ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20180620082101
Original scientific paper

Application of Video Scene Semantic Recognition Technology in Smart Video

Lele QIN, Lihua KANG

Abstract: Video behaviour recognition and semantic recognition understanding are important components of intelligent video analytics. Traditionally, human behaviour recognition has met problems of low recognition efficiencies and poor accuracies. For example, most existing behaviour recognition methods use the video frames obtained by even segmentation and fixed sampling as the input, which may lose important information between sampling intervals, fail to identify the key frames of the video segments and make use of the contextual semantics to understand current behaviour. In order to improve the semantic understanding capacity and efficiency of video segments, this paper adopts a 3-layer semantic recognition approach based on key frame extraction. First, it completes the segmentation for video recognition at the bottom layer, extracts the key frames in the video segments, primarily understands basic semantics of the persons' identifications, behaviours and environment, and then introduces the primarily acquired information into the middle layer for semantic integration, and through the integration of various semantics, adopts the loss function to learn the latent relationship between different modal semantics, to enhance the integrating capacity and the robustness of the character semantic integration, and finally, by overall fine tuning, semantic recognition and adjusting all the parameters of the network, completes the semantic recognition task of the video scenario. This method enjoys higher recognition accuracies based on certain datasets, capable of effectively recognizing the semantics of characters and behaviours in videos. Through practical testing, the adoption of the algorithm integrating key frame extractions with the video scene semantic recognition has improved the recognition accuracy and effect of the video character semantics.

Keywords: Convolutional Neural Network (CNN); deep learning; keyframe; semantic recognition; smart video

1 INTRODUCTION

In the world nowadays, there are so many hidden troubles of security due to large populations and complicated communications. Along with increasingly higher requirements for security and current scientific and technological development, surveillance cameras have witnessed an exponential increase in quantity and covered larger and larger areas. An enormous surveillance network with numerous cameras will generate a sea of data almost in an instant; then, how to extract information of value efficiently from such a sea of data is an urgent problem for smart video technology to solve. Specifically speaking, that is to enable cameras to be eyes, the video transmission network the neural network and the smart recognition technology and algorithms the brains, in a bid for comprehension and judgment of contents in monitored areas and automatic recognition and alarming of abnormal actions. Video action recognition is an important component of smart video analysis and its key lies in splitting the video into image sequences and sorting by spatio-temporal features. In recent years, thanks to great success of deep neural network in the field of image recognition, application of neural network to detection and recognition of video actions has achieved remarkable effects.

2 RELEVANT RESEARCH

As Convolutional Neural Network (CNN) technology made great success in image recognition, researchers began to switch their research emphasis to the field of video action recognition. After video segmentation, the current deep learning method is divided into two parts, i.e. video clip feature extraction and video level feature fusion extraction. In modern times, most technologies adopted are with CNN as the means of extracting video features; that is to divide a video into clips, extract one or more interested frames and input them into CNN for subsequent analysis, comprehension and fusion of clip features [1, 2]. With RBG video as research object, Tran [3] et al. used features of 3D convolutional extraction actions directly and achieved good effects. Yet due to limits on size of convolutional network, such 3D CNN failed to process videos of different lengths. Donahues et al. proposed LRCN structure and introduced LSTM (Long Short-Term Memory) to fuse features extracted by each frame on the basis of CNN's extracting features of independent frames. The foregoing two methods cost much training time and storage. Karen et al. [4]. brought forward Two-Stream method, which took dense optical flow as auxiliary input and employed two contrary CNNs to extract single-frame original image and multi-frame optical flow image features respectively, so as for fusion at final evaluation level. Application of two-stream technology improved the accuracy of action recognition remarkably and thus became the emphasis of research and use. In subsequent research, many cases were enhancement and improvement of deep-learning network on the basis of two-stream network. Temporal Segment Network (TSN), which was put forward by Wang [5], as a new video-based action recognition network structure, combines sparse temporal sampling statics with video-based surveillance and utilizes the entire video to support efficient learning; on the other hand, TSN contributes to learning of video data processing by CNN. Ji et al. [6] researched action recognition based on RGB video by constructing 3D CNN model, i.e. to first use a series of fixed kernel functions to generate multi-channel information for each frame, then capture motion information between multiple neighbouring frames by means of 3D CNN and eventually obtain final feature representation by combining information of all channels, so as to make a judgment on actions in the video.

In combination with relevant research achievements, this paper builds a character semantic action recognition model based on video scene deep learning and by use of keyframe extraction technology, so as to greatly improve the accuracy of action recognition. Main contributions of such research are as follows:

(1) Proposing the semantic recognition algorithm based on CNN deep learning; and

(2) Acquiring keyframe by means of video scene and clip segmentation, so as to improve the efficiency and accuracy of recognition.

3 VIDEO SEMANTIC ANALYSIS AND RELEVANT RESEARCH

As unstructured data, videos contain abundant semantic information and digging the internal relevancy of videos is of profound significance for improvement of precision ratio and recall ratio of video semantic inquiry. Semantic relevance refers to polysemy and synonymity of semantic conceptions among video data. For the sake of eliminating some associative architecture implied among semantics, it is required to analyse the relevancy of extracted semantics and decompose giant video database in addition to dimensionality reduction. Semantic information of tasks in a video can be specifically divided into character identify information, motion, facial expression, voice and so on [7-9]. Current methods for fusion of semantic themes express visual features of each image as a visual "Bag-of-words". For design of a probability model, it is acceptable to obtain potential semantic themes respectively from visual modality and text modality and fuse two kinds of semantic themes with a self-adaptive asymmetric learning method.

