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

MMTC Communications – Review

Vol. 14, No. 1, February 2023

IEEE COMMUNICATIONS SOCIETY

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the February 2023 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises four reviews that cover multiple facets of multimedia communication research including remote visual monitoring, multimodal emotion analysis, ambient backscatter communication, and vision transformer model compression. These reviews are briefly introduced below.

The first paper, published in IEEE Transactions on Multimedia and edited by Dr. Lucile Sassatelli，It is the definition of class prototypes obtained from supervised learning on labeled samples, and introduce several additional elements, including prototype refinement by averaging class samples with their nearest neighbors, a margin in the contrastive loss, and a self-paced weighting mechanism to progressively increment the importance of the contrastive loss in the total training loss.

The second paper, edited by Dr. Qin Wang, was also published in IEEE Multimedia. This paper is developed to learn fine-grained and coarse-grained know-ledge via a hierarchical and intra-modal semantic graph.

The third paper, edited by Dr. Wenming Cao, was published in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. It proposes a novel opportunistic ambient backscatter proposes a consistency sensitivity guided ensemble attack (CSEA) method for highly efficient search and estimate the victim model in the high-dimensional model space.

The fourth paper, published in the 2021 IEEE MultiMedia and edited by Dr. Zhiquan He, proposes a novel graph convolutional network (GCN) utilizing multiscale graphs derived from facial landmarks for facial expression recognition.

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

IEEE ComSoc MMTC Communications – Review Directors

Yao Liu
Rutgers University, USA
Email: yao.liu@rutgers.edu

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

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

Ye Liu
Macau University of Science and Technology, Macau, China
Email: liuye@must.edu.mo

A look into semi-supervised learning for multimedia communication

A short review for "Semi-supervised Contrastive Learning with Similarity Co-calibration"
Edited by Lucile Sassatelli

Y. Zhang, X. Zhang, J. Li, R. Qiu, H. Xu, and Q. Tian, "Semi-supervised Contrastive Learning with Similarity Co-calibration," IEEE Transactions on Multimedia, pp. 1–1, 2022, doi: 10.1109/TMM.2022.3158069.

Machine learning approaches have become central in tackling several problems in multimedia communication and analysis. This is specifically the case for immersive contents in the form of 360° videos meant to be watched and experienced in a Virtual Reality (VR) headset, or in the form of VR applications enabling a full-fledged embodied experience for the user able to freely move in six degrees of freedom [1]. These content modalities raise challenging questions on understanding and predicting the human behavior and attentional processes. For example, it is known that predicting head and gaze movement when a 360° video is streamed over the Internet is key to minimize the consumed data rate while maximizing the visual quality in the user's field of view, in order to focus the bandwidth budget on the pixels actually watched. This is however a challenging task where motion uncertainty must be taken into account into the machine learning model regressing the future coordinates [2]. The embodied experiences allowed by VR systems also open new opportunities for, e.g., patient rehabilitation and training. Wu et al. have shown the challenges of inferring user intentions from behaviors, which can however enable designing serious game environments for specific tasks considering possible user behaviors.

These questions require to have data annotated with user intentions, to in turn train deep learning models to infer user intentions, through extended experiments and questionnaires. As for other supervised machine learning tasks, manually annotated data are difficult to collect and this can impede the development of efficient models. To counteract this difficulty, one can be interested in labeling only a fraction of the data, while leveraging numerous unlabeled data to train a model and obtain better performance than those obtained if trained on the labeled data only. This corresponds to semi-supervised learning, one of the seminal works introducing pseudo-labeling [3]. Pseudo-labeling consists in supervising the model with the unlabeled data assigned to the classes predicted by the current model. While simple, such an approach can easily perform poorly and be trapped in local minima. Consistency regularization has been introduced to help supervise the model by augmenting the unlabeled data by perturbing these data samples, therefore creating new classes leveraged with contrastive learning in a pre-training phase. The model is then fine-tuned with few-shot labeled samples.

In their article [4], Zhang et al. show that performing these two training steps in a disconnected way is suboptimal. They instead propose a novel end-to-end training strategy where the contrastive learning should benefit from the class priors on the unlabeled data, and the class priors can be refined with an improved data representation learned contrastively. Their method is named Semi-supervised Contrastive Learning (SsCL), combining pseudo-labeling and contrastive learning.

