IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 23, 2019, accepted March 4, 2019, date of publication March 11, 2019, date of current version March 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2904046

Automatic Knowledge Discovery in Lecturing Videos via Deep Representation

JINJIAO LIN[1,2], CHUNFANG LIU[1], YIBIN LI[2], LIZHEN CUI[3], RUI WANG[1], XUDONG LU[3], YAN ZHANG[4], AND JIAN LIAN[4]

[1]School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250000, China
[2]School of Control Science and Engineering, Shandong University, Jinan 250061, China
[3]School of Software Engineering, Shandong University, Jinan 250061, China
[4]Department of Electrical Engineering and Information Technology, Shandong University of Science and Technology, Jinan 250000, China

Corresponding author: Jian Lian (lianjianlianjian@163.com)

This work was supported in part by the Teaching Reform Research Project of Undergraduate Colleges and Universities of Shandong Province under Grant Z2016Z036, in part by the Teaching Reform Research Project of Shandong University of Finance and Economics under Grant jy2018062891470, Grant jy201830, and Grant jy201810, in part by the Shandong Provincial Social Science Planning Research Project under Grant 18CHLJ08, in part by the Scientific Research Projects of Universities in Shandong Province under Grant J18RA136, in part by the SDUST Excellent Teaching Team Construction Plan under Grant JXTD20160512, in part by the Jinan Campus of SDUST Excellent Teaching Team Construction Plan under Grant JNJXTD201711, in part by the SDUST Young Teachers Teaching Talent Training Plan under Grant BJRC20160509, in part by the Teaching Research Project of Shandong University of Science and Technology under Grant JNJG2017104, and in part by the Scientific and Technological Planning Projects of Universities in Shandong Province under Grant J18KA328.

ABSTRACT In recent years, e-learning systems such as massive open online courses (MOOC) have been widely employed in academic institutions and have shown its power in enhancing the students' learning ability. Automatic detection and classification of knowledge points in lecture video would significantly enhance the performance of the online learning platform. Most of the previously presented approach for knowledge discovery focused on the text and audio documents, whereas the identification of knowledge points in videos still remains a challenge. To bridge this gap, we proposed a novel convolutional neural network which was designed for the characteristics of lecture video. It could both extract the temporal–spatial and semantic information from the multimedia record. To evaluate the performance of the proposed technique, we conducted comparison experiments between the state-of-the-art methods and ours. The experimental results demonstrated that the presented approach outperformed the state-of-the-art techniques and could be potentially invaluable for the accurate discovery of knowledge points within videos.

INDEX TERMS Knowledge acquisition, machine vision, neural network, optimization.

I. INTRODUCTION

Knowledge discovery has become a popular technique used for recognition and identification of productive information in various applications including bio-informatics [1], healthcare [2], data mining [3], and criminal investigation [4]. Meanwhile, a great deal of techniques have been presented to deal with the requirements of educational environments like E-Learning system, Massive Open Online Course (MOOC), and Modular Object-Oriented Dynamic Learning Environment (Moodle). To extract the knowledge from interaction of students with the e-learning system like Moodle, Lara *et al.* [5] proposed the use of knowledge discovery in databases (KDD). It could construct historical reference models of the students dropped out of and completed the course while the models would be exploited to discriminate one single student within the dropout or non-dropout group. Alfonseca *et al.* [6] presented a framework for establishing online information systems for linear texts in electronic format by using Adaptive Hypermedia and Natural Language processing algorithms. Guruler *et al.* [9] employed a decision tree classification algorithm to explore the meaningful patterns in large amount of University students data. Afterward, the influence factors including income level and type of registration to the University were revealed to be related to the students' success.

