Guest Editorial: Feature Learning from RGB-D Data for Multimedia Applications

Baochang Zhang¹ · Jungong Han² · Ling Shao²

Published online: 26 December 2016

© Springer Science+Business Media New York 2016

The emergence of Multi-channel Data has brought about a paradigm shift to many fields of data analytics such as multimedia analysis and computer vision. In real-world applications, we usually encounter data acquired from different modalities. Integrating useful information, such as data, semantic and knowledge, triggers opportunities to provide advanced and intelligent services. For example, different types of sensor data have been exploited for human action recognition: sensor data from visual cameras like RGB, time-of-flight (ToF), or infrared cameras, and non-visual sensor data like audio signals or inertial measurement unit (IMU) data. Although each modality has its own information and statistical properties, different modalities usually share similar high level concepts. Therefore, those modalities may provide complementary information about a common concept.

Feature learning technique learns a transformation from a raw input to a compact representation that can describe the scene. Compared to the traditional feature representation methods, it allows for learning expressive features directly from the raw data without manual annotations or hand-crafted heuristic rules. The set of used features could be continuously modified during the learning procedure, which is able to adapt to a changing and more realistic environment. In recent few years, such techniques (e.g., dictionary learning and deep learning) have been widely exploited to many applications including classification, segmentation, and scene understanding, etc. As the multiple-channel data becomes available, researchers are keen to develop feature learning algorithms in this new area.

✉ Baochang Zhang
bczhang@buaa.edu.cn; bczhang@139.com

Jungong Han
jungonghan77@gmail.com

Ling Shao
ling.shao@ieee.org

¹ School of Automation Science and Electrical Engineering, Beihang University, Beijing, China

² Northumbria University, Newcastle upon Tyne, UK

The aim of this special issue is to solicit the state-of-the-art feature learning approaches and technical solutions in the area of multimedia technology and applications using multiple-channel data, i.e., RGB-D. This special issue has attracted more than 30 manuscripts and the submissions have been strictly reviewed by over 80 reviewers, with 18 high-quality articles accepted in the end. These articles basically can be classified into three themes. The first theme is to explore the RGB-D data for scene recognition. The second theme is to explore multimodal data analysis for recognition tasks. The third theme is about applying multimodal models to some interesting applications, such as UAV and satellite. Below, we briefly summarize the highlights of each paper.

In *Real-Time Human Body Tracking Based on Data Fusion from Multiple RGB-D Sensors*. Juan C Nunez et al. present a human pose estimation method based on the skeleton fusion and tracking using multiple RGB-D sensors. The paper considers the skeletons provided by each RGB-D device and constructs an improved skeleton, taking into account the quality measures provided by the sensors at two different levels: the whole skeleton and each joint individually. Then, each joint is tracked by a Kalman filter, resulting in a smooth tracking performance. Experimental results performed on this dataset show that the system obtains better smoothness results than the most representative methods found in the literature.

In *Modality-specific and Hierarchical Feature Learning for RGB-D Hand-held Object Recognition*. Xiong et al. focus on a special but important area called RGB-D hand-held object recognition and propose a hierarchical feature learning framework for this task. First, the framework learns modality-specific features from RGB and depth images using CNN architectures with different network depth and learning strategies. Secondly, a high-level feature learning network is implemented for a comprehensive feature representation. Different with previous works on feature learning and representation, the hierarchical learning method can sufficiently dig out the characteristics of different modal information and efficiently fuse them in a unified framework. The experimental results on HOD dataset illustrate the effectiveness of our proposed method.

In *Filtered Pose Graph for Efficient Kinect Pose Reconstruction*. Pierre Plantard et al. propose a new pose reconstruction method based on modelling the pose database with a structure called Filtered Pose Graph, which indicates the intrinsic correspondence between poses. Such a graph not only speeds up the database poses selection process, but also improves the relevance of the selected poses for higher quality reconstruction. They apply the proposed method in a challenging environment of industrial context that involves sub-optimal Kinect placement and a large amount of occlusion. Experimental results show that our real-time system reconstructs Kinect poses more accurately than existing methods.