The main ingredient is the definition of class prototypes obtained from supervised learning on labeled samples. The unlabeled samples are then compared with every class prototype to identify the most likely class of each unlabeled sample. Once the class is identified, the K nearest unlabeled neighbors are selected to act as positive examples, and other K to act as negative examples. The contrastive loss therefore consists in minimizing the distances between the representation of the unlabeled data with the positive examples, while maximizing those with the negative examples. The resulting updated representation is in turn used to refine the classification of the labeled data, from which new prototypes are obtained to improve the selected positive and negative samples used in the contrastive loss involving the unlabeled samples.

To stabilize the training process, the authors introduce several additional elements, including prototype refinement by averaging class samples with their nearest neighbors, a margin in the contrastive loss, and a self-paced weighting mechanism to progressively increment the importance of the contrastive loss in the total training loss.

The authors show interesting improvements on the CIFAR-10 and ImageNet datasets. For example, on CIFAR-10, when using only 40 labels, error rates can be decreased by 3 percentage points compared to state-of-the-art.

This work spurs exciting questions for machine learning approaches for immersive media and embodied experiences. Beyond image data, considering time series for trajectory prediction will be key, as well as multimodal models leveraging physiological data such as heart rate or skin conductance, and the 3D environment described in a structured (high-level description of objects, their affordance and space occupancy) or unstructured (point clouds) way. Finally, how incorporating knowledge into this kind of training approaches is a promising challenge, where exploiting high-level knowledge on human interactions and behaviors could help benefiting from unlabeled data.

References:

[1] Hui-Yin Wu, Florent Robert, Théo Fafet, Brice Graulier, Barthelemy Passin-Cauneau, Lucile Sassatelli, and Marco Winckler. 2022. Designing Guided User Tasks in VR Embodied Experiences. Proc. ACM Hum.-Comput. Interact. 6, EICS, Article 158 (June 2022), 24 pages. https://doi.org/10.1145/3532208

[2] Quentin Guimard, Lucile Sassatelli, Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. 2022. Deep variational learning for multiple trajectory prediction of 360° head movements. In Proceedings of the 13th ACM Multimedia Systems Conference (MMSys '22). Association for Computing Machinery, New York, NY, USA, 12–26. https://doi.org/10.1145/3524273.3528176

[3] D.-H. Lee, "Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks", ICML, 2013.

[4] Y. Zhang, X. Zhang, J. Li, R. Qiu, H. Xu, and Q. Tian, "Semi-supervised Contrastive Learning with Similarity Co-calibration," IEEE Transactions on Multimedia, pp. 1–1, 2022, doi: 10.1109/TMM.2022.3158069.

Lucile Sassatelli is a full professor at Université Côte d'Azur, France. She obtained a PhD from Université of Cergy Pontoise, France, in 2008 and her professorial habilitation in 2019. She was a postdoctoral fellow at MIT, Cambridge, USA. She has been nominated a junior fellow of Institut Universitaire de France (IUF) in 2019. She is the scientific director of the French AI School project in Côte d'Azur. She leads several projects focusing on AI for multimedia transmission of immersive content, and analysis of gender representations in multimedia contents.
She has published papers in prestigious journals such as IEEE Transactions on Information Theory, IEEE Transactions on Wireless Communications, IEEE Transactions on Pattern Analysis and Machine

Deep Multigraph Hierarchical Enhanced Semantic Representation

A short review for "Deep Multigraph Hierarchical Enhanced Semantic Representation for Cross-Modal Retrieval"

Edited by Qin Wang

L. Zhu, C. Zhang, J. Song, S. Zhang, C. Tian and X. Zhu, "Deep Multigraph Hierarchical Enhanced Semantic Representation for Cross-Modal Retrieval," in IEEE MultiMedia, vol. 29, no. 3, pp. 17-26, 1 July-Sept. 2022.

