International Journal of Computer Vision
https://doi.org/10.1007/s11263-020-01351-w

GUEST EDITORIAL

Efficient Visual Recognition

Li Liu[1,2] · Matti Pietikäinen[2] · Jie Qin[3] · Wanli Ouyang[4] · Luc Van Gool[3]

Received: 18 June 2020 / Accepted: 21 June 2020

Visual recognition is the ability to recognize and localize visual categories such as faces, persons, objects, scenes, places, attributes, human expressions, emotions, actions and gestures, as well as object relations and interactions in images or videos, i.e. the ability to answer the basic and important question "What is Where", which is crucial for answering advanced reasoning questions such as: What is happening? What will happen next? What should I do? Visual recognition is the cornerstone of computer vision. Almost any vision task fundamentally relies on the ability to recognize and localize visual categories such as those mentioned above. Visual recognition thus touches many areas of artificial intelligence and information retrieval, such as image search, data mining, question answering, autonomous driving, medical diagnosis, robotics and many others.

The recent revival of interest in artificial neural networks, in particular deep learning, has brought tremendous progress in various computer vision problems (including visual recognition) and a broad range of fields beyond computer vision such as speech recognition and language translation. The beginning of deep learning in 2006 focused on the MNIST digit image classification problem and achieved the state of the art. Later in 2012, object recognition with the large scale ImageNet dataset achieved a significant breakthrough result by a Deep Convolutional Neural Network (DCNN) named AlexNet, which is arguably what reignited the field of artificial neural networks and triggered the recent revolution in artificial intelligence. Since then, research focus in visual recognition has begun to move away from feature engineering to feature learning. Recent advances in representation learning, especially deep learning, have opened up the possibility of visual recognition towards "large scale" and "in the wild", and many visual recognition algorithms have been made into products. Although visual recognition has made significant progress, especially in the past several years, there is continued need for vigorous research to solve many challenging problems towards highly efficient visual recognition including achieving energy efficiency and label/sample efficiency.

On the one hand, the high accuracy of various visual recognition tasks heavily depends on large scale Deep Neural Networks (DNNs) which require ultra high performance processors (*e.g.*, GPUs) with high computation capability. However we are in the era of post Moore's Law, and energy efficient sensing and computing is vital at all levels, from the smallest sensor like the chip to ultra high performance processors and systems like the cloud. In addition, with the ubiquity of mobile devices such as smartphones, Internet of Things (IoTs) and wearable devices which have very limited computing related resources (*e.g.*, power, memory, storage, CPUs, and bandwidth), recognizing efficiently on such devices is as critical as recognizing accurately. Therefore, there is pressing need for computational efficient algorithms to enable such devices to support a wide range of computer vision tasks. Edge intelligence is important to enable ubiquitous artificial intelligence over the next decade. On the other hand, the high accuracy of various visual recognition tasks heavily depends on massive amounts of labeled datasets which are painstakingly labeled by numerous workers or specialists. However, labeling instances is difficult,

Communicated by Li Liu, Matti Pietikäinen, Jie Qin, Wanli Ouyang, Luc Van Gool.

✉ Li Liu
li.liu@oulu.fi

Matti Pietikäinen
matti.pietikainen@oulu.fi

Jie Qin
qinjiebuaa@gmail.com

Wanli Ouyang
wanli.ouyang@sydney.edu.au

Luc Van Gool
vangool@vision.ee.ethz.ch

[1] College of System Engineering, National University of Defense Technology, Changsha, China

[2] Center for Machine Vision and Signal Analysis, University of Oulu, 90014 Oulu, Finland

[3] ETH Zürich, Zurich, Switzerland

[4] University of Sydney, Sydney, Australia

Published online: 10 July 2020

expensive, and time consuming, because it requires the efforts of experienced human annotators. In some applications like privacy sensitive applications and medical domain, obtaining labeled samples are even impossible.

Therefore, despite the great strides in visual recognition, there is urgent need for efficient visual recognition, namely developing efficient visual recognition techniques from three aspects: computationally efficient, label efficient, and sample efficient.

