

# Design of a home video behavior recognition system based on visual privacy security mechanism

D.M. Zhao<sup>1,2</sup>

<sup>1</sup>Academic Affairs Office (Laboratory Center), Dongguan City University,  
No. 1 Wenchang Road, Songshan Lake Avenue, Dongguan, 523419, China;

<sup>2</sup>Graduate School, University of Perpetual Help System Laguna,  
City of Biñan, Laguna, 4024, Philippines

## Abstract

The rapid development of the Internet and advanced technology has brought great convenience to people's lives; However, real-time video and other privacy information obtained from computers can be leaked, resulting in economic losses and not conducive to the construction of computer network security. In response to the above issues, this study introduces compressed perception theory and temporal adaptive modules to achieve visual shielding, and based on this, designs a home video behavior system based on visual privacy security mechanism. The research results show that in the comparison of measurement matrices at different levels, the Bernoulli random matrix has the highest recognition accuracy, with recognition accuracy rates of 100 %, 98.73 %, 98.76 %, and 85.62 % from the first layer to the fourth layer, respectively. In the recognition performance results of different video behavior recognition systems in the YouTube database, UCF Sports database, and Hollywood2 database, the average recognition accuracy of the proposed system is the highest in most cases, with 94.6 %, 73.5 %, and 77.1 %, respectively. In summary, the system proposed in the study can achieve accurate recognition of home video behavior after visual masking, and has good results in practical applications.

**Keywords:** visual privacy security, home videos, behavior recognition, time series adaptive network, compression perception.

**Citation:** Zhao, DM. Design of a home video behavior recognition system based on visual privacy security mechanism. *Computer Optics* 2025; 49(2): 263-272. DOI: 10.18287/2412-6179-CO-1456.

## Introduction

With the rapid development of Internet technology, the network has become an important part of people's daily life. However, with the network security problems, data leakage, network attacks and intrusions and other problems occur frequently, bringing great trouble to individuals and enterprises [1]. In addition, a large number of camera systems have been installed in homes, nursing homes, and public places, and computer vision based remote care services for the elderly, disabled, and pets have flourished [2]. But personal sensitive information is protected and cannot be disclosed in the public environment. When managing images in different scenes, the private information of the people captured in the images is inevitably threatened, and with the increasing popularity of diverse technologies such as the internet, the personal information security of ordinary consumers also faces risks [3]. Therefore, in line with the construction of computer network security, ensuring that user identities and sensitive information related to personal privacy are not violated is currently a hot research topic. The theory of image compression sensing (CS) can process various home video behavior information obtained by cameras, thereby achieving privacy and security effects of visual masking [4–6]. In order to tackle the problem of privacy leaks in home security footage, this research develops a Home Video Behavior Recognition It combines and detects the

video frames and optical flow sequences of visual shield CS (VSCS) state films. This is based on the Visual Privacy Security Mechanism (VPSM-HVBR), which incorporates a Timing Adaptive Module (TAM). By achieving feature extraction and recognition of blurry recordings, this research seeks to stop house invasions and the theft of private information by unscrupulous individuals. The home VSCS encoding proposed in the study achieves privacy and security by visually blocking the home behavior video, and then determines the optimal combination of the measurement matrix and the number of sub-pixel encoding layers after dimensionality reduction, thereby balancing the contradiction between visual privacy and security and the accuracy of video human behavior recognition. The innovation points of the research mainly include the following three points. Firstly, it designs a visual privacy security encoding on the ground of CS theory to determine the contradiction between the visual privacy security and the accuracy of video human behavior recognition; Secondly, it introduces TAM into two-dimensional deep convolutional residual networks to improve the recognition accuracy and efficiency of home VSCS state videos; Finally, a behavior fusion recognition method on the ground of video frames and optical flow sequences is proposed. The research structure is mainly separated into four parts. The first part is an introduction to the research background; The second part is a review of relevant research results; The third part is to construct

the VPSM-HVBR system; The fourth part is the validation of the effectiveness and feasibility of the proposed research methods; The final part is a summary of the research.

### 1. Related work

With the increasing attention paid to the fields of intelligent security and video surveillance, privacy breaches have become a hot topic of concern while enriching people's lives. Therefore, the visual privacy security has received widespread attention from many scholars. Wang et al. analyzed the generation of adversarial samples under white box and black box attack protocols to address privacy, security, and other issues in the application of artificial intelligence in the visual field. This helps readers discover the essence of adversarial samples and improve the robustness, security, and interpretability of learning models [7]. Zhang et al. studied and proposed a new online social network spatiotemporal access control model for the lack of effective social spatiotemporal access control theoretical models as guidance, as well as privacy security issues in this scenario. Visual validation was conducted, and the results verified the security and effectiveness of the model [8]. Liu et al. proposed a new method on the ground of fragment encoding using the idea of visual cryptography to address the issues of privacy information leakage and tampering in practical e-commerce applications. The effectiveness of this method was verified through performance evaluation of computing and communication overhead, and its ability to resist collusion attacks is very strong [9]. Gangwar et al. aimed to understand users' views on space, environment, behavior, and privacy security in courtyard design, and the data in their research comes from various parameters in courtyard design. The results show that people prefer courtyards because they can provide visual contact during summer rush hours and ensure user privacy and security [10].

Video behavior recognition has good development prospects in fields such as intelligent video real-time monitoring, and has become a hot topic for researchers in related fields due to its involvement in numerous economic benefits and market development prospects. Liu et al. designed a home behavior video recognition algorithm that combines CS theory and TAM module 2D deep residual network to address the issues in home surveillance videos. The results showed that this method can effectively recognize behaviors in compressed videos [11]. Compared with the above methods, the proposed method extracts two types of features from VSCS state videos: RGB video frames and optical flow sequences, and innovatively adjusts the ResNet-50 model to improve the accuracy of end-to-end behavior recognition for RGB video frames. Finally, the two features are pre trained and tested again, and the final recognition accuracy is improved after weighted fusion. Muhammad et al. proposed an attention mechanism on the ground of bidirectional short-term

memory for recognizing human actions in videos. The results show that the recognition rates of this method on the UCD11, UCF Sports, and J-HMDB datasets are 98.3%, 99.1%, and 80.2%, respectively [12]. Miao et al. proposed a video human motion recognition method on the ground of extreme learning machines to solve the problem of low accuracy in traditional RGB video human motion recognition algorithms. The experimental results indicate that this method has achieved good results in practical applications [13]. Li et al. designed a new human skeleton action recognition algorithm on the ground of the convolutional network model of spatiotemporal main graph to address the issue that existing human skeleton based action recognition algorithms cannot fully explore the spatiotemporal characteristics of motion. The results indicate that this algorithm can better meet the practical application requirements of human motion recognition in videos [14].

