The 14th International Conference on
Computer Science & Education (ICCSE 2019)
August 19-21, 2019. Toronto, Canada

A Students' Action Recognition Database In Smart Classroom

Xiaomeng Li
Depaetment of Educational Technology
Ocean University of China
Qingdao, China
lxm@stu.ouc.edu.cn

Min Wang
Depaetment of Educational Technology
Ocean University of China
Qingdao, China
1797255014@qq.com

Wei Zeng
Depaetment of Educational Technology
Ocean University of China
Qingdao, China
961549050@qq.com

Weigang Lu*
Depaetment of Educational Technology
and Department of Computer Science
and Technology
Ocean University of China
Qingdao, China
luweigang@ouc.edu.cn

***Abstract*—With the development of human action recognition, it is possible to automatically recognize students' actions in classroom, providing a new direction for classroom observation in teaching research. Training effective students' action recognition algorithms depends significantly on the quality of the action database. However, only a few existing action databases focus on learning environment. In this paper, we contribute to this topic from two aspects. First, a novel students' action recognition database is introduced. The spontaneous action database consists 15 action categories, 817 video clips of 73 students, which are collected in real smart classroom environment. Second, a benchmark experiment was conducted on the database using two kinds of recognition algorithms. The best result is achieved by Inception V3 with 0.9310 accuracy. Such a spontaneous database will help in the development and validation of algorithms for action recognition in learning environment.**

***Keywords*—database，smart classroom, action recognition, IDT, CNN**

I. Introduction

In advanced pedagogical methods, students are taken as the principal part and the center of the class, where the students' actions imply their learning state. For example, compared with using phone, hands up often reflect a higher engagement, it has been proved that student engagement contributes to more favorable outcomes for college students [1]. Observing and analyzing students' actions provide data for educational research. However, the traditional observation methods, such as scales and case study, is time consuming. Since the rising of computer-based qualitative analysis software, NVivo has been widely used to analyze video data recorded in classroom. Although using NVivo speeds up data management and analysis tasks, it still requires manually code data, far from the automatic recognition in classroom observation.

Action recognition is an increasingly popular research direction in computer vision with considerable applications in various fields. Numerous action databases ranging in size, scope and purpose are available for researchers in action recognition. In the early stage, many public basic human action databases have been published [2][3], capturing the most common daily actions in experimental environment. Over the years, especially the rapid development of deep learning, researchers have introduced much larger database [4][5][6], most of which are recorded in uncontrolled manner in real environment or download from the web or movies. These databases provide an open challenge to researchers to evaluate the accuracy of different approaches. Acquisition of an action database in particular location has advantages for certain areas of research, giving the researchers a direct control over the parameters of variability in the database. However, no adequate databases exist that provide students' spontaneous action videos of multiple views. The databases are either limited to frontal views, or have been acquired under controlled conditions. Hence, we introduce a novel database to overcome these limitations.

Due to the recent advancement in smart classroom, big digital-data are produced every day. Take Ocean University of China as example, 28 smart classrooms have been put into operation since April 2018 [7], such massive information's monthly data scale always achieves TB level. Analysis of students' behaviors in classes is becoming a trend in educational informatization. In order to make fully utilization of these video data, it is always the first step to build a students' action recognition database and provide a benchmark.

In this paper, a novel spontaneous students' action database is introduced, whose purpose is to provide training and testing data for action recognition algorithms in learning environment. Every action in this database is collected from real smart classroom environment, meaning that students' actions are spontaneous and authentic. Due to the novel design of the round table in the smart classroom, cameras can capture actions at all angles, thus provided a more real training and testing data for the development of action recognition in classroom scenarios. Access to the database can be requested by emailing the corresponding author.