Since 2017, we have organized five international workshops associated with top conferences (ICCV'17, ECCV'18, CVPR'19, ICCV'19 and CVPR'20), explicitly devoted to the topics "Compact and Efficient Feature Representation and Learning in Computer Vision (CEFRL)" and "Efficient Deep Learning in Computer Vision" which are closely related to the theme of this special issue "Efficient Visual Recognition". This is a clear signal of the growing interest in computer vision around these themes. The goal of this special issue has been to solicit and publish high quality papers addressing the "efficiency" of efficient visual recognition from different aspects, and identify future promising research directions.

As guest editors, we were happy that we received a great response to the Call for Papers of this special issue. In total, there were 95 submissions. Of these, 19 papers were accepted. With so many submissions, a large number of reviewers were required. Each paper was reviewed by at least three qualified reviewers. We appreciate the reviewers for their careful, insightful, and timely reviews, leading to the high quality of accepted papers. We also thank IJCV EICs for recognizing the widespread interest in this field, which warrants this issue. Finally, we thank the IJCV office for their helpful and efficient assistance to help this issue through its many stages toward publication. The accepted 20 papers can be grouped into five different main categories, as summarized in Table 1, and are briefly discussed below.

1. Surveys;
2. Network compression;
3. Network architecture optimization;
4. Multitask learning;
5. Learning hashing codes;

1 Surveys

The paper "*A Survey of Deep Facial Attribute Analysis*", by Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li and Ran He, gives a comprehensive survey of deep facial attribute analysis, covering two basic subproblems, *i.e.* facial attribute estimation and manipulation.

The general pipeline, a taxonomy of state of the art methods, datasets and evaluation metrics, real world applications, as well as challenges and promising future research directions for deep facial attribute analysis are discussed in the paper.

2 Network Compression

As opposed to conventional, hardware-agnostic network quantization algorithms, the paper "*Hardware Centric AutoML for Mixed Precision Quantization*" by Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin and Song Han proposes a hardware aware, fully automated, mixed precision quantization framework, which leverages the reinforcement learning to automatically determine the quantization policy, and includes the hardware accelerator's feedback in the design loop to gain latency, storage and energy benefits. To achieve "hardware in the loop", a hardware simulator is used to generate the direct feedback signals to the reinforcement learning agent. The authors find that the optimal policies on different hardware architectures (*i.e.*, edge and cloud architectures) under different resource constraints (*i.e.*, latency, energy and model size) are significantly different.

The paper "*Spatially Adaptive Filter Units for Compact and Efficient Deep Neural Networks*" by Domen Tabernik and Matej Kristan and Ales̆ Leonardis presents a new convolution filter composed of Displaced Aggregation Units (DAUs). DAUs learn spatial displacements and adapt the receptive field sizes of individual convolution filters to a given problem, thus reducing the dependency on handcrafting the convolution kernel size or depth of the network to achieve task-specific receptive field sizes. The authors show that DAUs can be used as substitution of convolutional filters in existing popular networks like AlexNet and ResNet, and can result up to $4\times$ more compact networks in terms of the number of parameters at similar performance. Through rigorous experiments, the authors show that DAUs can be applied to a number of computer vision tasks, such as image classification, semantic segmentation and blind image deblurring.

While most CNNs use homogeneous kernels for convolution, the paper "*HetConv: Beyond Homogeneous Convolution Kernels for Deep CNNs*" by Pravendra Singh, Vinay Kumar Verma, Piyush Rai and Vinay Namboodiri proposes a new type of convolution operation using heterogeneous convolution. The authors name their method Heterogeneous kernel based Convolution (HetConv). Unlike conventional homogeneous convolution using spatial filters of the same size for all input channels, HetConv adopts heterogeneous convolution which uses spatial filters of different sizes for different input channels. The authors claim that HetConv reduces the computation (FLOPs) and the number of parameters of CNNs without losing representational power, compared to conventional homogeneous convolution operation.

