MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://mmc.committees.comsoc.org/

MMTC Communications – Review

Vol. 12, No. 1, February 2021

IEEE COMMUNICATIONS SOCIETY

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the February 2021 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises five reviews that cover multiple facets of multimedia communication research including data classification, human pose estimation, node classification, and multi-view video compression. These reviews are briefly introduced below.

The first paper, published in IEEE Internet of Things Journal and edited by Dr. Jinbo Xiong, designed a distributed compressed sensing to establish the general representation of sensor data and classification model based on learning.

The second paper is published in IEEE Transactions on Multimedia and edited by Dr. Wenming Cao. It proposes a novel posture non-maximum suppression (NMS) algorithm specially designed for human posture estimation.

The third paper, published in IEEE Transactions on Neural Networks and Learning Systems and edited by Dr. Guitao Cao, investigates a variation of GCN called adaptive propagation GCN (AP-GCN) is proposed, which endows every node with an additional halting unit that can output a value controlling whether communication should continue for another step.

The fourth paper, published in IEEE Transactions on Image Processing and edited by Dr. Zhiquan He, studies a multi-view graph neural network (MV-GNN) to reduce compression artifacts in multi-view compressed images

The fifth paper, published in IEEE Transactions on Multimedia and edited by Dr. Rui Wang, researches action recognition that uses 2D CNNs, 3D CNNs and hybrid networks as well as the data set for action recognition.

All the authors, nominators, reviewers, editors, and others who contribute to the release of this issue deserve appreciation with thanks.

IEEE ComSoc MMTC Communications – Review Directors

Zhisheng Yan
Georgia State University, USA
Email: zyan@gsu.edu

Yao Liu
Binghamton University, USA
Email: yaoliu@binghamton.edu

Wenming Cao
Shenzhen University, China
Email: wmcao@szu.edu.cn

Phoenix Fang
California Polytechnic State University, USA
Email: dofang@calpoly.edu/span>

Data Classification and Anomaly Detection for IoT

A short review for "A Feature-Based Learning System for Internet of Things Applications"

Edited by Jinbo Xiong

D. Wu, H. Shi, H. Wang, R. Wang and H. Fang, "A Feature-Based Learning System for Internet of Things Applications," in IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1928-1937, April 2019.

With the rapid growth of IoT related applications and the advance of communication technology, massive data collected by sensor nodes for events in IoT has brought benefits to users [1]. Logical relationships of these events learned by reasoning can help users make decisions. Therefore, it is necessary to analyze the raw IoT data to find the internal relationship, and transform the raw data into useful information. As an important data analysis method, the data classification and anomaly detection methods based on machine learning provide more accurate and faster processing because of its self-learning capability. The unknown data can be classified by learning and detecting anomaly event quickly. Apparently, data classification lays the foundation for subsequent anomalous event detection, and the anomaly event detection is the ultimate goal of data classification.

In this paper, the authors propose three challenges to data classification and anomaly detection. Firstly, there are various sensor nodes in IoT, and each sensor node is a data source, and monitoring tasks of each sensor node may be different. As the collected sensor data are transmitted to the sink, the dimension of data received by the sink is high, and the data structure is complex and cannot be easily processed [2]. Secondly, sensor nodes in IoT have their own limitations in terms of the storage space, energy, computing power, communication range, etc. The existing data classification and anomaly detection cannot meet the real-time, accuracy and low-cost requirements [3]. Thirdly, the state change of things could be determined by many factors. It is difficult to analyze the massive sensor data, to detect the data anomaly and to extract the meaningful information [4].

To address these challenges, the authors propose a feature-based learning system for IoT applications. The learning system includes data classification and anomaly event detection process. Taking advantages of the nonparametric learning method, the learning system fully exploits the redundant state of temporal and spatial domains of multiclass sensor data to represented different classes of data. The authors devote their contributions in the following two aspects.

Firstly, the authors propose a distributed compressed sensing to establish the general representation of sensor data and classification model based on learning. Furthermore, the sink extracts the common sparse coefficient matrix and the unique sparse coefficient matrix by learning historical data and accurately identify the class label of data. It is worthy highlining that the proposed classification method-based learning does not need to update the common sparse coefficient matrix and the unique sparse coefficient matrix, thus reducing the complexity of the classification algorithm.