On the ground of the above research results, it can be found that most of the achievements are related to the visual privacy security and video human motion recognition, lacking research on spatiotemporal information extraction and precision recognition for fuzzy videos. In addition, to prevent the privacy behavior of users in the home environment from being stolen, a VPSM-HVBR system is proposed.

### 2. Design of a home video behavior recognition system based on the visual privacy security mechanism and time series adaptive network

At present, in terms of visual privacy and security, the balance between video behavior recognition accuracy and the visual privacy security cannot be achieved through most existing recognition algorithms for home videos (HV). Therefore, the study first constructs a visual privacy security encoding based on CS theory, and further introduces the use of home visual privacy security sub-pixel encoding to achieve privacy security processing in the visual shielding domain of home user behavior videos. Then, the final combination of measurement matrix and dimensionality reduction layer is determined to improve the recognition speed and accuracy of the system. Secondly, the research introduced the preprocessing process of VSCS state videos, including the adjusted ResNet-50 model, TAM model, and home video behavior feature extraction method. Finally, research the behavior recognition of optical flow sequences and design a fusion recognition method for video frames and optical flow sequences.

#### 2.1. Design of the visual privacy security encoding based on compressed perception theory

Traditional technology on the ground of computer vision behavior recognition records generally clear surveillance image data, but cameras capture all information of users in their living environment, which contradicts the privacy and security needs of users in real life [15, 16]. In

response to the above issues, this study introduces image CS technology into the VPSM-HVBR system, which preprocesses the home video behavior information collected by the camera to achieve privacy and security effects of visual masking. The specific operation of CS theory is as follows: assuming that there is a series of image temporal signals  $u$  with time  $M$  and a standard orthogonal basis vector  $\alpha_i (i=1, 2, \dots, m)$  in the conversion domain; So  $u$  can perform the calculation of Equation (1) for  $(\theta_1, \theta_2, \dots, \theta_m)$  on this orthogonal basis.

$$u = \sum_{i=1}^M \alpha_i \theta_i \text{ or } u = \alpha \theta. \quad (1)$$

The matrix sizes of  $u$  and  $\theta$  in Equation (1) are both  $M \times 1$ ;  $\alpha$  represents a coefficient dictionary, corresponding to a scale of  $M \times M$ . Assuming that the initial signal  $u$  is sparse on the  $Q$  term, then  $u$  is a linear combination of  $Q$  basis vector mappings, where  $Q \ll M$ . If the coefficients in Equation (1) are mostly minimal, then  $u$  can be dimensionally reduced. Assuming the existence of a  $K \times M$  order measurement matrix  $\beta$ ,  $\beta$  follows the Finite Isometric Criterion (FIC), the measurement matrix  $V$  for  $u$  acting on  $\beta$  can be obtained. The expression is shown in Equation (2).

$$\begin{cases} V = A\theta \\ A = \beta\alpha \end{cases} \quad (2)$$

In Equation (2),  $A$  represents the sensing matrix, with a scale of  $K \times M$  order. Due to  $u$  being sparse and  $\beta$  following FIC, the study used  $K$  measurements to obtain  $Q$  sparse projection values, and then used the sparse coefficient  $\hat{\theta}$  obtained from the minimum l-norm problem to reconstruct  $u$  using Equation (3).

$$\hat{u} = \alpha \hat{\theta} \quad (3)$$

In Equation (3),  $\hat{u}$  represents the final reconstructed signal. If home video behavior data is treated as a set of time images, to protect the privacy part of it, the image needs to be considered as the processing target. Therefore, the key to home video processing is to mask the visual effect of the video frame images. For the transmission and recognition of video images on the network, this study utilizes Blocked Compression Perception (BCP) technology to perform block processing and operations on all initial images and  $\beta$ , to reduce the time complexity of operations and the amount of data processed. The operation of BCP is as follows: assuming  $u_i$  is the  $i$ -th image sub block,  $\beta_b \in R^{k \times m}$  can be used for measurement and calculation to obtain the value of  $k$  dimension and the measurement value  $v_i$  of the  $i$ -th block, as shown in Equation (4).

$$\begin{cases} v_i = A_b \theta_i \\ k = \frac{K}{M} \times m \end{cases} \quad (4)$$

For the overall image,  $\beta$  is a diagonal matrix. To improve the operational efficiency of the system, we have studied the use of a block based approach to data pro-

cessing. In response to the shortcomings of CS technology in complex backgrounds, a study was conducted on the use of blind compression to directly replace CS sampling with  $\beta$  for the input video frame images. Meanwhile, BCP was used to sequentially divide the input video frame and  $\beta$  into  $2 \times 2$  matrix blocks, and then the corresponding matrix blocks were convoluted to achieve CS dimensionality reduction. This process involves Single Layer Subpixel Encoding (SLSE). SLSE not only improves the accuracy of recognition, but also performs well in reconstructing video frames. However, the level of security protection for visual information is relatively low. Therefore, a multi-layer subpixel encoding on the ground of SLSE is proposed in this study. The schematic diagram of the three-layer subpixel encoding process is shown in Figure 1.

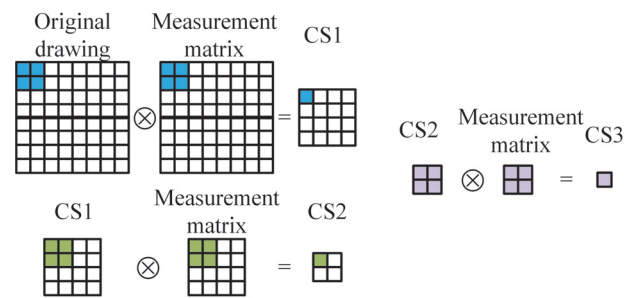


Fig. 1. Schematic diagram of three-layer subpixel encoding process

In CS,  $\beta$  plays an important role as it directly affects subsequent compression and signal processing tasks, so the selection of  $\beta$  is very important. At present, the mainstream measurement matrices include Gaussian random measurement matrix, Bernoulli random measurement matrix, Part Hadamard random measurement matrix, and Toeplitz matrix, which will be selected through facial recognition experiments in the future. The combination of the optimal number of dimensionality reduction layers and measurement matrix selected in the subsequent research is applied in end-to-end feature extraction and recognition. This process of privacy security for home video behavior is home privacy security VSCS encoding.