Table 1 A brief summary of accepted papers

"Efficiency" category	Paper title	Studied visual recognition problem
Survey	A survey to deep facial attribute analysis	Facial attribute analysis
Network compression	Hardware centric AutoML for mixed precision quantization	Image classification
	Spatially adaptive filter units for compact and efficient deep neural networks	Image classification, semantic segmentation
	HetConv: beyond homogeneous convolution kernels for deep CNNs	Image classification, object detection
	Learning an evolutionary embedding via massive knowledge distillation	Closed-set image classification and three open-set tasks: face recognition, vehicle reidentification and person reidentification
Network architecture optimization	Rectified wing loss for efficient and robust facial landmark localization with convolutional neural networks	Facial landmark detection
	SSN: learning sparse switchable normalization via SparsestMax	Image classification, object detection, semantic segmentation, action recognition, face recognition
Multitask learning	Multitask compositional network for visual relationship detection	Visual relationship detection, object/predicate/significance detection
	Disentangled representation learning of makeup portraits in the wild	Makeup invariant face Verification and makeup transfer
	Fine grained multihuman parsing	Multihuman parsing
Learning hashing codes	Deep hashing with hash consistent large margin proxy embeddings	Image retrieval
	Unified binary generative adversarial network for image retrieval and compression	Image retrieval, image compression
	Learning multifunctional binary codes for personalized image retrieval	Personalized image retrieval
	A general framework for deep supervised discrete hashing	Image retrieval
	Product quantization network for fast visual search	Image and video retrieval
	Weakly-supervised semantic guided hashing for social image retrieval	Social image retrieval
	Anchor-based self-ensembling for semisupervised deep pairwise hashing	Image retrieval tasks: similarity, ranking order and unseen categories
	Hadamard matrix guided online hashing	Image retrieval
Others	Tensorized multiview subspace representation learning	Multiview clustering, face and image clustering

The authors test the performance of HetConv using VGG, ResNet and MobileNet for image classification, as well as ResNet with faster RCNN for object detection.

In the paper "*Learning an Evolutionary Embedding* via *Massive Knowledge Distillation*" by Xiang Wu, Ran He, Yibo Hu and Zhenan Sun, the authors revisit and modify the formulation of the original knowledge distillation for open-set recognition problems, and propose an Evolutionary Embedding Learning (EEL) framework to learn a faster student network for open-set recognition problems via massive knowledge distillation. EEL is designed to enable fast and accurate student network development for knowledge distillation and uses a novel correlated embedding loss to match embedding spaces between the teacher and student network. The authors claim that EEL achieves better performance than state of the art for various large-scale open-set problems, including face recognition, vehicle reidentification and person reidentification.

3 Network Architecture Optimization

The paper "*SSN: Learning Sparse Switchable Normalization* via *SparsestMax*" by Wenqi Shao, Jingyu Li, Jiamin Ren, Ruimao Zhang, Xiaogang Wang and Ping Luo addresses the optimal normalizer selection problem in designing DNNs. To achieve this, the authors improves the Switchable Normalization (SN) which learns to select different normalizers for different convolution layers of a CNN and propose Sparse Switchable Normalization (SSN) which learns to select a

single normalizer for each normalization layer of a deep network to improve interpretability and inference speed over SN. The authors train SSN with a novel SparsestMax function that turns the sparse optimization problem into a simple forward propagation of a deep network. The authors test the effectiveness of SNN in multiple computer vision tasks including image classification, object detection, semantic segmentation, action recognition and face recognition.

Towards efficient and robust facial landmark localization, the paper "*Rectified Wing Loss for Efficient and Robust Facial Landmark Localization with Convolutional Neural Networks*" by ZhenHua Feng, Josef Kittler, Muhammad Awais, and Xiaojun Wu proposes a novel loss function dubbed Rectified Wing (RWing) loss for regression based facial landmark localization with CNNs. The proposed RWing loss is designed to improve the deep network training capability for small and medium range errors, which is based on a systematic investigation of different loss functions for facial landmark localization. Their experiments show that the regression based networks integrated with their proposed RWing loss achieve 2000 + FPS on GPU, with comparable or higher accuracy over the state of the art. In addition, a data augmentation strategy called pose based data balancing is proposed to achieve robustness against large pose variations. Lastly, a coarse-to-fine framework is proposed to further improve the performance.