Whereas video images contain a number of semantic conception relations, once researchers turned to the Concept Hierarchy Tree of Word Net, because there is one and only one path between two nodes in the Concept Tree and the length of such path can serve as a measurement of semantic similarity between these two concepts [10-13]. For video flow, Shot Boundary Detection is conducted first and the video is segmented into short takes with video segmentation algorithms, such as pixel algorithm, histogram algorithm, X2 histogram algorithm, X2 histogram block algorithm and contour a boundary ROC (Rate of Change) algorithm; then, the original motion trails of video object are extracted by use of moving object tracking algorithms, such as mean shift algorithm, object tracking based on Kalman filter, object tracking based on particle filter and algorithm based on modeling of moving object, with the longest trail to be processed and information extracted therefrom, including motion direction and slop of motion trail curve. At last, the said motion action will be marked by hand to extract the video verb semantic label [14-21]. Also, some other researchers proposed that semantic clews of multiple event recognitions should be fused by means of a deep-level learning strategy so that the issue of recognition would be solved by answering how to jointly analyse human actions, objects and scenes. That is to say, first, each type of semantic features is transmitted to an abstract path of multi-level features, with one fusion level to connect all different paths, accordingly to learn the mutually affecting relevancy of semantic clews via unsupervised trans-channel coding; lastly, the question of how semantic clews compose one event and a group of events is answered by fine tuning of large-amplitude objects on the architecture [21-24].

This paper adopts a 3-layer semantic recognition approach based on key frame extraction. First, it completes the segmentation for video recognition at the bottom layer, extracts the key frames in the video segments, primarily understands basic semantics of the persons' identifications, behaviours and environment, and then introduces the primarily acquired information into the middle layer for semantic integration, and through the integration of various semantics, adopts the loss function to learn the latent relationship between different modal semantics, to enhance the integrating capacity and the robustness of the character semantic integration, and finally, by overall fine tuning, semantic recognition and adjusting all the parameters of the network, completes the semantic recognition task of the video scenario. This method enjoys higher recognition accuracies based on certain datasets, capable of effectively recognizing the semantics of characters and behaviours in videos.

4 SCENE SEGMENTATION AND KEYFRAME ACQUISITION

Shot segmentation refers to short segmentation of video sequence, also called as short change detection. It is one of key technologies in video inquiry and also the first step of video processing, including pixel algorithm, histogram algorithm, X2 histogram algorithm, X2 histogram block algorithm and contour an boundary ROC (Rate of Change) algorithm. It is used to detect shot cut and fade-in/fade-out. This algorithm detects shot cut by calculating the difference between histograms of two consecutive images in the video. Besides shot cut, another way is fade-in/fade-out, i.e. the frame close to the junction of videos turns dark little by little and then bright again; therefore, the neighbouring pixel relevancy of each frame will go smaller first and then larger and the gradient between every two pixels is just the very representation of their relevancy. In the research herein the key frame in the clip is first to select [25, 26].

4.1 Clip Segmentation

Movements are distributed on different time varying from speed, resulting in unbalance of movement information content in time domain. For example, the movement to strike Ping Pong consists of arm withdrawal, readiness-to-strike, strike and so on. The phase from arm withdrawal to readiness-to-strike takes approximately 236 frames while that from readiness-to-strike to strike does approximately 186 frames, as shown in Fig. 1, with possible leftover of information content as consequences of sampling in terms of even segmentation [27, 28].

In the process of video detection, the subject occupies the most or vast majority space of frame; thus, it is acceptable to express the motion conditions of subject with the algorithm of dense optical flow. Dense optical flow is defined as below [24]:

$$T(x,\ y) = I(x+u,\ y+v) \tag{1}$$

Where: T - Past frame; I - Current frame; u, v - Offsets of pixel on x-axis and y-axis.

In this formula, u and v correspond to matrixes of coordinates x, y. When the subject in the frame was in obvious motion, the optical flow image should have a high absolute value, that is to say there often was a lot of action

information contained. Generally action information are distributed unevenly in the video; therefore, in order to make the motion information amounts in video clips tend to be even, in the case that the clip quantity NS is fixed, it is required to minimize the segment variance V_M. V_M is defined as follows [27]:

$$V_{\mathrm{M}} = \frac{\sum_i (M_i - \bar{M})}{N_{\mathrm{s}}} \tag{2}$$

Where: M_i - The motion information content of the ith clip; $\bar{M}$ - Average information content of clip.

Therefore, the information content of each clip, M, is shown as below:

$$M = \sum_i \sqrt{\sum_c \sum_{x,y} flow_i^2(x, y, z)} \tag{3}$$

Where: $flow_i(x, y, z)$ - Corresponding two-channel optical flow image in clip; z - Channel of optical flow.

In general, optical flow will fluctuate depending on changes of video quality, thus resulting in errors. In this case, it is acceptable to consider seeking local optimal solution in a sense, instead of considering overall optimal one. Therefore, in this paper, this problem is solved by introducing Greedy algorithm for solution of approximation.

