Guest Editorial Introduction to the Special Issue on Large Scale and Nonlinear Similarity Learning for Intelligent Video Analysis

Learning similarity and distance measures has become increasingly important for the analysis, matching, retrieval, recognition, and categorization of video and multimedia data. With the ubiquitous use of digital imaging devices, mobile terminals and social networks, there are massive volumes of heterogeneous and homogeneous video and multimedia data from multiple sources, views, and domains, e.g., news media websites, microblog, mobile phone, social networking, etc. Similarity and distance-based constraints can also be extended and incorporated to boost classification and relationship learning. Moreover, the spatio-temporal coherence among video data can also be utilized for self-supervised learning of similarity and distance metrics. This trend has brought several challenging issues for developing similarity and metric learning methods for large scale and weakly annotated data, where outliers and incorrectly annotated data are inevitable. Recently, scalability has been investigated to cope with lightweight and large scale metric learning, while nonlinear similarity models have shown their great potentials in learning invariant representation and nonlinear measures of video and multimedia data.

As guest editors of this special issue on "Large scale and nonlinear similarity learning for intelligent video analysis," we were happy to receive more than 120 submissions. Among them, 54 papers have been accepted in this issue, which can be grouped into nine major categories: (1) progress on distance and similarity metric learning, (2) subspace, sparse and low rank learning, (3) deep metric learning, (4) learning-based hashing, (5) face and human analytics, (6) visual tracking, (7) activity and gesture recognition, (8) image and video analytics, and (9) hardware-based high performance computing.

A. Progress on Distance and Similarity Metric Learning

This group includes 6 papers aiming at developing novel metric learning models as well as learning algorithms. The paper "An Information Geometry-Based Distance Between High-Dimensional Covariances for Scalable Classification" by Q. Wang, X. Lu, P. Li, Z. Gao, and Y. Piao investigates the geometrical structure of covariances in a Riemannian manifold. The proposed Information Geometry Based Distance (IGBD) embeds each zero-mean Gaussian distribution into a vector on the tangent space based on Fisher information metric. The authors further demonstrate its efficiency and effectiveness on image classification and video-based face recognition.

The paper "SLMOML: Online Metric Learning With Global Convergence" by G. Zhong, Y. Zheng, S. Li, and Y. Fu studies the online learning and convergence issues of metric learning. In particular, the proposed SLMOML adopt the passive-aggressive learning strategy, where the LogDet divergence is used for maintaining the closeness among successive metrics, and the hinge loss is introduced for enforcing discriminative ability. The global convergence of SLMOML is also given based on the Karush-Kuhn-Tucker (KKT) conditions. Extensive experiments show that SLMOML performs favorably on classification and retrieval applications.

The paper "A Unified Metric Learning-Based Framework for Co-Saliency Detection" by J. Han, G. Cheng, Z. Li, and D. Zhang presents a unified framework for joint learning of discriminative metric and co-salient object detector. The proposed method incorporates the metric learning regularization in the SVM-based model, and achieve state-of-the-art co-saliency detection results.

The paper "Mixture Statistic Metric Learning for Robust Human Action and Expression Recognition" by S. Dai and H. Man studies the metric learning problem on multiple statistics, including means, and covariance matrices, and parameters of Gaussian mixture. In the proposed method, multiple statistics are mapped to SPD Riemannian manifolds, and mixture of Mahalanobis metrics is learned for distribution-based discriminations. The results demonstrate its effectiveness for action and facial expression recognition tasks.

The paper "Learning Affine Hull Representations for Multi-Shot Person Re-Identification" by S. Karanam, Z. Wu, and R. J. Radke extends metric learning for measuring the distance between two sets of images using affine hulls. By incorporating metric learning with affine hull data modeling, the proposed method is effective in learning discriminative feature representation, and achieves notable performance improvement on multi-shot person re-identification.

The paper "Geometry-Aware Similarity Learning on SPD Manifolds for Visual Recognition" by Z. Huang, R. Wang, X. Li, W. Liu, S. Shan, L. Van Gool, and X. Chen extends metric learning to the Riemannian manifold of fixed-rank SPD matrices. And Riemannian Conjugate Gradient (RCG) algorithm is further presented to optimize the geometry-aware

Digital Object Identifier 10.1109/TCSVT.2018.2874080

SPD similarity learning (SPDSL). The results on visual classification demonstrate the merits of learning manifold-manifold transformation matrix.