Cross-modal retrieval aims to search instances of different modalities according to a query of a specific modality, which is applied in many scenarios, such as multimedia search, recommendation system [1], and VQA. The main challenge of cross-modal retrieval is how to efficiently narrow the semantic gap and heterogeneity gap between different modalities. By the aid of statistics-based approaches, e.g., canonical correlation analysis (CCA) [2] and latent Dirichlet allocation (LDA) [3], the traditional research works learn cross-modal correlation or global semantic relations [4]. How-ever, these approaches are constructed by shallow models, which are unable to effectively represent complex nonlinear semantic correlations across different modalities. Thanks to the flourishing of deep learning technology, considerable researchers recently applied this powerful tool to generate cross-modal/multiview representations. Compared with traditional techniques, deep models (CNN, RNN, etc.) have stronger semantic representation capability and larger parameter capacity, which can capture more high-level semantics from multi-modal content.

Although the existing research works have proeminently improved cross-modal retrieval via multifarious deep models, semantic gap and heterogeneous gap still need to be further narrowed. Specifically, on the one hand, the latest approaches focused on capturing high-level semantics, but ignored how to realize hierarchical multigrained knowledge discovery and semantic fusion. On the other hand, some studies considered fine-grained semantic information representation, but they reckon without the distribution of relations between semantic details in multi-modal instances. This deficiency makes it difficult to accurately align the cross-modal semantics, which hinders the further improvement of cross-modal retrieval performance. To this end, this paper investigates an efficient method to achieve two tasks: 1) explore more multigrained cross-modal semantic knowledge and 2) realize semantic relation distribution alignment efficaciously as well.

To capture multigrained semantic knowledge and reduce the cross-modal heterogeneity, authors propose a novel cross-modal representation technique, termed as deep multigraph-based hierarchical enhanced semantic representation (MG-HESR). Besides, to further narrow the semantic gap, a novel semantic relation alignment method is proposed, which aims to reduce the discrepancy of relation distribution between semantic details in cross-modal data. To realize this idea, authors construct three-level (scene, action, and object) semantic graphs from the details of images and texts to encode the hierarchical fine-grained semantic knowledge via a graph convolutional network (GCN). Moreover, intramodal semantic graphs are constructed via kNN method to model the coarse-grained semantic relationships. Then, these multigrained semantic knowledge imbedding's are fused via a novel fusion neural network. A new loss function, called semantic distribution alignment loss, is proposed to implement the fine-grained semantic relation alignment. In addition, cross-modal adversarial learning is integrated to produce modality-invariant representations.

In this paper, authors propose an end-to-end deep cross-modal representation model via integrating deep feature embedding, MG-HESR, and cross-modal adversarial learning. This model consists of three main components: 1) cross-modal feature embedding, 2) MG-HESR, and 3) cross-modal adversarial learning.

The first component aims at embedding images and texts into their feature subspaces via deep neural networks. In this work, authors use an Alex-

Net-like model that consists of five convolutional layers with max-pooling operation, and two layers of fully connected layers.

The second component is a novel multigraph-based enhanced semantic representation model. The main idea is to utilize multigraph models, i.e., two scene graph-based hierarchical semantic graphs and two intra-modal semantic graphs, to represent multigrained semantic knowledge, one graph per modality. Then, the semantic knowledge is fused for modality-invariant representation learning. It employs two types of semantic graph representation learning methods, i.e., a scene graph-based hierarchical semantic representation learning, and intra-modal semantic representation learning. The former is used to capture fine-grained knowledge from images and texts, and the latter is used to model coarse-grained semantic knowledge across the objects of the same modality. The multigrained knowledge is aggregated by GCN and then fused via a semantic fusion network. In addition, a novel cross-modal semantic distribution alignment method is proposed. Specifically, authors use a transformation function to transform the hierarchical semantic representation model of each modality semantic relation distributions into matrix. These matrices can represent the semantic relation between nodes in a scene graph-based model. A novel Kullback–Leibler divergence-based measurement, is proposed to measure the semantic relation distribution similarity between these matrices.

The third component consists of a cross-modal encoder with weight-sharing, two decoders, and three discriminators. It is used to realize intra-modal and inter-model adversarial learning to map cross-modal embeddings into a common representation subspace.

Authors compare the method MG-HESR with seven state-of-the-art methods. To evaluate the performance of our method comprehensively, two metrics, i.e., mean average precision (mAP) scores and precision–recall curve (PR-curve), are used. The results verifies that MG-HESR performs better than other approaches, which means that the combination of coarse-grained and fine-grained semantic knowledge can effectively improve the semantic representation learning.

