EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Video person reidentification based on neural ordinary differential equations and graph convolution network

Li-qiang Zhang[1], Long-yang Huang[2*] and Xiao-li Duan[3]

* Correspondence: longyanghuang@cafuc.edu.cn
[2]College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China
Full list of author information is available at the end of the article

Abstract

Person reidentification rate has become a challenging research topic in the field of computer vision due to the fact that person appearance is easily affected by lighting, posture and perspective. In order to make full use of the continuity of video data on the time line and the unstructured relationship of features, a video person reidentification algorithm combining the neural ordinary differential equation with the graph convolution network is proposed in this paper. First, a continuous time model is constructed by using the ordinary differential equation (ODE) network so as to capture hidden information between video frames. By simulating the hidden space of the hidden variables with the hidden time series model, the hidden information between frames that may be ignored in the discrete model can be obtained. Then, the features of the generated video frames are given to the graph convolution network to reconstruct them. Finally, weak supervision is used to classify the features. Experiments on PRID2011 datasets show that the proposed algorithm can significantly improve person reidentification performance.

Keywords: Person reidentification, Graph convolutional network, Neural ordinary equations

1 Introduction

In recent years, with the increasing attention to public safety and the development of video surveillance technology, more and more cameras are deployed in crowded places [1, 2]. However, the operation of large-scale video monitoring system produces a large number of monitoring data, which is difficult to quickly analyze and process only relying on human resources. Therefore, the use of computer vision technology to automatically complete the task of intelligent monitoring system came into being [3]. Although the current face recognition technology has been relatively mature, it is often unable to obtain effective face images in the actual monitoring environment. Therefore, it is very important to lock and search people with whole body information. This also makes human re-recognition technology gradually become a research hotspot in the field of computer vision, which has attracted people's extensive attention [4].

The purpose of character re-recognition is to accurately identify a person who appears in one camera, when he/she appears in other cameras again [4, 5]. Due to the influence of camera perspective, dramatic changes in the posture of moving human

body, illumination, occlusion, and chaos of background [6, 7], human re-recognition algorithm is still facing great challenges. At present, the research methods of human re-recognition are mainly divided into single-frame image-based and video-based human re-recognition [8].

Early video character detection methods are usually based on image detection, which can judge whether there is a person in each frame by extracting the static features of the image. However, with the wide application of depth model in the field of video detection, more and more attention has been paid to the temporal and dynamic characteristics of video information in recent years. Graph convolution network (GCN) and ordinary differential equation (ODE) are the latest achievements in machine learning. They apply unstructured and continuous models to various learning tasks. In this paper, the video continuous model is established by ordinary differential equation, and combined with the graph convolution network, a continuous time airspace personnel detection model based on video stream is proposed.

2 Methods

2.1 Graph convolutional network

Most of the graph neural network models use graph convolution, whose core is the convolution kernel parameter sharing in local areas. The same convolution kernel is used for the convolution operation of each graph node, which can greatly reduce the number of model parameters. Parameter update of the convolution kernel can be seen as learning a graph function $G = (\upsilon, \varepsilon)$, which respectively represent the connecting edge between vertices in the graph. The input is eigenmatrix $X \in R^{N \times D}$, N is the number of vertices, $D$is the characteristic dimension, and the matrix expression of the graph structure (usually expressed as adjacency matrix A). The output is $Z \in R^{N \times F}$ and F is the output dimension of the convolution layer of the graph. Each graph convolution layer can be represented as the following nonlinear function:

$$H^{(l+1)} = f\left(H^{(l)}, A\right) \tag{1}$$

where $H^{(0)} = X$, $H^{(l)} = Z$, and l is the number of convolution layers. For different models of the task, the appropriate convolutional function $f(\cdot, \cdot)$ will be selected and parameterized. This paper uses the same convolution function as Kipf et al. [8], whose basic form is:

$$f\left(H^{(l)}, A\right) = \sigma\left(AH^{(l)}W^{(l)}\right) \tag{2}$$

where $W^{(l)}$ is the weight parameter of the l-level neural network, and $\sigma(\cdot)$ is the nonlinear activation function, usually rectified linear unit (ReLU). After the above improvement, the graph convolution function can be calculated as:

$$f\left(H^{(l)}, A\right) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{3}$$

where $\hat{A} = A + I$, I is the identity matrix, and $\hat{D}$ is the diagonal vertex degree matrix of $\hat{A}$.