B. Subspace, Sparse and Low Rank Learning

This group includes 8 papers aiming at learning discriminative representation from the subpace, sparse and low rank perspectives. The paper "Localized LRR on Grassmann Manifolds: An Extrinsic View" by B. Wang, Y. Hu, J. Gao, Y. Sun, and B. Yin extends the low-rank representation model on Grassmann manifolds. Specifically, this method builds LRR locally in the tangent space at each Grassmannian point. The results demonstrate its effectiveness for image clustering on handwritten digital images and video clips.

The paper "Unified Sparse Subspace Learning via Self-Contained Regression" by S. Yi, Z. He, Y.-M. Cheung, and W.-S. Chen proposes a novel sparse PCA method by generalizing the self-contained regression-type framework. Structural sparse regularization is further incorporated to result in a joint sparse pixel weighted PCA method. Experiments are conducted to show its feasibility and effectiveness.

The paper "Alignment Distances on Systems of Bags" by A. Sagel and M. Kleinsteuber investigates the dissimilarity measure on Systems of Bags for modeling dynamic scenes and dynamic textures. It adopts a Jacobi-type method which guarantees to converge to a set of critical points. The results demonstrate that the alignment distance is effective for dynamic scene and dynamic texture classification.

The paper "BoMW: Bag of Manifold Words for One-Shot Learning Gesture Recognition From Kinect" by L. Zhang, S. Zhang, F. Jiang, Y. Qi, J. Zhang, Y. Guo, and H. Zhou extends sparse coding and dictionary learning to SPD manifold. In Bag of Manifold Words (BoMW), the Stein kernel is deployed to map SPD matrices into a vector space, and sparse coding and dictionary learning are further utilized for spatial pyramid BoW representation. The results show that BoMW is effective for RGB-D based gesture recognition.

The paper "Fast Grayscale-Thermal Foreground Detection With Collaborative Low-Rank Decomposition" by S. Yang, B. Luo, C. Li, G. Wang, and J. Tang presents a Collaborative Low-rank Decomposition (CLoD) model for grayscale-thermal foreground detection. In CLoD, collaborative low rank structure and modality weights are incorporated for improved and adaptive fusion, and a block-based iterative optimization algorithm is deployed for improving computational efficiency. Experiments show that CLoD performs favorably in comparison with the state-of-the-art methods.

The paper "Student's t-Hidden Markov Model for Unsupervised Learning Using Localized Feature Selection" by Y. Zheng, B. Jeon, L. Sun, J. Zhang, and H. Zhang incorporates subspace learning into Student's t-Hidden Markov Model (HMM) for improved clustering. It combines localized feature saliency (LFS) with Student's t-HMM for modeling hidden state observation emission distributions, and adopt Variational Bayesian (VB) for model learning. The results show that the proposed method performs robustly on both synthetic and real data sets.

The paper "Semi-Supervised Cross-View Projection-Based Dictionary Learning for Video-Based Person Re-Identification" by X. Zhu, X.-Y. Jing, L. Yang, X. You, D. Chen, G. Gao, and Y. Wang extends discriminative dictionary learning to the semi-supervised, multi-view, and video-based setting. Specifically, the SCPDL method jointly learns a pair of projection matrices as well as a pair of dictionaries for balancing discrimination and representation ability. The results demonstrate the effectiveness of SCPDL for video-based person re-identification.

The paper "Subspace Segmentation by Correlation Adaptive Regression" by W. Wang, B. Zhang, and X. Feng presents a correlation adaptive regression (CAR) model for better modeling the inter-cluster and intra-cluster characteristics. To this end, L_2-norm and L_{21}-norm are respectively adopted to constrain the coefficients corresponding to highly correlated and uncorrelated data points. Experimental results on face and hand-written digital images demonstrate the effectiveness of the CAR.

C. Deep Metric Learning

This group includes 7 papers aiming at extending metric learning for learning discriminative deep representation. The paper "Image-to-Video Person Re-Identification With Temporally Memorized Similarity Learning" by D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, and Z. Cai proposes a temporally memorized similarity learning neural network for image-to-video matching. Here, CNN is utilized for image representation, while an LSTM network is introduced after CNN to encode the temporal information for video representation. Then, both image representation and video representation are taken as input to the similarity sub-network. The results show that the learned deep metric is effective for image-to-video person re-identification.