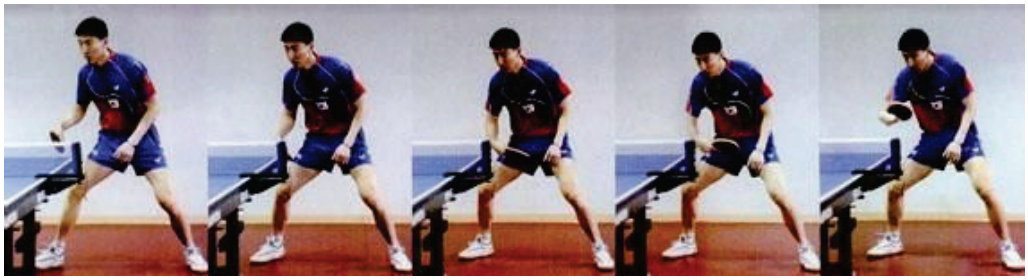

Figure 1 Ping Pong Strike Movement Breakdown

4.2 Keyframe Acquisition

It is very necessary to adopt CNN deep learning model for pre-use and in-use training so that the effects of video recognition would be improved greatly. Generally it is believed that the learning model learns the features of image, but an image could be either clear or fuzzy and a fuzzy image is possible to considerably affect final image recognition. Therefore, it is required to intensify the training on fuzzy image, accordingly to improve the acquisition degree of image information content. With the approach of information content evaluation, in this paper, keyframes are identified by finding image frames with maximum information content in the video clip [29, 30]. As research demonstrated, human vision pays more attention to the edge of observed object in the image, thus many detection methods based on image edge have achieved great successes. Ei may be defined as edge information content of the ith-frame image, shown as below:

$$E_i = \frac{\sum_{x,y} edge_i(x, y)}{W \times H} \tag{4}$$

Where: $edge_i(x, y)$ - Edge image obtained from video frame; W - Image width (pixel); H - Image height (pixel).

The edge extraction method raised by Dollar et al. has succeeded greatly and it is also employed in this paper. Meanwhile, the evenness of image gray scale also indicates how fuzzy the research object is. As compared in Fig. 2, it is believed that the left picture achieved higher scores under the condition that research objects maintain unchanged roughly. In the final video clip taken, the frame with most information content is keyframe [31, 32], as shown below:

$$KeyIndex = ragmax(E) \tag{5}$$

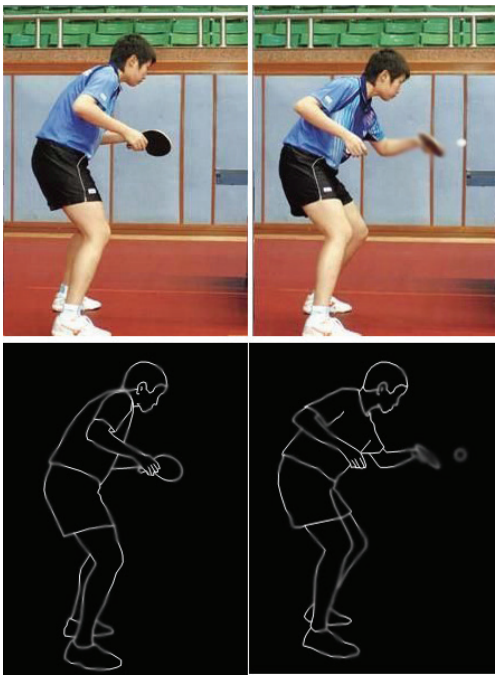

Figure 2 Comparison of Image Edge Conditions in Motion

4.3 Feature Fusion

Regarding spatial feature network design, in order for strong scale detail recognition ability of image network, in this paper the Inception structure is chosen for image sorting task, with Batch Normalization serving for more rapid and accurate convergence of that structure. Regarding temporal feature network design, a modified Inception structure is selected, and on the basis of original, the weight of first convolutional layer in network is

brought for channel expansion, in a bid to enable input to support more channels. Since continuous limited frame information is extracted from current optical flow network, it is possible to support extraction of video features from different temporal scales after an optical flow network with multiple temporal scales is introduced. Temporal feature network design still adopts the Inception structure and also makes necessary modifications to that structure, i.e. channel expansion is made to the first convolutional layer weight of network on the basis of original network. For only continuous limited frame information can be extracted from current optical flow network, it is acceptable to introduce the optical flow network with multiple temporal scales, so as to possess the feature of extracting different temporal scales. The implementation chart of spatio-temporal network learning machine training is as shown in Fig. 3. Frames selected from each clip serve as input and the network output is the output of last convolutional layer in corresponding network, with mean pooling to be used for fusion of both. Video clips are segmented on the basis of motion information content and video frames are selected from those clips randomly as input, so that each clip would be with spatio-temporal features; that is the input of convolutional layer. Mean pooling is indispensible and the results of pooling serve as input of loss function for the sake of loss calculation and later backpropagation [27, 31, 32].

