Research and Implementation of Video-Based Intelligent Analysis System for Children's Behavior

Yuli Min[1], Qingling Meng[2,3(✉)], and Hao Zhang[1]

[1] Zhaotong College, Zhaotong, Yunnan, China
[2] Hulunbuir College, Hailar, Inner Mongolia, China
mengql1019@qq.com
[3] Department of Education, Northeast Normal University, Changchun, Jilin, China

Abstract. Based on convolutional neural network technology and video key frame selection technology, we can complete the construction of human skeleton behavior recognition network of video sequence, and use Kivy framework to realize the intelligent analysis system of children's behavior in Python development environment. Under the online system, children's behaviors in local environment can be identified and analyzed from four functions: video decoding, behavior identification, analysis display and danger warning, which can help preschool teachers and parents to improve their ability to observe and understand children's behaviors, so as to carry out correct words and deeds education and make behavior corrections, avoid the occurrence of dangerous situations, and provide important basis for the educational strategies of children, teachers, educational environment and other aspects for subsequent development. Furthermore, it can further strengthen the communication between parents and teachers of young children, and it is practical and convenient for teachers to fully grasp the individual differences in children's development level, ability, experience, learning style and so on, so as to truly realize teaching students in accordance with their aptitude.

Keywords: Convolutional Neural Network Technology · Key Frame Selection · Human Behavior Recognition · Children Behavior

1 Introduction

Behavior is a variety of attitudes and activities produced by human beings under the influence of various stimuli. Because human beings have both biological and social attributes, human behavior can be divided into instinctive behavior and social behavior. Among them, social behavior is the behavior that people adapt to the surrounding environment, which is established through social processes such as learning, imitating, receiving education and interacting with others. Human's thinking cognition, emotion and conscious will provide direction and motivation for behavior. Similarly, human behavior becomes the explicit expression of human's spiritual world and conscious world. Since birth, in every stage of human growth, the development and change of

C. F. Peng et al. (Eds.): EIMT 2022, AHSSEH 3, pp. 354–363, 2023.
https://doi.org/10.2991/978-94-6463-012-1_40

behavior have shown unique characteristics of each stage. However, early childhood (3–6 years old) has become the key stage of concentrated cultivation of behavior habits and concentrated exposure of behavior problems with its unique age psychological characteristics and physical development level. Therefore, it is of great significance for children's follow-up growth to correctly treat their behavior problems and make a clear judgment on the reasons for their behavior, so as to cultivate good behavior habits, conduct correct words and deeds education for bad behaviors and make behavior corrections.

Children's behavior, as the external expression of children's development level, also meets the educational needs of norms and social requirements. Therefore, it is necessary to observe children's performance in daily life, games, study and work in a purposeful, planned and scientific way, including their words, expressions and behaviors, so as to analyze the rules and characteristics of children's psychological development, facilitate the understanding and recording of children's physical and mental development, and carry out targeted education on children's weak development areas to promote their healthy growth and good development. However, children's behaviors show different characteristics in different situations, and different children have different levels of physical and mental development and growing environment, which makes children's behaviors full of differences and uncertainties [6]. In addition, influenced by the current social and economic situation, cultural development, scientific and technological changes and other factors, children's behavior problems are increasing year by year, which greatly hinders children's normal psychological development and development. Therefore, the observation and analysis of children's behavior has become the premise for preschool teachers to teach students in accordance with their aptitude, and has gradually become a necessary ability for teachers. However, it is not difficult to find some difficulties in practical application, such as the unequal information between parents and teachers, the standardization of children's behavior analysis, the unity of behavior observation methods and the applicability of observation results. In view of this, this paper holds that the intelligent analysis system of children's behavior based on convolutional neural network technology and human skeleton behavior recognition technology in video sequence can realize accurate identification and quantitative analysis of children's behavior, obtain abundant information about children's growth, facilitate teachers to measure and evaluate children and find behavioral problems based on scientific data information, give reasonable educational suggestions in time, actively adjust teaching strategies, continuously optimize the educational environment and finally realize real individualized teaching [4].

