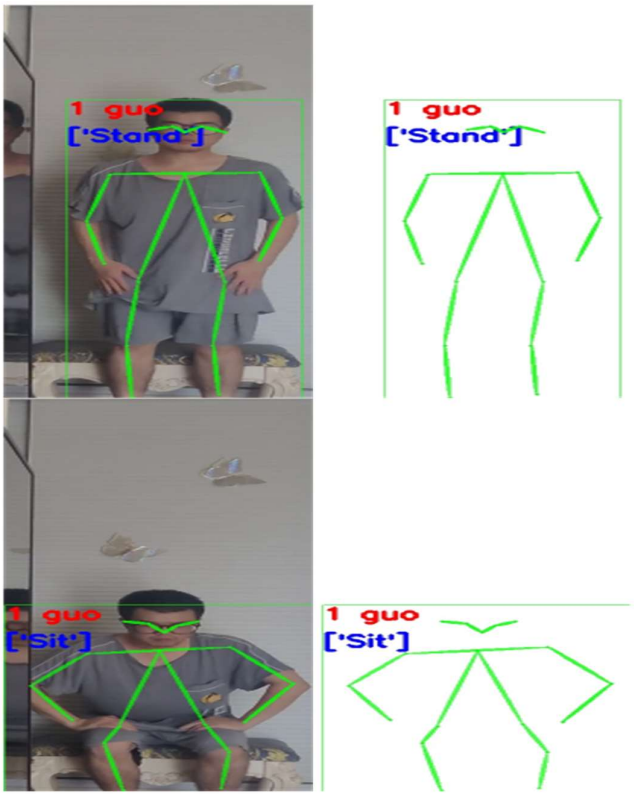

Fig. 13. Combination of skeleton information, identity recognition, and action recognition demonstration

5. Discussion

This study proposes a multimodal action and identity recognition system designed for bandwidth-constrained scenarios. By integrating RGB video and skeleton features through a dual-stream architecture, the system achieves efficient and robust action recognition. RGB features provide contextual background information, effectively complementing skeleton features that focus on capturing dynamic motions. The lightweight and structured nature of skeleton data significantly reduces transmission volumes, making it particularly suitable for constrained environments.

However, certain limitations persist. While frame interval extraction improves computational efficiency, it may result in the loss of critical motion details for continuous or complex actions. Future work could address this limitation through adaptive interval strategies. To address this, future research could explore adaptive frame interval strategies that dynamically adjust based on the complexity of the action, balancing computational efficiency and the preservation of motion details.

Another limitation lies in the size and diversity of the custom dataset. Although it encompasses a variety of typical surveillance scenarios, its scale may restrict the model's ability to generalize to more complex or unstructured environments. For example, in open, multi-person recognition scenarios, the dataset might not fully reflect the range of possible behaviors, limiting its effectiveness. Expanding the dataset's diversity and validating the system in broader, real-world settings could help overcome this limitation.

Despite these challenges, this study makes several significant contributions. Skeleton data encoding and frame interval extraction techniques successfully reduce transmission volumes by over 80% compared to raw RGB video, while simultaneously improving bandwidth utilization. These advancements make the system practical for remote monitoring applications in bandwidth-constrained environments. Nonetheless, achieving a balance between efficiency and performance in highly dynamic, real-time monitoring scenarios remains a topic for further investigation. Future research could explore distributed edge computing and more efficient skeleton data compression methods to enhance system scalability and real-time capabilities.

In conclusion, this study provides a practical solution for multimodal action and identity recognition in bandwidth-constrained scenarios. The integration of multimodal fusion and frame interval extraction improves recognition accuracy and efficiency while reducing data transmission demands, laying a solid foundation for future research. These findings offer not only a technical framework for further development but also valuable insights for applying such systems in real-world settings.

6. Conclusion

It therefore proposes a multi-modal action recognition that includes RGB and skeletal data, extraction of the technology of frame interval, optimizes transmission strategies, thus producing marked improvements in real-time performance at higher computational efficiency along with transmission efficiency. Experimental results demonstrate that multi-modal fusion can significantly provide improvements in the accuracy of action recognition in complicated scenarios and effectively address the limitations of single-modal approaches in terms of skeletal data combined with RGB video. The proposed paper puts forward a method of RGB-D multi-modal action recognition that combines skeletal information with RGB video. The novelty of this work involves proposing a multi-modal action recognition approach by combining skeletal information with RGB video. Besides, better performance of frame interval and text processing of skeletal data reduces the computational burden and bandwidth requirement on this system.

The proposed multi-modal action and identity recognition system is particularly suited for bandwidth-constrained environments, such as remote video surveillance, maritime monitoring, and real-time behavior analysis. By integrating skeletal data with RGB video streams, the system effectively

enhances the robustness and accuracy of action recognition. Moreover, through frame interval extraction and data compression techniques, it significantly reduces bandwidth requirements, enabling efficient real-time monitoring and behavior analysis under complex and resource-constrained conditions. This approach offers valuable applications in security monitoring, behavior recognition, and identity recognition, providing comprehensive support for intelligent surveillance systems.