### 2.2. VSCS state video preprocessing for VPSM-HVBR system

After the above VSCS encoding processing, the video can be directly used as the input part of the VPSM-HVBR system, that is, the input VSCS state video. However, most video classification models currently only focus on the initial video, so research needs to first preprocess the home VSCS state video. The initial video is an unprocessed high-definition original video, while the VSCS state video is a video that has undergone VSCS encoding processing. Therefore, research needs to first preprocess home VSCS state videos. The preprocessing operations for home VSCS state videos are as follows: first, cut the video into 32 frames, and then input it into the TV-L1 network to obtain the optical flow

sequence of VSCS state videos; Then it uses the finely tuned ResNet-50 model to pre train the RGB video frames of the home VSCS state video, and combines it with the TAM model to form a TANet network to pre

train the optical flow sequence of the home VSCS state video. The ResNet-50 model has 50 convolutional layers and fully connected layers, and its basic structure is shown in Figure 2.

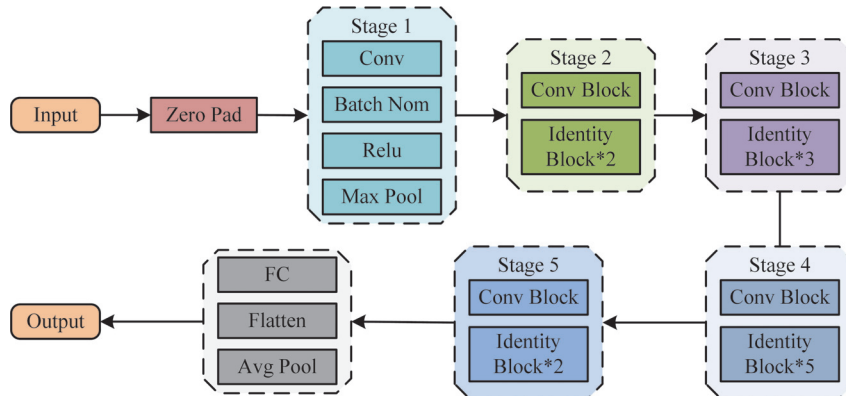


Fig. 2. Basic structure diagram of ResNet-50 model

The ResNet-50 model structure after fine-tuning in Figure 2 mainly consists of five stages. Stage 1 involves shallow feature operations on input image data, including convolution, normalization, activation function, and pooling. The sizes of convolution kernel, padding, and step size are  $7 \times 7$ , 3, and 2, respectively. The network structure of the remaining stages is the same, consisting of Conv Block and Identity Block, and is formed by adding the backbone network and the direct connection path. The latter has the same input and output dimensions as the backbone network, so it can be directly connected through the path. Before combining with the TAM module to form a TANet network, it is necessary to understand the TAM module. The temporal kernel parameters in TAM will be decomposed into position sensitive adaptive weights and irrelevant adaptive convolution kernels, dynamically learning temporal clues in a video adaptive manner. The TAM module consists of a local branch and a global branch. Compared with the global branch, the local branch uses short-term temporal information to generate importance weights related to position, which can observe the dynamic changes of temporal information in the short-term time dimension, accompanied by spatial noise information. Therefore, it is necessary to learn a position sensitive importance weight to capture the short-term temporal structure. The expression is shown in Equation (5).

$$I = \text{sigmoid} \left( \text{Conv1D} \left( \delta \left( \text{Conv1D} \left( \hat{U}, N, \frac{C}{\chi} \right) \right), 1, C \right) \right). \quad (5)$$

In Equation (5),  $I$  represents the importance graph of adaptive learning;  $C$  is the number of channels for the input tensor;  $C/\chi$  is the parameterization of the number of output channels;  $\delta$  represents the ReLU function;  $\text{Conv1D}(*, *, *)$  is temporal convolution;  $N$  represents the size of the convolutional kernel (CKE), set to 3, and  $\hat{U}$  represents the result of spatial compression of input  $U$  through average pooling. To match the size of  $U$ , the study

replicates  $\hat{I}$  to  $\hat{I} \in R^{C \times T \times H \times W}$  on the ground of spatial dimensions, where  $T, H, W$  represent the spatiotemporal dimension, and the calculation is shown in Equation (6).

$$\hat{I}_{c,t,i,j} = I_{c,t}. \quad (6)$$

In Equation (6),  $c, t, i, j$  are indices of channel, time, height, and width, respectively. The first Conv 1D in a local branch has a local field of view, corresponding to the same field size as the global branch adaptive convolution. The  $\hat{I} \in R^{C \times T}$  generated by local branches is a part of the temporal adaptive kernel parameter, which is sensitive to temporal position. Then perform timing enhancement operation through Equation (7).

$$LO = \hat{I} \odot u. \quad (7)$$

In Equation (7),  $LO$  represents the output feature map of the local branch;  $\odot$  represents multiplication by element. In the TAM model, the core is the global branch, which generates video related adaptive CKE on the ground of global temporal information. The main focus is on long-range temporal modeling and long-range dependencies in cargo video. To simplify the generation process of adaptive CKE, a channel by channel temporal CKE generation method is designed. At this time, only modeling temporal relationships are considered, and the generation process of CKE is shown in Equation (8).

$$\theta_c = \text{soft max} \left[ F \left( W_2, \delta \left( F \left( W_1, \hat{U}_c \right) \right) \right) \right]. \quad (8)$$

In Equation (8),  $\theta_c$  is the adaptive CKE of the  $C$ -th channel,  $F$  represents the fully connected layer, while  $W_1$  and  $W_2$  represent the convolutional kernels of the first and second layers, respectively. In the TAM model, both local and global branches use a double-layer design, which gives the model better fitting ability. Then there is temporal adaptive fitting, and the generated CKE learns the temporal structure information between

video frames through convolution. The calculation is shown in Equation (9).

$$V_{c,t,i,j} = \sum_n^n \theta_{c,n} \bullet LO_{c,t+n,i,j} \quad (9)$$

In Equation (9),  $\bullet$  represents scalar multiplication, and  $V \in R^{C \times T \times H \times W}$  represents the feature map obtained through temporal adaptive convolution. The feature map obtained from the optical flow sequence of the TAM model is shown in Figure 3.

In Figure 3, different CKE are generated at different times, which increases the computational complexity and optimization difficulty of the system. To solve the above problem, local branches are introduced. When performing adaptive convolution for global branches, the importance of corresponding time features can be evaluated through temporal position sensitive  $I$ , enabling the TAM model to better adapt to changes [17].