In summary, an MG-HESR method is developed to learn fine-grained and coarse-grained knowledge via a hierarchical and intra-modal semantic graph, which is able to capture multigrained semantic knowledge and the narrow semantic gap. Besides, a novel semantic relation distribution alignment method is proposed to reduce discrepancy of fine-gain semantic relation distribution. In addition, a cross-modal adversarial learning network is utilized to reduce heterogeneity. Com-pared with the state-of-the-art methods, MG-HESR effectively improves cross-modal retrieval performance.

References:

[1] J. Pang, D. Zhang, H. Li,W. Liu, and Z. Yu,“Hazy Re-ID: An interference suppression model for domain adaptation person re-identification under inclement weather condition,” in Proc. IEEE Int. Conf. Multimedia Expo., 2021, pp. 1–6.

[2] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” Neural Computation., vol. 16, no. 12, pp. 2639–2664, 2004.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[4] X. Zhan et al., “Product1M: Towards weakly supervised instance-level product retrieval via cross-modal pretraining,” in Proc. IEEE/CVF Int. Conf. Computation. Vis., 2021, pp. 11782–11791.

Qin Wang, Ph.D, is an Associate Professor at Nanjing University of Posts and Telecommunications (NJUPT), China. She received B.S. and Ph.D degrees from NJUPT, in 2011 and 2016. Prior to joining NJUPT, she was with the New York Institute of Technology (NYIT) between Feb. 2017 and Aug. 2020. From July 2018 to June 2020, she was a Postdoctoral Research Fellow at NJUPT. From 2015 to 2016, she was a visiting scholar at San Diego State University, USA. Her research interests include multimedia communications, multimedia pricing, resource allocation in 6G, and Internet of Things. She has published papers in prestigious journals such as IEEE Transactions on Vehicular Technology and IEEE Communications Magazine, in prestigious conferences such as IEEE INFOCOM SDP Workshops.

Consistency-Sensitivity Guided Ensemble Black-Box Adversarial Attacks in Low-Dimensional Spaces

A short review for "Optimizing Black attack performance by Exploiting a consistency and sensitivity guided ensemble attack (CSEA)method in a low-dimensional space"

Edited by Wenming Cao

Yuan J, He Z. Consistency-Sensitivity Guided Ensemble Black-Box Adversarial Attacks in Low-Dimensional Spaces[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7778-7786.

Deep neural networks are sensitive to adversarial attacks [1,2]. There are two types of adversarial attacks, white-box attacks, which have full access to the victim network, and black-box attacks, which have no knowledge about the network. This paper mainly discusses the challenge of black-box attacks on deep neural networks, which are known to be highly sensitive to adversarial attacks. A small amount of adversarial noise added to an input image can successfully fool a state-of-the-art classifier with a high probability.

Black-box attacks are a more challenging problem since the attacker can only query the victim network and obtain its output score for a given input image. In black-box attack research, the objective is to minimize the number of queries submitted to the victim network while achieving a high attack success rate. There are two major approaches that have been explored in the literature for black-box attacks: transfer-based [3,4] and query-based [5,6]. The transfer-based approach uses a trained surrogate network to generate attacks based on the white-box approach, hoping that this attack noise can be effectively transferred to the unknown target network. However, this approach often suffers from low success rates since the adversarial attack is a sophisticated error accumulation process depending on the specific parameter settings of the victim model and the input image. The query-based approach queries the target network continuously, searches or modifies the attack noise based on the query score feedback using gradient descent or other optimization methods. This approach often needs a larger number of queries since both the victim model and the input image have extremely high dimensions, and the gradient-based searches and attack noise optimization in such high-dimensional spaces involve a large number of search steps and queries. Currently, the average number of queries achieved by existing state-of-the-art methods remains high, often in the range of a few hundreds or even thousands.

The success of both targeted and untargeted attacks relies on significantly deviating the output score of the victim network from the correct score of the original image. Achieving a high attack success rate depends on how well the victim model is approximated and how the detailed network responses of the input image are characterized and exploited. However, high dimensionality presents a central challenge in black-box attacks. The adversarial noise in black-box attacks has the same dimension as the input image, and the search complexity typically increases exponentially with the number of dimensions. Meanwhile, the surrogate model must approximate the victim model in the black box, which has an unknown network structure and millions of model parameters.

