Image Processing, Electronics and Computers
L. Trajkovic et al. (Eds.)

doi:10.3233/ATDE240464

A Review of Behavior Recognition Research Based on Mobile Vision Devices

Yaqian WANG, Yuan CHAO[1], Zhen CAO, Huaiyang ZHU, Shuaishuai DU, Hengyu LU and Yijun ZHANG
School of Mechanical Engineering, Jiangsu University of Technology, Changzhou, China

Abstract. In view of the behavior recognition technology of mobile vision devices in security scenarios, this paper first describes the research and application progress of behavior recognition technology in security scenarios, and expounds the difficulties in its actual detection tasks, such as camera movement, behavior occlusion, illumination variation, background interference, multi-view variation, and interclass similarity. Then, according to the different network structures, the architecture composition and recognition characteristics of the model based on the two-stream convolutional architecture, the 3D convolutional architecture and the behavior recognition method based on the self-attention mechanism are detailed analyzed and elaborated. Then, the deployment and application of relevant behavior recognition algorithms are carried out on high-performance GPU and embedded microcomputer platforms. The accuracy, detection rate, parameter quantity and actual detection effect are compared and analyzed. Finally, based on the theoretical analysis and comparative experimental results, the limitations of the current behavior recognition algorithm are summarized, and the development direction of algorithm lightweight and adaptability improvement is pointed out.

Keywords. Security, mobile vision devices, deep learning, behavior recognition

1. Introduction

Nowadays society is becoming increasingly open, and various social security issues are emerged one after another. Traditional security models lack flexibility and have a high risk factor. To promote the innovation and development of security technology, a large number of mobile visual devices have emerged in the market. The excessive consumption of traditional security resources is compensated effectively based on the security method of mobile visual devices, which is characterized by real-time feedback, flexibility, portability and so on. At present, the main functions of the mobile security devices based on machine vision in the market are identity recognition and environmental detection [1,2], while behavior recognition can realize timely feedback on dangerous behaviors or malicious events. Therefore, nowadays the behavior recognition technology based on machine vision is one of the main directions for upgrading the functions of mobile security devices [3].

Traditional behavior recognition algorithms extract and reconstruct target features manually [4], and when deployed on mobile vision devices, they are greatly affected by

[1] Corresponding Author: Yuan Chao, chaoyuan@jsut.edu.cn

environmental changes and have poor accuracy and transferability [5]. The behavior recognition algorithms based on deep learning learn relevant action features autonomously to achieve a non-linear description of the recognized object [6], resulting in better recognition performance and less manual intervention. Therefore, compared to traditional security technologies, the research of behavior recognition algorithms based on mobile vision platforms and deep learning technology is more practical significance in terms of security intelligence and autonomy.

In order to further study the applicability and scalability of behavior recognition algorithms based on deep learning technology in the field of mobile security, this paper summarizes the current behavior recognition difficulties of mobile vision devices, commonly used behavior recognition algorithms and their recognition characteristics. On this basis, comparative experiments are carried out to summarize the advantages, limitations and applicable scenarios of various commonly used algorithms under the influence of different behavior recognition difficulties. Finally, the development direction of algorithm lightweight and adaptability improvement is pointed out. This is of great significance to improve the intelligence of mobile security.

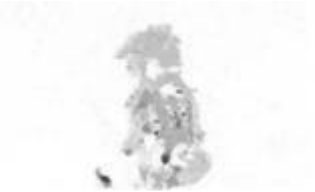
2. Difficulties in Mobile Vision Device Behavior Recognition Task

The detection and recognition of actual security behavior are affected by many uncertain factors based on mobile vision devices, which poses great challenges to their comprehensive recognition performance. Currently, there are some difficulties in mobile vision device behavior recognition tasks, such as camera movement, behavior occlusion, illumination variation, background interference, multi-view variation, and interclass similarity.

(1) Camera movement. Camera movement is an influential factor that cannot be avoided by mobile vision devices when performing behavior recognition tasks. The spatial and temporal information of the pedestrian's behavior in the image are disrupted by the basic movement of mobile visual devices or the dynamic movement of the camera itself, thereby affecting behavior judgment. As shown in Figure 1 [7], wherein, the original captured image is shown in Figure 1(a), the captured optical flow image is shown in Figure 1(b). And the optical flow image with camera movement removed is shown in Figure 1(c). From the figure, it can be clearly seen that the redundant and complex optical flow information caused by camera movement (i.e., the dark shaded part in Figure 1(b)) seriously affects the correct behavior judgment of the target by the behavior recognition model.