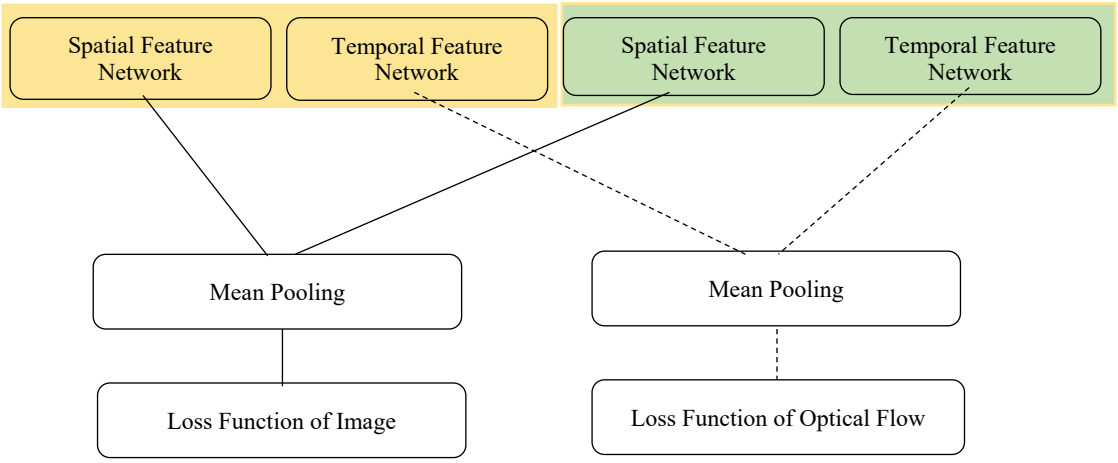

Figure 3 Implementation Chart of Network-integrated Learning Machine Training

In this way, it will be possible to extract keyframes from video clips successfully and then further complete shot segmentation.

5 TECHNOLOGICAL ARCHITECTURE OF VIDEO SEMANTIC RECOGNITION MODEL BASED ON DEEP LEARNING

The semantic understanding of image has hierarchical structure, including low-layer, middle-layer and high-layer. It respectively corresponds to the image processing layer (low-level visual features), image analysis layer (middle semantic features) and image recognition layer (high-level sampling semantics). It forms the bottom-up data-driving from low-layer to high-layer and top-down knowledge driving from high-layer to low-layer, of which the existence of middle layer aims to reduce the semantic gap between the low-layer and the high-layer. The semantic understanding of image scene must establish the mapping relationship between low-layer visual features and high-layer scene semantics, which belongs to the category of image semantic understanding.

Due to video data featured by complicated structure, abundant semantics and diversified data types etc., traditional data models like relational model and object model are incapable of playing such a vital role and a data model specific to video is required. Information in need of description in videos is distributed in three levels: (1) bottom features: characteristics extracted from original video data by use of automatic analysis technology, e.g. color, texture, shape, motion etc.; (2) spatial and temporal information of physical object: "what appears in the video", including objects extracted from the video, their motion trails and their spatio-temporal relationship, which information can be extracted from video data in an automatic or semi-automatic manner; (3) semantic information: "what happened in the video", information that people perceive while watching the video, which reflects people's comprehension of video contents [33, 34].

The video semantic recognition model based on scene deep learning consists of three levels: (1) middle-layer semantic feature extraction; (2) multi-channel semantic feature fusion; and (3) overall fine adjustment and semantic recognition. Firstly, CNN inputs keyframe images in all scenes into the extraction layer and extracts lower-level features of three channels, namely character semantics $S(0)P$, behavior semantics $S(0)B$ and context semantics $S(0)C$, in the set of these keyframe images. Then, by going through CNN in parallel, the reduction of lower-level feature vector dimension is learned and the training of middle-layer semantic feature extraction in three channels is accomplished, as shown in Fig. 4. Each process of channel semantic feature extraction consists of convolution, subsampling and full connection. In the process of semantic recognition, middle-layer semantic features are taken as input of multiple fusion layers, i.e. $I(n) = [S(n)P, S(n)B, S(n)C]$, and fusion of multi-channel semantic features is realized by means of CNN learning methods of multiple channels, i.e. $I(n+1) = = [S(n+1)P, S(n+1)B, S(n+1)C]$; therein, for relevancy of semantic features at each level, loss function is

introduced to learn adjustment parameters. Finally, the results of multi-channel semantic feature fusion act as input of recognition layer. Meanwhile, loss function of large margin is deployed to fine-adjust the parameters of entire network learning, resulting in mission of semantic recognition completed in the end.

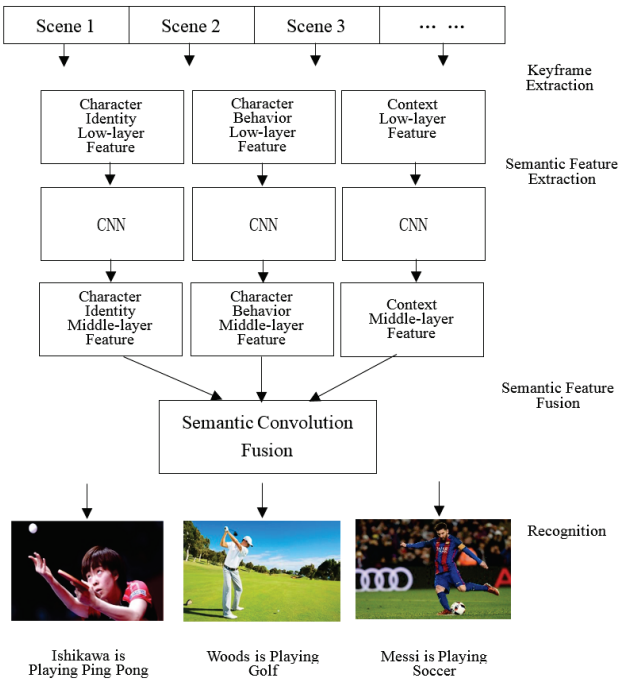

Figure 4 Technological Architecture of Video Semantic Recognition Model Based on Deep Learning

6 EXTRACTION AND FUSION OF CHANNEL SEMANTIC FEATURES

The semantic classification of image scene belongs to the category of overall scene semantic understanding, which is realized mainly by extracting visual features of the scene image, performing characteristic mapping, completing image contents description, designing classifier and finally completing image scene classification and recognition. The scene classification includes two key problems: image contents description and classification judgment. The image content description seeks to obtain the most judgment expression of the scene image, while the classification judgment sets modeling to get the calculation model of a scene category different from other scene categories through the study and training based on the image description of training sample set. The description of image contents includes feature extraction, visual dictionary generation, image feature mapping and intermediate semantic theme expression, etc. Classification judgment includes classifier design and classification. The semantic classification of semantic scenes can help improve the recognition efficiency and accuracy of scene semantics. The main steps of scene semantics classification are shown in Fig. 5.