Second, the authors propose to merge the radial basis function neural networks and back propagation neural network (RBF-BP) together to deal with anomaly event detection. Sensor data received by the sink are distributed to neurons in the input layer of RBF subnetwork. Hidden layer of RBF subnetwork later processes the sensor data, and stores the result into the output layer. The outputs of RBF subnetwork are transmitted to the hidden layer of BP subnetwork. In this regard, the training rate and the model generalization ability can be both guaranteed.

Thereafter, the advantage of the proposed data classification scheme is compared with the schemes under KNN, SVM and SRC. It is reported that the proposed scheme can achieve at least 7.598% gain in classification accuracy. Additionally, the proposed anomaly detection algorithm is compared with classical RBR and BP based approaches based on the standard data sample of fire database (SH1, SH4, and SH6) and Intel Lab Data. It is reported that the proposed scheme guarantees lower detection probability error.

In summary, the proposed feature-based learning system as well as the scheme therein is proved to be applicable to data classification and anomaly detection for IoT applications. Moreover, energy consumption and system scalability can also be realized by such system, which potentially benefits other resource-limited IoT applications.

References:

[1] L. Zhou, X. Wang, W. Tu, G.-M. Muntean, and B. Geller, “Distributed scheduling scheme for video streaming over multi-channel multi-radio multi-hop wireless networks,” IEEE J. Sel. Areas Commun., vol. 28, no. 3, pp. 409–419, Apr. 2010.

[2] S. Ji, D. Dunson, and L. Carin, “Multitask compressive sensing,” IEEE Trans. Signal Process., vol. 57, no. 1, pp. 92–106, Jan. 2009.

[3] T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos,“Distributed deviation detection in sensor networks,” ACM Special Interest Group Manag. Data, vol. 32, no. 4, pp. 77–82, 2003.

[4] A. Fawzy, H. M. O. Mokhtar, and O. Hegazy, “Outliers detection and classification in wireless sensor networks,” Egyptian Informat. J., vol. 14, no. 2, pp. 157–164, Jul. 2013.

Jinbo Xiong, is a Professor and Ph.D. supervisor with the Fujian Provincial Key Laboratory of Network Security and Cryptology and the College of Mathematics and Informatics at Fujian Normal University. He received the Ph.D. degree in Computer System Architecture from Xidian University, China, in 2013. He was a visiting scholar with the Department of Computer Science and Engineering at the University of North Texas, Denton, USA. His research interests include connected and autonomous vehicle, secure deep learning, cloud data security, mobile media management, privacy protection, and Internet of Things. He has published more than 100 papers in prestigious journals such as IEEE WCM, IEEE TII, IEEE TCC, IEEE IoT J, IEEE TNSE, FGCS and in major International conferences such as IEEE ICCCN, IEEE TrustCom, IEEE HPCC and IEEE ICPADS. He has applied 15 patents and two monographs in these fields. He is a member of IEEE.

Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation

A short review for "Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation"

Edited by Wenming Cao

Ning, Guanghan, Zhi Zhang, and Zhiquan He. "Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation" IEEE Transactions on Multimedia 20.5 (2017): 1246-1259.

Human pose estimation aims to restore the key points of the human body in images or videos, and depict the shape of the human body. These key points must satisfy a set of geometric constraints and interdependencies imposed by the human body model. Traditional models can capture important problem-specific dependencies between different output variables. However, the main disadvantage of these models is that they need to be designed manually. With the rapid development of convolutional neural networks, human pose estimation has recently made significant progress. However, the deep neural network itself is an algebraic computing system, which is not the most effective way to capture highly complex human knowledge, such as the interdependence between highly coupled geometric features and key points of human posture. In the high-dimensional feature space, the task of human pose estimation is still a very challenging nonlinear manifold learning process. This paper effectively characterizes external knowledge and injects it into the deep neural network to apply appropriate prior learning predictions to guide its training process.

The proposed framework adopts inception-resnet modules [1], [2], [3] and the stacked hourglass structure to construct a fractal network to regress human pose images into heatmaps with no explicit graphical modeling. The network is fractal in that it has the same network configuration at all levels of analysis and abstractions. This fractal network is designed to capture the multi-scale interdependence nature of human pose configuration and to represent these characteristics across different scales and resolutions. The inception network performs channel-wise concatenation of two tensors from different sources. This enforces the information represented by the features stored in these tensors to be complementary to each other. It encourages and directs these two sources to work on different concepts to produce a more robust union representation. The Resnet model performs pixel-wise addition of two tensors with the same number of channels.