2 Related Technical Introduction

2.1 Convolutional Neural Network

Neural network is also called Artificial Neural Networks (ANNs), which is a machine learning method that mimics the structure of human central nervous system. This kind of network relies on the complexity of the system, and achieves the purpose of processing information by adjusting the relationship between a large number of internal nodes. Its structure is shown in Fig. 1. The circle in the figure represents artificial nodes, also called neural units. The whole structure is divided into input layer, output layer and one or more

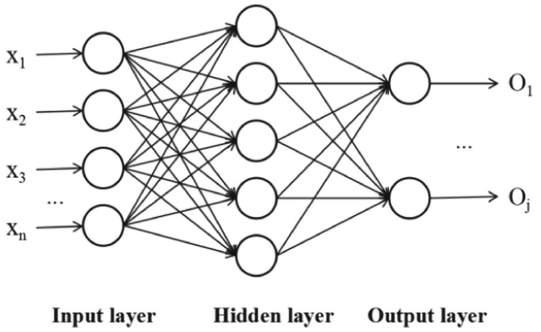

Fig. 1. Schematic diagram of artificial neural network structure

hidden layers. That is, Fig. 1 consists of four input layers of neural units, five hidden layers of neural units and two output layers of neural units. The connection strength of each nerve unit is expressed by weight to ensure that each nerve unit of the n-th layer is interrelated with the nerve unit of the n-1 layer. When the data passes through each unit of the input layer, each unit of the hidden layer is the weighted sum of each unit of the input layer. By analogy, the output layer is the final result calculated by one or more hidden layers. This algorithmic mathematical model of distributed parallel information processing, depending on its complexity, can achieve the purpose of processing information by adjusting the relationship between a large number of internal nodes [9].

Although artificial neural network can deal with complex nonlinear problems, its input layer only accepts vector input, but it has poor processing ability for data with certain spatial structure such as pixels and audio. Convolutional neural network can just make up for this deficiency. Convolutional layer components are used to extract the features of input data, and each element in convolution kernel is used to represent the weights of input nodes and output nodes of convolution layer. The input features of convolutional neural network are divided into subsets of convolution kernel size with nodes as the center. As shown in Fig. 2, convolution operation is performed for 5*5 input feature graph and 3*3 convolution kernel [5]. Through the calculation of convolutional artificial neural network, the key nodes in the picture or video can be represented as a new figure with certain rules, so as to highlight the core content and important features in the picture or video. As for the time information in the video, the 2D convolution method will be extended to 3D convolution, and the subset division of the video content space and time dimensions has been realized.

2.2 Video Key Frame Selection Technology

As a kind of continuous image stream data, video is difficult to directly complete operations such as truncation, segmentation and recognition because of its huge amount of data and unstructured characteristics. The basic unit of video consists of frames, and each frame can be understood as transforming video into a still image. In order to effectively utilize the time domain information of video sequences and reduce the computational complexity, before processing video sequences, it is necessary to extract the key frames

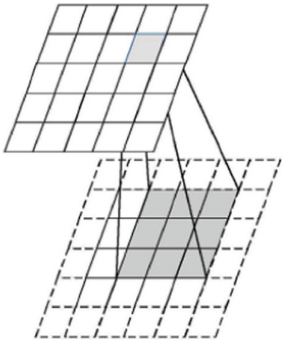

Fig. 2. Schematic diagram of convolutional operation

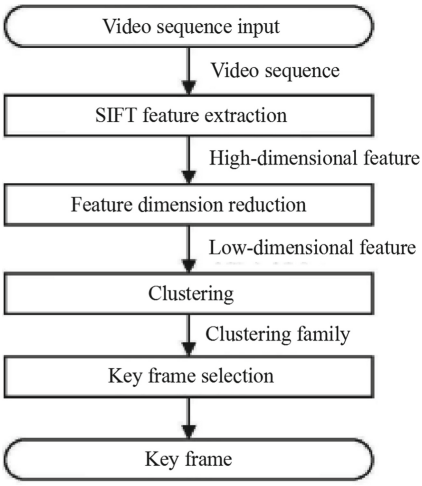

Fig. 3. Video key frame extraction process based on clustering

of video sequences, that is, the highly concentrated reflection of video content in a shot. Figure 3 shows the flow chart of key frame extraction algorithm based on clustering. Firstly, SIFT features are extracted from all frames of video content, that is, local features of images. SIFT features can remain unchanged under the conditions of image rotation, scaling, brightness change, viewing angle change, etc., and are rich in information, so as to facilitate fast and accurate matching. Then PCA algorithm is used to reduce the feature dimension, and finally clustering algorithm is used to select key frames from the obtained results.