In the research into channel semantic recognition model, the model provided by GAO et al. is well worthy of reference [35]. In this paper, that model is taken as reference and modified for research herein. Firstly, channel middle-layer semantic features are mainly convolution, sampling and full connection processes in CNN. Essentially, convolution is to make once or repeated non-linear changes on original feature vectors through one or multiple trainable convolution kernels [31]. Herein, it is feasible to describe neurons of the x^{th} layer with (N_x, b_x*b_x) and express the convolution operation of neurons between the x^{th} layer and the $(x-1)^{\text{th}}$ layer by means of $f(n*n)$ vector of one or multiple kernels and multiple connection tables (N_x*N_x-1). The input feature map (N_x, b_x*b_x) will be obtained by $f(n*n)$ vector of multiple trainable kernels convoluting an image of m*n dimension as input, then plus offset b, with N_x indicating the number of feature maps at the x-th layer and bx the dimension(s) of them. What is input into the first layer are images and that into subsequent stages subsets of convolutional feature image sets extracted from previous layer. Specifically to know exactly how many feature maps are sufficient for convolution to constitute one feature map of the following layer, it is required to preset a connection table between feature maps of two layers and such table shall record connection relations between those feature maps at two layers. In this way, the convolution layer formula of behavior semantic channel is as follows:

$$C_{A_{ij}k}^{(x)} = \sigma\left[\sum_{k\in D^x}\left(\sum_{i=1}^{n}\sum_{i=1}^{n} F_{A_{ij}}^{(x)} C_{A_{ij}k}^{(x-1)} + b^l\right)\right] \tag{6}$$

Where: $F_{A_{ij}}^{(x)}$ - Convolution kernel of the x^{th} layer on behavior semantic channel; C^x - Feature map set to e input at the x^{th} layer; b - Offset to be added after convolution operation

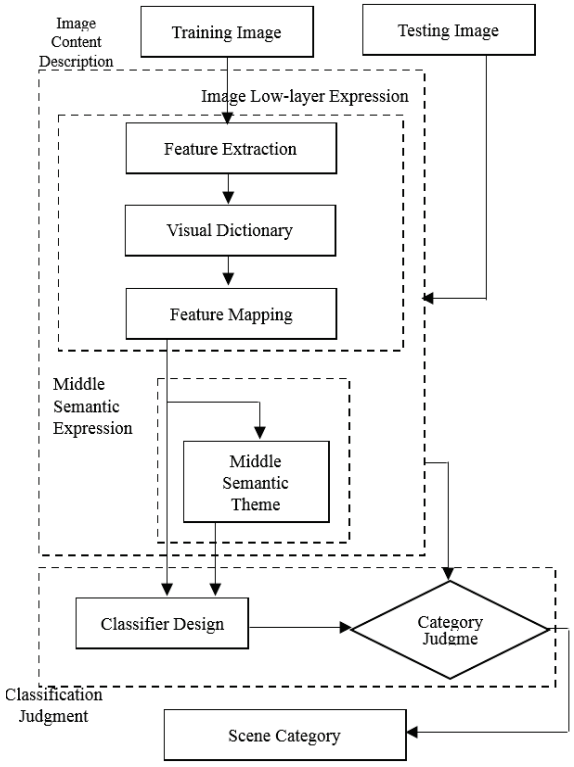

Figure 5 Flowchart of Scene Classification

The purpose of subsampling is to reduce the dimension of feature map obtained in convolution layer and the method applied is usually to sum up all pixels, of which the size is $n*n$ blocks, in input image, so that output image

would be reduced n times on two dimensions. The method for research herein is to sum up two pixels of each unrepeated areas in each feature map into one pixel, then weight with multiplying offset βz+1, add adding offset bz+1, and produce a feature map reduced twice by sigmoid activation function. Here, the convolution layer formula and the sampling layer formula of behavior semantic channel are as follows:

$$C_{A_{ij}}^{(x)} = f\left(\beta_{A_{ij}}^{(x)} * down\left(C_{A_{ij}}^{(x)}\right) + b_{A_{ij}}^{(x)}\right) \tag{7}$$

Where: *down* - Subsampling function.

Every output feature will correspond to a multiplying offset β and an adding offset b.

Full connection is to bring kernels to convolution operation at all feature maps of previous layer and reduce the feature vector to the vector of $1*n$ dimension, so that semantics at each channel would output $1*n$ vector features via respective full connection layers.

On the basis of completing extraction of middle-layer semantic features pertaining to character identity, behavior and context, middle-layer semantic features are taken as input of multi-channel semantic fusion layer, the vector matrix $V^x[C_p^{(x)}, C_A^{(x)}, C_S^{(x)}]$ is constructed as input of multi-channel semantic fusion and then features are extracted by means of multi-semantic CNN learning method. The convolution layer formula of semantic CNN at fusion layer is as follows [35]:

$$V^{(x+1)} = \delta\left(F^{(x+1)} Z^{(x+1)} + b^{(x+1)}\right) \tag{8}$$

Where: $Z^{(x+1)}$ - Middle-layer convolution output of three-layered fusion layer.