This work injects the geometric representation of knowledge into the heatmap layer of the network. Since the heatmaps to be predicted are correlated to each other as they share parameters on former layers, the constraint on one heatmap affects the parameters of these layers and therefore having an impact on the training of other heatmaps. It can be observed that intermediate layers in the network are low and mid-level visual features; higher-level semantic features are hard to locate and explicitly interpret. It is more effective to apply external knowledge and constraints upon to the predicted heatmaps. During the training process, the external knowledge and its visual representations are projected into the background and key point heatmaps using a projection matrix. This type of knowledge-guided learning inherently enforces long-range dependencies and configurations among body joints, while preserving the flexibility of visual representation of the network.

To translate the external knowledge into a higher-dimensional space, this work utilizes two fully-connected layers and three convolutional layers between the projection representation and the injected knowledge to learn linear projection, which will be removed during testing.

In this work, a novel posture non-maximum suppression (NMS) algorithm specially designed for human posture estimation is introduced. This method first finds the spot with the largest response, and then suppresses other spots from the same heat map, and is very close to the spots in other heat maps in the image coordinate system. Repeat this process until all spots are removed. The suppression is performed in the image coordinate system and the channel direction, so it is called the cross heat map NMS.

This paper uses the Percentage Correct Keypoints metric [4] for performance comparisons on the LSP dataset [5], and the PCKh measure [6], where the error tolerance is normalized with respect to the head size, for performance comparisons on the MPII Human Pose dataset [6]. The effectiveness of the proposed inception-resnet module and the benefit in guided learning with knowledge projection is evaluated on two widely used human pose estimation benchmarks.

In summary, this work encodes and injects external human knowledge into deep neural networks to guide its training process with learned projections for more effective human pose estimation. It adopts the stacked hourglass design and propose to use inception-resnet as the building block of our fractal network to regress human pose into heatmaps with no explicit graphical modeling. Utilizing a multi-resolution feature representation with guided learning, the network learns an empirical set of low and high-level features which are typically more tolerant to variations in the training set. Knowledge-guided learning is a generic scheme that can be potentially used to aid other deep neural network training tasks. The effectiveness of the proposed inception-resnet module and the benefit in guided learning with knowledge projection is evaluated on two widely used human pose estimation benchmarks.

References:

[1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for imagerecognition,” inProc. IEEE Conf. Comput. Vis. Pattern Recogn., 2016, pp. 770–778.

[2] C. Szegedyet al., “Going deeper with convolutions,” inProc. IEEE Conf.Comput. Vis. Pattern Recogn., 2015, pp. 1–9.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” inProc. IEEE Conf. Comput. Vis. Pattern Recogn., 2016, pp. 2818–2826.

[4] Y. Yang and D. Ramanan, “Articulated human detection with flexiblemixtures of parts,”IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2878–2890, Dec. 2013.

[1] S. Johnson and M. Everingham, “Clustered pose and nonlinear appearance models for human pose estimation,” inProc. Brit. Mach. Vis. Conf., 2010, pp. 12.1–12.11.

[2] M.Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D human poseestimation: New benchmark and state of the art analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recogn., 2014, pp. 3686–3693.

Wenming Cao received the M.S. degree from the System Science Institute, China Science Academy, Beijing, China, in 1991, and the Ph.D. degree from the School of Automation, Southeast University, Nanjing, China, in 2003. From 2005 to 2007, he was a Post-Doctoral Researcher with the Institute of Semiconductors, Chinese Academy of Sciences, Beijing, China. He is currently a Professor with Shenzhen University, Shenzhen, China. He has authored or coauthored over 80 publications in top-tier conferences and journals. His research interests include pattern recognition, image processing, and visual tracking.

Adaptive Propagation Graph Convolutional Network

A short review for "Adaptive Propagation Graph Convolutional Network"

Edited by *Guitao Cao*

Spinelli I , Scardapane S , Uncini A . Adaptive Propagation Graph Convolutional Network[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, PP.

Deep learning has achieved great success on many tasks, such as computer vision [1] and natural language processing [2]. But the data processed by these tasks are all Euclidean Structure Data. Then a major research question is how to generalize deep learning to other types of data that is Non-Euclidean Structure Data, by the implementation of novel differentiable blocks. Graph structure data is an important Non-Euclidean Structure Data and widely exists in real life, ranging from social networks, logistics transportation, recommendation system, fraud detection, biomedical applications and many others.

