A NEW INTELLIGENT SUPERMARKET SECURITY SYSTEM

Y. Zhang, S. Jin*, Y. Wu*, T. Zhao*, Y. Yan*, Z. Li*, Y. Li**

Abstract: With the rapid development of artificial intelligence in recent years, the application of intelligent security has become increasingly widespread. This paper presents a new intelligent system that uses Convolutional Neural Network (CNN) combined with a high-resolution camera to identify the theft behavior of customers. The CNN extracts relevant information from the theft and non-theft behavior of customers in supermarkets to establish a recognition model. Our results show that, by updating the data sets, the recognition model can be continuously optimized, and the average recognition accuracy finally reaches 83 %. The proposed system can independently identify the theft and non-theft behavior in video surveillance and sound alarm on the theft behavior in time. The advantages of the system are its low cost and high precision, which show excellent commercial value and application prospects.

Key words: *artificial intelligence, convolutional neural network, information extraction, sound alarm*

Received: April 8, 2019 **DOI:** 10.14311/NNW.2020.30.009
Revised and accepted: April 30, 2020

ACM CCS:
(1) Computing methodologies—Modeling and simulation—Model development and analysis—Model verification and validation
(2) Computing methodologies—Artificial intelligence—Computer vision—Image and video acquisition
(3) Computing methodologies—Machine learning—Machine learning approach
(4) Computer systems organization—Real-time systems—Real-time operating systems
(5) Mathematics of computing—Probability and statistics

1. Introduction

With a continued prosperity of the national economy and an increasing demand for household goods, the supermarket has become a core place for residents to purchase daily necessities. A series of surveys show that the primary business model of modern supermarkets is independent shopping [1]. This shopping mode inevitably

*Yiyi Zhang; Shangzhong Jin – Corresponding author; Yufeng Wu; Tianqi Zhao; Yongqiang Yan; Zenan Li; Yalan Li; College of Optical and Electronic Technology, China Jiliang University, Hangzhou, Zhejiang 310018, China, E-mail: zhangyiyi01130707@163.com, jinsz@cjlu.edu.cn, 1793519712@qq.com, 18a0402151@cjlu.edu.cn, 845113677@qq.com, lizenan9004@126.com, 173549662@qq.com

brings some management problems, the most serious of which is the occurrence of the theft in supermarkets. The Global Retail Theft Barometer released a survey of 222 retailers in 24 countries. The survey showed that global retailers lost $ 128 billion in 2013, of which US retailers accounted for $ 42 billion. The most significant cause of the losses comes from the theft [2,3]. Therefore, solving the theft problem becomes one of the essential things during the supermarket operation. At present, most supermarkets adopt the traditional surveillance mode, which monitors the video surveillance terminal through the human eye. It has many limitations, such as extended response time and a high error rate. Some supermarkets have an acoustic and magnetic detection system at entrance and exit [4], but none of them can effectively prevent the theft.

The development of intelligent security technology provides a new way to solve this problem [5]. For the intelligent security system, the camera is equivalent to the human eye, and the computer is equivalent to the human brain. It has not been widely used in most supermarkets because of its high cost and imperfect technological development. The system has only been tested prospectively in a small number of supermarkets. The Amazon Go convenience store opened in Seattle uses a variety of techniques, such as vision sensors, deep learning algorithms, and image analysis [6]. By sensing the relative position between people and shelves and the movement of goods on the shelves, the system can calculate who takes it, but the system can only cope with cases less than 20 people at present. In China, unmanned supermarkets have appeared in many cities, such as Yonghui supermarket, Jingdong supermarket, Alibaba's supermarket, and Amoy coffee [7, 8]. As far as Alibaba's supermarket is concerned, it adopts the latest security equipment of supermarket, Radio Frequency Identification (RFID). Then it uses intelligent surveillance technology to identify the goods and people's faces and actions to achieve intelligent security. However, these RFID tags also have some problems. For example, RFID tags cannot identify the products of unique materials such as glass. Besides, if the tag is held tightly, it will not be recognized by the machine [9, 10]. In summary, the current supermarket intelligent security system is still in the primary stage, its function is not perfect, and the cost is high. There is a long way to go to popularize intelligent security in future supermarkets.

This paper proposes a new intelligent security system in supermarkets that monitors the behavior of customers in time through the camera and then transmits the collected data to the personal computer (PC). The model that has been trained by the Convolutional Neural Network (CNN) in the personal computer (PC) identifies whether the customers steal the goods and if so, the video surveillance terminal gives an alarm. The CNN model can be continuously optimized by adding the corresponding data sets. After improvement, it presents a more accurate identification result of the theft. The system has many advantages, such as fast response, low cost, and high recognition accuracy, which significantly reduce the input of the workforce. The most important advantage of the system is that it holds the potential for further optimization. By using the intelligent security system, the supermarket clerks are delivered from the tedious and dull labor, which effectively reduce the economic loss and the operating cost of supermarkets [11]. In summary, the system has high commercial value and application prospects.

The remainder of this paper is as follows: Section 2 outlines the process of our system and introduces the detailed hardware about the system. Section 3 discusses the particular structure of CNN. Section 4 introduces how to collect the data sets for the training of the CNN model and proposes the model's optimization idea based on the personalities of the theft behavior. Section 5 presents the results of the CNN model on theft detection during the optimization process. Finally, conclusions and future works are discussed in Section 6.

2. Related work

The new intelligent security system, as mentioned in the paper, is shown in Fig. 1. The camera (OV2710, 1920×1080@25fps, USB720P Camera, China) monitors the behavior of customers in the supermarket in real-time, then it outputs video stream data and transmits the data to the PC through the network cable. The PC (Windows 10, i7-7700 HQ CPU @2.80 GHz, Lenovo, China) uses the Tensorflow (Tensorflow 1.0, Google, USA) framework in software PyCharm (Python 3.6, Guido van Rossum, Netherlands) which is simple and highly compatibility compared to other programming languages to preprocess the image. Then it recognizes and processes the image through the CNN model that has been trained. The model identifies whether the customers have theft, and if so, it outputs one, and the video surveillance terminal gives an alarm. Besides, the identification of theft is judged by the quality of pictures, so the pictures which are taken by cameras with different resolutions will affect the accuracy of the model's theft detection.