In order to lower semantic noise existing in the process of convolution, enhance the robustness of semantic fusion and allow the model to learn the relevancy among multi-channel semantics, the loss function of fusion semantics is put forward, with the formula as follows [35]:

$$L_{\overline{p}as} = \left\| V^{(x)} - V^{(x)} \overline{p}as \right\|^2 \tag{9}$$

Where: $L_{\overline{p}as}$ - Semantic feature pass.

$L_{\overline{p}as}$ merely fuses action behavior and context; in this way, $V^{(x)}\overline{p}as = [0, C_A^{(x)}, C_S^{(x)}]$. In order to improve the accuracy of complete loss function of semantic fusion, the weight w is introduced, and the formula is as follows:

$$L = wL_{pas} + w_p L_{\overline{p}as} + w_a L_{p\overline{a}s} + w_s L_{pa\overline{s}} \tag{10}$$

In this paper, relevant parameters of respective layers in the entire network are adjusted by automatic learning and supervised learning in a bid to complete the task of semantic recognition, and loss functions are built by adding the sorting of maximum margin into SVM sorter. Herein, a number of one-to-many models are trained, with one type corresponding to one model, with losses between real types $y \in \{1, -1\}$ and predicted types to be calculated for each model. Next, with fusion-layer feature vector V as training data of forward propagation and W as weight parameter between fusion layer and recognition layer, the maximum value is obtained and the formula is as follows [36]:

$$L(W) = \min_w \frac{1}{2} * W^{\mathrm{T}} * W + C\sum_{n=1}^{N} max(1 - W^{\mathrm{T}} * z * y, 0) \tag{11}$$

For the sake of simplifying the training process of multi-layer framework, the foregoing type-II will be expanded to multiple types, with corresponding matching l2-loss function as follows [36]:

$$\sum_{k\in Y}^{K} \sum_{y\in Y} max\left(1 - W_y^{\mathrm{T}} * z + W_k^{\mathrm{T}} * z, 0\right)^2 \tag{12}$$

Where: Y - Sample type set; W_k^{T} - Connecting weight of semantic K and fusion layer.

7 EXPERIMENT RESULTS

We selected an office area for testing, with the interface as shown in Fig. 6. Video data transmitted from many cameras display in polling in terms of four screens. Before this, according to the foregoing theories, the learning machine has undergone lots of learning about various work in multiple scenes, such as bank, ticket office and office hall, with corresponding video databases built. The system simulated multiple scenes:

(1) In simulation of normal work and shift handover, the computer is able to identify the current work status accurately; e.g. on the occasion of shift handover, both shifts swiped cards in the system and waited for handover completion before the system conducted the facial recognition according to their data and confirmed relevant information;

(2) When a staff swipes the card at the entrance guard before entering into office, the shooting system will compare the picture of staff in data-collecting area with that in personnel database and give alarm in case of any abnormality found; e.g. Staff A was the regular staff and entered by card swiping as required, followed by a stalker, the system gave alarm immediately and collected the picture information of stalker;

(3) When somebody, acting as a thief, stole a cell phone from another one, put the phone in the pocket and left quickly, the system gave alarm immediately, recorded the thief's facial image automatically and marked with a red box, as shown in Fig. 6;

(4) In simulation of abnormal behaviors in office, e.g. staff sleeping or chatting in work hours, the system will give message of abnormality.

Tab. 1 lists the video accuracy in certain scenarios of the model designed by this paper. For playing basketball and other sports, the actions were relatively obvious and the site semantic interpretations were more definite, so the recognition accuracies were all above 85%, while for the office environment and other more complex scenarios, the recognition accuracies slightly declined, the reasons lay in the lack of features for feature semantics and so on, resulting in compromised key frame extractions and limited semantic integration with surrounding

environment, which can be solved by increasing the learning machine learning, enriching the semantic cue materials, and improving the capacity utilization in the recognition.

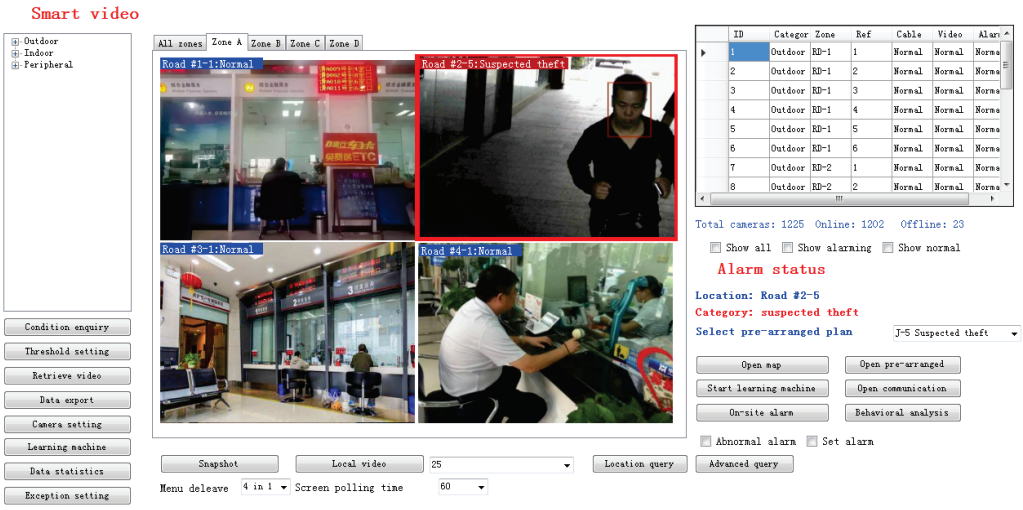

Figure 6 Testing Interface of Smart Video System

Table 1 Statistical table of experimental results of video scene dataset

Behavioral description	Recognition accuracy (%)
Playing basketball	85
Playing table tennis	86
Kicking shuttlecock	85
Normal office duty shifting behavior	75
Abnormal office duty shifting behavior (inconsistent shifting persons or procedures, etc.)	75
Follow the abnormal personnel and behavior entering the control area	69
Simulating theft behavior	70
Other activities during office hours (take snacks as an example)	68
Sleeping in the office	83
Having a too long talk in the office	66
…	...