To address the challenge of black-box attacks, this paper proposes a consistency sensitivity guided ensemble attack (CSEA) method for highly efficient search and estimate the victim model in the high-dimensional model space. The central idea is to construct an ensemble of surrogate models with diversified structures that perform collaborative search in the high-dimensional model space. The authors then estimate or approximate the victim model using a learned linear composition of these surrogate models.

Using random block masks on the input image, these surrogate models jointly construct and submit randomized and sparsified queries to the victim model. Based on the feedback results of these highly diversified queries, the surrogate models are able to effectively learn and evolve themselves in the model space. Guided by a consistency constraint, their learned composition is able to approximate the victim network very efficiently using a very small number of queries. Furthermore, the block-wise randomized and sparsified queries provide important information for the authors to estimate the attack sensitivity map for the input image. Using this sensitivity map, the authors can perform block-based local refinement of the attack to further increase its success rate.

The experimental results of the proposed CSEA approach show that it significantly reduces (by up to 50%) the number of needed queries to the victim network to achieve successful attacks compared to the state-of-the-art black-box attack methods. The authors attribute this success to the diversity of the surrogate models and the use of a consistency constraint to guide the model search process.

Overall, this work highlights the ongoing challenge of black-box attacks on deep neural networks and the need for more efficient methods to minimize the number of queries submitted to the victim network while achieving a high attack success rate. The proposed CSEA approach provides a promising solution to this challenge, and its success in reducing the needed queries to the victim network is significant in terms of its practicality for real-world applications.

References:

[1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.

[2] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C]. International conference on machine learning. PMLR, 2017: 214-223.

[3] Cheng S, Dong Y, Pang T, et al. Improving black-box adversarial attacks with a transfer-based prior[J]. Advances in neural information processing systems, 2019, 32.

[4] Huang Z, Zhang T. Black-box adversarial attack with transferable model-based embedding[J]. arXiv preprint arXiv:1911.07140, 2019.

[5] Moon S, An G, Song H O. Parsimonious black-box adversarial attacks via efficient combinatorial optimization[C]. International Conference on Machine Learning. PMLR, 2019: 4636-4645.

[6] Bhagoji A N, He W, Li B, et al. Practical black-box attacks on deep neural networks using efficient query mechanisms[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 154-169.

Wenming Cao, Ph.D, 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.

Facial Expression Recognition with Multiscale Graph Convolutional Networks

A short review for "Facial Expression Recognition with Multiscale Graph Convolutional Networks"

Edited by Zhiquan He

T Rao, J Li, X Wang, Y Sun, H Chen, "Facial Expression Recognition with Multiscale Graph Convolutional Networks," in IEEE MultiMedia, vol. 28, no. 2, pp. 11-19, 1 April-June 2021.

Emotions play an important role in any interpersonal communication and can express abundant information [1]. and VQA. The need of detecting a person's emotion through visual information, especially facial expression, has been increasing rapidly in various application fields, such as human–computer interface, animation, psychology analysis, and customer service [2]. Therefore, facial expression recognition (FER), aiming to recognize the emotion from the human face, is attracting increasing research interests in recent years. Recently, with the development of convolutional neural networks, the research of FER has gained a significant progress [4].

Although deep learning methods have achieved huge access in the field of FER, there still exists great room for improvement of the recognition performance. In deep learning methods, convolutional neural networks usually extract features equally from all parts of an image. However, in the case of FER, the recognized emotions are usually determined by few parts of the face, such as eyes and mouth, meanwhile, the other parts of the face only have little impact on the final result. What's more, the data bias existing in each dataset also largely affects the application of the methods. The different appearances of different people in different datasets may lead to various facial expressions for the same emotion. Developing more generalized methods in different facial expression dataset is also an emergency problem.