### 2.3. Construction of VPSM-HVBR system based on behavior fusion recognition

The research will input the optical flow sequence extracted from home VSCS state video into the network structure of optical flow sequence behavior recognition, as shown in Figure 4. The fine-tuned TANet model used in the study is set as follows: the convolution kernel size of the first layer of convolution is  $3 \times 3$ , and both the stripe and padding are set to 1.

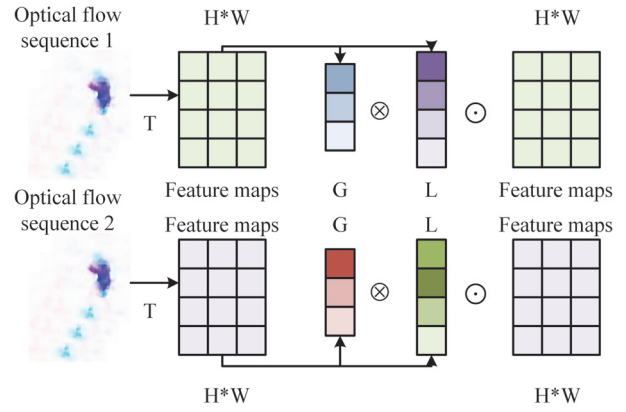


Fig. 3 Feature maps obtained from optical flow sequences using the TAM model

In Figure 4, the optical flow sequence is input into the TANet model for pre training and testing, and the feature vector is output in the last block of the fourth TA Block. Then, it is fully connected and processed with a softmax classifier to obtain the optical flow sequence recognition results for home VSCS state videos. And  $\oplus$  and  $\otimes$  are addition and convolution operators for elements, respectively. It sets  $U \in C \times T \times H \times W$  as the feature map of the video clip. In TAM, only the modeling in the time domain is focused, and the situation where the feature map is compressed by average pooling in the global space can be obtained as shown in Equation (10).

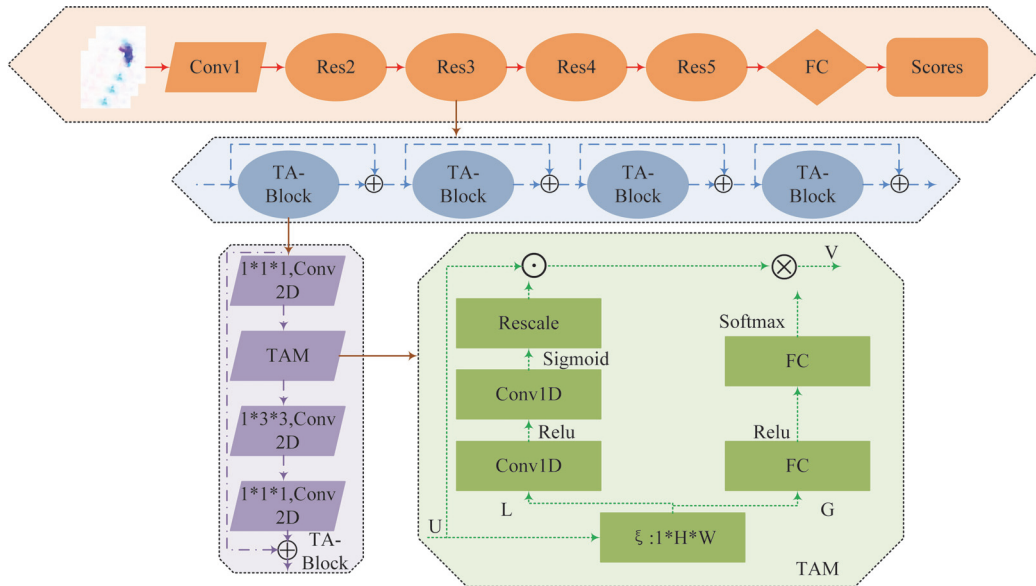


Fig. 4. Schematic diagram of network structure for optical flow sequence behavior recognition

$$\hat{U}_{c,t} = \zeta(U)_{c,t} = \frac{1}{H \times W} \sum_{i,j} U_{c,t,i,j} \quad (10)$$

In Equation (10),  $\zeta$  is the function that aggregates spatial information. In addition, the TAM model only focuses on the temporal variation patterns of learning video sequences, so it is necessary to first compress the spatial dimension of  $U$  using average pooling to obtain  $\hat{U}$ . The

TAM model consists of global branches and local branches, and the expression is shown in Equation (11).

$$V = G(\hat{U}) \otimes (L(\hat{U}) \odot U) \quad (11)$$

In Equation (11),  $G$  and  $L$  represent global and local branches, respectively. In addition, the study intercepted the video frames of home VSCS state videos and inputted

them into the network structure of video frame behavior recognition, as shown in Figure 5.

The video frames in Figure 5 are input into the ResNet model for training and testing, while the remaining operations are consistent with the behavior recognition of the optical flow sequence. The overall training process can be obtained as follows: first, the video frames extracted from the VSCS state home behavior video are input into the adjusted ResNet-50 model for training and testing. In the 50th layer, the output feature vectors are fully connected and processed with a softmax classifier to obtain the corresponding video frame recognition accuracy for the VSCS state home behavior video. Secondly, the extracted optical flow sequence is input into the pre trained TANet model. By retraining and testing the video frames in subsequent mainstream behavior recognition databases, the recognition accuracy of the dual stream

network can be obtained, and finally, weighted fusion is performed to obtain the final recognition accuracy. On the ground of the above content, the network structure diagram of the VPSM-HVBR system can be obtained, as shown in Figure 6.

**3. Analysis of the results of a home video behavior recognition system on the ground of visual privacy security mechanism and temporal adaptive network**

To verify the effectiveness and feasibility of the VPSM-HVBR system, the study first conducted facial recognition experiments to select a measurement matrix for the visual privacy security encoding, and then evaluated the effectiveness of the processing method for cropping video frame segments. Finally, it tests the performance of the proposed method in different categories.