The paper "Deep Metric Learning for Crowdedness Regression" by Q. Wang, J. Wan, and Y. Yuan extends deep metric learning to regression task for crowd counting. The proposed method integrates density related feature learning and metric learning based regression into an end-to-end network, and achieves favorable performance for crowd counting.

The paper "Deep Localized Metric Learning" by Y. Duan, J. Lu, J. Feng, and J. Zhou proposes a deep localized metric learning (DLML) for modeling heterogeneous datasets by learning multiple fine-grained deep localized metrics. K-Auto-Encoders is first used to obtain K local subspaces and one holistic subspace, and a fully connection network to each Auto-Encoder to learn a hierarchical nonlinear metric. Experiments show that DLML is effective for face recognition, person re-identification, and scene recognition.

The paper "Deep Multi-View Feature Learning for Person Re-Identification" by D. Tao, Y. Guo, B. Yu, J. Pang, and Z. Yu develops a deep multi-view feature learning (DMVFL) model for person re-identification. DMVFL exploits XQDA metric learning to combine handcrafted and deep representation, and achieve comparable performance on person re-identification.

The paper "Deep Co-Space-Sample Mining Across Feature Transformation for Semi-Supervised Learning" by Z. Chen, K. Wang, X. Wang, P. Peng, E. Izquierdo, and L. Lin exploits deep learning for pursuing feature transformation to select the reliable unlabeled samples in semi-supervised learning. To this end, category variation is adopted as an indicator of reliability measure and jointly updated along with feature transformation. The results validate its effectiveness for semi-supervised image classification.

The paper "Unconstrained Face Recognition Using a Set-to-Set Distance Measure on Deep Learned Features" by J. Zhao, J. Han, and L. Shao presents a deep Set-to-Set (S2S) metric learning model for measuring the similarity between two sets. Going beyond metric learning and deep representation, a kNN-average pooling operator is introduced for aggregating the similarity scores. The results demonstrate that the proposed method is effective and outperforms the baseline with a large margin for unconstrained face recognition.

The paper "Deep Deformable Patch Metric Learning for Person Re-Identification" by S. Bąk and P. Carr investigates the learning of deep distance metric robust to deformation of pedestrian images. To this end, it incorporates metric learning approaches with deformable models, and learn deep representation with CNNs. The results show that the proposed method performs favorably in comparison with the state-of-the-arts for person re-identification.

D. Hashing

This group includes 4 papers dedicating to high performance learning-based hashing in either supervised or unsupervised manner. The paper "Robust and Flexible Discrete Hashing for Cross-Modal Similarity Search" by D. Wang, Q. Wang, and X. Gao investigates multimodal hashing from a cross-modal similarity search perspective. The proposed robust and flexible discrete hashing (RFDH) model incorporates structural sparsity and learns modality-adaptive to improve the robustness and flexibility. Experiments show that RFDH exhibits promising performance for unsupervised multimodal hashing.

The paper "Scalable Discrete Supervised Multimedia Hash Learning With Clustering" by S. Zhang, J. Li, M. Jiang, and B. Zhang aims to relax the intrinsic discretion and pairwise/triplet similarity constraints in supervised deep hashing. To this end, the proposed Discrete Supervised Hashing (DISH) incorporates Asymmetric Low-rank Matrix Factorization with binary classifier learning, and propose the Fast Clustering-based Batch Coordinate Descent algorithm for efficient learning. Experiments validate the effectiveness of DISH for large-scale multimedia retrieval.

The paper "Weakly Supervised Multimodal Hashing for Scalable Social Image Retrieval" by J. Tang and Z. Li studies multimodal hashing under the weakly-supervised scenario, where orthogonal constraint and maximum entropy regularization. The results show that the proposed WMH is effective for social image retrieval.

The paper "Kernel Based Semantic Hashing for Gait Retrieval" by Y. Zhou, Y. Huang, Q. Hu, and L. Wang presents a kernel based semantic hashing (KSH) model to retrieve individuals with the gait biometric. Specifically, KSH is learned by minimizing triplet ranking loss based on sematic similarity score. The experiments demonstrate the effectiveness and efficiency of KSH.