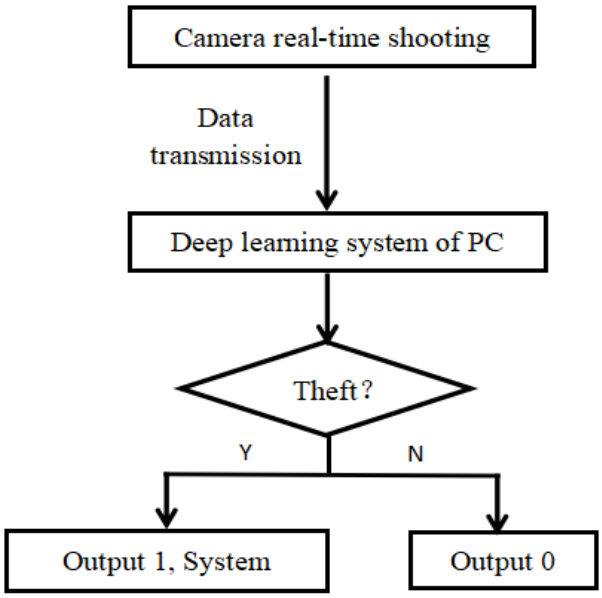

Fig. 1 *Flow chart of the system.*

3. Overview of the system

3.1 Conventional Convolutional Neural Network

The CNN is the preferred algorithm for current computer vision recognition. In the classic algorithms, the input is the rules and the batch data. The system outputs the answers according to the input rules and data. In the neural network, the input is the known data and the answers corresponding to the data. After training, the system outputs the rules based on the relationship between the data and the answers. These rules can be applied to new data to help the computer to generate answers autonomously [12]. For the neural network, the desired properties include more substantial computational power to enhance computational efficiency, more data to ensure perfect feature extraction, and better training methods to optimize the weights to make the prediction performance of the entire network more excellent. Among various neural networks, the CNN stands out because of its three unique advantages, local perception, weight-sharing network structure, and multiple cores. The local perception means the size of the convolution core is generally smaller than the size of the input image so that the feature extraction will pay more attention to the local area. Each neuron does only need to perceive the local area, such as a corner or a line in the picture, which is the basis of the animal's visual composition [13]. Then it integrates the local information at a higher level to obtain global information. The weight-sharing network structure is more similar to a biological neural network, which can reduce the complexity of the network model and correspondingly reduce the number of calculations. Because the parameters of a single core are determined, and the feature extracted from this core is single. The multiple cores can filter the input image by using multiple convolution cores and analyze the input image at multiple angles.

Based on the advantages of CNN, we compare the effects of multiple algorithms on the gender recognition of a human face. The accuracy of the CNN based on American Telephone & Telegraph (AT&T) face database for gender recognition has reached 99.38 % [14], which is much higher than the result of 93.00 % gender recognition by the combined algorithm of Principal Components Analysis (PCA) and Extreme Learning Machine [15]. The CNN's gender recognition is also over 98.00 % gender recognition through the combined algorithm of Continuous Wavelet Transform and Support Vector Machine (SVM) [16] and even exceeds the 98.93 % gender recognition through the combined algorithm of Discrete Cosine Transform (DCT) and Support Vector Machine (SVM) [17]. These results show the advantages of the CNN in-person feature recognition. What's more, the CNN has excellence in training large data sets and has gradually replaced Support Vector Machine (SVM) and Decision Trees [18]. The CNN is structurally stable, using forward-propagation to calculate the output value and adjusting the weight and offset value by back-propagation [19]. It is highly invariant to translation, scale, tilt, and other forms of deformation. The neural units between adjacent layers of the CNN are partially connected rather than fully connected, like the Back Propagation (BP) neural network [20]. Compared with standard forward neural networks with the same level, CNN has fewer neural connection units and fewer parameters, so the training is more straightforward and faster [21].

In summary, CNN is very suitable for intelligent security systems in supermarkets. However, the CNN still has some shortcomings, the selection of the initial parameters may substantially affect the network training, and the poor selection causes the network not to work or leads to the network training to fall into underfitting and over-fitting. Therefore, the selection of the initial parameters is particularly important, and multiple parameter adjustments are required to generate an ideal prediction model. Fig. 2 shows the typical LeNet-5 structure of the CNN [22]. The convolutional layers are labeled Cx, subsampling layers are labeled Sx, and fully-connected layers are labeled Fx, where x is the layer index. The INPUT is a picture, and the PC converts the picture to the matrix of the corresponding size. If the picture is a color image, it is equivalent to inputting the three matrices corresponding to the distribution of R, G, and B in the PC. The input picture is an $n \times n \times 3$ tensor, and $n \times n$ represents the image's dimensions, so the dimension of the convolution core defined in the corresponding convolutional layer must also be a tensor equal to 3. The convolution formula is as follows:

$$s(i,j) = (X \times W)(i,j) + b = \sum_{k=1}^{n\_in} X_k \times W_k(i,j) + b, \tag{1}$$

where n_in is the number of input matrices, X_k is the k-th input matrix, W_k is the kth sub convolution core matrix of the convolution core, and b is the paranoid value. Besides, $s(i,j)$ is the element value of the output matrix W corresponding to the convolution core at the corresponding position. The convolutional layer collects the critical data content from the input data, and the convolved output value is assigned by the Rectified Linear Unit (RELU) function to solve the vanishing gradient problem. The subsampling layer compresses each 2×2 element in the sub-matrix by one element to achieve the dimensionality reduction processing of the matrix and extract more essential features. Besides, it effectively controls the over-fitting. After multiple convolution operations, the fully-connected layer integrates highly abstracted features and normalizes the data to output a probability to each classification. The classification layer classifies the behavior according to the probability of full connection to achieve the final output [23].