II. Related Work

After years of efforts, several action databases have been proposed, some of those are widely used in baseline recognition. Early in its development, most of the databases were built by recording in a controlled manner, static background and environmental conditions, with few numbers of action classes. Weizmann [2] dataset introduced in 2005,

978-1-7281-1846-8/19/$31.00 ©2019 IEEE

is the most common benchmark for human action recognition. It was recorded with static background and fixed camera, all actions were acted by 9 actors, performing 10 basic human actions such as walking, running, and jumping, etc. Another commonly used database is KTH [3]. Compared with Weizmann, it has more difficulty in recognition due to the change of environment and illumination condition which consists of 6 activities performed by 25 actors, including walking, hand waving, and hand clapping, etc. Handcraft features are commonly used in training these databases, the most effective among which is Improved Dense Trajectory (IDT) with Fisher Vector (FV) [8], extracting local features from motions. However, one problem of IDT is the high computation complexity, limits the application for real-time recognition.

With the continuous improvement of algorithms, some of the highest recognition accuracies achieved nearly 100 percent in existing databases [9][10]. Compared with the algorithms, these databases are less changeling due to less action classes, subjects and controlled recording environmental conditions. Novel databases were proposed, represented more realistic action scenes, as well as cover larger application domains. Much effort has been devoted to the collection of realistic Internet video clips. HMDB51 [5] contains over 7000 clips collected from web sources, better capturing the diversity and complexity of human action. In addition to recorded databases. YouTube videos and movies are the most common data sources. UCF [11][12] and Hollywood [4][13] series database are typical of them, provided rich varieties of camera motion, view point, and video quality. Although those large databases have provided diverse human actions, only a little part of data related to educational field. YouTube 8M [6] contains a category named Classroom, it is the only category related to educational field, while no subclass is offered. In these large-scale databases, deep learning solutions have better performance than classical algorithms, using deep features is now becoming a trend in human action recognition. The majority of deep learning architectures are based on CNN due to its power of robust low to high features extraction from raw input data signal

It can be observed that no unique solution can apply to all application scenarios. To promote the application of the action recognition in the educational field, it is necessary to pay attention to introduce action database which concentrates on students' behavior. To the best of our knowledge, there are only few databases that focus on this field. One recent effort is BNU-LSVED2.0 [14], contains 2117 image sequences, 11 types of body language and facial expressions recorded in classroom environments, including tiny eye movement, hold the chin, finger touch, etc. All image sequences were annotated with categorical emotion labels, affective labels, and multi-dimensional Pleasure-Arousal-Dominance (PAD) labels. However, the database was recorded in traditional classes, focused on frontal views of actions, therefore do not captured real-world problems. With the implementation of information-based teaching, more and more smart classrooms are thrown in construction. While its cameras record big digital data every day, it is necessary to build a database to help train recognition methods specially for action recognition in smart classroom scenarios.

III. The Students' Action Recognition Database

In this section, datasets and implementation details in data collection and annotation will be introduced. Summary information about the database are provided in Table I.

TABLE I. DATABASE SUMMARY

Database Specifications	
Number of videos	817
Number of categories	15
Number of participants	73 (32 males, 41 females)
Clip duration	1-6 sec
Clip selection	Manual

A. *Data Source and Subjects*

Since the genuine and spontaneous actions outperform posed or stimulated ones in action recognition experiment, a smart classroom with 4 fixed cameras was chosen as the data source, capturing 12 classes per week. Fig. 1 shows the model diagram of the smart classroom from the top view. Classes are taken in both day and night, so the indoor illumination includes two conditions: sunlight and indoor light. Each camera has a resolution of 1280*720, frame rate of 30 fps. Cameras can move horizontally and vertically and rotate 360º, which makes the database a better reflection of students' activities.

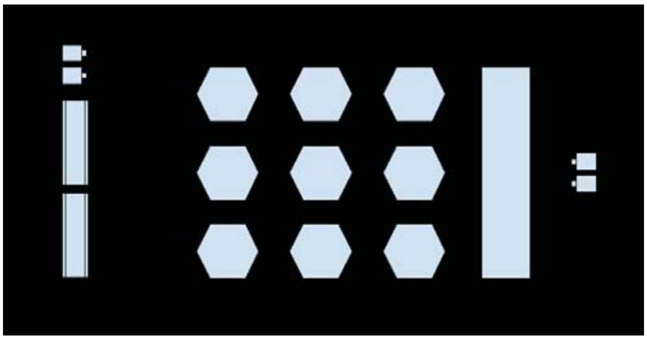

Fig. 1. Top view of the classroom

73 healthy college students, 32 males and 41 females, participated in the experiment. The students were of the age group ranging from 18 to 27 years and from different majors of the university. The database only includes the images of the students who signed the consent form agreeing to their action videos and images being used for research purpose. Of them, 54.7 percent wore glasses, 23.2 percent had facial hair. Because of the change of seasons, students' clothes changed from T-shirt to overcoat. Besides, the smart classroom employs several round tables, instead of "rows of seats and tables facing forward", to facilitate student interaction, which means the database has large variation in camera viewpoint. The differences within database have been proved to be very helpful to algorithm's robustness [15].