The currently presented knowledge discovery techniques have shown their performance in various types of information

The associate editor coordinating the review of this manuscript and approving it for publication was Nilanjan Dey.

management systems. However, most of them focused on extracting the information from text documents [6], whereas still failed in dealing with multimedia data, e.g., audio and video recordings. To transform large amount of raw data into useful information, data mining (DM) had been widely applied in cardiology and shown its benefits both for the patients, cardiologists, and nurses. Kadi *et al.* [7] presented a systematic review for the knowledge discovery in cardiology. Totally, 149 articles published between 2000 and 2015 were selected in their study. Ramos *et al.* [8] collected the temperature and relative humidity data from a part of 24 flats with homogeneous architecture and constant social strata during the heating season and a typical summer period. They revealed an energy poverty pattern that could prove the existence of discomfort during the heating season based on DM related techniques.

Meanwhile, deep learning, which has been employed in various machine learning tasks including image classification [10]–[13], image segmentation [14]–[16], image registrations [17]–[19], and video classification [20]–[22]. Among the state-of-the-art deep learning network architectures, the convolutional neural network (CNN) is widely conceived as one of the most vital one especially for feature extraction and classification. However, it remains a challenge to disclose the knowledge points from multimedia records.

Bearing the above analysis in mind, we propose a novel CNN architecture trained by **16,332** images along with the audio records, which were both sampled from 72 hours of lecturing video clips belonging to two different courses. The proposed CNN model is in the form of typical multi-channel structure. And each couple of the collected image and sound could be simultaneously fed into the presented CNN. Initially, the parameters including the convolution kernels and the pooling layers in the image channel and audio channel were trained by using the corresponding types of data, respectively. Thus, the output of the presented CNN model could be split into two parts with partial overlapping, which contained the shared information between the image and audio data in the same lecture video. And we assumed that the knowledge should be discovered in the shared portion of the output layer. To eliminate the likelihood of over-fitting and decrease the scale of the global parameters in the proposed CNN, we also introduced a constraint on the mutual information of two types of data. To produce the accurate outcome, we leveraged the alternate minimization approach to iteratively optimize presented objective function both in the training and testing procedures.

To validate the availability of the proposed method, we conducted comparative experiments on the collected samples between state-of-the-art techniques and the proposed approach. Experimental results demonstrated that the presented CNN architecture's superior performance over the state-of-the-art techniques. Generally, this study offers at least three contributions as follows.

- A novel CNN architecture was proposed to extract the shared information between the sampled pair of image and audio data from the same video clips
- We proposed a novel loss function for optimizing the presented CNN. It could simultaneously implement the maximization of intra-class similarity and the minimization of inter-class similarity for the classification of knowledge points and non-knowledge ones. Through integrating the softmax loss function and the proposed loss function, the presented algorithm had produced the highly discriminative subset of features to enhance the classification accuracy.
- Experiments on the practical data samples has demonstrated that the proposed technique is an potentially valuable tool for knowledge discovery in video with an impressive superiority.

The remainder of this paper is organized as follows. in Section II, We review the related work of deep learning techniques on video classification. In Section III, we present the details of the presented approach. We report the experimental results in Section IV and conclude in Section V.

II. RELATED WORK

Since visual and audio modalities are significantly correlated with each other, the information extracted from one modality could be exploited to improve the recognition of the other type of modality. Plenty of CNN-based techniques have been applied in fusing the visual and audio recordings for automated identification of video content.

As one promising solutions for speech recognition, Noda *et al.* [25] introduced an audio-visual speech recognition (AVSR) system especially useful for the noisy audios. They firstly presented that the appropriate selection of features was vital for sufficiently good performance. This study presented a hidden Markov model (HMM) technique for AVSR with noisy input. One deep denoising auto-encoder was used for extracting the audio features. Then, a CNN was exploited to extract visual features from the input images. In the final phase, a multi-stream HMM was used to integrate the audio and visual information.

To address the practical scene understanding task, Kojima *et al.* [26] presented a robot system by combining both the audio-visual and text information. The proposed technique focused on the information extraction relative to cooking scenes and included one CNN and one HMM.

To obtain the automatic prediction of continuously emotion states through the appropriate features, Basnet *et al.* [23] investigated the statistical performance of various features extracted from different convolutional neural networks. They proved that the features with minimal redundancy and maximal relatedness could be conserved by using feature selection paradigm, such as mutual information. The performance of frame-wise speculation of emotional state with the moderate length features was evaluated on spontaneous and naturalistic conversation between humans. Their experimental results

showed that the proposed framework could be exploited to classify the emotion state and outperformed the state-of-the-art visual-audio classification algorithms.