(a) Original captured

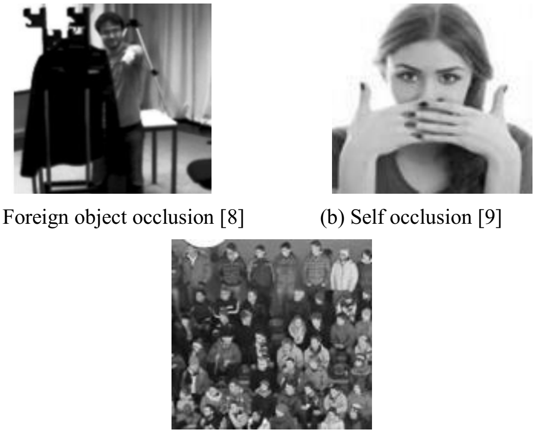
(b) Original optical flow



(c) Camera movement removed

Figure 1. Camera movement [7]

(2) Behavior occlusion. Behavior occlusion is a common problem in behavior recognition tasks, which mainly includes the occlusion of external foreign objects, the occlusion of human body parts, and the mutual occlusion between people in a crowd. As shown in Figure 2, in Figure 2(a) [8], it is impossible to determine whether the target has punched due to the hanger and clothing obscuration. In Figure 2(b) [9], it is impossible to determine whether the person is smiling or speaking due to hand occlusion. In Figure 2(c) [10], it is impossible to determine whether someone is operating the phone due to the mutual occlusion in a crowd. Extractable human body part feature information is reduced by occlusion, and the provided behavior judgments are inaccurate.

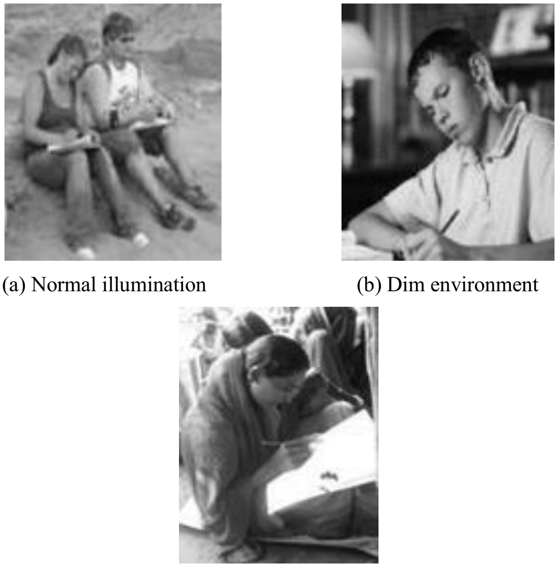


(a) Foreign object occlusion [8] (b) Self occlusion [9]

(c) Crowd occlusion [10]

Figure 2. Behavior occlusion

(3) Illumination variation. The impact of the changes in illumination environment on the presentation of pedestrian behavior and actions is uncontrollable. The blurred behavior performance is due to the changes in light intensity, which is another challenge for mobile vision devices when performing behavior recognition. As shown in Figure 3 [11], they are all writing actions, but obviously, the writing actions in Figure 3(a) are clear and natural under normal illumination, while the writing actions in Figure 3(b) and Figure 3(c) are relatively blurred under weak and strong light illumination. Therefore, how to make accurate behavior determinations under different degrees of illumination conditions is also a problem that needs to be further explored in the field of behavior recognition.

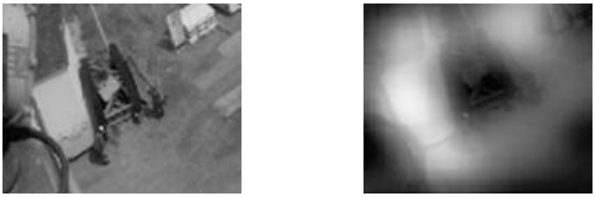


(a) Normal illumination (b) Dim environment

(c) Exposure environment

Figure 3. Illumination variation [11]

(4) Background interference. When mobile vision devices perform behavior recognition in actual security scenarios, the reliability of the mobile vision devices is affected seriously by the chaotic background information of the target object's environment. The image of workers working in the oil field captured by drones is shown in Figure 4 [12]. Wherein, the original video image is shown in Figure 4(a), while the irrelevant and complex background are weakened in Figure 4(b). The drone may overly focus on unnecessary background information (i.e. dark shaded part), thus ignoring whether the workers are safely carrying out the oilfield exploitation work, which greatly weakens the security supervision effect of the drone. Therefore, how to effectively handle redundant background information is the key to improve the performance of mobile vision devices.

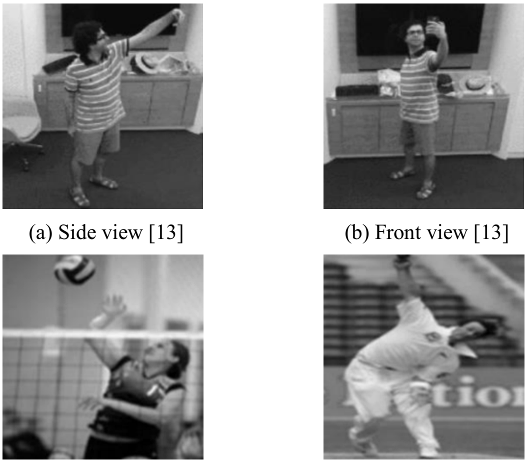


(a) Original captured image (b) Weakening of region-of-non-interest

Figure 4. Background interference [12]

(5) Multi-view variation and interclass similarity. In addition to the above task difficulties, multi-view variation and interclass similarity are also influence factors for mobile vision devices when performing behavior recognition. As shown in Figure 5, wherein, the side view selfie action captured by the moving probe is shown in Figure 5(a) [13], and the front view selfie action captured by the moving probe is shown in Figure 5(b) [13]. Obviously, there are significant differences in the presentation of the selfie actions under different perspectives. During the dynamic recognition process of mobile

vision devices, the diversity of action presentation is caused by the change of recognition viewpoint, and the effect of recognition is affected. The hitting action is shown in Figure 5(c) [14], and the throwing action is shown in Figure 5(d) [14]. It can be seen that the two actions have a high degree of similarity, which makes it difficult to judge the category accurately. Pedestrians are highly subjective as the target objects, many actions are presented with similarities, thus, the equipment judgement is also affected.