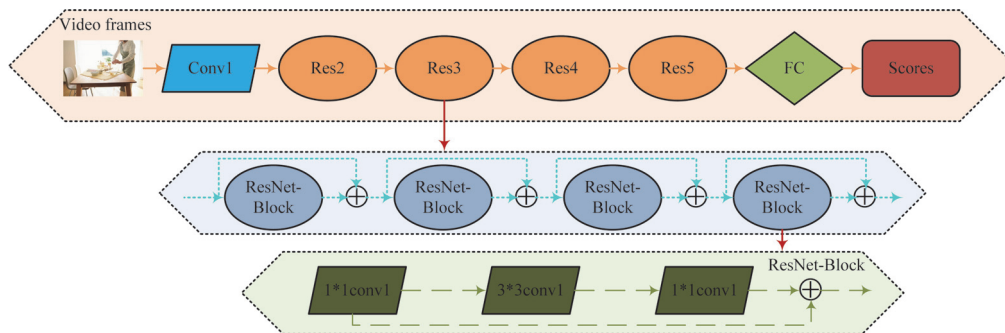


Fig. 5. Schematic diagram of network structure for video frame behavior recognition

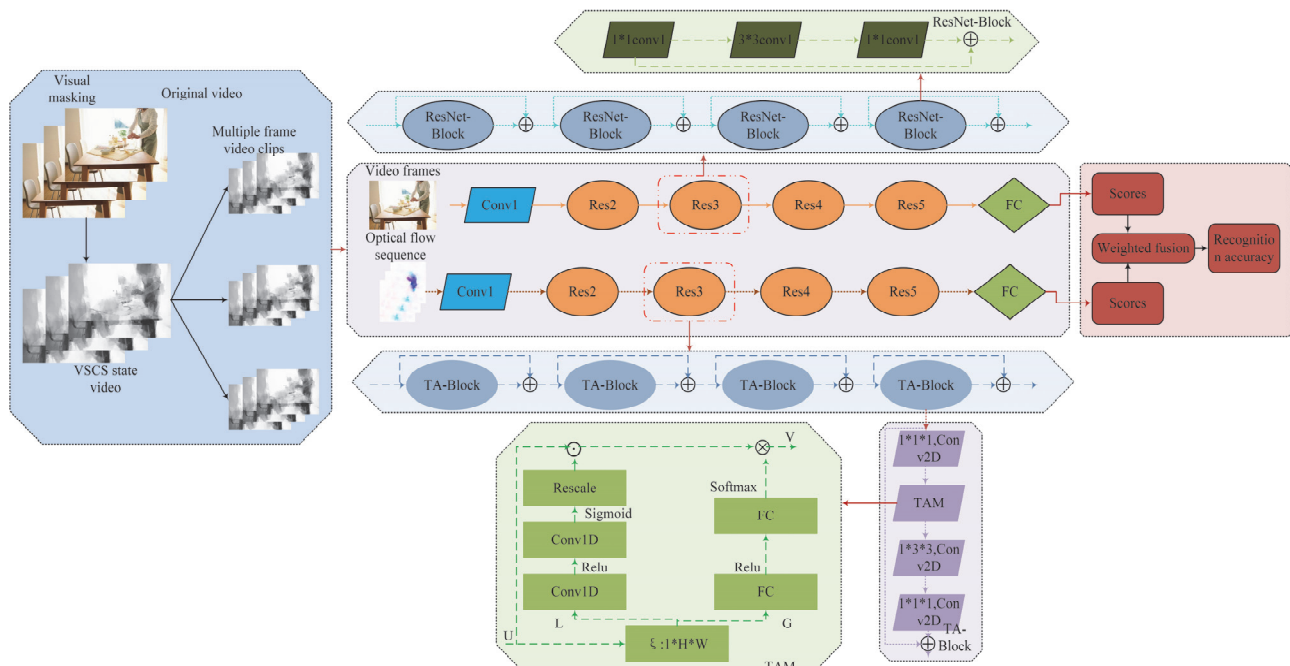


Fig. 6. Network structure diagram of VPSM-HVBR system

**3.1. Selection of measurement matrix for the visual privacy security coding based on CS theory**

The experimental platform selected 1 TB hard disk, 32 GB memory, and a 32 core Xeon processor, and the model training and testing were completed on software

PyTorch. The dataset was selected from the current mainstream public video recognition databases for experiments, namely the You Tube database, UCF Sports database, and Hollywood2 database. The YouTube database is derived from real-life video data, with a complex background that includes 11 types of actions and 12360 video

segments. The UCF Sports database has a total of 182 video sequences and 9 types of actions. The Hollywood2 database is sourced from movie videos, containing 12 types of actions and approximately 1800 video segments. 70% of the data in the database is trained, while the remaining data is used for testing. To select the optimal measurement matrix that meets the VPSM-HVBR system, this study conducted experiments using an ORL face database, with each image size of  $128 \times 128$ . Some faces in the database were lowered from the first layer to the fourth layer, corresponding to CS1, CS2, CS3, and CS4, with corresponding sizes of  $64 \times 64$ ,  $32 \times 32$ ,  $16 \times 16$ , and  $8 \times 8$  for each layer.

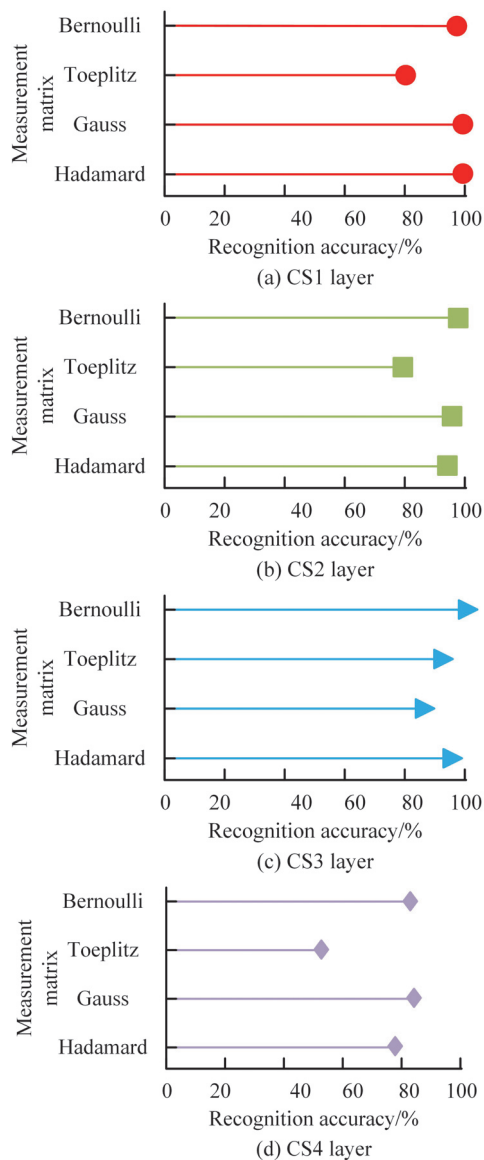


Fig. 7. Recognition rate results at different levels under different measurement matrices