2.2 Extraction of continuous hidden state characteristics based on The Constant differential equation

Ordinary differential equation network is a new branch of neural network. It makes the neural network continuous and uses ordinary differential equation solver to fit the neural network itself. The basic problem domain equation is as follows:

$$h_{t+1} = h_t + f(h_t, \theta_t) \tag{4}$$

$$\frac{dh(t)}{dt} = f(h(t), t, \theta) \tag{5}$$

$$h(T) = h(t_0) + \int_{T_0}^{T} f(h(t), t, \theta)dt \tag{6}$$

where h_t stands for the hidden state, and $f(\cdot)$ represents the nonlinear transformation of a monolayer neural network. Equation (5) represents the forward propagation process between the residual blocks in the standard residual network. The neural network of each layer fits the residual term, while in Eq. (6), the output of the neural network is regarded as the gradient of the hidden state. Then, the hidden state value of t can be obtained at any time by solving the equation integrally. The number of evaluation points can be considered as equivalent to the number of model layers of the discrete neural network. In this paper, basic applications of ODE network on various mainstream model structures are proposed. The convolutional neural network model and the implicit state model based on time span are referred.

The feature extraction of video pedestrian mainly includes two aspects. On the one hand, it is the static feature extraction of video frame images in regular space, including pedestrian edge, color, and other features. In this respect, the mainstream neural network has been able to obtain a high recognition rate. Experiments show that the static feature extraction of pedestrians does not need too deep network scale. On the other hand, it is also one of the difficulties of video pedestrian detection, which is the spatio-temporal dynamic characteristics of pedestrian in time span. Many scholars have proposed different methods to extract the dynamic characteristics of pedestrians. However, none of the current methods takes into account the continuous information lost between discrete video frames. From the perspective of continuous events, this paper attempts to fit the probability distribution of the hidden dynamic characteristics of person $\tilde{Z}$ through the hidden state model of ODE network, as shown in Fig. 1.

Firstly, the static feature vector X of a single frame is extracted by common convolutional network, such as residual network. If the feature of the image block is extracted, an additional layer of convolutional network is added to predict the category of the complete image of the frame. The feature vector of the complete image can be obtained from the sequentially arranged feature of the image block, and the pooling layer can also be used. Secondly, the static feature X is sampled in reverse chronological order, and the predicted initial hidden state $\tilde{Z}$ is obtained through the timing network (cyclic neural network is used in this paper). The hidden state probability distribution P($\tilde{Z}$) is obtained from the ODE network, and the hidden state value $\tilde{Z}_t$ can be predicted at any time. Finally, the implicit state value is converted to target feature vector X^* by the decoder.