(a) Side view [13]

(b) Front view [13]

(c) Hitting [14]

(d) Throwing [14]

Figure 5. Multi-view variation [13] and interclass similarity [14]

In summary, the difficulties in behavior recognition tasks that mobile vision devices need to overcome when performing actual security detection are complex and various. For this reason, how to select or improve the behavior recognition algorithm to adapt to the actual security detection task is a key direction that needs to be further researched and explored.

3. Behavior Recognition Algorithm

The core of behavior recognition algorithms is that whether mobile vision devices can make accurate judgments and respond promptly to abnormal or dangerous behaviors in security scenarios [15]. The behavior recognition methods based on deep learning have higher application flexibility, higher autonomy, and less manual intervention compared to traditional algorithms, making them the mainstream behavior recognition algorithms at present.

In the research of video behavior understanding, the behavior recognition algorithms based on deep learning have strong spatio-temporal modeling or temporal learning capabilities, which can effectively monitor and recognize behaviors and are widely used [16]. Usually, according to the different network structures, common behavior recognition algorithms are mainly divided into three categories: behavior recognition algorithms based on two-stream convolutional architecture, 3D convolutional architecture, and self-attention architecture.

3.1. Two-Stream Convolutional Architecture

The Two-Stream Convolutional Networks [17] is a deep learning algorithm, which is widely applied in the field of security [18]. Currently, most of the behavior recognition algorithms based on two-stream convolutional architecture are classified into three categories according to different information processing mechanisms: the behavior recognition algorithms based on two-stream convolutional networks, TSN (Temporal Segment Networks) [19] and two-stream fusion recurrent networks.

The behavior recognition algorithm based on two-stream convolutional networks processes optical flow information and RGB information separately through two independent convolutional networks, and has good robustness against changes in illumination and perspective [20]. Wang et al. [21] proposed a security detection algorithm for abnormal behavior determination based on an improved two-stream convolutional neural network. The algorithm has good stability in environments with variable viewing angle and illumination, but poor occlusion robustness.

The two-stream convolutional neural network algorithm framework is simple and clear, but it consumes a huge amount of time and computational resources. TSN adopts a sparse sampling sequence processing method, which reduces the computational consumption of two-stream architecture networks [19], and is suitable for some security scenarios with relatively limited computational resources [22]. Wu et al. [23] proposed a detection algorithm for detecting pedestrian fall behavior on escalators based on improved TSN. The recognition is accurate and real-time, but the recognition effect on similar behaviors is poor.

The above two algorithms are unable to recognize behaviors with large time spans and coherence. At this time, the method of the two-stream fusion LSTM (Long Short Term Memory) [24] is used to strengthen the long-term dependency relationship between video frames, and the problems can be effectively improved, but the computational cost is high. To address this problem, Islam et al. [25] proposed an improved dual-stream convolutional neural network, which mimics the network architecture of LRCN [26] and fuses separable LSTMs to achieve the recognition of long-term sequence violence, with good robustness and generalization ability. It also reduces the high computational cost of the model and is suitable for deployment on mobile vision devices.

Table 1. Summary of two-stream convolutional architecture network algorithms

Two-stream architecture behavior recognition algorithms	Advantages	Limitations
Two-stream convolutional neural network [17]	Good robustness to changes in perspective and illumination variation	High consumption of computational resources and poor occlusion robustness
TSN [19]	Relatively low consumption of computational resources	Poor interclass similarity recognition performance
Two-stream fusion LSTM [24]	Good at recognizing behavior actions with large time spans	High computational cost, redundant information, and poor real-time performance

The summary of the behavior recognition algorithms based on two-stream convolutional architecture is shown in Table 1. The principles of the algorithms are basically the same, but the advantages and limitations are different. The behavior recognition algorithms based on two-stream convolutional network have good robustness to perspective and illumination variation, but they consume large computational resources and have poor occlusion robustness. The behavior recognition algorithms based on TSN have a relatively low computational cost, but the algorithms have difficulty in identifying similar behavior categories. The behavior recognition algorithms based on two-stream fusion LSTM are good at identifying coherent behaviors, but the algorithms have high computational costs and poor real-time performance. For mobile vision devices, the behavior recognition algorithms based on two-stream convolutional architecture are generally too complex to compute, and how to improve them in combination with the actual security task requirements is a worthy direction for further research.

3.2. 3D Convolutional Architecture

The behavior recognition algorithms based on the 3D convolutional architecture [27] directly extract temporal and spatial features from the original input video through 3D convolutional layers, which reduce the complexity of data preprocessing. Among them, the most representative algorithms are C3D (Convolutional 3D) [28], I3D (Inflated 3D ConvNet) [29], R (2+1)D [30] and Slowfast [31].