The recognition accuracy results at different levels under different measurement matrices are shown in Figure 7. Figure 7 shows that the Bernoulli random matrix has the highest recognition accuracy among different levels, with recognition accuracy rates of 100%, 98.73%,

98.76%, and 85.62% for CS1-CS4 layers, respectively; The Toeplitz matrix has the lowest recognition accuracy at each layer, while the Gauss random matrix and Hadamard random matrix have little difference in recognition accuracy at each layer; Therefore, this study chose the Bernoulli random matrix as the optimal measurement matrix for the VPSM-HVBR system. In addition, considering the recognition accuracy and efficiency, as well as the visual shielding information that meets user privacy security, the dimensionality reduction effect of the second and third layers is similar, but the data volume of the third layer dimensionality reduction is smaller; Therefore, this study selected CS3 layers as the optimal number of dimensionality reduction layers and combined them with Bernoulli random matrices.

### 3.2. Result analysis of VPSM-HVBR system

To investigate the effectiveness of different processing methods for cropping video clips, validation experiments were conducted on the YouTube database, UCF Sports database, and Hollywood2 database. The corresponding FPS for the experimental videos was 30. According to previous research, it has been found that when each video clip in the input network is edited into 16 non overlapping frames, some relatively complex or similar 16 frame behavior small segments will ignore most of the corresponding behavior information, making it difficult to accurately reflect the behavior actions, resulting in misjudgment. At the same time, the system training data proposed in the study are all larger than 32 frames. Therefore, the study improved the frame rate of small segments of the input image and cut the video into 32 frames for comparative experiments. The study cropped all input home behavior videos into small clips of 16 frames and 32 frames, using twenty small batches of clips. The parameter settings are as follows: the initial learning rate is 0.0001. After 6 iterations, the learning rate will decrease to the original 0.9, and fine-tuning will terminate after 34 iterations.

The comparison results of recognition accuracy between 16 frames and 32 frames on different databases are shown in Figure 8. Figure 8 illustrates that the recognition rate of employing 32 short video clips is greater than that of 16 frames on the YouTube, UCF Sports, and Hollywood2 databases (4.67%, 4.12%, and 4.35%, respectively). Accordingly, research indicates that the optimal outcome is to chop the incoming video into 32 frames. The study sets all video domains and optical flow sequences of home VSCS state movies with privacy protection to  $112 \times 112$  for ease of later feature extraction and recognition. Due to the lack of a comprehensive database on privacy behavior, to facilitate more effective analysis in research, this study selectively established a dataset of privacy behavior in a home environment on the ground of three types of databases. To ensure the authenticity of the experiment, the study invited 20 experts and 60 students from related majors to rate the behavior in the database.

The score is 1–10, and the higher the score, the greater the likelihood that the behavior requires privacy and security measures. The study set a total of 2 rounds of scoring, one on the ground of semantics, and then on the ground of this, a second round of scoring is conducted for the content of the video. The results of two rounds of scoring on three databases are shown in Table 1.

According to Table 1, finally, ShavingBear (SB), ApplyEyeMakeup (AEM), ApplyLipstick (AL), and JumpingJack (JJ) were selected from the YouTube database. It selects Hug (HG), Drink (DK), Kiss (KS), and Situp (SU) in the UCF Sports database. It selects KS, Hug person (HP), Eat (ET), and AnswerPhone (AP) from the Hollywood2 database, all of which require privacy and security processing. For options that do not require secure privacy processing, four categories were randomly selected from three databases. PlayingGuitar (PG), HandstandPushups (HP), BoxingSpeedBag (BSB), and PlayingCello (PC)

were selected from the YouTube database. It selects brush hair (BH), check (CW), fly flac (FF), and pullup (PP) in the UCF Sports database. It selects GetOutCar (GOC), HandShake (HS), Sitdon (SN), and StandUp (SDU) in the Hollywood2 database. 200 movies from each category were chosen for the study's following trials, for a total of 1600 videos. Of these, 960 were utilized for training and the remaining 200 for testing. In the experiment, an 8-fold cross test approach with 1000 iterations was chosen to increase the generalization level of the model for small sample data. In addition, the Momentum optimizer used in the study serves as a tool for parameter updates, and the loss function is partially implemented through small operators. To verify the performance of home VSCS state videos under TAM, a visual presentation of video adaptive CKE generated by global branches was studied. And it was tested on AEM in the YouTube database, and an I3D model was selected for comparison.

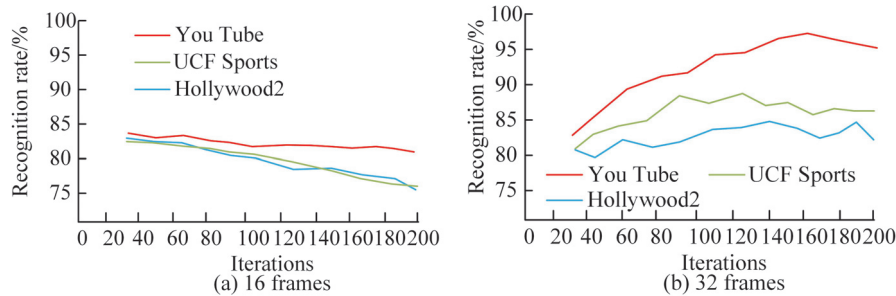


Fig. 8. Curve of recognition accuracy changes between 16 frames and 32 frames on different databases

Tab. 1. Two rounds of scoring results on three databases

Scoring content	Database	Number of types	Require privacy and security processing/Average score	No need for privacy and security processing
Score based on semantics	You Tube	103	25	78
	UCF Sports	53	16	37
	Hollywood2	14	6	8
Rate based on content	You Tube	26	TaiChi/8.8, HulaHoop/8.9, Skilling/9.0, JumpingJack/9.5, ShavingBeard/9.7, ApplyEyeMakeup/9.7, Archery/9.1, ApplyLipstick/9.6, BlowDryHair/9.4, BlowingCandles/9.4	15
	UCF Sports	17	Kiss/9.8, Hug/9.4, Eat/9.1, Jump/8.9, Drink/9.4, Situp/9.2	10
	Hollywood2	7	Kiss/9.7, Hug person/9.5, Eat/9.3, AnswerPhone/9.5	2

The visualization results of video convolution kernels generated by different models are shown in Figure 9. As seen in Figure 9, the distribution shapes and scales produced by the TANet model's global branch CKE are more uniform and diversified than those produced by the I3D model. In conclusion, the TANet model's adaptive CKE can enhance both computational effectiveness and recognition performance. Consequently, while modeling video sequences, it is appropriate and required to apply video adaptive modeling techniques.