Hou *et al.* [24] proposed a speech enhancement technique for reducing the noise in speech data. Since the previous techniques focused on the audio format. Inspired by multi-modality learning and wide applications of CNN, the authors presented an audio-visual deep CNN. It incorporated both the audio and visual information into one unified architecture. They also proposed a multi-task learning structure for establishing the association between audio and visual data within the output layer. Generally speaking, the multi-channel CNN first handled the audio and visual input separately and integrated at the output layer. An end-to-end strategy was adopted and the parameters were optimized by using back propagation mechanism.

III. METHODOLOGY

In recent years, audio-visual recognition has achieved significant enhancement with the applications of deep learning mechanism especially CNN. Inspired by the application of CNN, we propose one CNN-based knowledge discovery algorithm for lecturing videos. The details of the proposed technique are listed as follows.

A. NETWORK ARCHITECTURE

For the audio-visual recognition applications, CNN-related technique usually employ a multi-channel network structure and entails three primary elements differing from the other instances of CNNs, which are weight sharing, local receptive field (LRF), and pooling operator. CNN prefers the local information rather than the global ones. It arises from the similarity with human eyeballs that could capture the local area of the images through confining every neuron in the brain, which only relates to its neighboring neurons. Furthermore, the shared weights between different neurons significantly correlates to local information within the image. Pooling layers is primarily leveraged to reduce the extracted feature dimension. For instance, only the maximal value is taken as the outcome for each feature map for max-pooling layer.

To address the challenge of knowledge discovery in lecture videos for online education platforms, we propose a novel CNN architecture. The proposed CNN architecture and its parameters were firstly trained on the publicly available audio-visual datasets presented in [23]–[26]. Then, the CNN framework was fine-tuned with the manually collected images and audio recordings. As shown in Fig. 1, the proposed CNN contains 7 convolutional layers with 7 max-pooling layers and 2 fully-connected (FC) layers for each input pair of image and audio data.

This multi-channel CNN architecture has 2 separate input channels and they shared the same setting of parameters including feature maps. As shown in Fig. 1, the visual stream is fed into the top channel and the audio stream is fed into the bottom channel. After the data has been input into the CNN, 7 continuous convolutional layers and 7 supporting max-pooling layers are used to extract both the shared information from the input audio-visual data and the unique features from each input stream. Since the two channels are the same as each other, we only supply the details of one channel as following. To note that all of the strides for the convolutional layers are 2.

- One convolutional layer (24 19 × 19 kernels) with its matched max-pooling are presented to deal with the input data (audio or visual).
- One convolutional layer (24 17 × 17 kernels) with its matched max-pooling are presented.
- One convolutional layer (48 15 × 15 kernels) with its matched max-pooling are presented.
- One convolutional layer (96 13 × 13 kernels) with its matched max-pooling are presented.
- One convolutional layer (192 11 × 11 kernels) with its matched max-pooling are presented.
- One convolutional layer (384 9 × 9 kernels) with its matched max-pooling are presented.
- One convolutional layer (512 7 × 7 kernels) with its matched max-pooling are presented.
- There are two FC layers (with 1024 neurons) and the ReLU operators at the end of the proposed CNN.

B. LOSS FUNCTION

Commonly, the softmax loss function, which has been widely exploited by various CNNs, could be formulated as:

$$L_s = \sum_{i=1}^{m} log \frac{e^{W_{y_i}^T X_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T X_i + b_j}} \tag{1}$$

where $X_i \in \mathscr{R}^d$ represents the extracted feature corresponding to the i_{th} input data belonging to the category of $y_{i_{th}}$. $W_j \in \mathscr{R}^d$ denotes the column j_{th} of the matrix $W \in \mathscr{R}^{d \times n}$ at the second FC layer and $b \in \mathscr{R}^n$ denotes the error. And m stands for the batch size and n is the identity's quantity.