C3D adds the time dimension directly to the two-dimensional convolution for capturing dynamic features on the spatio-temporal domain, and the algorithm demonstrates comprehensive feature extraction capability although there is no significant improvement in accuracy [32]. Therefore, it is often used as a feature extractor in security behavior detection. However, in the actual detection process, it is susceptible to light changes and background interference.

I3D obtains a well-performing 3D convolutional model by inflationary expansion, which reduces the training cost while retaining efficient performance. I3D has good robustness to variation of illumination, background and perspective in the environment, it can also recognize behaviors with high fine-grain or fast speed [33]. Gao et al. [34] proposed a public place security behavior detection algorithm based on I3D, which achieved efficient recognition of covert behaviors such as theft and arson. However, the computational complexity of I3D is relatively high, and the deployment and application of edge devices are not ideal.

R(2+1)D splits 3D ResNet (Residual Network) [35] into 2D spatial convolution and 1D temporal convolution, which reduces the computational complexity while guaranteeing the model generalization ability, and it is a preferred choice for actual security tasks with computational speed requirements. However, the ‘2D+1D’ network model structure sacrifices the network's spatio-temporal modeling ability to a certain extent, and the accuracy is not as good as the full 3D convolutional network with the same model volume [36].

Slowfast adopts a two-path structure to process the high-frequency and low-frequency information of the video data separately in order to obtain a comprehensive and detailed behavior representation. Therefore, Slowfast is able to effectively process action information with different speeds and complexities, but it suffers from high computational complexity. Gankhuyag et al. [37] proposed a lightweight traffic police gesture recognition algorithm based on improved Slowfast, which could be deployed on autonomous vehicles. It improved the recognition speed of the algorithm while ensuring recognition accuracy. However, the accuracy of the algorithm decreased when detecting at a long distance, indicating that Slowfast has significant potential for the improvement of small target behavior recognition.

The summary of behavior recognition algorithms based on 3D convolutional architecture is shown in Table 2. Among them, C3D has good comprehensive feature extraction ability, but it is sensitive to environmental changes. I3D has better robustness to environmental changes and can identify small actions, but it has high computational complexity. R(2+1)D reduces the computational complexity of the model and is suitable for security scenarios with higher demand for speed. Slowfast is good at identifying behaviors with varying speed and complexity, but its recognition performance for detecting small target actions is poor. The behavior recognition algorithms based on 3D convolutional architecture have their own characteristics and limitations. For behavior recognition tasks with varying situations in security scenarios, it is worth further exploring how to select suitable algorithms and make adaptive improvements to them.

Table 2. Summary of 3d convolutional architecture network algorithms

Typical 3D convolutional architecture algorithms	**Advantages**	**Limitations**
C3D [28]	Comprehensive feature extraction capability	Easy to be affected by illumination and background
I3D [29]	Good robustness to environmental changes, able to recognize fine-grained behaviors	Relatively high computational complexity
R(2+1)D [30]	Fast calculation speed	The recognition accuracy is not as good as that of full 3D convolutional models with the same volume
Slowfast [31]	The recognition effect is better for behaviors with varying speed and complexity	Poor performance in detecting small target actions

3.3. Self-Attention Architecture

The behavior recognition algorithm based on self-attention architecture is one of the emerging research directions in the field of video understanding in recent years. Self-attention [38] strengthens the global information connection during the feature processing of the network, and it also has certain advantages for security recognition tasks with complex and varying situations. Among them, the behavior recognition algorithms incorporating Non-local neural networks [39] and the behavior recognition algorithms based on transformer [40] are the most dominant self-attention behavior recognition algorithms nowadays.

Non-local neural networks are a type of non-local self-attentive neural network module, which has outstanding advantages in security recognition tasks with large spatial-temporal spans. Mi et al. [41] proposed a behavior recognition algorithm that integrates Non-local module for the safety control of school students. This algorithm achieved accurate and stable recognition in situations where the target object's position changed, there was background interference, or partial occlusion occurred. However, the Non-local module has poor interpretability and poses high application debugging difficulties for some platforms that are convenient for deploying mobile vision [42].

The behavior recognition algorithms based on transformer are another type of neural network algorithms based on self-attention architecture, which focus more on global feature association. Mazzeo et al. [43] proposed a dangerous behavior detection algorithm for public safety surveillance based on 3D convolutional fusion transformer. The algorithm had strong robustness to the interference of illumination variation and background redundancy, but its recognition effect is not good at security tasks with relatively high fine-grained requirements [44]. Liu et al. [45] proposed a behavior recognition algorithm dedicated to detecting the falling action of the elderly based on TimeSformer [46], which achieved real-time security monitoring of the elderly living alone. The algorithm had a fast response speed, but the parameter quantity is too large, making it unfriendly for the deployment on the front-end devices [47].