By testing, the application of semantic recognition learning technology based on video scene in smart video and corresponding algorithm model are completely successful.

8 CONCLUSIONS

This paper accomplishes the recognition of task actions in smart video well by use of video scene semantic recognition learning technology. Firstly, the said technology utilizes video scene and clip segmentation to obtain keyframe, thus with both efficiency and accuracy of action recognition improved. Then, it deploys CNN to extract and fuse channel information like character identity, behavior and context, introduces loss function, digs potential relevancy among semantics of different channels, adjusts the learning parameters of entire network and realizes the recognition of task action by dint of SVM sorter.

By testing, the application of semantic recognition learning technology based on video scene in smart video and corresponding algorithm model are completely successful.

Acknowledgements

This work is financially supported by the Scientific Research Project of Hebei Science and Technology Department, China (No. 16214707), the Teaching Research Project of Polytechnic College of Hebei University of Science & Technology (No. 2018Z01).

9 REFERENCES

[1] Qin, L. L. & Kang, L. H. (2016). Technical framework design of safety production information management platform for chemical industrial parks based on cloud computing and the internet of things. *International Journal of Grid and Distributed Computing, 9*(6), 299-314. https://doi.org/10.14257/ijgdc.2016.9.6.28

[2] Kang, L. H., Hou, J. H., Han, B. G., et al. (2017). Design of management information platform of smart architecture project based on Cloud & BIM. *Hebei Journal of Industrial Science and Technology, 34*(6), 459-464.

[3] Tran, D., Bourdev, L., Fergus, R., et al. (2015). Learning spatiotemporal features with 3D convolutional networks. *2015 IEEE International Conference on Computer Vision (ICCV)*, 4489-4497. https://doi.org/10.1109/iccv.2015.510

[4] Simonyan, K. & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems (S1049-5258), 1*(4), 568-576.

[5] Wang, L., Xiong, Y., Wang, Z., et al. (2016). Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. *ACM Transactions on Information Systems, 22*(1), 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[6] Ji, S. W., Xu, W., Yang, M., et al. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 221-231. https://doi.org/10.1109/TPAMI.2012.59

[7] Qin, L. L., Yu, N. W., & Zhao, D. H. (2018). Applying the Convolutional Neural Network Deep Learning Technology to Behavioural Recognition in Intelligent Video. *Tehnicki vjesnik, 25*(2), 528-535. https://doi.org/10.17559/tv-20171229024444

[8] Concolato, C., Feuvre, J. L., Denoual, F., et al. (2017). Adaptive streaming of HEVC tiled videos using mpeg-dash. *IEEE transactions on circuits and systems for video*

technology, *99*, 1-2. https://doi.org/10.1109/tcsvt.2017.2688491

[9] Vasenev, A., Hartmann, T., & Dorée, A. G. (2013). Employing a virtual reality tool to explicate tacit knowledge of machine operators. *Proceedings of the 30th International Symposium on Automation and Robotics in Construction and Mining (ISARC 2013)*, 248-256. https://doi.org/10.22260/isarc2013/0027

[10] Zeng, M. & Zhou, Y. L. (2015). Simulation of pedestrian detection based on deep learning model. *Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 35*(6), 111-116.

[11] Zhang, D. (2017). High-speed Train Control System Big Data Analysis Based on Fuzzy RDF Model and Uncertain Reasoning. *International Journal of Computers, Communications & Control, 12*(4), 577-591. https://doi.org/10.15837/ijccc.2017.4.2914

[12] Zhang, D., Sui, J., & Gong, Y. (2017). Large scale software test data generation based on collective constraint and weighted combination method. *Tehnicki vjesnik, 24*(4), 1041-1049. https://doi.org/10.17559/TV-20170319045945

[13] Huang, K. Q., Chen, Z. T., Kang, Y. F. (2015). Intelligent visual surveillance: a review. *Chinese Journal of Computers, 38*(6), 1093-1118.

[14] Liu, Y. (2017). Video Monitoring System Based on Optical Flow Field Analysis and Deep Learning. *Journal of Xiangnan University, 38*(2), 18-22.

[15] Sumi, L. & Ranga, V. (2016). Sensor enabled internet of things for smart cities. *Fourth International Conference on Parallel.Distributed and Grid Computing (PDGC), Waknaghat,* India, 295-300. https://doi.org/10.1109/pdgc.2016.7913163

[16] Shao, H. & Wang, N. Y. (2018). A new method for moving objects detection in wisdom territory video surveillance. *Modern surveying and mapping, 41*(2), 51-53.