Many graph neural networks (GNNS) models have been proposed to deal with some tasks such as node classification and edge prediction in recent years. Graph convolutional neural (GCN) [3] network is a simple and effective method of graph processing. GCN are built by interleaving vertex-wise operations with a communication step. In practice, a single GCN layer provides a weighted combination of information across neighbors, representing a localized 1-hop exchange of information. Taking the GCN layer as a fundamental building block, several research questions have received vast attention lately, most notably:(i) how to design more effective communication protocols, able to improve the accuracy of the GCN and potentially better leverage the structure of the graph [4]–[6]; and (ii) how to trade-off the amount of local (vertex-wise) operations with the communication steps [7].

In this paper, a variation of GCN called adaptive propagation GCN (AP-GCN) is proposed, which endows every node with an additional halting unit that can output a value controlling whether communication should continue for another step (combining the information from neighbors farther away), or should stop. In this way, the models can select the number of communication steps independently for every node and this number is adapted and computed on-the-fly during training. And Adaptive Propagation GCN (AP-GCN) is the only model in the literature combining these two properties. In order to implement this adaptive unit, the previous work on adaptive computation time in recurrent neural networks [8] is leveraged to design a differentiable method to learn this propagation strategy.

Thus the authors' major contribution is to propose an adaptive propagation GCN model, which allowed to vary the number of communication steps independently for each vertex to improve the performance of GCN. However, the vast majority of proposals has considered a single, maximum number of communication steps that is shared for all the nodes in the graph.

AP-GCN separates the node-wise operations from the propagation step. The former is implemented with a generic NN applied on a single node. This embedding is then used as the starting seed for a propagation step. Key to our proposal, the number of propagation steps depends on the index of node i and it is computed adaptively while propagating. The mechanism to implement this is inspired by the adaptive computation time in RNNs [8]. First, we endow each node with a linear binary classifier acting as a 'halting unit' for the propagation process. After the generic iteration k of propagation, we compute node-wise. In order to ensure that the number of propagation steps remains reasonable, following we adopt two techniques. Firstly, we fix a maximum number of iterations T. Secondly, we use the running sum of the halting values to define a budget for the propagation process. In the end, the number of propagation steps can be controlled by the definition of a propagation cost.

Extensive experiments demonstrate the improved performance of the proposed adaptive propagation GCN (AP-GCN). The results show that model can achieve superior or similar results to the best proposed models so far on a number of benchmarks, while requiring a small overhead in terms of additional parameters.

In summary, the proposed adaptive propagation GCN (AP-GCN) selects automatically the number of propagation steps performed for each node across the graph. We showed experimentally that the method performs favorably or better than the state-of-the-art, that it is robust to the training set size and, in most cases, it can adapt its behavior to the dataset more or less robustly depending on the hyper-parameter's choice.

References:

[1] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. Densely connected convolutional networks. IEEE Conference on Computer Vision and Pattern Recognition, 2017

[2] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015

[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," Proc. International Conference on Learning Representations (ICLR), 2017.

[4] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, " Cayleynets: Graph convolutional neural networks with complex rational spectral filters," IEEE Transactions on Signal Processing, vol. 67, no. 1, pp. 97–109, 2018.

[5] F. M. Bianchi, D. Grattarola, L. Livi, and C. Alippi, "Graph neural networks with convolutional ARMA filters," arXiv preprint arXiv:1901.01343, 2019.

[6] J. Klicpera, S. Weißenberger, and S. G¨unnemann, "Diffusion improves graph learning," in Advances in Neural Information Processing Systems, 2019, pp. 13 333–13 345.

[7] J. Klicpera, A. Bojchevski, and S. G¨unnemann, "Predict then propagate: Graph neural networks meet personalized pagerank," in Proc. 2019 International Conference on Learning Representations (ICLR), 2019.

[8] A. Graves, "Adaptive computation time for recurrent neural networks,"arXiv preprint arXiv :1603.08983, 2016.

Guitao Cao obtained her Ph.D. in 2006 from Shanghai Jiao Tong University with a focus on pattern recognition. She is currently a professor of Software Engineering Institute, East China Normal University (ECNU), Shanghai. ECNU is the top tier university in China with a high rank (Level A) in Software Engineering in China. She was also a visiting researcher with University of Missouri Columbia. She has published decades of peer reviewed papers in top venues including IEEE Transactions on Cybernetics, IEEE Transactions on Multimedia, and IEEE Transactions on Biomedical Engineering. Prof. Cao is also the Principal Investigator for many research funding with major sponsors including the National Science Foundation of China, Ministry of Industry and Information Technology of the People's Republic of China, and Science Foundation of Shanghai. Her research interests include pattern recognition, image processing and machine learning.