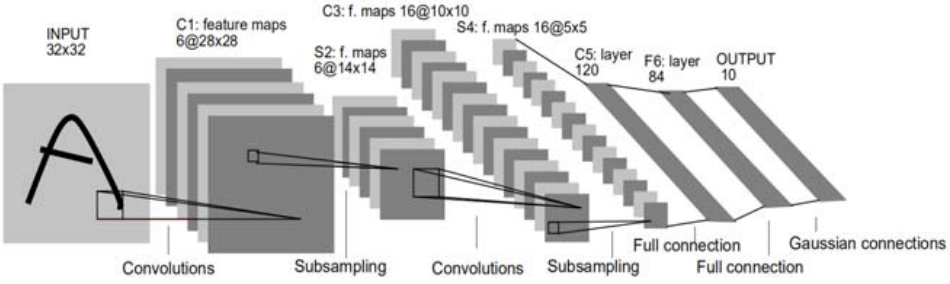

Fig. 2 *The LeNet-5 structure.*

The LeNet-5 structure has seven layers of neural networks, including two convolution layers, two pooling layers, and three fully-connected layers. These layers ensure the feature extraction and the pattern classification, which are performed

and generated at the same time in training. The LeNet-5 filters the input features by adding pooling layers and extracts image features through the convolution layer. In the LeNet-5 structure, Sigmoid is used as an activation function to introduce non-linear factors that can improve the neural network's ability for outputting predictive models and solve problems that cannot be solved by linear models. The LeNet-5 structure also reduces computational complexity through sparse connections between layers, and it is suitable for character recognition [24] and human motion recognition [25].

3.2 Improved Convolutional Neural Network of the system

The new CNN structure proposed is optimized based on the LeNet-5, which uses ReLU as activation function instead of the Sigmoid function to solve the vanishing gradient problem when optimizing the neural networks. Also, it speeds up the training of the model. The new CNN structure increases the number of combined layers of the convolutional layer and the subsampling layer to extract more striking features with smaller parameters, which can enhance the classification performance. Therefore, it can better identify theft and non-theft behavior, and it has better computing performance. Fig. 3 shows the new structure of the CNN, which comprises eleven layers, including four convolution layers, four pooling layers, and three fully-connected layers. All of these layers contain trainable parameters. The CNN unifies the size of the theft and non-theft data sets to 100×100 for input and operates through four sets of convolutional layers and subsampling layers. After calculating each layer of convolutional layers, the activation function RELU is used to assign values. Then the subsampling layers compress the matrix element

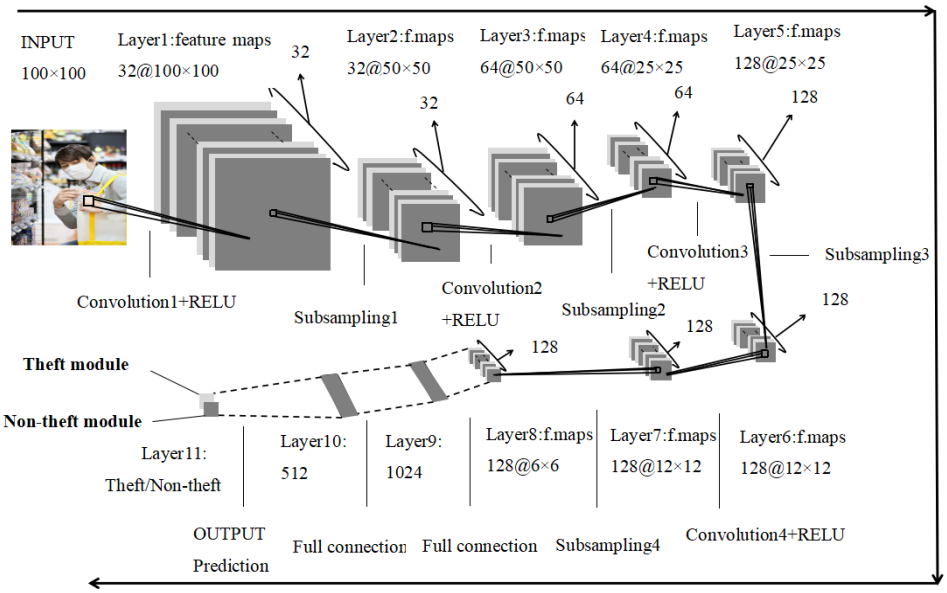

Fig. 3 *The new CNN structure of intelligent security system.*

and transmit it to the fully-connected layer. After integrating features through the two fully-connected layers, the third layer of fully-connected layers (Output layer) finally outputs the theft and non-theft features.

Tab. I shows the detailed information of the 11 layers. In Tab. I, the Con. means the convolutional layer, the Sub. means the subsampling layer, and the Full. means the fully-connected layer. The Full. 3 can also be called the output layer. The ReLU means whether to use the function ReLU or not, 1 represents used, 0 represents not used. The Matrix size equals the size of the output matrix multiplies the number of the filter.

	Con. 1	Sub. 1	Con. 2	Sub. 2	Con. 3	Sub. 3	Con. 4	Sub. 4	Full. 1	Full. 2	Full. 3
Filter	5×5	2×2	5×5	2×2	3×3	2×2	3×3	2×2	/	/	/
Step	1	2	1	2	1	2	1	2	/	/	/
ReLU	1	0	1	0	1	0	1	0	/	/	/
Matrix size	100×100 $\times 32$	50×50 $\times 32$	50×50 $\times 64$	25×25 $\times 64$	25×25 $\times 128$	12×12 $\times 128$	12×12 $\times 128$	6×6 $\times 128$	1024	512	Theft/ Non-theft

Tab. I *Information for each layer of the new CNN structure.*

Based on the CNN mentioned above, the back-propagation and forward-propagation algorithms are used to repeatedly train the input data sets to construct an intelligent security recognition model.