B. *Data Collection and Annotation*

The whole process of obtaining single action video clips is shown in Fig.2. In order to collect students' actions that represent everyday classroom behavior, firstly, the segment is started by watching videos from daily classes, annotating any video clips that can represent a single non-ambiguous action. Especially those actions related to learning environment or particularly arise in smart classroom. Take categories Chat and Talk as examples, in general college classroom, few students show their point of view, whereas in modern teaching idea, sharing opinions is believed to benefit learning effect [16]. Collecting these kinds of data contribute to specific application scenarios.

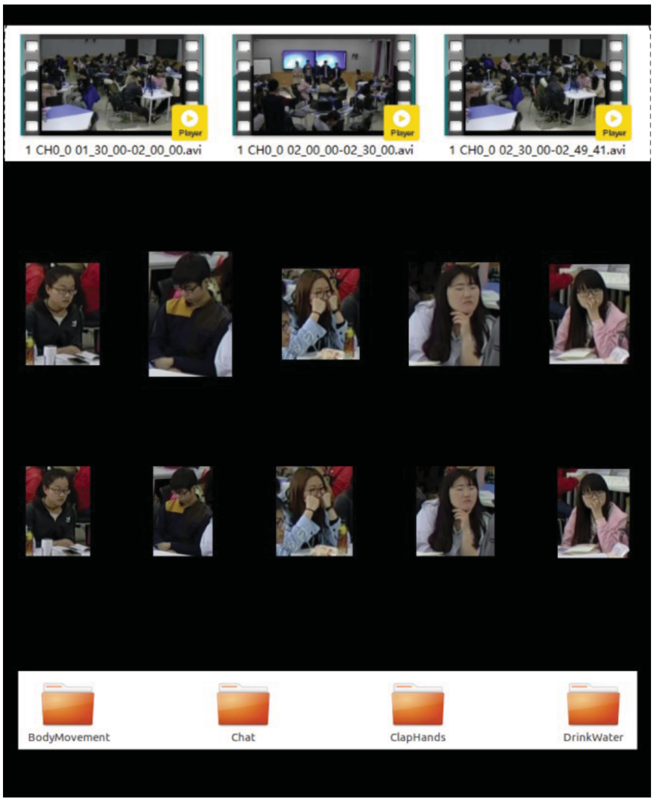
Player
1 CH0_0 01_30_00-02_00_00.avi
1 CH0_0 02_00_00-02_30_00.avi
1 CH0_0 02_30_00-02_49_41.avi
BodyMovement
Chat
ClapHands
DrinkWater

Fig. 2. Process of data collection

In order to control the quality of single clip, each clip is limited by a minimum of 60 pixels in height for the main student, minimum contrast level, minimum 1 second of clip length and acceptable compression artifacts. The segment used FormatFactory 4.5.0.0.

After the first step, an original database was generated with 18 action categories of students' common actions in learning environment. Since all the spontaneous actions were collected in a real learning environment rather than simulated by induction videos, it is commonly seen in original videos where actions are expressed continuously, shown in Fig 3. In this case, a single clip was cut into two or more actions, while the first action keeps the connection with latter action. Then the multiple actions in one video clip were cut into two non-ambiguous clips. It is because this kind of condition is quite normal in real-life situations, while our database is aimed at studying action recognition in realistic, unconstrained environments.