2.3 Human Behavior Recognition Technology

Human behavior recognition is the key frame extraction of the previous step. There are two parts in the process of human behavior recognition. The first part is the detection of human key points based on CPN (Cascaded Pyramid Network) structure. The second

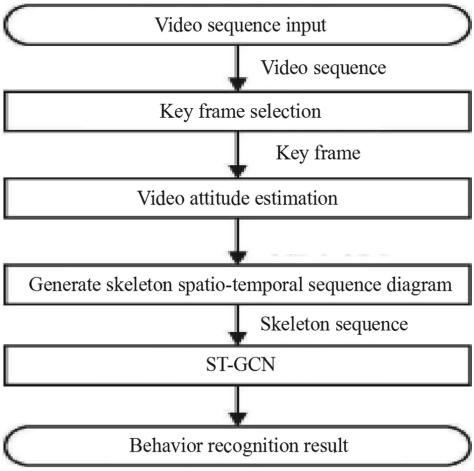

Fig. 4. Algorithm flow of human skeleton behavior recognition

part is the human behavior recognition network ST-GCN based on skeleton sequence spatiotemporal map. The CPN framework adopts the top-down detection strategy. First, the pedestrian detection framework is used to detect the candidate frame. Then, the GolbalNe module is used to detect the key points of human body, such as eyes, shoulders, elbows, knees, ankles and other parts, for each detected character candidate frame, and at the same time, 1 * 1 image convolution operation is performed. The RefineNet module can repair the key points with occlusion, unclear or complicated background and complete the result output [8]. As shown in Fig. 4, after the key points are detected, the spatio-temporal sequence diagram of human bones is generated according to the posture estimation, and the local features of adjacent joints in space are completed by graph convolution neural network (GCN), and then the temporal features need to be superimposed to determine the local features of joint changes in time. Time convolution (TCN) is selected in ST-GCN. Because the shape of characters in key frames is basically fixed, the traditional convolution layer can be used to complete the time convolution operation, that is, 3D convolution can complete the processing of skeleton spatio-temporal sequence diagram.

2.4 Development Environment

According to the design requirements of the intelligent analysis system for children's behavior, the development environment of the system has certain requirements for hardware equipment, in which the CPU of the system selects Intel Core i7-9700H and the graphics card selects two GTX1080 graphics cards to ensure the system's ability to handle a large number of video files. As for the software system, Ubuntu 16.04 under Linux is selected as the underlying operating system, the overall development environment is VSCode, Python 3.6.1 is selected as the development language, and Python 1.1 is selected as the deep learning framework. In the training stage of the system, the NTU RGB+D data set XSUB protocol is selected to train and verify the skeleton behavior

```
workon torch
python -m pip install docutils pygments pypiwin32 kivy _deps.
sd2==0.1.22 kivy_deps.glew-=0.1.12
pyhon -m pip insll kivy deps gstreamer==0.1.17
pyton -m pip istall kivy=.1.11.1
pyhon -m pip isall kivy_examples==1.11.1
pytbon torch\share\kivy-examples\demo\showcas\main.py
```

Fig. 5. Kivy framework installation and deployment key code

recognition network based on video sequence by using 44,810 samples containing only one person, so as to ensure the recognition accuracy of video content in the subsequent production and use of the system [2]. The development architecture of the system chooses Kivy framework, which is light in volume and simple in configuration and deployment. As shown in Fig. 5, it is the key code for installation and deployment, that is, after activating torch virtual environment, complete the installation dependency of Python, and then complete the installation of Kivy and Kivy routines.

Compared with the traditional image interface design framework, Kivy framework has better cross-platform performance, and users can get a good experience on both personal PC and mobile devices. In addition, Kivy also supports the separation of interface design file and implementation code, and directly imports the design file from the code, which makes the code clearer.

3 Requirements Analysis

3.1 System Requirements Analysis

The system supports preschool teachers' users to complete the analysis and identification of children's behaviors and the judgment of children's behaviors and habits through content analysis of video content in the form of online monitoring, offline caching, equipment shooting, etc., so as to further deepen their comprehensive understanding of children's information, and based on this, timely adjust children's education strategies to realize special education and guidance. At the same time, it can strengthen the communication and exchange between preschool teachers and parents, improve the inequality of preschool education development information, and promote the healthy growth of preschool children in the school and family environment.

The functional requirements of the system are mainly divided into four parts, namely, video decoding, behavior recognition, analysis display, and danger warning. The video decoding function can decode and play video files of various formats, such as AVI, MP4, MPG, DAV and MOV. The behavior recognition function can recognize children's behaviors during video playback, and distinguish normal behaviors from dangerous behaviors. Under the analysis display function, the original video and the comparison results of behavior recognition analysis can be displayed intuitively. However, the danger warning function can give corresponding tips according to the results of behavior identification and analysis.