Assume that I_k and I_l is the input image and audio signal, respectively. The entire feature extraction procedure could be formulated as:

$$X_i = C(I_i, \theta_i)(i \in \{k, l\}) \tag{2}$$

where $C(.)$ denotes the convolution operator for extracting the features, O_i represents the output features, and θ stands for the set of parameters including the weights in the feature maps within the proposed CNN architecture. To note that there are both the unique and shared parts of the audio and visual input. Therefore, we introduce U, V_k, and V_l as the feature vectors to denote both the unique and shared feature subsets in the output feature vector within the output layer, and this process could be formulated as:

$$F_i = \begin{bmatrix} F_{share} \\ F_{unique} \end{bmatrix} = \begin{bmatrix} SX_i \\ U_i X_i \end{bmatrix} \quad (i \in \{k, l\}) \tag{3}$$

where SX_i denotes the sharing between audio and visual input and $U_i X_i, i \in \{k, l\}$ denotes the unique feature from each input channel. To make use of the mutual information

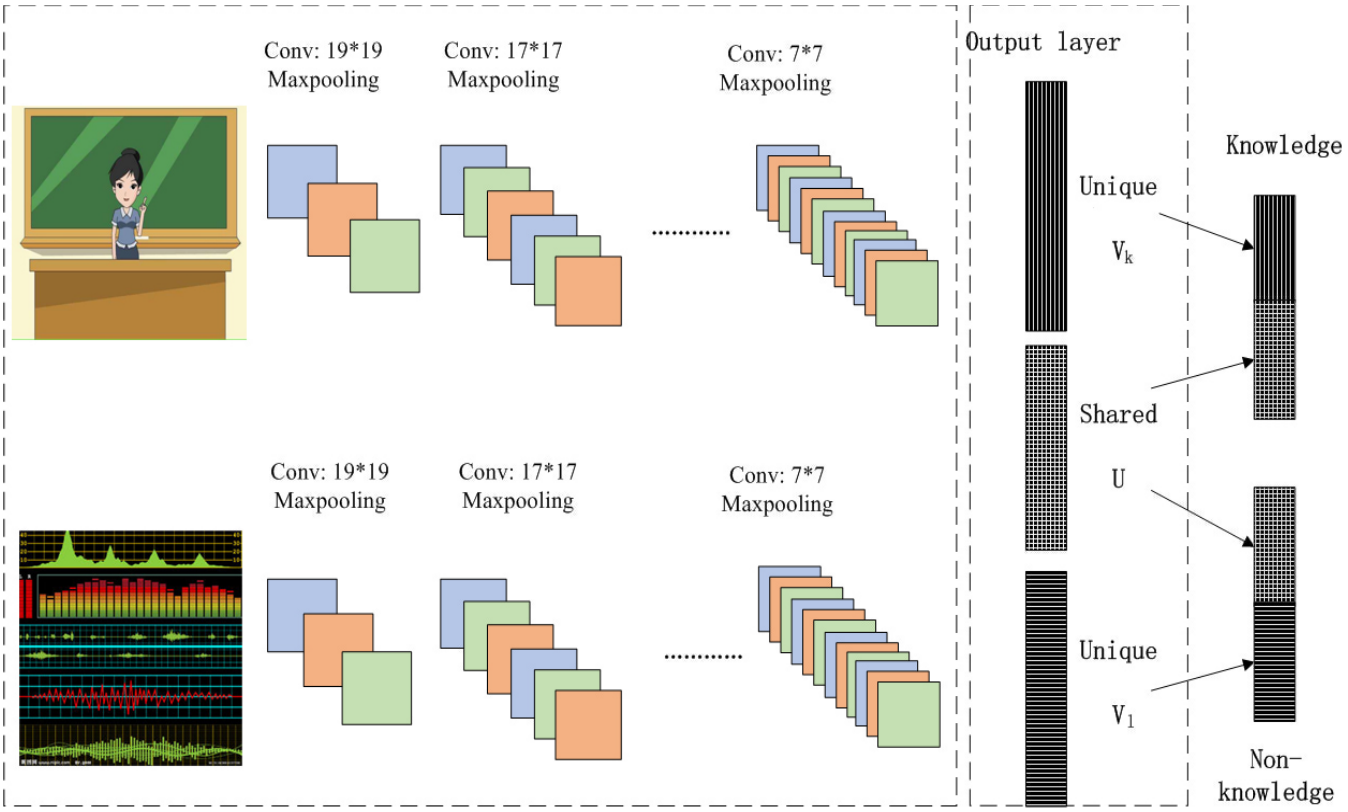


FIGURE 1. The proposed CNN architecture. The max-pooling layers are used to reduce the dimensionality of the initially extracted features and avoid over-fitting. Both the features from the audio signals and the corresponding visual parts could be jointly acquired from the unique and shared parts in the output layer.

between the shared and unique information, we introduced one mutual information regularized term combined with the original softmax loss. It could be formulated as:

$$\mathscr{L}(F, c, \theta, S, U) = \sum_{i\in\{P,Q\}} softmax(F_i, c, \theta, S, U_i)$$
$$s.t.\ MI(S, U_i) = 0 \quad (i \in \{k, l\}) \tag{4}$$

where c denotes the category of the discovered knowledge (knowledge or non-knowledge) and $MI(.)$ [28] is used to calculate the mutual information of the input audio-visual data.