E. Face and Human

This group includes 6 papers aiming at the recognition and analytics (e.g., landmark detection, pose estimation, crowd counting) of face and pedestrian images. The paper "Driver Facial Landmark Detection in Real Driving Situations" by M. Jeong, B. C. Ko, S. Kwak, and J.-Y. Nam investigates the face landmark detection for face images in real driving scenario. The proposed method adopts an ensemble of local weighted random forest regressor for local modeling, and further incorporates global shape model based on the landmark spatial relation. The results validate its feasibility in real time driver-state monitoring systems.

The paper "Learning Bidirectional Temporal Cues for Video-Based Person Re-Identification" by W. Zhang, X. Yu, and X. He integrates convolutional neural networks (CNNs) and bidirectional recurrent neural networks (BRNNs) for spatio-temporal representation of videos. The classification-based identification and verification-based doublet losses are incorporated for end-to-end training of the whole network. The experiments validate the effectiveness of the proposed method for video-based person re-identification.

The paper "P2SNet: Can an Image Match a Video for Person Re-Identification in an End-to-End Way?" by G. Wang, J. Lai, and X. Xie investigates the end-to-end learning of CNN model for the image-to-video person re-identification task. To this end, the proposed P2SNet involves a k-nearest neighbor triplet module in to CNN for joint feature and metric learning. The results show that P2SNet is effective for image-to-video re-identification.

The paper "Video-Based Person Re-Identification With Accumulative Motion Context" by H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang and S. Yan, and J. Feng presents an Accumulative Motion Context (AMOC) network, which aggregates long-range motion context for video-based person re-identification. The results demonstrate the effectiveness of exploiting motion context, and AMOC performs favorably in comparison to the state-of-the-arts.

The paper "ADORE: An Adaptive Holons Representation Framework for Human Pose Estimation" by L. Dong, X. Chen, R. Wang, Q. Zhang, and E. Izquierdo investigates human 2D pose estimation by exploiting local and global cues from a 2D still image. The proposed ADORE framework consists of two parts: (i) Independent Losses Pose Nets (ILPNs) for global level localization, and (ii) Convolutional Local Detectors (CLDs) for local refinement. Experiments demonstrate the effectiveness of ADORE.

The paper "An Efficient Method of Crowd Aggregation Computation in Public Areas" by M. Xu, C. Li, P. Lv, N. Lin, R. Hou, and B. Zhou studies the crowd aggregation behavior by developing comprehensive collective crowd descriptors. The proposed descriptor is computationally efficient and effective in modeling spatial and the temporal motion context, and can be deployed to the evolution analysis of the group movement and the crowd abnormal detection.

F. Visual Tracking and Detection

This group includes 7 papers dedicating to improve appearance and matching models in visual tracking and their applications to human tracking and video-based object detection. The paper "Visual Tracking via Nonlocal Similarity Learning" by Q. Liu, J. Fan, H. Song, W. Chen, and K. Zhang incorporates similarity learning for modeling nonlocal interactions of target appearance. Here, polynomial kernel feature map is utilized to characterize nonlocal similarity, and linear regression is deployed to update tracker in the particle filtering framework. The results demonstrate the effectiveness of similarity learning in visual tracking.

The paper "Robust Likelihood Model for Illumination Invariance in Particle Filtering" by B. Al Delail, H. Bhaskar, M. J. Zemerly, and M. Al-Mualla investigates the similarity measure and likelihood estimation issues in visual tracking. Concretely, novel likelihood estimator and enhanced template dictionary updating are presented to improve particle filtering based tracking. The results show that notable improvement can be achieved on video sequences with illumination variations.

The paper "Robust Visual Tracking via Multi-Scale Spatio-Temporal Context Learning" by W. Xue, C. Xu, and Z. Feng suggests to improve tracking performance by exploiting spatio-temporal contexts. In particular, fast perceptual hash algorithm is presented to update long-term historical targets and the medium-term stable scene, and fusion salient sample detection is deployed to combine short-term overall samples. Experiments on tracking benchmark demonstrate the effectiveness of the proposed MSTC model.

The paper "Tracking With Static and Dynamic Structured Correlation Filters" by S. Wang, D. Wang, and H. Lu investigates structured correlation filters for improving tracking performance in case of partial occlusion. The proposed tracker involves the static and dynamic models for initial estimation and final refinement, respectively. The results show that the proposed tracker performs effectively on sequences with occlusion and scale variation.