4 Multitask Learning

The paper "*Multitask Compositional Network for Visual Relationship Detection*" by Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao studies the problem of visual relationship detection. In order to address visual relationship detection, the authors firstly propose a novel subtask, i.e. the significance detection which refers to the task of identifying object pairs with significant relationships. Then, they propose a framework called Multitask Compositional Network (MCN) to jointly learning three closely related tasks: object detection, predicate detection, and significance detection. Experiments on two datasets Visual Relationship Detection (VRD) and Visual Genome (VG) show that MCN improves the performance of visual relationship detection, while simultaneously outputting results for object detection, predicate detection, and significance detection.

The paper "*Disentangled Representation Learning of Makeup Portraits in the Wild*" by Yi Li, Huaibo Huang, Jie Cao, Ran He and Tieniu Tan proposes a disentangled feature learning approach to simultaneously address two tasks, i.e., makeup-invariant face verification and makeup transfer, in a single generative network. The authors propose to decompose a makeup portrait into three components: makeup which intends to capture the makeup style, identity which intends to preserve the source identity, and geometry which helps to handle face misalignment in unpaired data. In addition, a newly collected makeup portraits dataset called Cross Makeup Face (CMF) is provided in the paper. Extensive experimental results on three existing datasets and the proposed CMF dataset verify that the proposed method can improve the learning of the two tasks studied.

The paper "*Fine Grained Multihuman Parsing*" by Jian Zhao, Jianshu Li, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng studies the MultiHuman Parsing (MHP) task and presents a new large-scale and fine-grained MHP benchmark dataset for understanding humans in crowded scenes. The authors also propose a Nested Adversarial Network (NAN) model to improve the performance of MHP by decomposing the original MHP task into three granularities (Semantic Saliency Prediction, Instance Agnostic Parsing, and Instance Aware Clustering) and adaptively imposing a prior on the specific process, each with the aid of a GAN based subnet. The authors claim that the proposed NAS is effective and efficient, significantly outperforming previous state of the art in MHP.

5 Learning Hashing Codes

In the paper "*A General Framework for Deep Supervised Discrete Hashing*" by Qi Li, Zhenan Sun, Ran He and Tieniu Tan, based on the assumption that the learned binary codes should be ideal for classification, both the similarity and classification information are used to learn hash codes within one stream framework. Notably, the outputs of the last layer are constrained to be binary codes directly, which is rarely investigated in deep hashing algorithms. Three publicly accessible image retrieval datasets are used to test the effectiveness of their approach.

The paper "*Learning Multifunctional Binary Codes for Personalized Image Retrieval*" by Haomiao Liu, Ruiping Wang, Shiguang Shan and Xilin Chen introduces a supervised Dual Purpose Hashing (DPH) model to jointly preserve two kinds of similarities in both high level semantics and visual attributes, while most of existing supervised hashing methods only consider one single type of semantic similarity. With such a framework, the binary codes of novel images can be readily obtained by quantizing the outputs of a specific CNN layer, and different retrieval tasks can be achieved by using the binary codes in different ways.

The paper "*Unified Binary Generative Adversarial Network for Image Retrieval and Compression*" by Jingkuan Song, Tao He, Lianli Gao, Xing Xu, Alan Hanjalic and Hengtao Shen proposes a label efficient algorithm named a Binary Generative Adversarial Network (BGAN+) to jointly learning two tasks: image retrieval and image compression. Instead of learning a single binary code serving both tasks

as in most existing works, the authors propose to simultaneously learn two binary representations for each image. The authors claim that the proposed framework is more effective than the state-of-the-art supervised approaches, despite the fact that the binary codes are learned in an unsupervised fashion. Extensive experimental results show that BGAN+ outperforms existing retrieval methods with significant margins and achieves promising performance for image compression, especially for low bit rates.