In *RGB-D Datasets Using Microsoft Kinect or Similar Sensors: A Survey*. Ziyun Cai et al. systematically survey popular RGB-D datasets for different applications including object recognition, scene classification, hand gesture recognition, 3D-simultaneous localization and mapping, and pose estimation. They provide the insights into the characteristics of each important dataset, and compare the popularity and the difficulty of those datasets. The survey is to give a comprehensive description about the available RGB-D datasets and thus to guide researchers in the selection of suitable datasets for evaluating their algorithms.

In *Depth completion for Kinect v2 sensor*. Wanbin Song et al. propose a depth completion method, which is designed especially for the Kinect v2 depth artifacts, and exploit the position information of the color edges extracted from the Kinect v2 sensor to guide the accurate hole-filling around the object boundaries. Experimental results demonstrate the effectiveness of the proposed depth image completion algorithm for the Kinect v2 in terms of completion accuracy and execution time.

In *Development of an Automatic 3D Human Head Scanning-Printing System*. Longyu Zhang et al. introduce an automatic 3D human head scanning- printing system, which provides a complete pipeline to scan, reconstruct, select, and print out physical 3D human heads. They evaluated the accuracy of the proposed system by comparing our generated 3D head models, from both standard human head model and real human subjects. Computational cost is also provided to further assess our proposed system.

In *A survey of depth and inertial sensor fusion for human action recognition*. Chen Chen et al. provide an overview of the recent investigations where both vision and inertial sensors are used together and simultaneously to perform human action recognition more effectively. The thrust of this survey is on the utilization of depth cameras and inertial sensors as these two types of sensors are cost-effective, commercially available, and more significantly they both provide 3D human action data. An overview of the components necessary to achieve fusion of data from depth and inertial sensors is provided. In addition, a review of the publicly available datasets that include depth and inertial data which are simultaneously captured via depth and inertial sensors is presented.

In *Indoor Scene Recognition via Multi-task Metric Multi-kernel Learning from RGB-D Images*. Yu Zheng et al. propose a multi-task metric multi-kernel learning algorithm that exploits the inter-source similarities and complementarities between color images and depth images to conduct the indoor scene recognition. Specifically, they utilize multi-task metric learning to learn a Mahalanobis metric for RGB-D images. Multi-task metric learning can extract the common properties from color images and depth images to learn better metrics. The experimental results have demonstrated that the proposed method can lead to better indoor scene recognition.

In *Semi-direct Tracking and Mapping with RGB-D Camera for MAV*. Shuhui Bu et al. present a novel semi-direct tracking and mapping (SDTAM) approach for RGB-D cameras which inherits the advantages of both direct and feature based methods, and consequently it achieves high efficiency, accuracy, and robustness. The proposed approach achieves real-time speed which only uses part of the CPU computation power, and it can be applied to embedded devices such as phones, tablets, or micro aerial vehicles (MAVs).

In *Cauchy Estimator Discriminant Learning for RGB-D Sensor-based Scene Classification*. Dapeng Tao et al. propose a new dimensional reduction algorithm called Cauchy estimator discriminant learning (CEDL). CEDL simultaneously addresses two goals: (1) to decrease negative influences to some extent when there is noise in the input samples; (2) to preserve the local and global geometry structure of the input samples. Experiments with the frequently used NYU Depth V1 dataset suggest the effectiveness of CEDL compared with other state-of-the-art scene classification methods.

In *Gender Classification using 3D Statistical Models*. Wankou Yang et al. present an effective gender classification based on 3D face model based on 3D principal components analysis (3D Eigenmodels) and 3D independent components analysis (3D ICmodels). The experimental results on BU_3DFE database demonstrate that the proposed method can achieve good performance.

In *Landmark-based multimodal human action recognition*. Stylianos Asteriadis et al. propose a novel, multi-modal human action recognition, in which each action is represented by a basis vector and spectral analysis is performed on a matrix of new action feature vectors. Evaluation on three publicly available datasets demonstrates the potential of the approach.

In *A Novel Multiple-channels Scheduling Algorithm Based on Timeslot Optimization in the Advanced Orbiting Systems*. Ye Tian et al. focus on multiple-channels data from satellite, by

considering that the performance of the existing multiple-channels scheduling algorithm is not good enough and the buffer size is rarely considered, a novel multiple-channels scheduling algorithm based on timeslot optimization is presented and its performance under the finite buffer size is also studied. The multiple-channels scheduling algorithm for the finite buffer size is extensively studied and the upper bound of rate of frame-lost timeslots of each asynchronous VC is concluded.