Considering the abovementioned problems, in this article, the authors propose a novel graph convolution network (GCN) based on multiscale landmark graphs extracted from facial images for facial expression recognition. GCN can process data with non-Euclidean structure like landmark graph extracted from the human's face. Emotional information in facial expression images can be represented by the facial landmark graph. Compared to CNN features widely used in traditional methods, extracting facial landmarks can discard the negative impact of holistic features extracted from the whole facial image, such as appearance difference and irrelevant facial parts. The authors first extract facial landmarks using practical facial landmark detector (PFLD) model [3]. Then generate undirected graphs that can represent the facial image following a defined principle that each part of the face is linked to the nose. Through a multi-neighborhood partition (MNP) strategy, subgraph are further generated from coarse to fine based on the facial landmark graph. After that, the generated graphs are sent to multiscale GCN to extract graph representations separately, which are then combined to classify the emotional label of the facial expression images.

This paper aims to use facial landmarks to represent facial expression images and recognize the emotion types of facial expressions. To better represent facial expression through facial landmarks, the author proposes the multi-neighborhood partition of the facial landmark graph to extract the multiscale landmark graph for GCN. The method mainly consists of three parts: (1) facial landmark extraction, (2) multiscale graph generation, and (3) GCN module. Given a facial expression image, the authors first extract the 98 landmarks using the PFLD model. Moreover, a multiscale subgraph can be generated using an MNP strategy for landmark graph to represent different facial parts that impact facial expression from coarse to fine. Finally, the multiscale GCN is employed to deal with the facial landmark graph and its subgraphs and recognize the emotion type of facial expression.

Multiscale Graph Generation is used to construct the facial expression image into a graph structure. Facial images can be represented by facial landmarks, which contain the fiducial points of a human's face. Linking the facial landmark in a predefined order can generate the contour of the facial organs and represent the location and shape of them. Therefore, the facial expression can be described through a graph. the facial expression image can be represented by an undirected graph $G=(V,E)$ with N nodes, where the landmarks $V=\{V_i\}_{i=1}^{N}$ are defined as a node set and $\{V_i,V_j\}\in E$ is defined as the edge sets. The coordinates and probability of each landmark is the value of each node. The linkage of each node is the value of the edge sets. The nose is linked to the other facial organs and the facial contour is linked to the eyes to fix the positions of different parts of the human's face. $A\in\mathbb{R}^{N\times N}$ is the adjacency matrix that describes the link edge of E.

GCN module is used to deal with multiscale graphs. The graph representation of facial expression can generate K different scale subgraphs $G_k(g_k|0\leq k\leq K)$ using MNP. These subgraphs are used as GCN filters to extract the graph representation of the facial parts. The GCN operates parallel above the input K hop multi-neighborhood subgraphs on the spatial dimension. Then, the authors accomplish the graph convolution process on the spectral domain for each branch of the network by a graph Fourier transform.

In this article, the authors proposed a GCN utilizing multiscale graphs derived from facial landmarks for FER. Facial landmarks extract the fiducial points that can describe the location and shape of different facial parts. They can efficiently represent facial expressions without the influence of image variations, such as face occlusion (mostly with hand), partial faces, low-contrast images, and eyeglasses, and data bias brought from the different appearances of people. Utilizing the multi-neighborhood strategy can effectively detect these multiscale facial parts from facial landmark graphs. Combining with multiscale GCN, graph representations for facial expression can be extracted for classification. Moreover, the method shows more robust results when using different training datasets. This will decrease the demand for reliable training data, which helps to utilize a huge amount of images.

References:

[1] S. Zhao et al., "Discrete probability distribution prediction of image emotions with shared sparse learning," IEEE Trans. Affect. Comput., vol. 11, no. 4, pp. 574–587, Oct.–Dec. 2020.

[2] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," IEEE Trans. Affect. Comput., vol. 9, no. 4, pp. 526–540, Oct.–Dec. 2018.

[3] Guo X, Li S, Yu J, et al. PFLD: A practical facial landmark detector[J]. arXiv preprint arXiv:1902.10859, 2019.

[4] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse regression," IEEE Trans. Multimedia, vol. 19, no. 3, pp. 632–645, Mar. 2017.

Zhiquan He, received the M.S. degree from the Institute of Electronics, Chinese Academy of Sciences, in 2001, and the Ph.D. degree from the Department of Computer Science, University of Missouri-Columbia, in 2014. He is currently an Assistant Professor with the College of Information Engineering, Shenzhen University, China. His research interests include image processing, computer vision, and machine learning.

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.