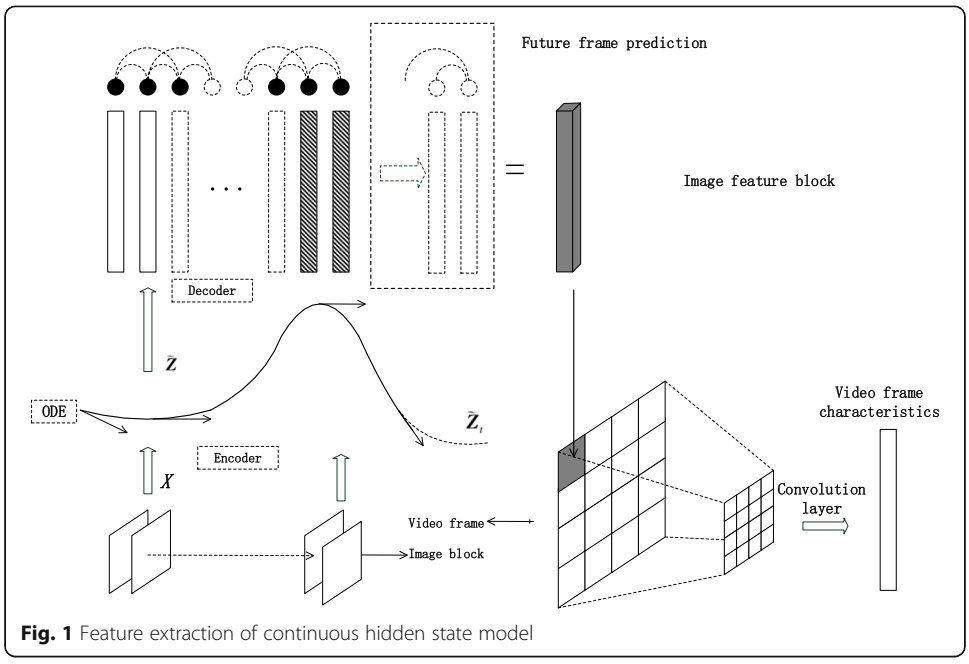

Fig. 1 Feature extraction of continuous hidden state model

2.3 The video person reidentification based on ODE and GCN

The video frames firstly are considered on the time span of contact with $G_k = (\upsilon_k, \varepsilon_k)$ graph model, the window size is $2k + 1$, at the current moment as the center. With the current moment as the center, each video frame has k entry and exit edges and a self-ring edge, a total of $2k + 1$. And the undirected graph is used for the reason of considering the relevance of before and after the event simultaneously. Each layer of graph convolutional network can contain $\boldsymbol{n}$ such windows, depending on the size of the network and usually determined by the length of the video block. Thus, the state update equation of each node in the middle layer of the graph convolution network can be expressed as:

$$X_t^{l+1} = \sigma\left(\sum_{t_i=t-k}^{t+k} \frac{1}{\tilde{Z}} X_{t_i}^l W_{\mathrm{t}_i}^l\right) k = 1, 2, \cdots \tag{7}$$

where $\tilde{Z}$ is the normalization factor, the same as Eq. (7). l represents the number of layers in the graph convolution network, and σ is the activation function.

Video detection still boils down to classification. In the classification task, there are mature full supervision algorithms. However, a large amount of high-quality data is required, while the acquisition of high-quality video in real scenes is a difficulty, and relevant real data is lacking at this stage. The transfer learning for small datasets and the weak supervision algorithm with low requirements for the tag quality of data samples show outstanding advantages.

Input: Graph model function $G(\upsilon, \varepsilon)$, graph function parameter W, video block window size k, initialize the graph model adjacency matrix A_0, the input feature of the node in the figure X_0, the margins of positive and negative samples in triplet losses were margin.

Output: X^* can distinguish target characteristics, characteristics of person sample space center C_p, the non-person sample C_n feature space center:

1) Initialize $A = A_0$, $C_P = 0$, $C_n = 0$
2) Randomly sampling Triplet $\hat{X}_i = Triplet(X_a, X_p, X_n)$ from input feature sample X_0, where X_a is the anchor point, X_p is the same random sample point as the anchor point category, and X_n is the opposite of the anchor point category
3) Repeat
4) Forward transmission
5) For X in Triplet $Triplet(X_a, X_p, X_n)$ do:
6) For Gall layers do:
7) Generate its diagonal node degree matrix $\boldsymbol{D}$ from $\boldsymbol{A}$
8) To calculate the normalized coefficient $\frac{1}{Z} = \tilde{D}^{-\frac{1}{2}}\hat{A}\tilde{D}^{-\frac{1}{2}}$
9) For X_t in X do:
10) Node status update $X_t = \sigma(\sum_{X_{ti} \in neighbor^k{}_{X_t}} \frac{1}{Z} X_{ti} W_{ti})$
11) End for
12) Update the adjacency matrix A from the new graph node state
13) End for
14) Return X, update triplet set $\hat{X}$
15) End for
16) Return $\hat{X}$ as X^*
17) Back propagation
18) For $Triplet(X_a, X_p, X_n)$ in X^* do:
19) Computing triple loss$L = \max(\|X_a - X_p\|_2^2 - \|X_a - X_n\|_2^2 + m \arg in, 0)$
20) End for
21) Calculate the average loss $\overline{L}$
22) Back propagation $\overline{L}$ gradient using Adam algorithm
23) Until convergence or reach the maximum number of training

3 Experimental

The algorithm is tested on two video character datasets prid2011 and ilids vid [9]. The prid2011 dataset contains video captured by two static cameras. A camera recorded 385 people's video information, and B camera recorded 749 people's video information, of which A and B cameras collected 20 people at the same time. Each person's video subset contains 5 to 675 video frames. In order to ensure the effectiveness of spatio-temporal features, 178 person's video frames are selected. All the videos in the dataset are shot in the outdoor environment with less occlusion and no crowding. Everyone has a wealth of walking posture images. Figure 2 shows some examples of personal video frames.