Table 3. Summary of self-attention architecture algorithms

Self-attention architecture algorithm	Advantages	Limitation
Integrating Non-local neural networks [39]	Have robustness to situation with target varied position, background interference and partial occlusion	Poor interpretability and inflexible application debugging
Based on Transformer [40]	Strong anti-interference ability against illumination variation and background redundancy	Unstable recognition of fine-grained behavior action, large number of parameters, not conducive to front-end device deployment

The summary of the behavior recognition algorithms based on self-attention architecture is shown in Table 3. Among them, the behavior recognition algorithms that integrate the Non-local module can maintain good recognition performance in situations with target varied position, background interference and partial occlusion, but its interpretability is poor, which is not conducive to application debugging on mobile vision platforms. The behavior recognition algorithms based on Transformer have strong anti-interference ability against illumination and background variation, but the parameter quantity is too large, which is not conducive to the deployment on the front-end devices, and at the same time, the recognition effect on subtle actions is not stable. The behavior recognition algorithms based on self attention architecture generally have strong anti-interference ability, but each has its own characteristics. How to select suitable algorithms based on specific security situations to achieve maximum recognition effect is a direction worthy of further research and exploration.

4. Deployment and Application Comparison of Behavior Recognition Algorithms

In this paper, the current status of security applications of mainstream behavior recognition algorithms based on deep learning are respectively elaborated and analyzed in detail and summarized. However, in terms of the actual application of the algorithms, there is little literature about their deployment and application on mobile vision devices. In order to further study and verify the performance of different algorithms applied on different platforms, this section carries out the deployment and application of relevant behavior recognition algorithms on high-performance GPUs and embedded microcomputers, and analyses and compares the accuracy , detection rate, parameter quantity and the actual detection effect.

4.1. Comparison-of Algorithm-Performance on-Different-Deployment-Platforms

Deep learning behavior recognition algorithms generally suffer from the problem of excessive computation, therefore, it is necessary to conduct feasibility analysis on the actual deployment of various behavior recognition algorithms on mobile visual edge devices. In this paper, the representative algorithms mentioned in Section III are selected for comparison in terms of recognition accuracy on the public dataset UCF101.

Meanwhile, a mobile vision platform equipped with Jetson Xavier NX as the core, which is selected as a representative of embedded microcomputer, and a PC with NVIDIA GeForce RTX 3070 graphics card with 8G video memory is selected as a representative of high-performance GPU for the comparison in terms of actual detection rate. As shown in Figure 6, the Jetson Xavier NX control board, USB camera, WiFi antenna, and mobile power supply are used to construct the removable embedded vision platform. In this paper, TensorRT is further used to accelerate the computation of each algorithm, which is easy to compare and analyze. The actual performance of each algorithm is shown in Table 4, wherein, FPS (Frames Per Second) is the unit of detection rate and M (Million) is the statistical index of network parameter quantity.

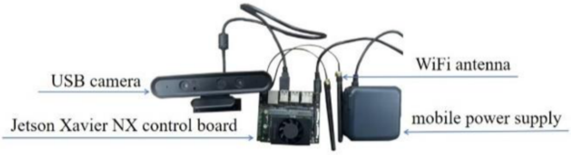


Figure 6. Algorithm deployment platform

Table 4. Comparison of actual performance of algorithms

Algorithms	The accuracy on UCF101(%)	Detection rate(FPS)		Parameters(M)
		RTX3070	Jetson Xavier NX	
Two-Stream [17]	81.9	12.3	1.7	116.7
TSN [19]	90.7	16.2	3.7	23.7
LRCN [26]	74.2	11.3	1.4	84.7
C3D [28]	85	15.0	2.7	78.0
I3D [29]	93.2	18.4	5.0	28.0
R(2+1)D [30]	93.4	29.0	6.3	63.7
Slowfast [31]	91.8	22.7	5.9	33.8
TimeSformer [46]	91.5	30.4	6.5	121.4

(1) The accuracy. In actual security scenarios, the behavior recognition accuracy of mobile vision devices is the key to determine the abnormal behavior of suspicious individuals. As shown in Table 4, the accuracy of behavior recognition algorithms based on 3D convolutional architecture is generally above 90%, which is significantly better than behavior recognition algorithms based on two-stream convolutional architecture with an accuracy of 70%~90%, indicating that the behavior recognition algorithms based on two-stream convolutional architecture have certain effectiveness in separate processing of RGB and optical flow information. However, in comparison, the behavior recognition algorithms based on 3D convolutional architecture can better capture the spatio-temporal dependencies between consecutive frames in videos, and the recognition accuracy is generally higher. The behavior recognition algorithms based on self-attention also have a high recognition accuracy. Although it does not significantly surpass 3D convolutional networks, it also reflects the outstanding ability of the self attention mechanism to integrate global information when processing video sequences.

From a detailed perspective, I3D and R (2+1) D have strong capabilities in capturing and integrating complex spatio-temporal information in videos, with the most significant recognition accuracy. Slowfast combines two complementary convolutional streams to capture the speed changes of behaviors, with unique adaptability and a recognition accuracy of 91.8%. TimeSformer adopts a self-attention mechanism to process spatio-temporal sequences, and has significant competitiveness in global information extraction, with an accuracy of 91.5%. However, the network architectures of two-stream convolutional network, C3D and LRCN are not as advanced as the other networks, and thus the accuracy is not high.