In *Sparse Multiple Instance Learning as Document Classification*. Shengye Yan et al. focus on multiple instance learning (MIL) with sparse positive bags (which we name as sparse MIL). A structural representation is presented to encode both instances and bags. This representation leads to a non-i.i.d. MIL algorithm, miStruct, which uses a structural similarity to compare bags. The methods achieve significantly higher accuracies and AUC (area under the ROC curve) than the state-of-the-art in a large number of sparse MIL problems, and the document classification analogy explains their efficacy in sparse MIL problems.

In *Real-time Recognition of Medial Structures within Hand Postures through Eigen-space and Geometric Skeletal Shape Features*. Lalit Kane et al. apply intuitive Eigen-space based Principal Components of Symbolic Structure (PCSS) and geometric Equi-Polar Signature (EPS) features to accomplish the recognition task. Both PCV and EPS process the skeleton globally without sections without associating contour information. Recognition accuracy up to 94% is obtained on a 22 posture dataset comprising of 10,560 depth frames with 480 samples for each posture. Depth sensor based acquisition is employed to meet the real-time requirements.

In *Plant Identification via Multipath Sparse Coding*. Authors propose a novel plant identification method based on multipath sparse coding using SIFT features, which avoids the need of feature engineering and the reliance on botanical taxonomy. And they evaluate the proposed method on several plant datasets and find that multi-organ is more informative than single organ for botanist. Experimental results also validate that the proposed method outperforms the state-of-the-art methods.

In *Real-time Visual Tracking Based on Improved Perceptual Hashing*. Mengjuan Fei et al. proposed a perceptual hashing based template-matching method for object tracking to efficiently track objects in challenging video sequences. They propose Laplace-based Hash (LHash) and Laplace-based Difference Hash (LDHash). By qualitative and quantitative comparison with some representative tracking algorithms, experimental results show that our improved perceptual hashing-based tracking algorithms perform favorably against the state-of-the-art algorithms under various challenging environments in terms of time cost, accuracy and robustness.

In *Modality Identification for Heterogeneous Face Recognition*. Muhammad Khurram Shaikh et al. propose a novel image sharpening based modality pattern noise technique for modality identification. The proposed system has been evaluated on three challenging benchmarks of heterogeneous face databases. The proposed technique has produced outstanding results and will open new avenues of research for automated HFR methods in future.

These 18 selected contributions basically can reflect the new achievements in the field of RGB-D and multimodal learning and we hope they can provide a solid foundation for future new approaches and applications.

Finally, we would like to thank all authors for their contributions, the reviewers for reviewing these papers, and Editor-in-Chief, Prof. Borko Furht, for his support and guidance throughout the process.

Baochang Zhang received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, 2001, and 2006, respectively. From 2006 to 2008, he was a Research Fellow with the Chinese University of Hong Kong, Hong Kong, and with Griffith University, Brisbane, Australia. Currently, he is an associate professor with the Science and Technology on Aircraft Control Laboratory, School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. He held a senior postdoc position in PAVIS department, IIT, Italy. He was supported by the Program for New Century Excellent Talents in University of Ministry of Education of China. His current research interests include pattern recognition, machine learning, face recognition, and wavelets.

Jungong Han is currently a Senior Lecturer with the Department of Computer Science and Digital Technologies at Northumbria University, Newcastle, UK. He received his Ph.D. degree in Telecommunication and Information System from Xidian University, China, in 2004. During his Ph.D study, he spent one year at Internet Media group of Microsoft Research Asia, China. Previously, he was a Senior Scientist (2012–2015) with Civolution Technology (a combining synergy of Philips Content Identification and Thomson STS), a Research Staff (2010–2012) with the Centre for Mathematics and Computer Science (CWI), and a Senior Researcher (2005–2010) with the Technical University of Eindhoven (TU/e) in Netherlands.

Ling Shao (M'09-SM'10) is currently a Professor and Head of the Computer Vision and Artificial Intelligence Group with the Department of Computer Science and Digital Technologies at Northumbria University, Newcastle upon Tyne, UK. His research interests include Computer Vision, Image/Video Processing, Pattern Recognition and Machine Learning. He is an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Image Processing, and several other journals.