Fig. 4 shows the flow chart of the intelligent security system. The parameters are

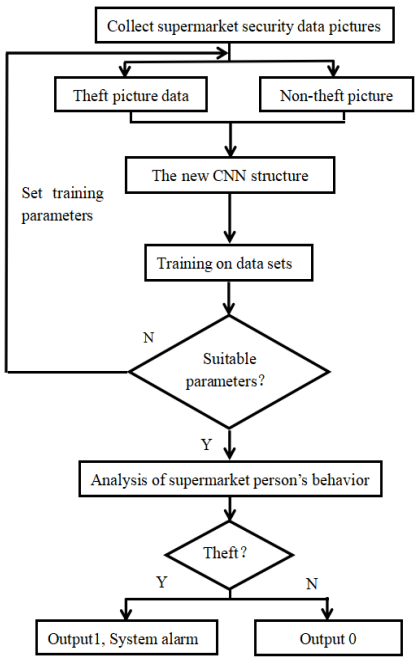

Fig. 4 *Flow chart of identification module.*

continuously adjusted through the back-propagation of the output to avoid under-fitting or over-fitting until a suitable parameter is found. When the parameter is appropriate, the CNN model is successfully built by training the theft and non-theft data sets, which can identify the behavior of people in the surveillance camera. If the CNN model identifies behavior to be theft, it outputs one, and the PC gives an alarm. If not, it outputs zero, and the PC does not respond.

4. Research methodology

4.1 Data collection

The training of the CNN structure and the judgement of intelligent recognition of theft or non-theft behavior are based on extensive data collection. Supermarket theft behavior is diverse, and the characteristics of the thieves are also different. Therefore, we selected corresponding different persons for data sets to exclude the influence of personal factors (including height, gender, clothes, and other factors) in training. Besides, some people might carry a backpack when shopping in supermarkets, and thieves also use backpacks as their tools to hold stolen goods, so backpack is one of the influencing factors in data collection.

According to the above analysis, we had collected theft and non-theft data from some supermarkets. Fig. 5 and Fig. 6 display the data collection process of

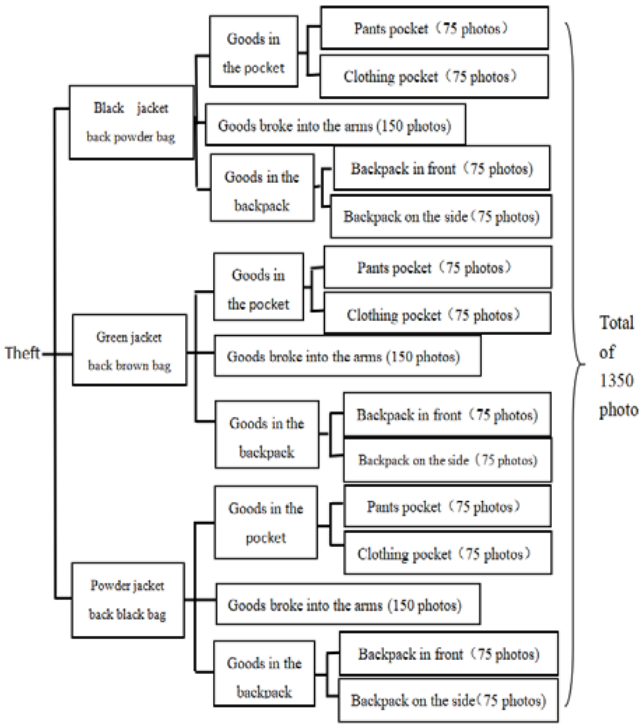

Fig. 5 *Theft module data collection.*

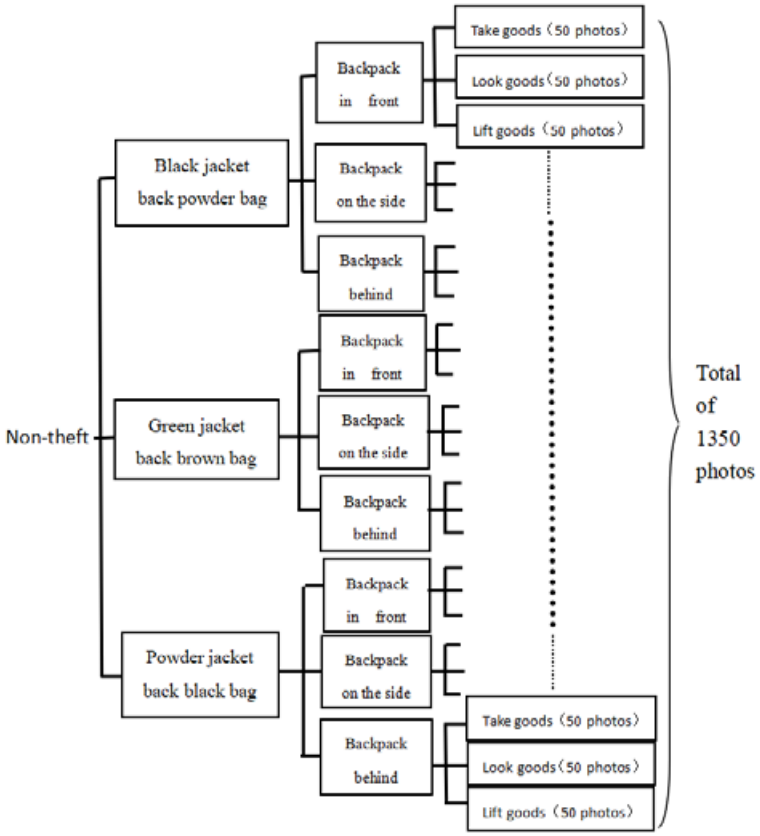

Fig. 6 *Non-theft module data collection.*

the theft and non-theft movement. Firstly, the theft and non-theft testers were the same persons. Then the data were collected through each person wearing three suits of clothes with significantly different colors to eliminate the interference of clothes. What's more, three different positions where the backpack placed on the human body helped exclude the impact of the backpack factor on theft identification. Besides, the theft and non-theft behavior mentioned above were repeated by four types of persons in different body shapes: tall male, short male, tall female, and short female. Each person should randomly select the product to eliminate the interference of the product factor. Therefore, the experiment needed to collect a total of 1350 theft action data sets from each individual who was supposed to wear three suits of clothes in significantly different colors. Then the sum of the theft data sets from four persons was 5400, and the same was true for the non-theft data sets.