The comparison results of time complexity between the initial state and VSCS state on different datasets are

shown in Figure 10. Figure 10 shows that the red VSCS data is significantly lower than the blue initial state data. Specifically, compared to the initial data, the VSCS data size has decreased by 37%, but the time complexity has decreased by only 94.63% on different datasets. To further verify the comparative experiment of the recognition performance of VPSM-HVBR system and existing systems for privacy and security processing of videos on initial images, namely the human behavior recognition system based on video surveillance, the artificial intelligence video behavior recognition system, and the intelligent video behavior recognition system, where RP represents

the need for privacy and security processing, and NRP represents the need for privacy and security processing [18–20].

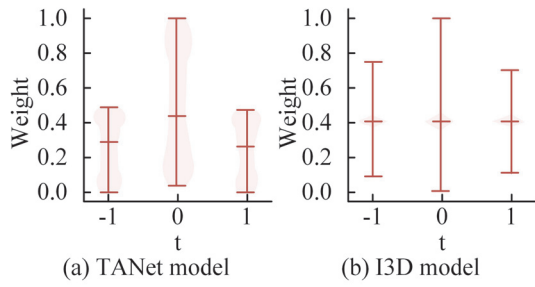


Fig. 9. Visualization results of video convolution kernels generated by different models

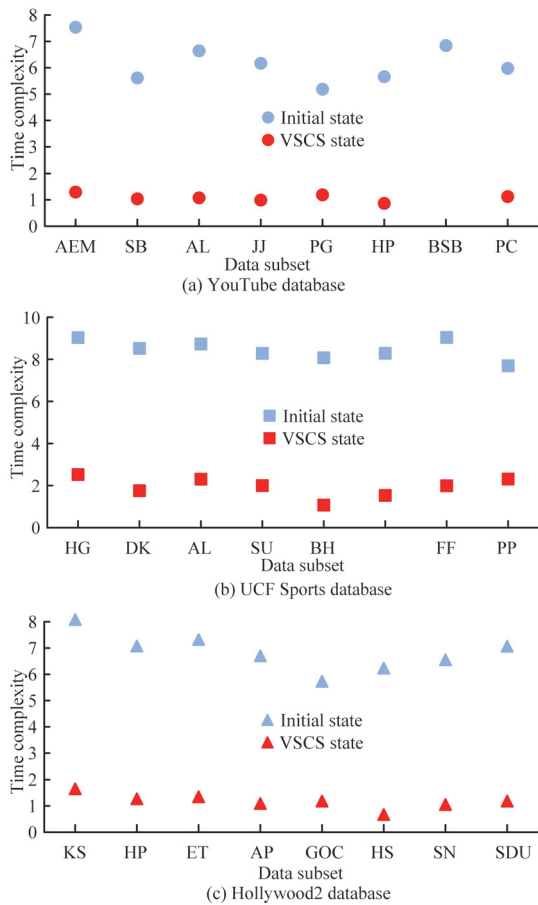


Fig. 10. Comparison of time complexity between initial state and VSCS state on different datasets

The performance of different systems in identifying behaviors that require privacy and security processing on different datasets is shown in Figure 11. The video recognition performance of various systems for RP is not very varied, as Figure 11 illustrates, making it better than most other approaches. The VPSM-HVBR system's average recognition accuracy is 94.6%, 73.5%, and 77.1%, respectively, in the performances of the YouTube, UCF Sports, and Hollywood2 databases. With a 35% reduction in overall data volume, the VPSM-HVBR system outperforms previous approaches for NRP films.

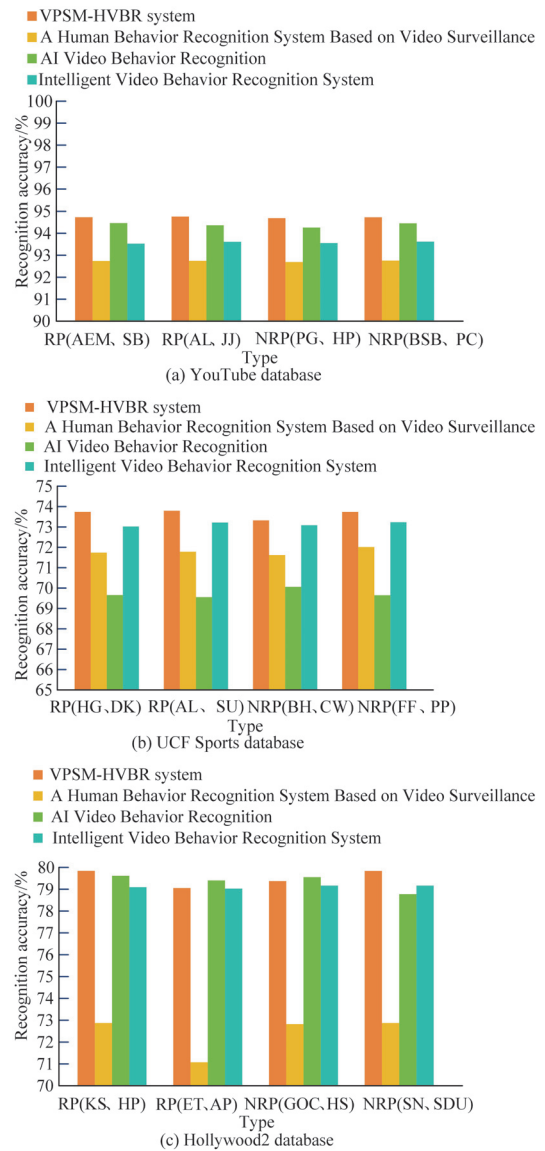


Fig. 11. Performance of behavior recognition requiring privacy and security processing on different datasets by different systems

### Conclusion

Recently, video surveillance technology has been widely applied in various fields, posing a great threat to user privacy and security. In response to the above issues, the study introduces CS theory and TAM, first proposes a visual privacy security encoding scheme, and then designs a VPSM-HVBR system to solve the privacy leakage problem of users in home surveillance videos. The experimental results show that the Bernoulli random matrix has the highest recognition accuracy among different levels, with recognition accuracy rates of 100%, 98.73%, 98.76%, and 85.62% for CS1-CS4 layers, respectively. In the recognition performance results of the YouTube database, UCF Sports database, and Hollywood2 database, compared with other systems, the average recognition accuracy of the VPSM-HVBR system is the highest in most cases, with 94.6%, 73.5%, and 77.1%, respec-

tively; For NRP videos, the VPSM-HVBR system performs better with a 35% reduction in total data volume. In summary, the method proposed in the study is reasonable and can achieve good results in different datasets, demonstrating good practicality. However, there are still shortcomings in research, and there is currently a lack of databases related to home video behavior recognition. In future research, it is necessary to build a scientifically realistic database of home video behavior to achieve more objective evaluation.