C. OPTIMIZATION

According to the Lagrangian multiplier(s), the output of the proposed CNN could be expressed as the following objective function.

$$\mathscr{L}(F, c, \theta, S, U_i) = \sum_{i\in\{k,l\}} softmax(F_i, c, \theta, S, U_i) + \lambda \sum_{i\in\{k,l\}} MI(S, U_i) \tag{5}$$

where λ denotes the trade-off parameter used to balance the weights of softmax loss and the presented MI loss. We then leveraged the alternating minimization algorithm as well as back-propagation strategy to optimize the objective function, iteratively. Notable that the gradients of S and U_i could be expressed as:

$$\frac{\partial \mathscr{L}}{\partial S} = \sum_{i\in\{k,l\}} \frac{\partial softmax(F_i, c, \theta_i, S, U_i)}{\partial S} + \sum_{i\in\{k,l\}} \frac{\partial MI(S, U_i)}{\partial S} \tag{6}$$

$$\frac{\partial \mathscr{L}}{\partial U_i} = \sum_{i\in\{k,l\}} \frac{\partial softmax(F_i, c, \theta_i, S, U_i)}{\partial U_i} + \sum_{i\in\{k,l\}} \frac{\partial MI(S, U_i)}{\partial U_i} \tag{7}$$

where the values of all of the variables should be optimized by using alternating minimization algorithm at a level of γ.

$$\theta^{(t+1)} = \theta^{(t)} - \gamma \frac{\partial \mathscr{L}}{\partial \theta^{(t)}} \tag{8}$$

$$S^{(t+1)} = S^{(t)} - \gamma \frac{\partial \mathscr{L}}{\partial S^{(t)}} \tag{9}$$

$$U_i^{(t+1)} = U_i^{(t)} - \gamma \frac{\partial \mathscr{L}}{\partial U_i^{(t)}} \tag{10}$$

We initialized the parameters including the feature maps with random values.

IV. RESULTS AND DISCUSSION

To evaluate the performance of the presented CNN architecture, we conducted comparing experiments on the manually collected dataset between state-of-the-art CNNs and ours.

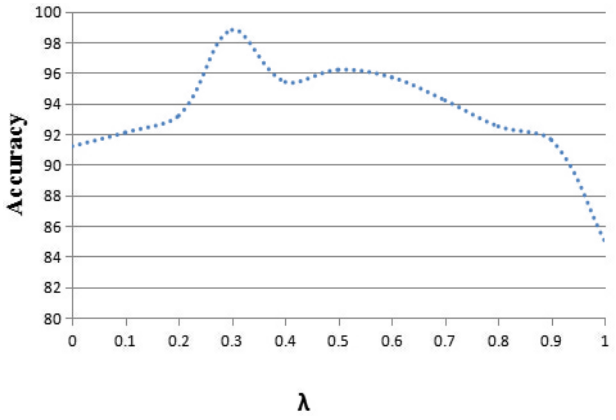


FIGURE 2. Performance of the proposed CNN with different λ.