The paper "Online-Learning-Based Human Tracking Across Non-Overlapping Cameras" by Y.-G. Lee, Z. Tang, and J.-N. Hwang studies the multi-object tracking and segmentation of human across multiple non-overlapping cameras. Here, intra- and inter-camera tracking models are respectively designed to develop a scalable human tracker. The results show that the proposed method achieves state-of-the-art performance on the benchmark under real-world camera network scenarios.

The paper "A Stochastic Attribute Grammar for Robust Cross-View Human Tracking" by X. Liu, Y. Xu, L. Zhu, and Y. Mu proposes a stochastic attribute grammar model for jointly tracking and re-identifying of humans across camera views. Complementary and discriminative human attributes are leveraged for enhancing the above joint parsing task, and an alternating method is employed for both top-down and bottom-up inference. The results demonstrate the proposed method is effective in tracking humans under significant changes of camera viewpoints.

The paper "T-CNN: Tubelets With Convolutional Neural Networks for Object Detection From Videos" by K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, and W. Ouyang extends image-based object detection to videos by incorporating temporal and contextual information from tubelets (i.e. T-CNN). Specifically, high performance tracking is used to globally revising detection confidences, and contextual information is exploited to suppress low-confidence detection scores. T-CNN won the object-detection-from-video (VID) task in ILSVRC2015.

G. Activity and Gesture

This group includes 5 papers dedicating to enhance activity and gesture recognition by improving feature representation and classification models. The paper "Cross-Agent Action Recognition" by H. Wang and L. Wang investigates the task of action recognition training on source agents while testing on target agents. To this end, the proposed cross-agent action recognition adopts Sequential Transfer Sparse Coding (STSC) and considers three scenarios: single source and single target, multiple sources and single target, multiple sources and multiple targets. The experiments show that STSC significantly improves the cross-agent action recognition and surpasses the baseline with a large margin.

The paper "3D Human Action Recognition Using a Single Depth Feature and Locality-Constrained Affine Subspace Coding" by C. Liang, L. Qi, Y. He, and L. Guan studies the problem of human action recognition based on depth videos. Concretely, Locality-constrained Affine Subspace Coding (LASC) is presented on Depth Motion Maps (DMMs). The results demonstrate the effectiveness of LASC for 3D human action recognition.

The paper "Activity Recognition in Sensor Data Streams for Active and Assisted Living Environments" by F. Al Machot, A. Haj Mosa, M. Ali, and K. Kyamakya presents a windowing algorithm for recognizing complex daily activities. The proposed method first identifies the best fitting sensor and then extracts statistical spatio-temporal features for activity recognition. The experiments demonstrate the effectiveness and robustness to active and assisted living environments.

The paper "First-Person Daily Activity Recognition With Manipulated Object Proposals and Non-Linear Feature Fusion" by M. Wang, C. Luo, B. Ni, J. Yuan, J. Wang, and S. Yan presents a novel framework for first-person daily activity recognition. It first leverages motion cues and R-CNN for manipulated object detection, and then exploits non-linear feature fusion to combine object and motion features. Experiments show that it performs favorably on the first-person daily activity benchmark.

The paper "Large-Scale Gesture Recognition With a Fusion of RGB-D Data Based on Saliency Theory and C3D Model" by Y. Li, Q. Miao, K. Tian, Y. Fan, X. Xu, R. Li, and J. Song investigates the task of gesture recognition based on RGB-D videos. To suppress gesture-irrelevant factors, saliency features are integrated with the C3D features for improving recognition performance. The results demonstrate its promising performance on the Chalearn LAP Large-scale Gesture Recognition Challenge 2016.

H. Image and Video Analytics

This group includes 9 papers aiming at developing and applying similarity learning to more image and video analytic tasks. The paper "A Coarse-to-Fine Algorithm for Matching and Registration in 3D Cross-Sourced Point Clouds" by X. Huang, J. Zhang, Q. Wu, L. Fan, and C. Yuan investigates the task of 3D point cloud matching in the presence of density/scale variation and noise/outliers. The proposed method adopts a coarse-to-fine scheme, where ESF descriptor and scale embedded GMM are respectively adopted in the two stages. The results demonstrate the effectiveness and robustness of the method for outdoor 3D object detection and indoor 3D scene matching.