(2) Detection rate. In addition to the accuracy, the behavior detection speed of mobile vision devices is also an important performance factor in actual security scenarios. According to Table 4, the recognition speed of each algorithm on PC can meet the recognition speed requirement in actual security scenarios [48]. Among them, TimeSformer reduces the computational cost by strengthening the dependency relationship between video frames in the global long sequence video, achieving a recognition rate of 30.4 FPS. R(2+1)D splits the convolution of spatio-temporal domain, which reduces the unnecessary computational consumption, and the recognition rate reaches 29.0 FPS. Slowfast efficiently processed video information through feature extraction at different scales, and achieves a rate of 22.7 FPS. The depth of C3D and I3D models is large, requiring high-level computing power support, and the recognition speed is limited. The behavior recognition algorithms based on two-stream convolutional architecture need to process optical flow information, which consumes more computational resources and results in a lower recognition rate.

Similar comparative differences can be derived from the comparison of the recognition rates among the algorithms on the Jetson Xavier NX. However, due to the significant difference in computational power between embedded microcomputers and high-performance GPUs, the recognition rate of each algorithm on Jetson Xavier NX is significantly lower than that on PC. In real security scenarios, mobile vision devices with autonomous behavior judgment capabilities can significantly reduce data transmission latency and at the same time reduce the consumption of hardware resources. Therefore, one of the important directions for the future development of the security industry is how to simplify the model or optimize and upgrade the algorithm to make it suitable for the deployment of mobile vision devices.

(3) Network parameter. The parameter quantity of the algorithm is also an important factor that directly affects the application performance of mobile vision devices. As shown in Table 4, although TimeSformer has outstanding performance in terms of the accuracy and recognition rate, its parameter quantity is much larger than that of other networks, posing a significant challenge to the memory space of mobile vision devices. The shunt information processing mechanism of two-stream convolutional networks also has a large number of parameters, while the overall structure of the rest network is relatively simple, so the number of parameters is relatively small. It can be seen that there is no direct relationship between the model size and the recognition effect. How to improve and obtain an algorithm model with appropriate size and excellent performance is one of the core issues that urgently need to be solved when deploying behavior recognition algorithms on mobile devices.

In summary, the comprehensive performance of behavior recognition algorithms is limited by the scale of computational and memory resources. Therefore, how to balance the relationship among the model size, the network depth of algorithms, the detection speed and recognition accuracy is the key to determine the comprehensive performance of mobile vision devices in behavior recognition tasks.

4.2. The Recognition Performance of Algorithms under Various Interferences

In addition to inherent performance factors such as recognition accuracy, detection rate, and parameter quantity, the detection performance of mobile vision devices in actual security scenarios will be affected by many external uncontrollable factors. Therefore, in this paper, under the interferences of camera movement, behavior occlusion, illumination variation, background interference, and multi-view variation, 100 tooth brushing action detection samples with the length of 1~3 seconds are each made by self for actual detection and accuracy statistics. 50 tooth brushing and lipstick application action video clips with the length of 1~3 seconds are each made by self as the detection samples under the interference of interclass similarity for actual detection and accuracy statistics. The detection results are shown in Table 5.

According to Table 5, affected by various interferences, the recognition accuracy of each algorithm is generally not high. However, in comparison, it can be seen that different behavior recognition algorithms have different advantages and shortages when dealing with different complex situations. On the whole, the recognition accuracy of each algorithm is generally not high under the interference of various factors. The behavior recognition algorithms based on 3D convolution and self-attention mechanism generally perform better. The former is mainly due to their superior spatio-temporal joint modeling ability and diverse architecture design, while the latter benefits from their strong global information capture ability. The behavior recognition algorithms based on the two-stream convolution architecture perform relatively poorly, on the one hand, due to the lack of progressiveness of the model structure, and on the other hand, due to the limitations of the shunting information processing mechanism.

Table 5. The recognition performance of each algorithm under different recognition task difficulties

Algorithms		**Recognition accuracy and relative effectiveness**					
		Camera movement	***Behavior occlusion***	***Illumination variation***	***Background interference***	***Interclass similarity***	***Multi-view variation***
Two-stream	Two-Stream [17]	29% worse	12% worse	59% better	42% average	39% average	57% better
	TSN [19]	41% average	24% average	38% average	40% average	30% worse	40% average
	LRCN [26]	22% worse	13% worse	20% worse	24% worse	23% worse	19% worse
3D	C3D [28]	37% average	27% average	33% worse	31% worse	40% average	35% average
	I3D [29]	57% better	30% average	61% better	58% better	57% better	59% better
	R(2+1)D [30]	39% average	29% average	41% average	63% better	56% better	59% better
	Slowfast [31]	58% better	47% better	43% average	59% better	61% better	36% average
Self-attention	TimeSfor-mer [46]	56% better	43% better	58% better	61% better	41% average	37% average