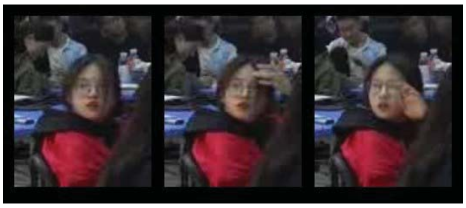

Fig. 3. Connection in actions

To reduce the imbalance of video clips' number, some action categories in small number were forsaken, such as pick up books. Besides, for some actions in a small amount which were considered to appear frequently, some more extra videos were segmented and annotated, such as write and Use laptop. Finally, a database includes 15 categories was established, each action category has at least 26 clips, less than 6*s* long. The categories and numbers of video clips are shown in Table II.

TABLE II. NUMBER OF VIDEO CLIPS OF EACH CATEGORY

Action Categories	***Number***
Drink water	26
Push glasses	27
Touch hair	31
Hand up	35
Chat	38
Nod	41
Write	44
Use laptop	52
Clap	59
Look up	66
Turn around	67
Stand up	71
Sit down	74
Talk	83
Use phone[a]	103
Total	**817**

[a] Use phone includes using phone and tablet.

In order to provide a more precise evaluation of existing methods, each clip was annotated with meta information except for action category labels. The meta information comprises viewpoint and visible parts. Viewpoint describes the camera's position relative to the students, including front, back, left, and right. Visible body parts include head, upper body, lower body, and the full body is visible (denoted as upper, lower, full). For example, the first action from category drink water recorded from left, upper body visible is annotated as DrinkWater_01_left_upper.

C. *Video Normalization*

The video clips extracted from original videos differ in size. In order to ensure the consistency across the database, all clips were scaled to 240 pixels, the width was scaled accordingly to maintain the original aspect ratio. All the video clips were compressed by the DivX 10.1 codec with the ffmpeg video library.

D. *Features*

For the specialty of record equipment and environment, the database is characterized by the followings:

- Recorded in unconstrained real smart classroom environment, there is much diversity in viewpoint, lighting, clutter and occlusion.
- The database contains various actions which former databases aiming in learning environment do not include, such as use laptop and clap, which are actually commonly seen in classes.
- The cameras are fixed on the wall and the tables are designed round for students to freely communicate and discuss in class, the special arrangement ensures the cameras can record student's actions from several angles, consequently obtaining more comprehensive action information and features.

- The collection process lasted for a long period of time, students' dress, varies from overcoat to T-shirt, providing another changing factor in classification.
- Since communication and discussion are one of the most common parts in student-centered classes, the database includes one action category named Chat to cover more human interaction.

All the above features enable the database reflect the real environment better in a more comprehensive perspective, meanwhile, more difficult for the current algorithm to classification.

IV. Experiment and Results

Action recognition has been an active research area for the past years. A recent review over different databases and methods can be found in [17]. After analyzing how a set of methods that are already established for action recognition in other databases can be applied to our database, 4 different combinations of algorithms are used to evaluate the usability of the database. One is IDT based on the extraction of local space-time information from videos, which performed better than other handcrafted feature representation methods. For classification, K-nearest neighbors (KNN) and Support Vector Machine (SVM) combined with IDT are implemented. Another is Convolutional Neural Network (CNN), the heart of the deep architecture. Two implementations of CNN are used, VGG-16 and Inception V3 respectively. Deep learning approaches have been proved a good choice in industrial and commercial applications for the past two decades. Currently, human action recognition in videos with deep features becomes a trend in computer vision.

The training process is performed on a workstation with an Intel Core i5-4590 CPU, NVIDIA GeForce GTX TITAN X GPUs and 32GB RAM.

A. IDT Combined With SVM and KNN

IDT representation was proposed in [18] that achieved breakthrough in hand-crafted features. By computing the dense optical flow and densely sample feature points, it extracted dense trajectories and computed trajectory-aligned descriptors. The description of the feature descriptors computed are as follows:

1) HOG: The histogram of oriented gradient (HOG) [19] was first proposed for human detection and subsequently applied to many visual tasks.

2) HOF: Reference [13] introduced the Histogram of Optical flow (HOF) spatiotemporal descriptor, a 3D flowbased version of HOG, which gave a robust feature to localise the action motion.

3) MBH: The MBH (Motion Boundary Histograms) [20] feature calculates a histogram of the optical flow image gradient, and calculates optical flow information in the X and Y directions of the optical flow image, respectively.