[17] Atan, O., Andreopoulos, Y., Tekin, C., et al. (2014). Bandit framework for systematic learning in wireless video-based face recognition. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing, 9*(1), 180-194. https://doi.org/10.1109/icassp.2014.6853687

[18] Venugopalan, S., Hendricks L. A., Mooney R., Saenko, K. (2016). Improving lstm-based video description with linguistic knowledge mined from text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1961-1966. https://doi.org/10.18653/v1/d16-1204

[19] Dai, Y., Wu, W., Zhou, H.B., Zhang, J., et al. (2018). Numerical Simulation and Optimization of Oil Jet Lubrication For Rotorcraft Meshing Gears. *International Journal of Simulation Modelling, 17*(2), 318-326. https://doi.org/10.2507/IJSIMM17(2)CO6

[20] Dai, Y., Zhu, X., Zhou, H., Z. Mao, et al. (2018). Trajectory Tracking Control for Seafloor Tracked Vehicle by Adaptive Neural-Fuzzy Inference System Algorithm. *International Journal of Computers Communications & Control*, *13*(4), 465-476. https://doi.org/10.15837/ijccc.2018.4.3267

[21] Dai, Y., Zhu, X., & Chen, L. S. (2016). A Mechanical-Hydraulic Virtual Prototype Co-Simulation Model for a Seabed Remotely Operated Vehicle. *International Journal of Simulation Modelling, 15*(3), 532-541. https://doi.org/10.2507/IJSIMM15(3)CO11

[22] Everingham, M., Eslami, S. M. A., Gool, L. V., et al. (2015). The Pascal, Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision, 111*(1), 98-136. https://doi.org/10.1007/s11263-009-0275-4

[23] Kim, H., Kim, J., Oh, T., & Lee, S. (2017). Blind Sharpness Prediction for Ultra-high-definition Video Based on Human Visual Resolution. *IEEE Transactions on Circuits & Systems for Video Technology*, *27*(5), 951-964. https://doi.org/10.1109/tcsvt.2016.2515303

[24] Chorianopoulos, K., Giannakos, M., Chrisochoides, N., & Reed, S. (2014). Open service for video learning analytics. *Proceedings of the 14th International Conference on Advanced Learning Technologies.* Athens, Greece, 28-30. https://doi.org/10.1109/icalt.2014.19

[25] Donahue, J., Hendricks, L. A., Guadarrama, S. (2015). Long-term recurrent convolutional networks for visual recognition and description. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625-2634. https://doi.org/10.1109/cvpr.2015.7298878

[26] Shilbayeh, S. & Vadera, S. (2014). Feature selection in metalearningframework. *Proceedings of the 2014 Science and Information Conference,* London, England, 269-275. https://doi.org/10.1109/SAI.2014.6918200

[27] Li, M. X., Geng, Q. C., Mo, H., et al. (2018). A method of key-frame based video action recognition. *Journal of System Simulation*, 4, 1-8.

[28] Ma, C. Y., Chen, M. H., Kira, Z., et al. (2017). TS-LSTM and Temporal-Inception: Exploiting Spatiotemporal Dynamics for Activity Recognition. *arXiv preprint arXiv*, 1703.10667.

[29] Pesce, M., Munaretto, D., & Zorzi, M. (2014). A Markov decision model for source video rate allocation and scheduling policies in mobile networks. *Proceedings of the 13th Annual Mediterranean Ad Hoc Networking Workshop*, Piran, Slovenia: Institute of Electrical and Electronics Engineers, 119-125. https://doi.org/10.1109/medhocnet.2014.6849113

[30] Kuen, J., Lim, K. M., Lee, C. P. (2015). Self-taught learning of a deep invariant representation for visual tracking via temporal slowness principle. *Pattern Recognition, 48*(10), 2964-2982. https://doi.org/10.1016/j.patcog.2015.02.012

[31] Doulamis, N. & Doulamis, A. (2014). Semi-supervised deep learning for object tracking and classification. *Proceedings of the 2014 IEEE International Conference on Image Processing*. Pairs, France, 848-852. https://doi.org/10.1109/icip.2014.7025170

[32] Karpathy, A., Toderici, G., Shetty, S., et al. (2014). Large-scale video classification with convolutional neural networks. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, USA, 1725-1732. https://doi.org/10.1109/cvpr.2014.223

[33] Hong, S., You, T., Kwak, S., et al. (2015). Online tracking by learning discriminative saliency map with convolutional neural network. *arXiv preprint arXiv*, 150206796.

[34] Zhang, J. M., Ma, S. G., & Sclaroff, S. (2014). MEEM: Robust tracking via multiple experts using entropy minimization. *Proceedings of the 2014 ECCV - European Conference on Computer Vision*, Zurich, Switzerland, 188-203. https://doi.org/10.1007/978-3-319-10599-4_13

[35] Gao, X., Chen Z., Yue, W. J., & Gong, K. (2018). Human Semantic Recognition Model Based on Video Scene Deep Learning. *Computer Technology and Development, 28*(6), 53-58.

[36] Zhang, X. S., Zhang, H. W., Zhang, Y. D., et al. (2016). Deep Fusion of Multiple Semantic Cues for Complex Event Recognition. *IEEE Transactions on Image Processing, 25*(3), 1033-1046. https://doi.org/10.1109/tip.2015.2511585

Contact information:

Lele QIN, Associate Research Fellow (Corresponding author)
School of Economic Management of Hebei University of Science & Technology,
No. 70 East Yuhua Road, Shijiazhuang, Hebei 050018, China
E-mail: Mr_qin@163.com

Lihua KANG, Lecturer
School of Civil Engineering of Hebei University of Science & Technology,
No. 70 East Yuhua Road, Shijiazhuang, Hebei 050018, China
E-mail: 369639442@QQ.com