The paper "Second- and High-Order Graph Matching for Correspondence Problems" by R. Zhang and W. Wang studies the graph and hypergraph matching problem in a unified framework. Specifically, K-Nearest-Neighbor-Pooling Matching (KNNPM) and cell-algorithm-based MCMC with subpattern structure are respectively proposed for second- and high-order graph matching. The experiments show the robustness and improvements of the proposed algorithms.

The paper "A New Accurate and Fast Homography Computation Algorithm for Sports and Traffic Video Analysis" by S. Liu, J. Chen, C.-H. Chang, and Y. Ai investigates the homography computation task to meet the speed and accuracy requirement of sports and traffic video analysis. The proposed method includes quasi-optimal solutions for correspondence initialization as well as genetic algorithm based feature correspondence. The results demonstrate its effectiveness and efficiency.

The paper "An End-to-End Compression Framework Based on Convolutional Neural Networks" by F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, and D. Zhao suggests a deep CNN model which can be seamlessly integrated into the existing image codecs. Compact CNN is used to learn a compact representation before putting it to an image codec. Given the decoding result, reconstruction CNN is then deployed to remove the compression artifacts for refinement. Experiments show that it can be incorporated with JPEG, JPEG2000 and BPG for improved reconstruction.

The paper "Learning From Web Videos for Event Classification" by N. Chesneau, K. Alahari, and C. Schmid leverages the correlation between web videos and the associated textual information for event classification. By removing the irrelevant videos, the event classifiers can be learned without manually annotated data and achieve state-of-the-art results.

The paper "Learning Affective Features With a Hybrid Deep Model for Audio-Visual Emotion Recognition" by S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian presents a hybrid deep model, i.e. CNN+DBN, for audio-visual emotion recognition. A two-stage learning scheme is adopted to train the hybrid network. The results demonstrate its effectiveness.

The paper "Latent Bi-Constraint SVM for Video-Based Object Recognition" by Y. Liu, M. Hoai, M. Shao, and T.-K. Kim studies the task of recognizing object based on video by exploiting the temporal context. To this end, two datasets are provided, and a Latent Bi-constraint SVM (LBSVM) is developed. Experiments verify the benefit of introducing set and video for object recognition.

The paper "Copy and Paste: Temporally Consistent Stereoscopic Video Blending" by Z. Wang, X. Chen, and D. Zou studies the temporal consistency issue for stereoscopic video blending. It introduces a temporally-coherent mask propagation mechanism as well as a temporal blending algorithm. The results demonstrate the effectiveness and efficiency of the proposed method.

The paper "Learning a General Assignment Model for Video Analytics" by X. Zhang, Z. Zhu, Y. Zhao, and D. Chang studies several video analytics tasks, e.g., motion segmentation and activities recognition, from general assignment perspective. Furthermore, it present a novel general assignment model by incorporating structural dissimilarity and consistency boosting. The results demonstrate that the proposed method performs favorably in comparison to the state-of-the-arts.

I. Hardware-Aware Efficient Learning

Light-powered and embedded implementation of learning models also received upsurging interests along with the popularity of smart phones and cameras. This group includes 2 papers on this topic. The paper "A Light-Powered Smart Camera With Compressed Domain Gesture Detection" by A. Amaravati, S. Xu, N. Cao, J. Romberg, and A. Raychowdhury investigates the light-powered implementation of gesture detection. To this end, gesture feature extraction and efficient classification are conducted from the compressed measurements. Energy-efficient implementation is also provided on a low power MCU and powered by a solar powered DC-DC converter and regulator with MPPT. The system can achieve satisfying accuracy with 95mJ of energy per frame.

The paper "A Hardware Architecture for Cell-Based Feature-Extraction and Classification Using Dual-Feature Space" by F. An, X. Zhang, A. Luo, L. Chen, and H. J. Mattausch develops a dual-feature-based object recognition coprocessor. Hardware-friendly implementation is provided to exploit the complementarity of the two feature domains, and a prototype chip fabricated in 65 nm SOI CMOS is reported.

J. Remarks

This special issue aimed at promoting the scaling-up of metric learning for massive video and multimedia data, and the development and applications of nonlinear and deep similarity learning models. We believe it will offer a timely collection of information which is of both theoretical contribution and high practical value to the broad community of circuits and systems for video technology.