After extracting features, FV encodes both first and second order statistics between the video descriptors and a Gaussian Mixture Model (GMM). Evaluations show an improved performance over bag of features, an often-used encoding method, for both image and action classification [20]. The features encoded by FV were then fed into the following classifiers:

1) Linear SVM [21]: As previous experiments in different databases have shown that this type of SVM performed superior for action recognition than its polynomial and radial basis function-based variants. The features are standardized before SVM classification to obtain better results. The parameters of the linear kernel were $C \in \{0.1,1,10,100,1000\}$ $C \in \{0.1, 1, 10, 100, 1000\}$, using 10-fold cross-validation on the training set.

2) KNN: KNN is classified by measuring the distance between different features, for which preliminary tests have shown that K=3 and feature standardization gave the best results.

B. CNN Implement By VGG-16 and Inception V3

VGG-16 [22]is employed as one of the implementations of CNN, using Keras, TensorFlow backend. It is adopted due to its promising results in various vision tasks and is applied to represent individual frames and average the obtained features for each frame in the video to form the final encoding. The architecture is pre-trained on ImageNet [23].

Inception V3 is employed as another implementation for CNN, using Keras, TensorFlow backend. Inception V3 is the third generation of GoogleLeNet [24]. Although its network topology is more complicated, the character of parameter amounts, memory, and computing resources are much smaller due to its special design pattern comparing with the traditional network, or even the same period. It is implemented by classifying one frame at a time, ignoring the temporal features of video and attempting to classify each clips with in one single frame, pre-trained on ImageNet. Transfer learning is employed to retrain Inception on our data. we fine-tuning the top dense layers for 10 epochs at 10240 images per epoch to retain as much of the previous learning as possible.

For deep learning methods always requires a large amount of data, data augmentation (DA) plays an important role in CNN training. In order to make the most of our training examples, data augment was done by a number of random transformations. In the experiment, random rotation, zoom, sheer, fill are employed. By DA, CNN model would never see twice the exact same picture. This helps prevent overfitting and helps the model generalize better. Examples of the result of data augmentation are shown in Fig. 4.

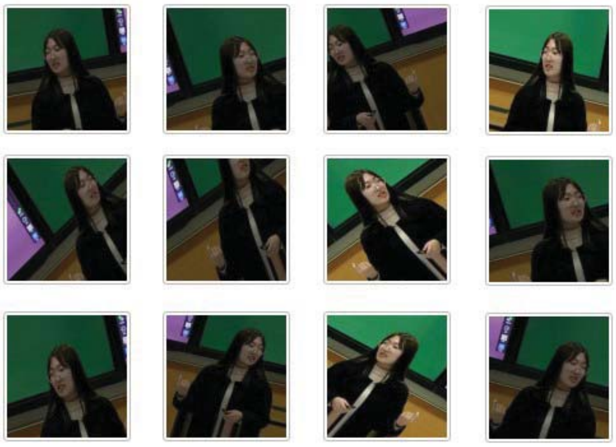

Fig. 4. Examples of data augmentation

C. Results

The performances of different algorithms are shown in Table III, where the accuracy for each algorithm is provided.

TABLE III. RESULTS OF DIFFERENT METHODS

Method	***Accuracy***
IDT+SVM	71.63%
IDT+KNN	53.66%
VGG-16	91.95%
Inception V3	**93.10%**

As observed from Table III, the best accuracy was achieved by Inception V3. VGG-16 also achieved relatively good performance. However, Inception V3 has much higher computational efficiency than VGG-16, due to its fewer parameters, suitably factorized convolutions, and aggressive regularization. The graph of accuracy using Inception V3 is shown in Fig. 5, using relative time as the horizontal axis, default 0.6 smoothing. It can be observed that training stopped early before the training loss not improved to speed up the learning process. This is due to the small scale of the database, EarlyStopping was employed to stop training callbacks early.