Both the experimental results and the analysis are described in the following.

A. DATASET

We firstly trained the proposed CNN model with the publicly available presented in [23], [25], and [26], we also collected 16,322 lecture images and the corresponding audio streams to train our proposed CNN. The images and audio signals were collected from 72 hours of lecture video clips at one online learning platform. Furthermore, we leveraged the data augmentation techniques to increase the diversity of the raw material including both the images and audio signals since the performance of CNNs heavily rely on the input amount of data samples. Generally, we enlarged the input data with image translation and audio signal segmentations. Afterwards, the images are resized from 1600×900 into 512×512 with overlapping. It is notable that the preprocessing was implemented by using Matlab 2013b.

B. THE CHOICE OF λ

To find the optimal value of the trade-off parameter in Eq. (5), we firstly conducted the classification experiments with different settings of λ from 0 to 1 with step of 0.01, and the results is shown in Fig. 2.

Notable, the accuracy performance is the optimal when λ is set at 0.3 while it is worse no matter if the value of λ is greater or less than 0.3. Therefore, we set λ as 0.3 in the training, testing, and evaluation procedures of the proposed CNN architecture.

C. TRAINING AND EVALUATING

We labeled the audio-visual samples as knowledge and non-knowledge according to the inner content, respectively. In general, we chose 60 percent of the samples into the training set, 30 percent into the evaluation set, and the other samples in the testing set. The proposed CNN framework was fine-tuned with back propagation mechanism. The practical execution was implemented on high performance Graphics Processing Unit (GPU) and the deep learning platform Tensorflow. For the Tensorflow system, the learning rate is set at 0.01. It took 10^4 iterations at most and each iteration was about 0.2 seconds.

TABLE 1. Face verification performances on LFW. (the entries in the column images represent the number used to train the face verification methods, respectively).

Methods	Number of Loss Functions	Accuracy (%)
Noda *et al.* [25]	One	87.23
Kojima *et al.* [26]	One	90.55
Basnet *et al.* [23]	One	92.68
Hou *et al.* [24]	One	93.37
Our Method	Two	**98.49**

D. EXPERIMENTS

To evaluate the performance of the proposed mutual information loss based CNN, we conducted the comparison experiments between the state-of-the-art techniques [23], [25], [26] and ours on our manually sampled dataset.

As shown in Table. 1, the proposed technique achieved superior performance over the state-of-the-art audio-visual recognition techniques, while the number of loss functions used in the training of the presented CNN is two and the number of loss functions used in the other techniques is one.

E. ANALYSIS

From the comparing experimental results on the manually collected data samples, we could notice the availability of the introduced mutual information regularization term. Through incorporating both the shared information from the input audio-visual data streams and unique information extracted from each input channel, the combined loss function could produce better accuracy over the single softmax loss function. We also revealed that the shared information could reveal knowledge that can be discovered from the lecturing video.

The proposed CNN could significantly improve the classification of knowledge discovered from the data through combing both the softmax and MI. To note that we also conducted the comparing experiments on different values of λ in Eq. (5), which is leveraged to implement the trade-off between the softmax and MI loss. The experimental results demonstrated that the optimal value of λ_i should be set at 0.3. Since the introduced MI loss could bring about the complementary information from both the audio-visual pairs.

V. CONCLUSION

The accurate recognition of the difference between knowledge points and non-knowledge points in lecture videos is an potentially valuable tool for supporting the teachers and students. A large amount of studies have been carried out for this task and have shown their usefulness in the classification. However, most of the state-of-the-art techniques focused on the text document instead of the multimedia data. Therefore, we presented a novel two-channel CNN architecture with one MI loss function. It offers a unique algorithm in an automatic and non-invasive manner. Within the proposed CNN, two parameter-sharing channels are presented to address the visual image and audio signal, respectively.