Multi-View GNN for Reducing Compression Artifacts in Multi-view Compressed Images

A short review for "MV-GNN: Multi-View Graph Neural Network for Compression Artifacts Reduction"

Edited by Zhiquan He

He, X., Liu, Q., & Yang, Y. (2020). MV-GNN: Multi-view graph neural network for compression artifacts reduction. IEEE Transactions on Image Processing, 29, 6829-6840.

Multi-view video (MVV) [1] has been adopted as the fundamental data representation in many three dimension (3D) and interaction oriented visual applications, which can provide immersive perceptual viewing experience of a scene [2]. However, a large amount of redundant data has been a challenge for those multi-view video based systems, which need efficient compression technology for transmission. Inevitable compression artifacts in multi-view video (MVV) can clearly degrade the quality of experience in many interaction-oriented 3D visual applications. Under the framework of asymmetric coding, low quality images can be enhanced with high-quality images from the neighboring viewpoints considering the similarity among different views.

With the success of deep learning in many computer vision and image processing tasks, learning-based algorithms have been proposed to reduce the artifacts in the compressed images.
Recently, graph neural networks (GNN) [3] have been widely applied since they have a strong inductive bias and capability for relational reasoning. The input data in GNN is constructed as a graph and the hidden representation of all nodes in graph are recurrently updated based on the aggregation and update mechanism. Since graph is good at describing the relationship between nodes, the correlation among different viewpoints can be effectively exploited by using recursive message passing to iteratively propagate information over the graph. Suitable graph construction and propagation model should be designed for the compressed MVV artifacts reduction.

Compression artifacts and warping error cause different cross-view quality gaps for various sequences, and thus the contribution of cross-view priors can hardly be located and considered in previous works. As for the multi-image methods, the inter-view and inter-frame priors can be exploited to reduce compression artifacts [4]. In this paper, a multi-view graph neural network (MV-GNN) is proposed to reduce compression artifacts in multi-view compressed images. In this work, the low-quality image from current viewpoint and warped images from neighboring viewpoints are separately processed by the feature extraction net with multiscale residual blocks. Then a *K*-neighbor graph is constructed based on the feature similarity in the dynamic GNN-based fusion net. The image feature vector is taken as the initial representation for each node, which is iteratively updated using a recurrent function. After several steps of propagation, the final hidden state of each node is used to derive the enhanced result using the reconstruction net.

The authors' major contribution is proposed a multi-view graph neural network (MV-GNN) to reduce compression artifacts in multi-view compressed images. They dedicate to design a fusion mechanism which can exploit contributions from neighboring viewpoints and meanwhile suppress the misleading information. In their method, a GNN-based fusion mechanism is designed to fuse the cross-view information under the aggregation and update mechanism of GNN.

Cross-view priors are not always helpful due to compression artifacts and warping error, especially more unpredictability are created under different coding conditions. The existing method cannot be directly utilized for compressed MVV since it cannot adaptively utilize the valuable priors and meanwhile suppress the misleading priors from different distorted multi-view images. Since GNN is good at relational reasoning and inductive bias, it can be used to exploit more valuable cross-view priors by making adaptive adjustments under different coding conditions. The proposed of GNN-based fusion mechanism can fuse multi-view feature maps better. The author first adaptively utilizes the multi-view priors to reduce compression artifacts in the compressed MVV images by designing an end-to-end residual network. In this network, multi-scale residual blocks are employed to detect feature maps for the HEVC-compressed images [5] in different scales, which can provide more representative information for the following feature fusion. Besides of the adaptation, a GNN-based fusion mechanism is proposed for better reconstruction of multi-view priors in MVV. In this mechanism, a *K*-neighbor graph is constructed on the multi view fused feature maps. Then the aggregation and update mechanism of GNN is utilized to adaptively exploit the useful priors from compressed MVV and meanwhile suppress misleading priors.

Experimental results showed that the method outperforms other methods over all sequences both in average PSNR and SSIM gain. Meanwhile, better visual performance has been achieved by MV-GNN. In summary, compression artifacts in MVV can be properly reduced by the proposed MV-GNN. Good quality images can be obtained for the end-user even with low bandwidth, which provides a promising solution for future remote 3D visual applications.