Fig. 7(A-C), (D-F), (G-I), (J-L) represent the theft behavior of four types of people. The theft behavior includes the following three situations: taking goods into the arms, stuffing goods into the pocket, and putting goods into the backpack, which are framed in red in Fig. 7(G), (H), (I), respectively. Fig. 8(a-c), (d-f), (g-

i), (j-l), represent the non-theft behavior of four types of people. The non-theft behavior includes the following three situations: taking goods, looking for goods, and carrying goods, which are framed in red in Fig. 7(g), (h), (i), respectively.

In summary, a total of 10,800 theft and non-theft data sets were collected and input into the new CNN as training data sets.

Fig. 7 *Theft behavior.*

Fig. 8 *Non-theft behavior.*

4.2 Optimization process of the CNN in the system

In this paper, the CNN was progressively optimized through the principle of the CNN feature extraction and the collected data sets. Here, the tall male is called Male No. 1, and the short male is called Male No. 2. Correspondingly, the tall female is called female No. 1, and the short female is called female No. 2. Fig. 9 shows the training process which uses the method of controlling variables to gradually eliminate the effects of backpack position, clothes, gender, and height factors in each phase of the experiment. From these collected data sets, the influence of the backpack is first eliminated by continuously changing the position of the backpack, which is shown in Fig. 5 and Fig. 6. In the experimental phase 1, four types of people built their own CNN models from a suit of clothes. According to Fig. 5 and Fig. 6, the data collection process contains 450 theft data sets and 450 non-theft data sets.

In the experimental phase 2, four types of people built their own CDD models from two suits of clothes for a total of 1800 data sets. In the experimental phase 3, four types of people built their own CNN models from three suits of clothes for a total of 2700 data sets. By calculating the average of the recognition accuracy

of the three experimental phases, we analyzed the recognition accuracy of different suits of clothes and eliminated the influence of clothes.

In the experimental phase 4, considering the height and gender of persons in training, three groups of different height and gender were tested, respectively. The data sets were up to 5400. The average recognition accuracy of the experimental phase 4 was compared with the experimental phase 3 to analyze the influence of height and gender, and we eliminated the influence of height and gender.

In the experimental phase 5, a total of 10800 data sets were comprehensively trained to construct the final CNN model. We analyzed the changing in the recognition accuracy of the entire optimization process by comparing the results of the five experimental phases.

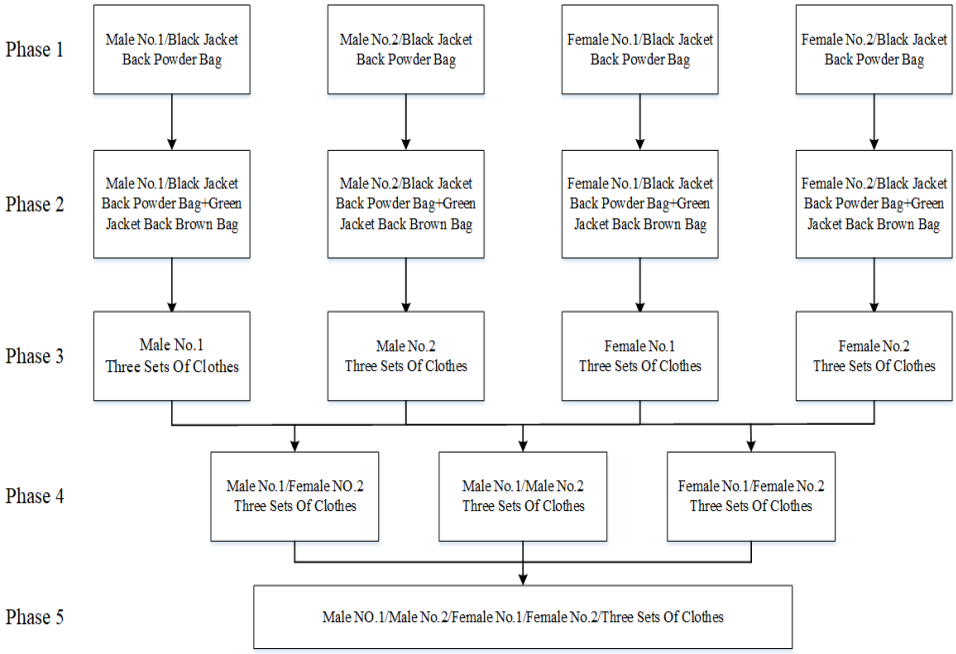

Fig. 9 *Training level of the CNN model.*

5. Result

5.1 The results of system detection accuracy

Tab. II and Tab. III show the theft and the non-theft recognition in each experimental phase, respectively. The experimental phase 1-3 reflects the influence of the number of clothes suits on the experimental results. The experimental results display that the recognition accuracy of the theft and non-theft behavior for four types of persons fluctuated with the increasing number of clothes suits. However, the overall theft and non-theft recognition accuracy is increasing. The experimental phase 4 reflects the influence of the contour of the human body, and the

	Experimental process Behavior recognition	Total No. of Theft identification	Unidentified number of theft	Theft recognition [%]	Average Theft recognition [%]
Phase 1	Male No. 1 Black Jacket Back Powder Bag	100	29	71	
	Male No. 2 Black Jacket Back Powder Bag		74	26	
	Female No. 1 Black Jacket Back Powder Bag		17	83	
	Female No. 2 Black Jacket Back Powder Bag		87	13	
Phase 2	Male No. 1/Black Jacket Back Powder Bag + Green Jacket Back Brown Bag	100	42	58	50.50
	Male No. 2/Black Jacket Back Powder Bag + Green Jacket Back Brown Bag		65	35	
	Female No. 1/Black Jacket Back Powder Bag + Green Jacket Back Brown Bag		28	72	
	Female No. 2/Black Jacket Back Powder Bag + Green Jacket Back Brown Bag		63	37	
Phase 3	Male No. 1/Three Sets of Clothes	100	62	38	50.25
	Male No. 2/Three Sets of Clothes		60	40	
	Female No. 1/Three Sets of Clothes		22	78	
	Female No. 2/Three Sets of Clothes		35	65	
Phase 4	Male No. 1/Female No.2 Three Sets of Clothes	100	51	49	66.67
	Male No. 1/Male No. 2 Three Sets Of Clothes		26	74	
	Female No. 1/Female No. 2 Three Sets Of Clothes		23	81	
Phase 5	Male No. 1/Male No. 2/Female No. 1/ Female No. 2 Three sets of clothes	100	19	81	81.00