We will investigate more sophisticated multi-modal fusion strategies in our future work to further improve the adaptiveness of our system, with better accuracy for a wider range of application scenarios, especially for combined usage of identity and action recognition. Future systems will be able to handle various complex monitoring scenarios using more intelligent fusion strategies and much more efficient multi-target recognition, contributing to further development of intelligent surveillance systems.

Funding

This work was supported by Key R&D Program of Shandong Province: Action plan to boost scientific and technological innovation in rural revitalization (No: 2023TZXD051).

Disclaimer/Publisher's Note

References

[1] Liu, A. A., Xu, N., Nie, W. Z., Su, Y. T., Wong, Y., & Kankanhalli, M. (2016). Benchmarking a multimodal and multiview and interactive dataset for human action recognition. IEEE Transactions on cybernetics, 47(7), 1781-1794.

[2] Shabaninia, E., Nezamabadi-pour, H., & Shafizadegan, F. (2024). Multimodal action recognition: a comprehensive survey on temporal modeling. Multimedia Tools and Applications, 83(20), 59439-59489.

[3] Tu, C. H., Yang, C. Y., & Hsu, J. Y. J. (2019, May). Idennet: Identity-aware facial action unit detection. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-8). IEEE.

[4] Ankalaki, S. (2024). Simple to Complex, Single to Concurrent Sensor based Human Activity Recognition: Perception and Open Challenges. IEEE Access.

[5] Liu, H., & Wang, L. (2018). Gesture recognition for human-robot collaboration: A review. International Journal of Industrial Ergonomics, 68, 355-367.

[6] Yang, C., Chen, D., & Xu, Z. (2021, December). Action recognition system for security monitoring. In International Conference on Artificial Intelligence, Virtual Reality, and Visualization (AIVRV 2021) (Vol. 12153, pp. 62-67). SPIE.

[7] Dhiman, C., & Vishwakarma, D. K. (2019). A review of state-of-the-art techniques for abnormal human activity recognition. Engineering Applications of Artificial Intelligence, 77, 21-45.

[8] Diraco, G., Rescio, G., Caroppo, A., Manni, A., & Leone, A. (2023). Human action recognition in smart living services and applications: context awareness, data availability, personalization, and privacy. Sensors, 23(13), 6040.

[9] Diraco, G., Rescio, G., Siciliano, P., & Leone, A. (2023). Review on human action recognition in smart living:

Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing. Sensors, 23(11), 5281.

[10] Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., & Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition, 108, 107561.

[11] Yin, Y., Xie, L., Jiang, Z., Xiao, F., Cao, J., & Lu, S. (2024). A Systematic Review of Human Activity Recognition Based On Mobile Devices: Overview, Progress and Trends. IEEE Communications Surveys & Tutorials.

[12] Al-Faris, M., Chiverton, J., Ndzi, D., & Ahmed, A. I. (2020). A review on computer vision-based methods for human action recognition. Journal of imaging, 6(6), 46.

[13] Ke, Q., Liu, J., Bennamoun, M., An, S., Sohel, F., & Boussaid, F. (2018). Computer vision for human–machine interaction. In Computer vision for assistive healthcare (pp. 127-145). Academic Press.

[14] Chakraborty, B. K., Sarma, D., Bhuyan, M. K., & MacDorman, K. F. (2018). Review of constraints on vision‐based gesture recognition for human–computer interaction. IET Computer Vision, 12(1), 3-15.

[15] Dao, T., Roy-Chowdhury, A., Nasrabadi, N., Krishnamurthy, S. V., Mohapatra, P., & Kaplan, L. M. (2017, October). Accurate and timely situation awareness retrieval from a bandwidth constrained camera network. In 2017 IEEE 14th international conference on mobile ad hoc and sensor systems (MASS) (pp. 416-425). IEEE.

[16] Gad, G., Gad, E., Fadlullah, Z. M., Fouda, M. M., & Kato, N. (2024). Communication-efficient and privacy-preserving federated learning via joint knowledge distillation and differential privacy in bandwidth-constrained networks. IEEE Transactions on Vehicular Technology.

[17] Luo, D., Zhang, Y., Sun, G., Yu, H., & Niyato, D. (2024). An Efficient Consensus Algorithm for Blockchain-Based Cross-Domain Authentication in Bandwidth-Constrained Wide Area IoT Networks. IEEE Internet of Things Journal.

[18] Canel, C., Kim, T., Zhou, G., Li, C., Lim, H., Andersen, D. G., ... & Dulloor, S. (2019). Scaling video analytics on constrained edge nodes. Proceedings of Machine Learning and Systems, 1, 406-417.