WANGMENG ZUO, *Professor*
School of Computer Science and Technology
Harbin Institute of Technology
China

LIANG LIN, *Professor*
School of Data and Computer Science
Sun Yat-sen University
China

ALAN L. YUILLE, *Professor*
Johns Hopkins University
USA

LEI ZHANG, *Principal Researcher*
and *Research Manager*
Microsoft Research
USA

HORST BISCHOF, *Professor* and
Vice Rector of Research
Graz University of Technology
Austria

FATIH PORIKLI, *Professor*
Research School of Engineering
Australian National University
Australia

Wangmeng Zuo received the Ph.D. degree in computer application technology from HIT in 2007. In 2004, from 2005 to 2006, and from 2007 to 2008, he was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. From 2009 to 2010, he was a Visiting Professor with Microsoft Research Asia. His current research interests include discriminative learning, image modeling and low-level vision, and biometrics. He has published over 50 papers in top tier academic journals and conferences, including IJCV, IEEE T-PAMI, T-IP, T-NNLS, T-IFS, *Pattern Recognition*, CVPR, ICCV, ICML, ACM MM, ECCV, and NIPS. He is an Associate Editor for the *IET Biometrics* and the Guest Editor for the *Neurocomputing* Special Issue on Smart Computing for Large Scale Visual Data Sensing and Processing and the Pattern Recognition Special Issue on Compositional Models and Structured Learning for Visual Recognition.

Liang Lin received the B.S. and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 1999 and 2008, respectively. From 2006 to 2007, he was a joint Ph.D. Student with the Department of Statistics, University of California at Los Angeles (UCLA). He was a Post-Doctoral Research Fellow with the Center for Vision, Cognition, Learning, and Art, UCLA. His research focuses on new models, algorithms, and systems for intelligent processing and understanding of visual data, such as images and videos. He has published over 50 papers in top tier academic journals and conferences, including the PROCEEDINGS OF THE IEEE, T-PAMI, T-IP, T-CSVT, T-MM, *Pattern Recognition*, CVPR, ICCV, ECCV, ACM MM, and NIPS. He was supported by several promotive programs or funds for his works, such as the Program for New Century Excellent Talents of Ministry of Education, China, in 2012, and the Guangdong Natural Science Funds for Distinguished Young Scholars in 2013. His Ph.D. dissertation received the China National Excellent Ph.D. Thesis Award Nomination in 2010. He received the Best Paper Runners-Up Award from ACM NPAR 2010, the Google Faculty Award in 2012, and the Best Student Paper Award from ICME 2014.

He has served as a Special Session Chair for ICIG 2013 and ICME 2014. He has served as an Associate Editor for *Neurocomputing* and as the Guest Editor for the *Pattern Recognition* Special Issue on Compositional Models and Structured Learning for Visual Recognition.

Alan L. Yuille received the B.A. degree in mathematics from the University of Cambridge in 1976. His Ph.D. on theoretical physics, supervised by Prof. S. W. Hawking, was approved in 1981. He was a Research Scientist with the Artificial Intelligence Laboratory, MIT, and the Division of Applied Sciences, Harvard University, from 1982 to 1988. He has served as an Assistant Professor and an Associate Professor at Harvard University until 1996. He was a Senior Research Scientist with the Smith-Kettlewell Eye Research Institute from 1996 to 2002. In 2002, he joined the University of California at Los Angeles as a Full Professor with a joint appointment in statistics and psychology. He obtained a joint appointment in computer science in 2007. He moved to John Hopkins Unverisity in 2015. His research interests include computational models of vision, mathematical models of cognition, and artificial intelligence and neural networks.

Horst Bischof received the M.S. and Ph.D. degrees in computer science from the Vienna University of Technology in 1990 and 1993, respectively, and the Habilitation (venia docendi) degree in applied computer science in 1998. He is currently the Vice Rector for Research with the Graz University of Technology, Austria, where he is also a Professor with the Institute for Computer Graphics and Vision. He has published over 630 peer-reviewed scientific papers. His research interests include object recognition, visual learning, motion and tracking, visual surveillance and biometrics, medical computer vision, and adaptive methods for computer vision. He is a member of the Scientific Board of Joanneum Research. He is also a Board Member of the Fraunhofer Institute für Graphische Datenverarbeitung.