The overall design of the system needs to follow the principle of accuracy, and the accurate identification of children's behavior is the basis of follow-up analysis and

targeted education. At the same time, the system also needs some real-time functions to realize the identification, analysis and feedback of online surveillance video. In addition, in the process of practical application, in order to effectively overcome the factors such as illumination, brightness change, complex background or foreground, picture vibration and jitter, the robustness of the system needs to be increased to maintain the stability of the analysis and recognition performance of the system [7].

3.2 Global Design

Children's behavior intelligence analysis system adopts C/S architecture as a whole, that is, desktop application system. The whole system is developed in Python language, and PyCharm is selected as the editor. In the overall layout of the main functional interface of the system, based on Kivy Widget class, the design and development of the basic framework of the system are completed with KV files. Then, each function module is deployed under the function interface, and the display area at the top of the page is the analysis display function module, which can display the original video file and the behavior identification analysis diagram at the same time. The middle part is the operation area, which provides buttons for kindergarten teachers to perform specific operations, such as playing, pausing, stopping, and the results of behavior recognition analysis. The lower end is the setting area, which can realize the setting of some functional parameters of the system. The overall operation and operation flow of the system is shown in Fig. 6, in which the video decoding module is responsible for the data input of the overall system, and the video file connection or local address can be obtained directly through the text box on the system interface. The behavior recognition module is the core of the system, which samples the video according to the sampling rate set by the system to obtain key frames and complete the subsequent behavior recognition analysis. Danger warning can perform corresponding actions according to the results of behavior identification analysis, distinguish daily or normal behaviors, bad behaviors and dangerous behaviors, and record them. When the frequency of dangerous behaviors cumulatively exceeds

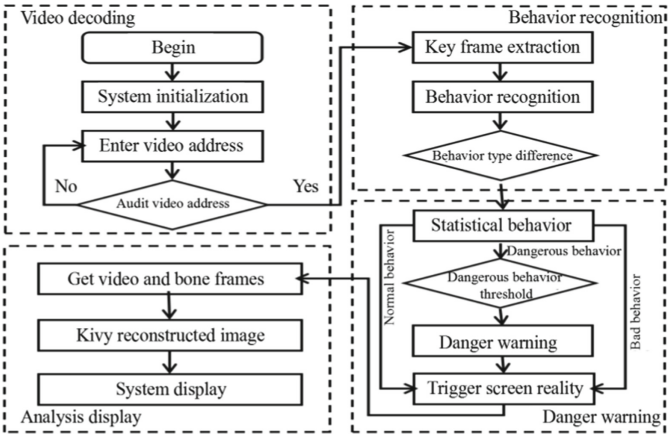

Fig. 6. Overall operation and operation flow of intelligent analysis system for children's behavior

the critical threshold of the system, an early warning signal will be sent to judge the occurrence of dangerous behaviors. The display module will complete the result and content of key frame behavior recognition and refresh according to the video stream.

4 Functional Implementation

4.1 Video Decoding

Under this function module, the user can input the URL address of the video source through the text box, including the server IP address, signal port, file name and other information. The system will automatically check the input address, and if it fails, it needs to be re-entered and checked again. After the system is approved, the stream of online monitoring video file is acquired, and the VideoCapture module of OpenCV is called by Python code to complete video decoding and playing. For the video files that are cached offline or shot by mobile devices, you can play them by directly entering the physical address of the video in the text box. In this system, the function of text box is completed by TextInput control under Kivy framework. After the user completes the input, the on_text_validate event can be triggered by clicking or pressing the Enter key with the mouse to obtain the input content.

4.2 Behavior Recognition

Under this function module, the system will extract the key frames of online monitoring video and local video files by using two behavior recognition methods for two different video files according to the video decoding module. For local or offline video files, the video key frames are obtained by the video key frame selection method based on clustering, and the key frames are used instead of the original video files to identify children's behavior. For online surveillance video, it is necessary to cut out the corresponding character scenes through OpenCV, complete video sampling in these scenes, and complete behavior recognition once every 16 frames [3]. The recognition process is to detect the key points of children's human posture in key frames through CPN network, so as to form the spatio-temporal sequence diagram of human bones, and to complete the behavior recognition of children's human skeleton sequence Shi Kongtu by ST-GCN. For example, in early childhood, children aged 2–3 years old are a little clumsy, grabbing, scratching, biting and other behaviors, and depending on toys and articles. Children around 3 years old have social attributes in their behaviors, have preliminary behaviors of daily life, play and study, enhance the overall coordination of behaviors, have large-scale movements such as running and jumping, and walk naturally and rhythmically. The behavior of children aged 4–5 years is mainly manifested in the obvious improvement of self-care and labor behavior. They are flexible and have enhanced control ability, their walking speed is basically the same as that of adults, their limb balance ability is greatly improved, they can have more complicated limb movements or adopt certain sports skills, and clearly distinguish the differences of their own behaviors in work, games and sports.