This paper offers several contributions. Firstly, a novel deep CNN was presented to realize the classification of knowledge and non-knowledge. Then, to our best knowledge this is the first attempt to introduce the mutual information loss into CNN architecture. Finally, the proposed CNN outperforms the state-of-the-art techniques.

In the near future, we would continue to research on various types of CNNs and the applications in various applications, e.g., medical image processing [27], [28].

DECLARATIONS

Ethical Approval and Consent to participate: Approved. Consent for publication: Approved. Availability of supporting data: We can provide the data.

COMPETING INTERESTS

These no potential competing interests in our paper. And all authors have seen the manuscript and approved to submit to your journal. We confirm that the content of the manuscript has not been published or submitted for publication elsewhere.

AUTHOR CONTRIBUTIONS

All authors took part in the discussion of this study.

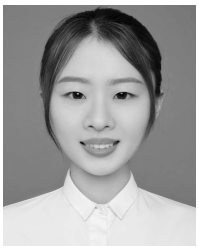
ACKNOWLEDGMENT

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

REFERENCES

[1] T. Oyama, K. Kitano, K. Satou, and T. Ito, "Extraction of knowledge on protein–protein interaction by association rule discovery," *Bioinformatics*, vol. 18, no. 5, pp. 705–714, 2002.
[2] J. Dallongeville, N. Marécaux, D. Cottel, A. Bingham, and P. Amouyel, "Association between nutrition knowledge and nutritional intake in middle-aged men from Northern France," *Public Health Nutrition*, vol. 4, no. 1, pp. 27–33, 2001.
[3] E. T. Wang and G. Lee, "An efficient sanitization algorithm for balancing information privacy and knowledge discovery in association patterns mining," *Data Knowl. Eng.*, vol. 65, no. 3, pp. 463–484, 2008.
[4] J. Schroeder, J. Xu, and H. Chen, "CrimeLink explorer: Using domain knowledge to facilitate automated crime association analysis," in *Intelligence and Security Informatics*. Berlin, Germany: Springer-Verlag, 2003.
[5] J. A. Lara, D. Lizcano, M, A. Martínez, J. Pazos, and T. Riera, "A system for knowledge discovery in E-learning environments within the European higher education area—Application to student data from Open University of Madrid, UDIMA," *Comput. Educ.*, vol. 72, pp. 23–36, Mar. 2014.
[6] E. Alfonseca, P. Rodríguez, and D. Pérez "An approach for automatic generation of adaptive hypermedia in education with multilingual knowledge discovery techniques," *Comput. Educ.*, vol. 49, no. 2, pp. 495–513, 2007.
[7] I. Kadi, A. Idri, and J. L. A. Fernandez-Aleman "Knowledge discovery in cardiology: A systematic literature review," *Int. J. Med. Inform.*, vol. 97, pp. 12–32, Jan. 2017.
[8] N. M. M. Ramos, R. M. S. F. Almeida, and M. L. Simões, and P. F. Pereira, "Knowledge discovery of indoor environment patterns in mild climate countries based on data mining applied to in-situ measurements," *Sustain. Cities Soc.*, vol. 30, pp. 37–48, Apr. 2017.
[9] H. Guruler, A. Istanbullu, and M. Karahasan, "A new student performance analysing system using knowledge discovery in higher educational databases," *Comput. Educ.*, vol. 55, no. 1, pp. 247–254, 2010.
[10] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3642–3649.
[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[12] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[13] M. Z. Alom *et al.* (2018). "The history began from AlexNet: A comprehensive survey on deep learning approaches." [Online]. Available: https://arxiv.org/abs/1803.01164
[14] P. Moeskops *et al.* "Deep Learning for Multi-task Medical Image Segmentation in Multiple Modalities," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2016, pp. 478–486.
[15] Y. Lin, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, *Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation*. Berlin, Germany: Springer-Verlag, 2017.
[16] H. Pan, W. Bo, and J. Hui "Deep learning for object saliency detection and image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, 2015.
[17] X. Yang, R. Kwitt, M. Styner, and M. Niethammer, "Quicksilver: Fast predictive image registration—A deep learning approach," *NeuroImage*, vol. 158, pp. 378–396, Sep. 2017.
[18] G. Wu, M. Kim, Q. Wang, B. C. Munsell, and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 7, pp. 1505–1516, Jul. 2016.
[19] G. Wu, M. Kim, Q. Wang, Y. Gao, S. Liao, and D. Shen, "Unsupervised deep feature learning for deformable registration of MR brain images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2013, pp. 649–656.
[20] A. J. Lipton, H. Fujiyoshi, and R. S. Patil, "Moving target classification and tracking from real-time video," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Oct. 1998, pp. 8–14.
[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, and R. Su, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
[22] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4694–4702.
[23] R. Basnet, M. T. Islam, T. Howlader, S. M. M. Rahman, and D. Hatzinakos, "Statistical selection of CNN-based audiovisual features for instantaneous estimation of human emotional states," in *Proc. Int. Conf. New Trends Comput. Sci. (ICTCS)*, Oct. 2017, pp. 50–54.
[24] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.
[25] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Appl. Intell.*, vol. 42, no. 4, pp. 722–737, 2015.
[26] R. Kojima, O. Sugiyama, and K. A. Nakadai , "Audio-visual scene understanding utilizing text information for a cooking support robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep./Oct. 2015, pp. 4210–4215.
[27] X. Ren *et al.*, "Drusen segmentation from retinal images via supervised feature learning," *IEEE Access*, vol. 6, pp. 2952–2961, 2017.
[28] L. Jian, Y. Zheng, W. Jiao, F. Yan, and B. Zhao, "Deblurring sequential ocular images from multi-spectral imaging (MSI) via mutual information," *Med. Biol. Eng. Comput.*, vol. 56, no. 6, pp. 1107–1113, 2018.