Tab. II *The recognition accuracy of theft in each experimental phase.*

	Experimental process Behavior recognition	Total No. of non-theft identification	No. of error alarms	Non-theft recognition [%]	Average Non-theft recognition [%]
Phase 1	Male No. 1 Black Jacket Back Powder Bag	100	44	56 %	63.25
	Male No. 2 Black Jacket Back Powder Bag		17	83	
	Female No. 1 Black Jacket Back Powder Bag		75	25	
	Female No.2 Black Jacket Back Powder Bag		11	89	
Phase 2	Male No. 1/Black Jacket Back Powder Bag + Green Jacket Back Brown Bag	100	28	72	72.00
	Male No. 2/Black Jacket Back Powder Bag + Green Jacket Back Brown Bag		17	83	
	Female No. 1/Black Jacket Back Powder Bag + Green Jacket Back Brown Bag		49	51	
	Female No. 2/Black Jacket Back Powder Bag + Green Jacket Back Brown Bag		18	82	
Phase 3	Male No. 1/Three Sets Of Clothes	100	23	77	76.25
	Male No. 2/Three Sets Of Clothes		15	85	
	Female No. 1/Three Sets Of Clothes		38	62	
	Female No. 2/Three Sets Of Clothes		19	81	
Phase 4	Male No. 1/Female No. 2 Three Sets Of Clothes	100	21	79	77.67
	Male No. 1/Male No. 2 Three Sets Of Clothes		22	78	
	Female No. 1/Female No. 2 Three Sets Of Clothes		24	76	
Phase 5	Male No. 1/Male No. 2/Female No. 1/ Female No. 2 Three Sets Of Clothes	100	15	85	85.00

Tab. III *The recognition accuracy of non-theft in each experimental phase.*

experimental results show that height is the main influencing factor. Combining a pair of persons with clear outlines, we find out that the overall theft and non-theft behavior recognition accuracy is higher than the experimental phase 1-3. In this experimental phase 4, we remove the height difference of the male No. 1/female No. 2 combination, which has no noticeable difference in the combination. The experimental phase 5 reflects the overall integration of all collected data sets. The recognition accuracy is rising.

Fig. 10 represents the change of the overall recognition accuracy of the experimental phase 1-4. The experimental results display that the overall recognition accuracy of the system for non-theft behavior is higher than the theft behavior in the experimental phase 1-3. The curves of the theft are more stable than the curves of the non-theft, and the system optimization shows an upward trend. In the experimental phase 4, the recognition accuracy reaches more than 70 %.

Fig. 11 represents the overall recognition accuracy of each experimental phase. The experimental results show that the recognition accuracy of theft behavior, the recognition accuracy of non-theft behavior, and the average recognition accuracy are steadily increasing with the optimization of the system, and the average recognition accuracy reaches 83 % in experimental phase 5.

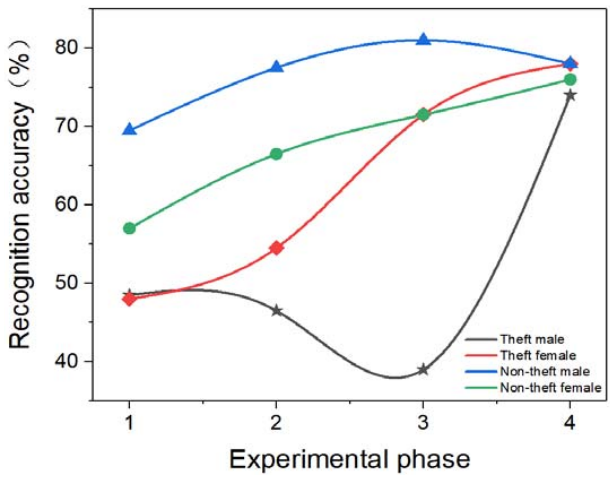

Fig. 10 *Comparison of male and female identification accuracy in experimental phase 1-4.*

5.2 Actual inspection of the system

In the experimental phases of this paper, we selected the accuracy-test videos about theft and non-theft behavior from people outside the data sets. The video was captured as images by frame, and the system would judge the behavior of people in the image in turn. In the video, people showed theft behavior many times, such as taking goods into the arms, stuffing goods into the pocket, and putting goods into the backpack, and correspondingly showed non-theft behavior many times, such as taking goods, looking for goods, and carrying goods. Then the

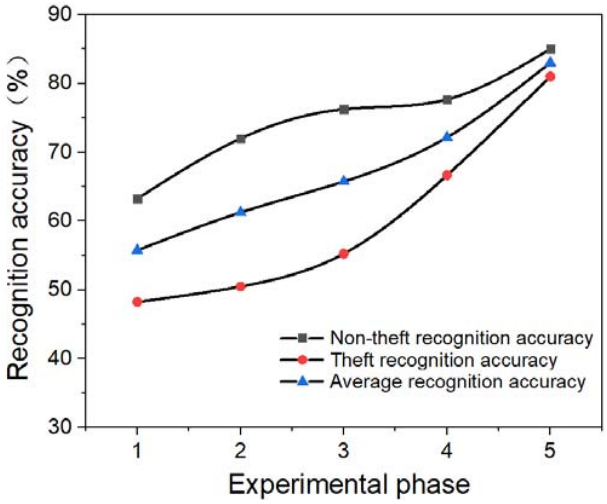

Fig. 11 *Comparison of recognition accuracy in each experimental phase.*

system performed behavioral determination through a well-trained neural network system to record the number of unidentified times in the theft and the number of false alarms in non-theft behavior.