References:

[1] K. Muller et al., “3D high-efficiency video coding for multi-view video and depth data,” IEEE Trans. Image Process., vol. 22, no. 9, pp. 3366–3378, Sep. 2013.

[2] Smolic et al., “3D video and free viewpoint video technologies, applications and MPEG standards,” in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2006, pp. 2161–2164.

[3] F. Scarselli, M. Gori, A. Chung Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” IEEE Trans. Neural Netw., vol. 20, no. 1, pp. 61–80, Jan. 2009.

[4] Y. Xie, J. Xiao, T. Tillo, Y. Wei, and Y. Zhao, “3D video super-resolution using fully convolutional neural networks,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2016, pp. 1–6.

[5] G. J. Sullivan, J. M. Boyce, Y. Chen, J.-R. Ohm, C. A. Segall, and A. Vetro, “Standardized extensions of high efficiency video coding (HEVC),” IEEE J. Sel. Topics Signal Process., vol. 7, no. 6, pp. 1001–1016, Dec. 2013.

Zhiquan He is currently working as assistant professor in College of Information Engineering, Shenzhen University, China. He received his M.S. degree from Institute of Electronics, Chinese Academy of Sciences in 2001, and the PhD degree from the department of Computer Science, University of Missouri-Columbia in 2014. His research area is in the areas of image processing, computer vision and machine learning.

Convolutional Networks With Channel and STIPs Attention Model for Action Recognition

A short review for "Convolutional Networks With Channel and STIPs Attention Model for Action Recognition in Videos"

Edited by Rui Wang

Wu, Hanbo, Xin Ma, and Yibin Li. "Convolutional Networks With Channel and STIPs Attention Model for Action Recognition in Videos." IEEE Transactions on Multimedia VOL. 22, NO. 9, SEP.2020.

Action recognition has received more and more attention in the field of computer vision in recent years. It is widely used in many fields such as intelligent surveillance, video retrieval and elderly care and so on. With the significant improvement in deep learning, convolutional neural networks have attracted widespread attention and been applied to many deep learning tasks including action recognition. The major research of action recognition tends to use 2D CNNs, 3D CNNs and hybrid networks and the data set for action recognition is mainly based on RGB data or RGB-D data.

However, there are still some problems though the method based on CNN has achieved remarkable success in action recognition. First, CNNs lack the ability to model the long-term temporal dependence of the entire video. Second CNNs usually perform within a limited field of view which leads that CNNs lack the ability to focus on the information about the area of motion of the human body. However, the importance of the action is different in different time and space. Under the same attention mechanism, the time and area that are not very relevant to the action will affect the accuracy of action recognition.

To solve these problems, a novel video-based action recognition framework is proposed in this paper. The video with dynamic image sequences (DISs), which effectively describe the video by modeling local spatiotemporal dynamics and correlation is presented to model the long-term temporal dependence of the entire video. In addition, a channel based on CNNs and spatiotemporal interest point attention model (STIPs) is proposed to focus on the discriminative channel and the spatial movement area of human actions in the network.

To effectively learn the spatiotemporal dynamics of long-term action videos, an image segmentation algorithm containing only a few dynamic frames is presented in which use depth CNNs to process the generated dynamic image sequence frames in this paper. Inspired by [1], new temporal pooling which use a paired linear sorting machine to learn a linear function. Its parameters can encode the frame order in the video and use it as a new video representation. The frames in each segment are aggregated to generate a single dynamic image, and the temporal pooling is directly applied to the pixels of the image to describe the local spatial-temporal dynamics. The output of the convolutional layer is used as the input of the channel and STIPs attention model.

To enhance the feature learning ability of CNNs, the author proposes a channel and STIPs attention model (CSAM). The model contains both channel attention (CA) for discovering the discriminative channels and STIPs attention (SA) to focus on the regions for efficient action recognition. In the Channel Attention Module, global average pooling across the spatial dimension of each feature map is used to the features of the last convolutional layer in the dynamic image sequence. In the STIPs Attention Module, background suppression and local and temporal constraints [2] are applied to obtain the more effective STIPs. The spatial attention weight is generated by learning the distribution of STIPs projection in the feature map space.