He was the Program Co-Chair of ECCV2006 and the area chair of all major vision conferences several times. He was the General Chair of CVPR 2015. He is currently an Associate Editor for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition*, *Computing and Informatics*, and the *Journal of Universal Computer Science*. He was a Local Organizer for ICPR 1996. He was the Chairman of the DAGM/ÖAGM Conference in 2012 and the Co-Chairman of international conferences, including ICANN and DAGM.

Dr. Bischof is a member of the European Academy of Sciences. He has received 19 awards, including the 29th Pattern Recognition Award in 2002, the Main Prize of the German Association for Pattern Recognition DAGM in 2007 and 2012, the Best Scientific Paper Award from the BMCV 2007, the BMVC Best Demo Award 2012, and the Best Scientific Paper Awards from the ICPR 2008, ICPR 2010, PCV 2010, AAPR 2010, and ACCV 2012.

Lei Zhang (M'04–SM'11) was with Microsoft Research Asia for 12 years as a Senior Researcher, leading a research team working on visual recognition, image analysis, and large-scale data mining. He is a Senior Researcher with Microsoft Research, working with the Cloud Computing and Storage Group on visual recognition and exploring the solutions that can leverage the power of cloud computing. His years of work on large-scale search-based image annotation have generated many practical impacts in multimedia search, including a highly scalable solution of duplicate image clustering for billions of images. From 2013 to 2015, he moved to Bing Multimedia Search as a Principal Development Manager, helping develop cutting-edge solutions for web-scale image analysis and recognition problems, including image caption generation and high-precision image entity linking. He is the author or co-author of over 100 published papers in fields, such as multimedia, computer vision, web search, and information retrieval. He holds over 40 U.S. patents for his innovations in these fields. He is a Senior Member of the ACM. He has served as an Editorial Board Member for *Multimedia Systems*, and as the program co-chair, the area chair, or committee member for many top conferences.

Fatih Porikli (M'01–SM'04–F'14) received the Ph.D. degree from New York University, New York, NY, USA, in 2002. He has served as a Distinguished Research Scientist with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. Before joining MERL in 2000, he developed satellite imaging solutions at HRL, Malibu, CA, USA, and 3-D display systems at AT&T Research Laboratories, Middletown, NJ, USA. He is currently a Professor with the Research School of Engineering, Australian National University (ANU), Canberra, ACT, Australia. He is also the Computer Vision Group Leader with NICTA, Australia.

He has contributed broadly in the areas of object and motion detection, tracking, image-based representations, and video analytics. He was a recipient of the R&D 100 Scientist of the Year Award in 2006. He has received four best paper awards at premier IEEE conferences, including the Best Paper Runner-Up of the IEEE CVPR in 2007. He has also received five other professional prizes: the Most Popular Scientist Award by Popular Science, Turkey, in 2007, the Superior Invention Award by MELCO, Japan, in 2008, the MELCO Presidents Award, Japan, in 2007, the MERL Directors Award, USA, in 2008, and the MELCO-PUS Research Excellence Award, Japan, in 2009. His technological contributions have been transferred into many products.

He has authored over 130 publications and has over 66 patents. He is the co-editor of two books. He is the sole author of a pioneering paper on fast histogram computation, which is one of the most-cited papers in this area with over 600 citations in the last six years. Integral histogram technique has become the de facto standard for histogram-based image descriptors (HOG and so on) demonstrating speeds up by several orders of magnitude and enabling computationally efficient implementations.

Dr. Porikli was the General Chair of the AVSS 2010 and WACV 2014 and the Program Chair of the WACV 2015 and AVSS 2012. He has organized over 20 IEEE Computer Society cosponsored workshops, including the IEEE Workshop on Online Learning for Computer Vision in the last six years. He has served on the organizing committees of several flagship conferences, including the ICCV, ECCV, CVPR, ICIP, AVSS, ICME, ISVC, and ICASSP. He served as the Area Chair of ICME 2006, IV 2008, CVPR 2009, and ICPR 2010. He has served as the Associate Editor for five premier journals for the past eight years, including the *IEEE Signal Processing Magazine*, the *SIAM Journal on Imaging Sciences*, the *Journal of Image & Video Processing* (EURASIP), the *Journal on Machine Vision Applications* (Springer) from 2012 to 2015, and the *Journal on Real-time Image & Video Processing* (Springer).