From a detailed perspective, I3D has the most outstanding comprehensive performance, with stable recognition and strong generalization ability, especially in terms of illumination variation, background interference, high interclass similarity, etc. It is suitable for behavior recognition in complex environments such as public places and warehouses. R(2+1)D and Slowfast have better performance in situations with background interference and high interclass similarity, and especially Slowfast has good robustness in camera movement and behavior occlusion. They are more suitable for application in crowded places with complex backgrounds and diverse types of activity. TimeSformer has strong global information learning ability and performs well in terms of camera movement, behavior occlusion, illumination variation, etc, which is slightly insufficient when dealing with recognition tasks with high interclass similarity. It is suitable for deploying in scenario with varying illumination and perspectives, and certain occlusion situations. The shunting mechanism of two-stream convolutional networks has better robustness in the case of illumination variation and multi-view variation, but its recognition performance is poor under other interferences. Therefore, it is suitable for fixed places with illumination variation and multiple viewpoints. The actual performance of TSN, LRCN and C3D is average, so their application scenarios are relatively limited.

In summary, different behavior recognition algorithms have different adaptation scenarios according to different recognition characteristics. 3D convolutional architecture algorithms are suitable for complex security scenarios such as public places, warehouses, etc. Two-stream convolutional architecture algorithms are more different from each other, and the applicable scenarios need to be decided by the specific situation. The behavior recognition algorithms based on self-attention mechanism are suitable for complex intelligent traffic monitoring and nighttime monitoring scenarios. How to choose algorithms to meet the needs and challenges of specific security scenarios, and to ensure the accuracy and efficiency of recognition is an important direction for the development of intelligent security behavior recognition.

5. Conclusion

In this paper, for the recognition difficulties such as camera movement, behavior occlusion, illumination variation, background interference, multi-view variation and interclass similarity, which exist in actual security detection of mobile vision devices, the commonly used representative behavior recognition algorithms such as two-stream convolutional neural network, C3D, TimeSformer, etc, are analyzed and elaborated in detail from multiple perspectives, such as model structure, computational complexity, parameter quantity, and actual detection effect, and their applicable scenarios, advantages and limitations are summarized. At the same time, the algorithms are deployed on high-performance GPU and embedded microcomputer platforms, and experimental comparative analyses are carried out. It can be seen that some progress has been made in the research of behavior recognition based on mobile vision devices, but there is still significant research space on improving algorithms in the direction of model simplification and lightweight. The specific descriptions are as follows.

(1) For the problem that the computational resource consumption of behavior recognition algorithms is generally large, which is not conducive to the deployment and application of mobile visual devices, it can be optimized from the structure of the network. How to carry out effective model quantization and compression methods to reduce energy consumption while ensuring recognition accuracy, in order to conform to the computational and storage limitations of embedded microcomputers, is the main direction worthy of further research.

(2) To address the issue of how to select and improve appropriate behavior recognition algorithms to adapt to security needs in different specific contexts, it is possible to consider selecting algorithms that perform well under specific interference factors according to the specific characteristics of the actual security scenario, such as illumination, viewpoint, occlusion, background changes and other various environmental factors. In addition, targeted behavior datasets can be tailored to the characteristics of the security task, and behavior samples under various complex environments or occlusion conditions can be added to improve the adaptability of the algorithm in practical applications. Alternatively, new algorithm training strategies can be explored and

developed to enable the model to quickly adapt to new scenarios and improve the algorithm's anti-interference ability.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 51905235), the Open Fund of Jiangsu China-Israel Industrial Technology Research Institute (Grant No. JSIITRI202101), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No. SJCX22_1485).

References

[1] Wu X, Xu J, Wang J, Li W, Guo, Y. Identity authentication on mobile devices using face verification and ID image recognition. Procedia Computer Science, 2019, 162: 932-939.

[2] Yan F, Zhou H, Peng ZH, Chen M, Xiong Y, et al. An Industrial Environmental Security Monitoring System in Mobile Device Based on Soft PLC//2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE, 2020: 711-715.

[3] Zhang D, Zhang H, Duan S, Luo Y, Jia F, et al. Deep action: A mobile action recognition framework using edge offloading. Peer-to-Peer Networking and Applications, 2022: 1-16.

[4] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.

[5] Khanday NY, Sofi SA. Taxonomy, state-of-the-art, challenges and applications of visual understanding: A review. Computer Science Review, 2021, 40: 100374.

[6] Wu D, Sharma N, Blumenstein M. Recent advances in video-based human action recognition using deep learning: A review//2017 International joint conference on neural networks (IJCNN). IEEE, 2017: 2865-2872.

[7] Wang H, Schmid C. Action recognition with improved trajectories//Proceedings of the IEEE international conference on computer vision. 2013: 3551-3558.

[8] Weinland D, Özuysal M, Fua P. Making action recognition robust to occlusions and viewpoint changes//Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11. Springer Berlin Heidelberg, 2010: 635-648.

[9] Urooj A, Borji A. Analysis of hand segmentation in the wild//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4710-4719.

[10] Mahmood A, Al-Maadeed S. Action recognition in poor-quality spectator crowd videos using head distribution-based person segmentation[J]. Machine Vision and Applications, 2019, 30(6): 1083-1096.

[11] Qi T, Xu Y, Quan Y, Wang Y, Ling H. Image-based action recognition using hint-enhanced deep neural networks. Neurocomputing, 2017, 267: 475-488.