Fig. 12 represents a judgement of CNN, which depends on the behavior of the person in the video and the display result. The result is a determination vector with two elements. The first element represents the weight value of the non-theft

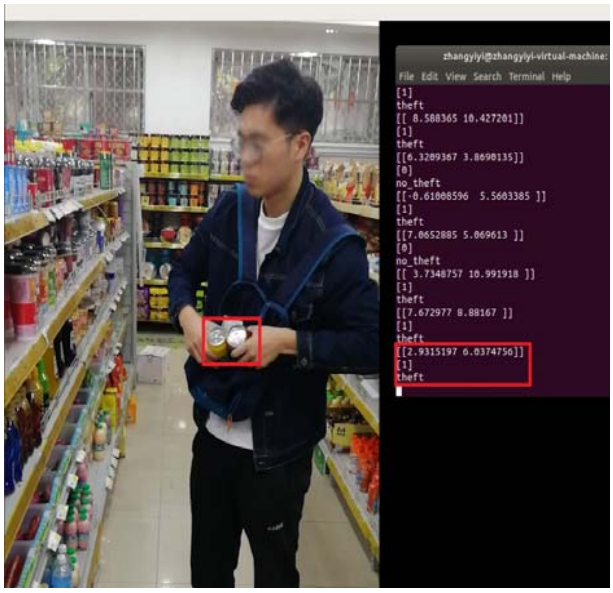

Fig. 12 *Theft behavior recognition judgement.*

behavior, and the second element represents the weight value of the theft behavior. When the second element value of the determination vector is higher than the first element value, the CNN recognizes that the theft behavior weight ratio is higher than the non-theft behavior. The PC judges that it is theft behavior and issues an alarm. Fig. 13 shows the value of the first element is higher than the second element. The CNN recognizes that the non-theft behavior weight ratio is higher than the theft behavior. The PC decides it is non-theft behavior. After integrating all the data sets, we found that the average recognition accuracy of the system reaches 83 %.

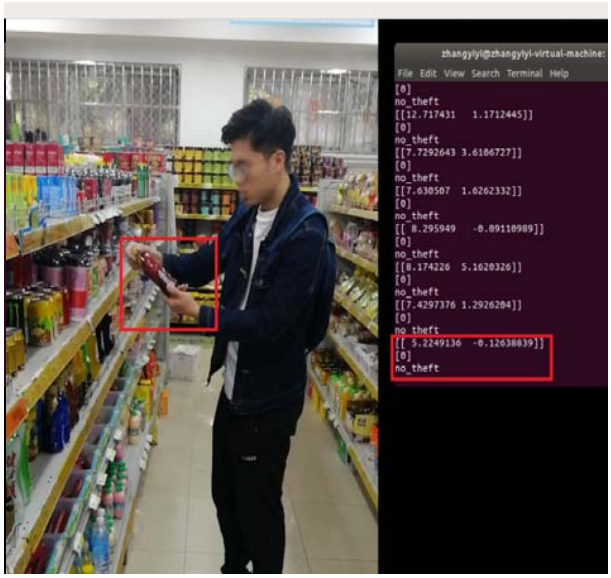

Fig. 13 *Non-theft behavior recognition judgement.*

6. Conclusions and future work

In summary, the traditional video surveillance mode in supermarkets has excellent limitations. Simple workforce monitoring can't effectively prevent supermarket theft behavior. In the aspect of artificial intelligence, the function of the supermarket intelligent security system is not perfect, and the cost is high. However, this paper proposes a new supermarket security system, which can achieve the automatic identification and alarm.

In this paper, the positions of backpack, the color of clothes, and the outline of the human body are regarded as factors to identify the accuracy of the theft and non-theft behavior. By collecting the same number of theft data sets and non-theft data sets, we constructed models for four types of people in five experimental phases. With the continuous advancement of the five experimental phases, the system performance was continuously optimized, and the recognition accuracy was steadily improved. The increase in the accuracy of theft identification is particu-

larly obvious among them. The experimental results showed that factors such as clothes suits and body shapes had a significant influence on recognition accuracy. With the constant addition of various clothes suits and various aspects of human body shapes, the recognition accuracy of theft and non-theft behavior were continuously improved. After training all the data sets, the average recognition accuracy of the system reached 83 %. Therefore, the new intelligent supermarket security system realizes real-time intelligent security monitoring, which has the advantages of low cost and high recognition precision and shows excellent commercial value and application prospects.

The CNN can be continuously evolved and optimized, so it has excellent potential for improvement. Based on the experiments done so far, we have two directions in future work. On the one hand, we can continue to increase the amount of experimental data to improve the recognition rate of the CNN. We can also try to upgrade the hardware. For example, we can use higher resolution cameras to detect whether the recognition accuracy will be further improved or not. On the other hand, we are able to learn from VGGNet's structural model [26] and then optimize our own CNN structure. We can further deepen the network structure to improve performance, thereby improving the recognition rate of the CNN structure proposed in the paper from a theoretical perspective.

Acknowledgement

Project supported by the National Natural Science Foundation of China (Grant No. 2018YFF0214904), the National Natural Science Foundation of China (Grant No. 61975182), the Science and Technology Plan Project of Zhejiang Province (Grant No. 2020C03095) and Key Research and Development Project of Zhejiang Lab (Grant No. 2019DE0KF01).

References

[1] SANO N. Estimation of customer behaviour in sales areas in a supermarket using a hidden Markov model. *International Journal of Knowledge Engineering and Soft Data Paradigms*. 2016, 5(2), pp. 135–145, doi: 10.1504/IJKESDP.2016.075981.

[2] TAYLOR E. Supermarket self-checkouts and retail theft: The curious case of the SWIPERS. *Criminology & criminal justice*. 2016, 16(5), pp. 552–567, doi: 10.1177/1748895816643353.

[3] DEYLE E. In: Deyle E, ed. – *Global retail theft barometer* [online], MarketWatch, 2015 [viewed 2019-03-25]. Available from: http://literature.puertoricosupplier.com/079/F078635.pdf.

[4] LI L. Narrow width acousto-magnetic anti-theft marker having multiple resonators. *U.S. Patent No. 9,189,935,* 2015. Available from: https://patents.google.com/patent/US9189935B2/en.