### Acknowledgements

The research was supported by Design and Implementation of RFID-based Experimental Equipment Management Platform (No. 2022QJY001Z), “Young Teachers Development Fund” of Dongguan City University.

### References

- [1] Yang Y, Song X. Research on face intelligent perception technology integrating deep learning under different illumination intensities. *Journal of Computational and Cognitive Engineering* 2022; 1(1): 32-36. DOI: 10.47852/bonviewJCCE19919.
- [2] Lei Y. Research on microvideo character perception and recognition based on target detection technology. *Journal of Computational and Cognitive Engineering* 2022, 1(2): 83-87. DOI: 10.47852/bonviewJCCE19522514.
- [3] Dey P, Jana DK. Evaluation of the convincing ability through presentation skills of pre-service management wizards using AI via T2 linguistic fuzzy logic. *Journal of Computational and Cognitive Engineering* 2022; 2(2): 133-142. DOI: 10.47852/bonviewJCCE2202158.
- [4] Hui Z, Li J, Gao X, et al. Progressive perception-oriented network for single image super-resolution. *Inf Sci* 2021; 546(2): 769-786. DOI: 10.1016/j.ins.2020.08.114.
- [5] Fante KA, Abdurahman F, Gameda MT. An ingenious application-specific quality assessment methods for compressed wireless capsule endoscopy images. *Trans Environ Electr Eng* 2020; 4(1): 18-24. DOI: 10.22149/tee.v4i1.139.
- [6] Chi Z. Research on satellite remote sensing image fusion algorithm based on compression perception theory. *J Comput Methods Sci Eng* 2021; 21(2): 341-356. DOI: 10.3233/JCM-204411.
- [7] Wang W, Dong J, Ziwen HE, et al. A brief introduction to visual adversarial samples. *J Cyber Secur* 2020; 5(2): 39-48. DOI: 10.19363/J.cnki.cn10-1380/tn.2020.02.04.
- [8] Zhang L, Zhang Z, Zhao T. A novel spatio-temporal access control model for online social networks and visual verification. *Int J Cloud Appl Comput* 2021; 11(2): 17-31. DOI: 10.4018/IJCAC.2021040102.
- [9] Liu R, Song J, Huang Z, et al. EQRC: A secure QR code-based E-coupon framework supporting online and offline transactions. *J Comput Secur* 2020; 28(5): 577-605. DOI: 10.3233/JCS-191416.
- [10] Gangwar G, Kaur P, Taylor G. User's perception of the relevance of courtyard designs in a modern context: A case of Traditional Pol Houses, Ahmedabad. *Civ Eng Archit* 2020; 8(3): 379-389. DOI: 10.13189/cea.2020.080323.
- [11] Liu J X, Zhang M, Sun N, et al. Home video behavior recognition based on convolutional residual network with visual privacy protection mechanism and temporal adaptive module. *2021 6th Int Conf on Communication, Image and Signal Processing (CCISP) 2021*: 150-154. DOI: 10.1109/CCISP52774.2021.9639335.
- [12] Muhammad K, Mustaqeem, Ullah A, et al. Human action recognition using attention based LSTM network with dilated CNN features. *Future Gener Comput Syst* 2021; 125: 820-830. DOI: 10.1016/j.future.2021.06.045.
- [13] Miao A, Liu F. Application of human motion recognition technology in extreme learning machine. *Int J Adv Robot Syst* 2021; 18(1): 4-18. DOI: 10.1177/1729881420983219.
- [14] Li YZ, Yuan JZ, Liu HZ. Human skeleton-based action recognition algorithm based on spatiotemporal attention graph convolutional network model. *J Comput Appl* 2021; 41(7): 1915-1921. DOI: 10.11772/j.issn.1001-9081.2020091515.
- [15] Qi W, Wang X, Liu Z, et al. Visual recognition of ortho-xylene based on its host-guest crystalline self-assembly with  $\alpha$ -cyclodextrin. *J Colloid Interface Sci* 2021; 597(21): 325-333. DOI: 10.1016/j.jcis.2021.03.024.
- [16] Moskovsky AD, Burgov EV, Ovsyannikova EE. An approach to scene recognition based on the local interaction of a group of robots. *Mekhatronika Avtomatizatsiya Upravlenie* 2021; 22(2): 94-103. DOI: 10.17587/mau.22.94-103.
- [17] Nougrara Z, Hachemi K. A new method of drainage network extraction using SAR radar images: A case study of Djanet (Algeria). *J Taibah Univ Sci* 2021; 15(1): 1101-1107. DOI: 10.1080/16583655.2021.2018816.
- [18] Pu Y, Jiang Y, Hu HM. HR-HAR: A hierarchical relation representation for human activity recognition based on Wi-Fi. *IET Commun* 2023; 17(1): 29-44. DOI: 10.1049/cmu2.12497.
- [19] Lyu Y, Yang Z, Liang H, et al. Artificial intelligence-assisted fatigue fracture recognition based on morphing and fully convolutional networks. *Fatigue & Fracture of Engineering Materials and Structures* 2022; 45(6): 1690-1702. DOI: 10.1111/ffe.13693.
- [20] Herman M, Wagner J, Prabhakaran V, et al. Pedestrian behavior prediction for automated driving: requirements, metrics, and relevant features. *IEEE Trans Intell Transp Syst* 2022; 23(9): 14922-14937. DOI: 10.1109/TITS.2021.3135136.

### Authors' information

**Dongmin Zhao** (b. 1991) graduated from Central South University in 2018 with a major in Software Engineering. Currently, he serves as the Deputy Director of the Experimental Teaching Service Center of Dongguan City College Experimental Center. Research interests: computer software and computer applications, computer hardware technology, and higher education. E-mail: [zdm20230801@126.com](mailto:zdm20230801@126.com)

*Received November 08, 2023. The final version – June 14, 2024.*