[12] Sun Y. ATSN: Attention-based temporal segment network for action recognition. Tehnički vjesnik, 2019, 26(6): 1664-1669.

[13] Zhang P, Lan C, Xing J, Zeng W, Xue J, et al. View adaptive neural networks for high performance skeleton-based human action recognition. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(8): 1963-1978.

[14] Akila K, Chitrakala S. Highly refined human action recognition model to handle intraclass variability & interclass similarity. Multimedia Tools and Applications, 2019, 78: 20877-20894.

[15] Chen Y, Zheng B, Zhang Z, Wang Q, Shen C, et al. Deep learning on mobile and embedded devices: State-of-the-art, challenges, and future directions. ACM Computing Surveys (CSUR), 2020, 53(4): 1-37.

[16] Yao H, Hu X. A survey of video violence detection. Cyber-Physical Systems, 2023, 9(1): 1-24.

[17] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos//Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1. 2014: 568-576.

[18] Hongchao S, Hu Y, Guoqing Z, Chuyue Z. Behavior Identification based on Improved Two-Stream Convolutional Networks and Faster RCNN//2021 33rd Chinese Control and Decision Conference (CCDC). IEEE, 2021: 1771-1776.

[19] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, et al. Temporal segment networks: Towards good practices for deep action recognition//European conference on computer vision. Springer, Cham, 2016: 20-36.

[20] Sun Z, Ke Q, Rahmani H, Bennamoun M, Wang G, et al. Human Action Recognition from Various Data Modalities: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3200-3225.

[21] Hao W, Zhang R, Li S, Li J, Li F, et al. Anomaly event detection in security surveillance using two-stream based model. Security and Communication Networks, 2020, 2020.

[22] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, et al. Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(11): 2740-2755.

[23] Wu Y, Wu SY, Yan Z. Research on Pedestrian Fall Action Recognition from Escalators//Application of Intelligent Systems in Multi-modal Information Analytics: 2021 International Conference on Multi-modal Information Analytics (MMIA 2021), Volume 2. Springer International Publishing, 2021: 277-286.

[24] Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation, 1997, 9(8): 1735-1780.

[25] Islam Z, Rukonuzzaman M, Ahmed R, Kabir MH, Farazi M. Efficient two-stream network for violence detection using separable convolutional lstm//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[26] Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, et al. Long-term recurrent convolutional networks for visual recognition and description//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.

[27] Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 221-231.

[28] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

[29] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[30] Tran D, Wang H, Torresani L, Ray J, LeCun Y, et al. A closer look at spatiotemporal convolutions for action recognition//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.

[31] Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211.

[32] Yao L, Torabi A, Cho K, Ballas N., Pal C, et al. Describing videos by exploiting temporal structure//Proceedings of the IEEE international conference on computer vision. 2015: 4507-4515.

[33] Liu G, Zhang C, Xu Q, Cheng R, Song Y, et al. I3d-shufflenet based human action recognition[J]. Algorithms, 2020, 13(11): 301.

[34] Gao J, Cheng X, Liu X, Shi F, Zhao M, et al. Feature Enhancement Based Multi-Feature Fusion Network for Video Anomaly Detection in Offshore Surbeillance//2023 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE, 2023: 550-557.

[35] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[36] Jin H, Yang J, Zhang S. Efficient Action Recognition with Introducing R (2+ 1) D Convolution to Improved Transformer//2021 4th International Conference on Information Communication and Signal Processing (ICICSP). IEEE, 2021: 379-383.

[37] Gankhuyag G, Yae H, Shim Y, Park C, Min K. A Lightweight Traffic Police Action Recognition Deep Learning Network for Edge Device//2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia). IEEE, 2022: 1-3.

[38] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 1980, 36(4): 193-202.

[39] Wang X, Girshick R, Gupta A, He K. Non-local neural networks//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.

[40] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[41] Mi J, Yuan J. Human action recognition on campus monitor by integrating non-local attention mechanism//5th International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM 2023). IET, 2023, 2023: 178-183.

[42] Li Y, Jin X, Mei J, Lian X, Yang L, et al. Neural architecture search for lightweight non-local networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10297-10306.

[43] Mazzeo PL, Spagnolo P, Fasano M, Distante C. Human Action Recognition with Transformers//International Conference on Image Analysis and Processing. Cham: Springer International Publishing, 2022: 230-241.

[44] He J, Chen JN, Liu S, Kortylewski A, Yang C, et al. Transfg: A transformer architecture for fine-grained recognition//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(1): 852-860.

[45] Liu H, Mu J, Zhang Z. Fall Detection for Surveillance Video Based on Deep Learning//International Conference in Communications, Signal Processing, and Systems. Singapore: Springer Nature Singapore, 2022: 123-129.

[46] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding?//ICML. 2021, 2(3): 4.

[47] Abdali AR. Data efficient video transformer for violence detection//2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT). IEEE, 2021: 195-199.

[48] Li J, Kang X, Jin L, Wu Y, Hai D, et al. A foreground-focused action recognition algorithm for intelligent unmanned systems//2020 3rd International Conference on Unmanned Systems (ICUS). IEEE, 2020: 500-504.