The ilids vid dataset consists of 300 different pedestrians who are viewed through two disconnected cameras in the public open space. The dataset is created from two non-overlapping pedestrian camera views observed in the i-lids multi-camera tracking scene (MCTS) dataset captured under the multi Camera CCTV network in the arrival hall of the airport. It consists of 600 image sequences from 300 different individuals, each with a pair of sequences from two camera views. Each image sequence has a

Fig. 2 Example of PRID2011

variable length, from 23 frames to 192 frames, with an average of 73 frames. Figure 3 shows some examples of personal video frames.

The experiment is based on the pytorch deep learning framework. The hardware configuration is 32 GB memory, Intel (R) Core (TM) i7-4790k processor and NVIDIA gtx1080 8GB graphics card. Each experiment randomly generated training set and test set, and repeated 10 experiments under the same conditions. The average value of 10 experimental results is taken as the final result of this experiment. The experimental results evaluate the performance of the algorithm by the recognition rate.

On the video dataset, every 5 frames are trained, and all the data are tested. At the same time, the detection rate and false alarm rate of the hidden state model in the untrained frame are tested. All images are trained and all frames are tested. A small batch of 128 frames is used for training, the number of training iterations is 2000, the initial learning rate is 0.0001, and the gradient descent method is used. The hidden dimension of ODE network convolution model is 64, and the specific structure is shown in Table 1. In this paper, group normalization is used for all normalization layers, and the maximum number of groups is 32. In classification training, cross-entropy loss and triple loss are used for full supervised training and weak supervised training, and the models are CGN UXe and CGN uwk, respectively. The encoder of implicit state model adopts long short memory network (LSTM). The decoder is a fully connected layer, the hidden layer depth of the

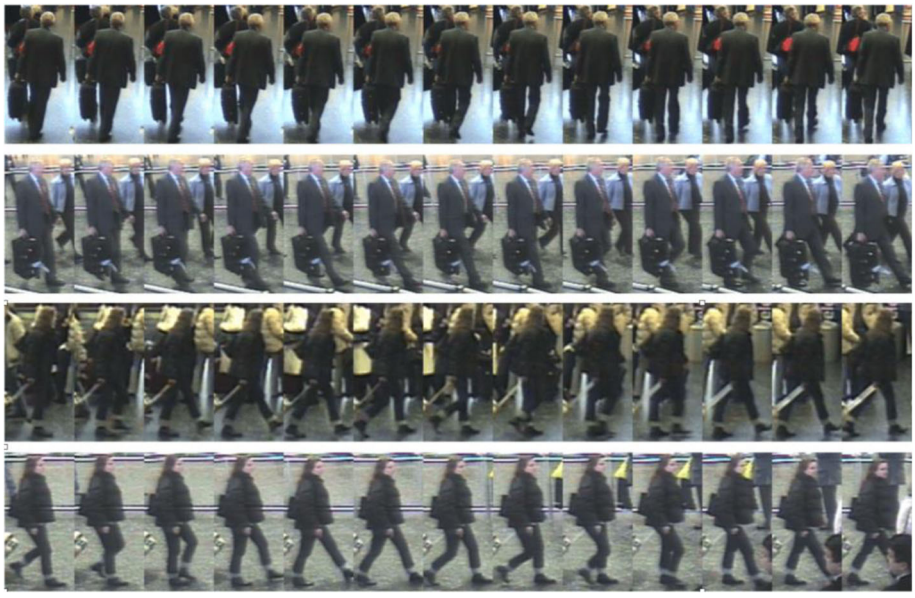

Fig. 3 Example of iLIDS-VID

model is 128, and the window value k is 2. The implicit state sampling adopts Monte Carlo method, and the sampling points of each video clip are 100 + 50, which are the detection of the current period and the prediction of the future period, respectively.

This paper uses cross-validation for image datasets. Batch number 32, hidden dimension 32, number of cycles 100, initial learning rate 0.001. The Adam method is used for back propagation gradient. At the same time, cross-entropy and triple loss are applied to the full supervised and weak supervised classification training, respectively, and the best performance of each comparison index is obtained.