[5] CHEN Z.Y., ZHAI Y.M., TIAN L. Design of Intelligent supermarket cashier service system. In: *2016 4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEEECS 2016),* Jinan, China. Atlantis Press, 2016, pp. 293–296, doi: 10.2991/iceeecs-16.2016.62.

[6] WINGFIELD N., MOZUR P., CORKERY M. In: GIULIA M, ed. – *Retailers Race Against Amazon to Automate Stores* [online] The New York Times, 2018 [viewed 2019-03-25]. Available from: https://www.nytimes.com/2018/04/01/technology/retailer-stores-automation-amazon.html.

[7] LU X.Y., FANG J.S. "New Retail": Innovating the Development Model of China's Fresh E-commerce. *DEStech Transactions on Social Science, Education and Human Science*, 2019, 657(2), pp. 185–189, doi: 10.12783/dtssehs/aems2019/33542.

[8] TANG S. Investors' Psychological Expectation of Retailers under the New Retail Model. In: *4th International Conference on Humanities Science, Management and Education Technology (HSMET 2019)*, Singapore. Atlantis Press, 2019, pp. 756–765, doi: 10.2991/hsmet-19.2019.141.

[9] RATHINASABAPATHY G., RAJENDRAN L. RFID Technology and Library Security: Emerging Challenges. *Journal of Library, Information and Communication Technology*. 2015, 1(1), pp. 34–43, Available from: http://escienceworld.in/index.php/jlict/article/view/102/102.

[10] ZANG K., XU H., ZHU F., LI P. Analysis and Design of Group RFID Tag Security Authentication Protocol. In: *Conference on Complex, Intelligent, and Software Intensive Systems,* Berlin, Germany. Springer, Cham, 2019, pp. 637–645, doi: 10.1007/978-3-030-22354-0_57.

[11] LEE S.H., LEE D.W. A study on ICT technology leading change of unmanned store. *Journal of Convergence for Information Technology*, 2018, 8(4), pp. 109–114, doi: 10.22156/CS4SMB.2018.8.4.109.

[12] CHOLLET F. *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*. MITP-Verlags GmbH & Co. KG, 2018.

[13] HUBEL D.H., WIESEL T.N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 1962, 160(1), pp. 106–154, doi: 10.1113/jphysiol.1962.sp006837.

[14] LIEW S.S., HANI M.K., Gender classification: a convolutional neural network approach. *Turkish Journal of Electrical Engineering & Computer Sciences*, 2016, 24(3), pp. 1248–1264, doi: 10.3906/elk-1311-58.

[15] JIAO Y., YANG J., FANG Z., XIE S., PARK D. Comparing studies of learning methods for human face gender recognition. In: *Chinese Conference on Biometric Recognition,* Berlin, Heidelberg, Germany. Springer, 2012, pp. 67–74, doi: 10.1007/978-3-642-35136-5_9.

[16] BASHA A.F., JAHANGEER G.S.B. Face gender image classification using various wavelet transform and support vector machine with various kernels. *International Journal of Computer Science Issues (IJCSI)*, 2012, 9(6), pp. 150–157. Available from: https://search.proquest.com/openview/6fbb981e646162367f8854f32e27fbe8/1?pq-origsite$=$gscholar{&}cbl$=$55228

[17] BERBAR M.A. Three robust features extraction approaches for facial gender classification. *The Visual Computer*, 2014, 30(1), pp. 19–31, doi: 10.1007/s00371-013-0774-8.

[18] LECUN Y., KAVUKCUOGLU K., FARABET C. Convolutional networks and applications in vision. In: *Proceedings of 2010 IEEE international symposium on circuits and systems,* Paris, France. IEEE, 2010, pp. 253–256, doi: 10.1109/ISCAS.2010.5537907.

[19] LECUN Y., BOSER B.E., DENKER J.S., HENDERSON D., HOWARD R.E., HUBBARD W.E., JACKEL L.D. Handwritten digit recognition with a back-propagation network. In: *Advances in neural information processing systems*, Denver, CO. Morgan Kaufmann, 1990, pp. 396–404. Available from: http://papers.nips.cc/paper/293-handwritten-digit-recognition-with-a-back-propagation-network.pdf.

[20] JARRETT K., KAVUKCUOGLU K., RANZATO M.A., LECUN Y. What is the best multi-stage architecture for object recognition? In: *2009 IEEE 12th international conference on computer vision*. Kyoto, Japan. IEEE, 2009, pp. 2146–2153, doi: 10.1109/ICCV.2009.5459469.

[21] LEE H., GROSSE R., RANGANATH R., NG A.Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: *Proceedings of the 26th annual international conference on machine learning,* Montreal, QC, Canada. ICML, 2009, pp. 609–616, doi: 10.1145/1553374.1553453.

[22] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, 86(11), pp. 2278–2324, doi: 10.1109/5.726791.

[23] CIREGAN D., MEIER U., SCHMIDHUBER J. Multi-column deep neural networks for image classification. In: *2012 IEEE conference on computer vision and pattern recognition*, Providence, RI, USA. IEEE, 2012, pp. 3642–3649, doi: 10.1109/CVPR.2012.6248110.

[24] YUAN A., BAI G., JIAO L., LIU Y. Offline handwritten English character recognition based on convolutional neural network. In: *2012 10th IAPR International Workshop onDocument Analysis Systems*, Gold Cost, QLD, Australia. IEEE, 2012, pp. 125–129, doi: 10.1109/DAS.2012.61.

[25] VALLE E.A., STAROSTENKO O. Recognition of human walking/running actions based on neural network. In: *2013 10th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, Mexico City, Mexico. IEEE, 2013, pp. 239–244, doi: 10.1109/ICEEE.2013.6676005.

[26] JUN H., SHUAI L., JINMING S., YUE L., JINGWEI W., PENG J. Facial Expression Recognition Based on VGGNet Convolutional Neural Network. In: *2018 Chinese Automation Congress (CAC)*, Xi'an, China. IEEE, 2018, pp. 4146–4151, doi: 10.1109/CAC.2018.8623238.