Comparisons and ablation study is experimented in this paper in three dataset —— the SDUFall dataset, the SBU Kinect interaction dataset and the NTU RGB + D dataset. The author improved the important influence of different components of DIS[3]，CSAM and LSTM in ablation study section. In addition, the author compared the accuracy with different method and come to the conclusion that the method of this paper can achieve the accuracy of 98.82 which is than other pervious method.

To brief the main contribution of the author is to a channel attention model for improving the action recognition performance in depth videos based on CNNs which can solve the problem of the lack of ability to model the long-term temporal dependence of the entire video of CNNs and focus on the information about the area of motion of the human body. The influence of different parts is improved in experiment and the method of paper can achieve higher accuracy than before.

References:

[1] H. Bilenet al., “Dynamic image networks for action recognition,” inProc. IEEE Conf. Comput. Vision Pattern Recognit., Las V egas, NV , USA, 2016, pp. 3034–3042.

[2] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 3, pp. 257–267, Mar. 2001.

【3】 L. Wang, Y. Qiao and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors", Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 4305-4314, 2015.

[4] R. Poppe, "A survey on vision-based human action recognition", Image Vision Comput., vol. 28, no. 6, pp. 976-990, Jun. 2010.

Rui Wang received his B.S. and Ph.D. degree both in Electronics and Information Engineering from Xidian University, Xi'an, Shaanxi, China, in 2004 and in 2009, respectively. Now he is currently the full professor in School of Communication and Information Engineering, Shanghai University, China, and the senior member of IEEE, and Chinese Institute of Electronics. Recently, he serves as technical project evaluation expert for Shanghai Municipality, National Natural Science Foundation of China. He was awarded outstanding teacher award of baosteel in China, 2017.

Dr. Wang's research interests include Internet of Things, Intelligent Information Processing, Pattern Recognition. His researches are supported by National Natural Science Foundation of China (NSFC), Key projects of the ministry of education of China, China Academy of Railway Sciences, Foundation of Science and Technology Commission of Shanghai Municipality, Shanghai academy of environmental sciences, and so on. He firstly introduced the theory of geometric algebra to the field of multi-dimensional signal processing. Until now, he has published 5 books, approximately 40 professional journals and more than 20 scientific conferences including IEEE ACCESS, SCIS, PRL, IJRS, IJDSN, AACA, FITEE, JSPS, CPL, CMC, ICME, ICPR, ICSP, DataCOM, ICALIP, obtained 3 patent applications and registered 4 software copyrights in China.

Paper Nomination Policy

Following the direction of MMTC, the Communications – Review platform aims at providing research exchange, which includes examining systems, applications, services and techniques where multiple media are used to deliver results. Multimedia includes, but is not restricted to, voice, video, image, music, data and executable code. The scope covers not only the underlying networking systems, but also visual, gesture, signal and other aspects of communication. Any HIGH QUALITY paper published in Communications Society journals/magazine, MMTC sponsored conferences, IEEE proceedings, or other distinguished journals/conferences within the last two years is eligible for nomination.

Nomination Procedure

Paper nominations have to be emailed to Review Board Directors: Zhisheng Yan (zyan@gsu.edu), Yao Liu (yaoliu@binghamton.edu), Wenming Cao (wmcao@szu.edu.cn), and Phoenix Fang (dofang@calpoly.edu). The nomination should include the complete reference of the paper, author information, a brief supporting statement (maximum one page) highlighting the contribution, the nominator information, and an electronic copy of the paper, when possible.

Review Process

Members of the IEEE MMTC Review Board will review each nominated paper. In order to avoid potential conflict of interest, guest editors external to the Board will review nominated papers co-authored by a Review Board member. The reviewers' names will be kept confidential. If two reviewers agree that the paper is of Review quality, a board editor will be assigned to complete the review (partially based on the nomination supporting document) for publication. The review result will be final (no multiple nomination of the same paper). Nominators external to the board will be acknowledged in the review.

Best Paper Award

Accepted papers in the Communications – Review are eligible for the Best Paper Award competition if they meet the election criteria (set by the MMTC Award Board). For more details, please refer to http://mmc.committees.comsoc.org/.

MMTC Communications – Review Editorial Board

Multimedia Communications Technical Committee Officers

MMTC examines systems, applications, services and techniques in which two or more media are used in the same session. These media include, but are not restricted to, voice, video, image, music, data, and executable code. The scope of the committee includes conversational, presentational, and transactional applications and the underlying networking systems to support them.