4 Result and discussion

In the experiment, 300 people were randomly selected to form a training set, and the rest 300 people were selected to form a test set. The experimental results are compared with other typical algorithms (dynamic RNN-CNN network [10], cumulative motion context (AMOC) network [11], algorithm using matrix shared attention [12],

Table 1 Gradient estimation model architecture

The module	Gradient estimation model
Subsampling module (optional)	[Normalized layer, 3×3 convolutional layer, ReLU layer]×1 [Normalized layer, 4×4 convolutional layer (step size 2), ReLU layer]×2
The ODE module	[Normalized layer, ReLU layer, 3×3 convolutional layer]×2 [Normalized layer]×1
Convolution model Tacit module	[Linear transformation layer, ReLU layer]×3
Fully connected module	[Normalized layer, ReLU layer, global maximum pooling layer, linear transformation layer]

Note: the step size of unmarked convolution layer is 1 by default. The number after the multiplication symbol in brackets indicates the number of times the submodule in brackets is repeated

Table 2 Person reidentification rates of different methods of PRID2011

Algorithm	Rank1	Rank5	Rank10	Rank20
Paper [10]	**58.0**	**84.0**	**91.0**	**96.0**
Paper [11]	68.7	94.3	98.3	99.3
Paper [12]	62.0	86.0	94.0	98.0
Paper [8]	76.0	94.0	97.0	98.0
Proposed	**80.6**	**95.5**	**98.4**	**99.4**

rearrangement application method based on matrix shared attention [8]). See Table 2 for the results of the personnel re-identification rate.

It can be seen from the data in Table 2 that the recognition rate of this algorithm has been significantly improved compared with the existing algorithms. Rank1 is 80.6%, which is 4.4% higher than the method proposed in literature [8]. Rank5 and rank20 algorithms have some improvements compared with other algorithms [13–17].

From the experimental data of ilds-vid dataset in Table 3, it can be seen that this method has higher recognition rate than the existing mainstream methods. Experiments further verify the effectiveness of the algorithm.

5 Conclusion and future work

Human recognition is an important research topic in the field of computer vision. In order to improve the performance of video character recognition, a video character recognition algorithm based on ode and graph convolution network is proposed. Firstly, the ode tacit model hidden in the video distribution is used to supplement the information lost between frames. Then, through the digital convolution learning network to connect the continuity and interval between video frames, the relationship between unstructured features is established, and the positive and negative samples are divided. Finally, the feature using a full join layer or directly calculating the center distance of positive and negative samples is selected to get the classification results. Experimental results show that this method can significantly improve the performance of video character recognition, which is of great significance to the research of video character recognition.

Our main work in the future is to continue to improve the recognition accuracy, reduce the complexity of the algorithm, and reduce the time consumption.

Table 3 Person reidentification rates of different methods of iLIDS-VID

Algorithm	Rank1	Rank5	Rank10	Rank20
Paper [10]	**70.1**	**89.6**	**96.2**	**96.3**
Paper [11]	83.7	95.7	98.4	97.3
Paper [12]	78.1	95.0	98.0	99.1
Paper [8]	83.0	96.2	98.2	98.0
Proposed	**87.5**	**97.6**	**99.2**	**99.8**

Abbreviations
ODE: Ordinary differential equation; GCN: Graphic convolutional network; MCTS: Multi-camera tracking scene; LSTM: Long- and short-term memory network; ReLU: Rectified linear unit; AMOC: Accumulative motion context

Acknowledgements
Not applicable

Authors' contributions
Li-qiang Zhang and Long-yang Huang as the primary contributor, completed the analysis, experiments and paper writing. Xiao-li Duan helped perform the analysis with constructive discussions. The author(s) read and approved the final manuscript.

Authors' information
Zhang Li-qiang received the B.S. degrees at the School of Computer Science and Technology, Nanjing University of Technology, Nanjing, China, in 1999. His research interests include image/video processing, computer vision, and super-resolution.
Huang Longyang received the B.S. degree at the School of Electronic Engineering, Changchun University Of Science and Technology, Changchun, China, in 1996; the M.S. degree in Telecommunications Engineering from University of Electronic Science and Technology of China, Chengdu, China, in 2004; and the Ph.D. degree in Circuits & Systems from Beijing University of Posts and Telecommunications, Beijing, China, in 2009. He is an associate professor at the Civil Aviation Flight University of China. His research interests include signal processing, aeronautical telecommunications, and air traffic management.
Duan Xiaoli received the B.S. degree at the School of Electronic Engineering, Zhongyuan University Of Technology, Zhengzhou, China, in 2002; the M.S. degree in Electronic Engineering from Xiamen University, Xiamen, China, in 2006; and the Ph.D. degree in Electronic Engineering from Sichuan University, Chengdu, China, in 2019. She is a professor at the Zhengzhou Normal University. Her research interests include signal processing.