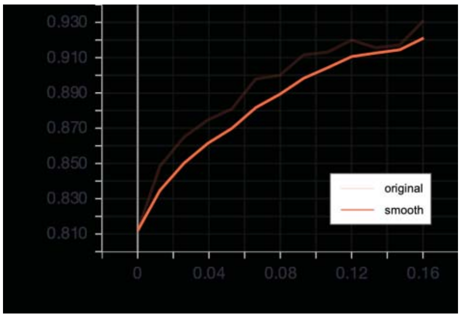

Fig. 5. The accuracy using Inception V3

Moreover, experiment results show that based on same features extracted by IDT, SVM significantly outperformed KNN in action classification. The poor performance of KNN in our database was probably because of the imbalanced dataset. In addition, the computational cost of KNN is relatively high compared to SVM, in real-time applications, SVM may be used since its computational complexity is low to other classifiers.

V. Conclusion and Future Work

In this paper, a novel spontaneous students' action recognition database recorded in real smart classroom environment is introduced for further training and testing of different action recognition methods. It has been described in details of the database's video clips acquisition procedure and its contents. Afterwards, different recognition methods were used to evaluate database baseline.

In the near future, we will establish a full action recognition pipeline for students' action recognition based on our database, and evaluate its capabilities by employing it for educational purpose.

Acknowledgment

This work has been supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 16YJC880057).

References

[1] Carini R M, Kuh G D, Klein S P. Student engagement and student learning: Testing the linkages[J]. Research in higher education, 2006, 47(1): 1-32.

[2] Blank M, Gorelick L, Shechtman E, et al. Actions as space-time shapes[C]//Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. IEEE, 2005, 2: 1395-1402.

[3] Laptev I, Caputo B. Recognizing human actions: a local SVM approach[C]//null. IEEE, 2004: 32-36.

[4] Marszałek M, Laptev I, Schmid C. Actions in context[C]//CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2009: 2929-2936.

[5] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition[C]//2011 International Conference on Computer Vision. IEEE, 2011: 2556-2563.

[6] Abu-El-Haija S, Kothari N, Lee J, et al. Youtube-8m: A large-scale video classification benchmark[J]. arXiv preprint arXiv:1609.08675, 2016.

[7] Ocean University of China smart classroom put into trial operation. (2018). Retrieved from http://www.ouc.edu.cn/dd/38/c10639a187704/page.htm

[8] Sánchez J, Perronnin F, Mensink T, et al. Image classification with the fisher vector: Theory and practice[J]. International journal of computer vision, 2013, 105(3): 222-245..

[9] Wang P, Li W, Gao Z, et al. Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring[C]//Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015: 1119-1122.

[10] Wang P, Li W, Gao Z, et al. Action recognition from depth maps using deep convolutional neural networks[J]. IEEE Transactions on Human-Machine Systems, 2016, 46(4): 498-509.

[11] Reddy K K, Shah M. Recognizing 50 human action categories of web videos[J]. Machine Vision and Applications, 2013, 24(5): 971-981.

[12] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

[13] Laptev I, Marszałek M, Schmid C, et al. Learning realistic human actions from movies[C]//CVPR 2008-IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2008: 1-8.

[14] Wei Q, Sun B, He J, et al. BNU-LSVED 2.0: Spontaneous multimodal student affect database with multi-dimensional labels[J]. Signal Processing: Image Communication, 2017, 59: 168-181.

[15] Wang H, Ullah M M, Klaser A, et al. Evaluation of local spatio-temporal features for action recognition[C]//BMVC 2009-British Machine Vision Conference. BMVA Press, 2009: 124.1-124.11.

[16] Gilboy M B, Heinerichs S, Pazzaglia G. Enhancing student engagement using the flipped classroom[J]. Journal of nutrition education and behavior, 2015, 47(1): 109-114.

[17] Singh T, Vishwakarma D K. Video benchmarks of human action datasets: a review[J]. Artificial Intelligence Review, 2018: 1-48.

[18] Wang H, Oneata D, Verbeek J, et al. A robust and efficient video representation for action recognition[J]. International Journal of Computer Vision, 2016, 119(3): 219-238.

[19] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//international Conference on computer vision & Pattern Recognition (CVPR'05). IEEE Computer Society, 2005, 1: 886-893.

[20] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103(1): 60-79.

[21] Vapnik V, Vapnik V. Statistical learning theory Wiley[J]. New York, 1998: 156-160.

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[23] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[24] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.