The paper "*Deep Hashing with Hash Consistent Large Margin Proxy Embeddings*" by Pedro Morgado, Yunsheng Li, Jose Costa Pereira, Mohammad Saberian and Nuno Vasconcelos proposes to use a fixed set of proxies (weights of the CNN classification layer) to eliminate the rotational ambiguity issue of proxy embeddings which encourages nonbinary embeddings, and a method to design such proxies. The authors show that the Hash-Consistent Large Margin (HCLM) proxies obtained by their method can encourage the saturation of hashing units, leading to highly discriminative hashing codes. In addition, a semantic extension (sHCLM) is proposed with an aim to improve hashing performance in a transfer scenario.

The paper "*Weakly Supervised Semantic Guided Hashing for Social Image Retrieval*" by Zechao Li, Jinhui Tang, Liyan Zhang and Jian Yang addresses the efficient nearest neighbor image search problem by achieving label efficiency. A novel Semantic Guided Hashing (SGH) method coupled with binary matrix factorization is proposed to simultaneously explore the weakly-supervised, imperfect labeled information and the underlying data structures. The binary matrix factorization model is responsible for learning the binary features of images, to address the problem of imperfect tags. The underlying data structures are discovered by adaptively learning a discriminative data graph. The authors claim that their method is the first work that incorporates the hash code learning, the semantic information mining and the data structure discovering into one unified framework.

The paper "*Hadamard Matrix Guided Online Hashing*" by Mingbao Lin, Rongrong Ji, Hong Liu, Xiaoshuai Sun, Shen Chen and Qi Tian proposes a sample efficient online hashing method named Hadamard Matrix Guided Online Hashing (HMOH). HMOH introduces the Hadamard Matrix into hashing and considers each column of the Hadamard matrix as the target code for each class, which by nature satisfies several desired properties of hashing codes. Extensive experiments on four widely-used benchmarks demonstrate the superior accuracy and efficiency of HMOH over various existing methods.

The paper "*Anchor based Selfensembling for Semisupervised Deep Pairwise Hashing*" by Xiaoshuang Shi, Zhenhua Guo, Fuyong Xing, Yun Liang, Lin Yang proposes a label efficient deep hashing method by leveraging unlabeled data to learn hash functions. The authors develop an anchor based solution for semisupervised hashing, by preserving the pairwise similarity relationship among both labeled and unlabeled samples as well as the semantic similarity information hidden in unlabeled data. The superior performance of the proposed method over recent state of the art methods is demonstrated in multiple retrieval tasks.

Different from the papers which focus on learning hashing codes introduced above in this category, the paper "*Product Quantization Network for Fast Visual Search*" by Tan Yu, Jingjing Meng, Chen Fang, Hailin Jin and Junsong Yuan studies the problem of efficient image retrieval and proposes a method called Product Quantization Network (PQN). The authors propose the differentiable soft-assignment quantization, which can be integrated as a layer in CNNs to construct PQNs. In addition, two PQN variants: Residual PAQ (RPQN) and Temporal PQN (TPQN) are presented. Comprehensive experiments conducted on multiple public benchmarks demonstrate the state of the art performance of the proposed series of PQNs in fast image and video retrieval.

6 Others

The paper "*Tensorized Multiview Subspace Representation Learning*" by Changqing Zhang, Huazhu Fu, Jing Wang, Qinghua Hu, Xiaochun Cao and Wen Li studies the problem of multiview clustering and proposes an algorithm termed as Tensorized Multiview Subspace Representation Learning (TMSRL) by simultaneously taking advantages of multiple views and prior constraint. To explore the correlations within each view and across multiple views, TMSRL enforces the low-rankness of the 3D tensor consisting of the subspace representation matrices of different views. To exploit prior, a constraint matrix is devised to guide the subspace representation learning within a unified framework. The subspace representation tensor equipped with a low rank constraint captures the complementary information among different views and reduces redundancy of subspace representations, leading to higher accuracy of subsequent tasks. The effectiveness of TMSRL is validated with extensive experiments on multiple multiview datasets.