4.3 Analysis Display

Through behavior identification, children's behaviors are simply divided into three categories: normal behaviors, bad behaviors and dangerous behaviors. Under the analysis display function, Kivy displays the Video or Image files read by OpenCV through the image and video controls. There are two types of videos: RGB video and skeleton action video, which represent the original video sequence and the human skeleton behavior video sequence generated by the pose estimation network respectively. Under the Kivy framework, Texure module is used to reconstruct the video sequence, and then the content is displayed in the display area. According to the displayed content, the user completes the analysis of children's behavior, such as the thinking characteristics of children of different ages, the correlation between emotion and behavior, the psychological growth and the development of social attributes. In the face of some bad behaviors of young children, such as poor self-care ability, sluggish behavior and hyperactivity. This situation should be communicated with parents of young children in time, and corresponding solutions should be put forward, which should be implemented simultaneously in both family and school environments to correct the bad behavior of young children.

4.4 Danger Warning

Under this function module, the system will complete the judgment distribution and early warning of dangerous behaviors. For local video, the system generally does not turn on the warning behavior, mainly for online surveillance video. When dangerous behaviors occur in monitoring, the system will automatically record the names and times of behaviors, and set the warning threshold to control the error rate of the warning system [1]. When the number of behaviors exceeds the threshold value, the system automatically sends out an early warning of dangerous behaviors, so that teachers can take timely measures to protect and help young children. Common dangerous behaviors of young children are: falling, twisting caused by physical pain or spasm, aggressive behavior.

5 Conclusions

The intelligent analysis system of children's behavior realizes the identification and analysis of children's skeletal behavior by convolution neural network technology and video key frame extraction technology. The system can select the key frames of the online surveillance video and local video files, and on the basis of human posture evaluation and the spatio-temporal sequence diagram of skeletal behavior, it can recognize children's behavior by ST-GCN, and analyze and apply it according to the recognition results. With the help of the current high-tech forces, we can further improve our ability to observe and understand children's behaviors with comprehensive data and information. It is convenient for teachers to fully grasp the individual differences in children's development level, ability, experience, learning style, etc., better teach children in accordance with their aptitude in early childhood education, and help children grow up healthily. At the same time, it is also a new attempt to further promote the reform of educational informatization.

Acknowledgements. Fund Project: The 2022 Yunnan Provincial Department of Education Scientific Research Fund Project "Research on the Development Dilemma and Optimization Path of Zhaotong Rural Preschool Education under the Background of Rural Revitalization", project number 2022J0963; 2021 Inner Mongolia Autonomous Region Hulunbuir College Discipline Construction Special Project "From the Perspective of Transformation and Development" "Research on the Construction of 'New Liberal Arts' for Preschool Education Majors in Colleges and Universities", Project No. 2021XKPT035.

References

1. Fei Fan. 2014. *Research and implementation of automatic detection and recognition algorithm of abnormal behavior of moving human body in intelligent video surveillance.* Nanjing University of Posts and Telecommunications.
2. Li Chaolong. 2019. *Research on human motion recognition based on graph convolution neural network*. Southeast University.
3. Liu Chen. 2021. *Design of human behavior identity for intelligent nursing robot*. Jiangnan University.
4. Ma Limin. 2020. *Significance and methods of teachers' observation and analysis of children's behavior*. Survey of Education.
5. Sheng, Wenshun, and Yanwen Sun. 2019. Application of convolution neural network in image recognition. *Software Engineering.*
6. Wang Jing. 2019. *Behavior education and guidance exploration framework for children.* Hundred Prose Schools (New Chinese Backpage).
7. Wu Yanchun. 2019. *Online human motion analysis based on deep learning.* University of Jinan.
8. Xiao Tianzi. 2019. *Research and implementation of human key point detection algorithm based on video features.* Beijing University of Posts and Telecommunications.
9. Zhang Chi et al. 2021. Review of development and application of artificial neural network model. *Computer Engineering and Applications.*