Funding
Scientific research projects of the Civil Aviation Air Flight University of China, Research on accurate track prediction and real-time collision detection technology (J2010-33); Special Research Project on University Integrity in Henan Province in 2020 (2020-LZZD03): a study on the Evaluation of Financial Disclosure Transparency in Henan Province—based on the Analysis of Financial Disclosure Data of 38 Provincial Administrative Colleges.

Availability of data and materials
The labeled dataset used to support the findings of this study are available from the corresponding author upon request.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Electrical Engineering Department, Zhengzhou Technical College, Zhengzhou 450121, Henan Province, China. [2]College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China. [3]School of Economics & Management, Zhengzhou Normal University, Zhengzhou 450044, China.

Received: 7 May 2021 Accepted: 16 June 2021
Published online: 05 August 2021

References
1. J. Wells, "This video call may be monitored and recorded": video visitation as a form of surveillance technology and its effect on incarcerated motherhood [J]. Screen Bodies **4**(2), 76–92 (2019)
2. Cai Z, Li D, Deng L, et al. Smart city framework based on intelligent sensor network and visual surveillance [J]. Concurrency and Computation: Practice and Experience **33**(12), 1–10 (2019). https://doi.org/10.1002/cpe.5301.
3. C.R. Karanam, B. Korany, Y. Mostofi, *Tracking from one side: multi-person passive tracking with WiFi magnitude measurements [C]//Proceedings of the 18th International Conference on Information Processing in Sensor Networks* (2019), pp. 181–192
4. Y.C. Chen, X. Zhu, W.S. Zheng, et al., Person re-identification by camera correlation aware feature augmentation [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence **40**(2), 392–408 (2018)
5. Y. Fu, Y. Wei, Y. Zhou, et al., Horizontal pyramid matching for person re-identification [C]//Proceedings of the AAAI Conference on Artificial Intelligence. **33**, 8295–8302 (2019)
6. Y. Lin, L. Zheng, Z. Zheng, et al., Improving person re-identification by attribute and identity learning [J]. Pattern Recogn. **95**, 151–161 (2019)

7. J.K. Kang, T.M. Hoang, K.R. Park, Person re-identification between visible and thermal camera images based on deep residual CNN using single input [J]. IEEE Access **7**, 57972–57984 (2019)
8. Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [EB/OL].[2019-05-07]. https://arxiv.ovg/pdf/1609.02907.pdf.
9. Z. Wang, L. He, X. Gao, et al., Multi-scale spatial-temporal network for person re-identification [C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2052–2056 (2019)
10. Y. Cho, Dynamic RNN-CNN based malware classifier for deep learning algorithm [C]//2019 29th International Telecommunication Networks and Applications Conference (ITNAC). IEEE, 1–6 (2019)
11. H. Liu, Z. Jie, K. Jayashree, et al., Video-based person re-identification with accumulative motion context [J]. IEEE transactions on circuits and systems for video technology **28**(10), 2788–2802 (2017)
12. Khatun A, Denman S, Sridharan S, et al. A deep four-stream Siamese convolutional neural network with joint verification and identification loss for person re-detection [J]. 2018.
13. Wei W, Yang XL, Zhou B, Feng J, Shen PY. Combined energy minimization for image reconstruction from few views. Mathematical Problems in Engineering, vol.2012, article number:154630, doi: https://doi.org/10.1155/2012/154630,2012.
14. W. Wei, X. Fan, H. Song, H. Wang, Video tamper detection based on multi-scale mutual information. Multimed. Tools Appl. **78**(19), 27109–27126 (2019). https://doi.org/10.1007/s11042-017-5083-1
15. Wei, Wei, Dawid Polap, Xiaohua Li, Marcin Woźniak, and Junzhe Liu, Study on remote sensing image vegetation classification method based on decision tree classifier, Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018, p 2292-2297, July 2, 2018, ISBN-13: 9781538692769
16. L. Wu, Y. Wang, S. Ling, M. Wang, 3-D PersonVLAD: learning deep global representations for video-based person reidentification. IEEE Transactions on Neural Networks and Learning Systems **30**(11), 3347–3359 (2019)
17. Kansal Kajal, Venkata Subramanyam, Prasad Dilip K., Kankanhalli Mohan. CARF→Net: CNN attention and RNN fusion network for video→based personreidentification [J]. Journal of Electronic Imaging, **28**(2), 023036.1-023036.12 (2019).

Publisher's Note