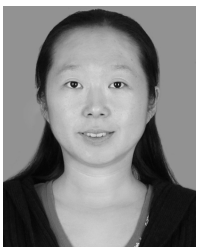
JINJIAO LIN received the Ph.D. degree in computer software and theory from Shandong University. She is currently an Associate Professor with the School of Management and Engineering, Shandong University of Finance and Economics. Her current research interests include smart learning, education big data, knowledge discovery, software engineering, software adaptability, management information systems, and decision support systems.



CHUNFANG LIU is currently pursuing the master's degree with the School of Management and Engineering, Shandong University of Finance and Economics. Her research interests include knowledge discovery, data mining, and management information systems.

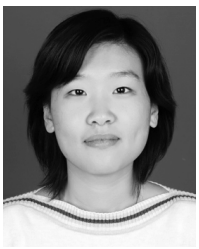
YIBIN LI received the Ph.D. degree from the School of Electrical Engineering and Automation, Tianjin University. He is currently a Professor with the School of Control Science and Engineering, Shandong University, China. His research interests include intelligent robots, intelligent vehicles, and intelligent control.

LIZHEN CUI is currently a Professor with the School of Software, Shandong University, Jinan, China. His research interests include big data, deep learning, and intelligent data analysis.



RUI WANG received the M.S. and Ph.D. degrees in computer science and technology from Shandong University. She is currently an Assistant Professor with the Shandong University of Finance and Economics. She has authored two books and more than 20 papers. Her research interests include active service and methods in recommendation systems of smart learning.

XUDONG LU received the Ph.D. degree from Shandong University, where he is currently a Lecturer and a Supervisor of master's degree programs with the School of Software. His research interests include software engineering, data science and engineering, and intelligent data analysis.



YAN ZHANG is currently pursuing the Ph.D. degree with the Shandong University of Science and Technology. Her research interests include machine learning, machine vision, and image analysis.

JIAN LIAN is currently an Instructor with the Shandong University of Science and Technology. His interests include machine learning and image processing.