Guest Editors' Introduction to the Special Section on Learning with Shared Information for Computer Vision and Multimedia Analysis

Trevor Darrell, Christoph Lampert, Nicu Sebe, *Senior Member, IEEE*, Ying Wu, *Senior Member, IEEE*, and Yan Yan

In the real world, a realistic setting for computer vision or multimedia recognition problems is that we have some classes containing lots of training data and many classes containing a small amount of training data. Therefore, how to use frequent classes to help learning rare classes for which it is harder to collect the training data is an open question. Learning with Shared Information is an emerging topic in machine learning, computer vision and multimedia analysis. There are different levels of components that can be shared during concept modeling and machine learning stages, such as sharing generic object parts, sharing attributes, sharing transformations, sharing regularization parameters and sharing training examples, etc. Regarding the specific methods, multi-task learning, transfer learning and deep learning can be seen as using different strategies to share information. These learning with shared information methods are very effective in solving real-world large-scale problems.

As guest editors of this special section on "Learning with Shared Information for Computer Vision and Multimedia Analysis", we were happy to receive 27 submissions to our special section. Among them, 12 papers have been accepted in this issue. The accepted 12 papers in this special section can be grouped into three different main categories: (i) Multi-modal approach; (ii) Domain adaptation method; (iii) Novel applications.

1 Multi-Modal Approach

The first group of 5 papers are centered on multimodal analysis for sharing information.

The paper "Learning Compositional Sparse Bimodal Models" by S. Kumar, V. Dhiman, P. Koch, and J. Corso, proposes a novel model representing bimodal percepts that exploits the compositional structure of language. The compositional sparse learning approach jointly learns the over-complete dictionaries, sparse bases, and cross-modal linking matrix. In contrast to prior work in bimodal modeling which is primarily discriminative in nature, the compositional sparse learning approach is generative and hence transparent. The authors demonstrate the effectiveness of sparsity and compositionality by both qualitative and quantitative evaluations.

The paper "Deep Multimodal Feature Analysis for Action Recognition in RGB+D Videos" by A. Shahroudy, T. Ng, Y. Gong, and G. Wang, presents a new deep learning framework for a hierarchical shared-specific component factorization, to analyze RGB+D features of human action videos. Each layer of the proposed network is an autoencoder based component factorization unit, which decomposes its multimodal input features into common and modality-specific parts. The authors further extend the deep factorization framework by applying it in a convolutional setting.

The paper "Hetero-manifold Regularisation for Cross-modal Hashing" by F. Zheng, Y. Tang, and L. Shao, introduces the concept of hetero-manifold for integrating the uni- and cross-modal similarities of multi-modal data in a global view. Both types of similarity are represented in the Laplacian matrix corresponding to the hetero-manifold. The Laplacian matrix is smooth when the Hamming distance is replaced by the Euclidean distance, which hints that no hash functions could be learned without all uni- and cross-modal similarities being defined on the hetero-manifold. Therefore, the proposed framework of hetero-manifold regularised hash function learning could benefit from the view of treating multi-modal data as a whole.

The paper "BreakingNews: Article Annotation by Image and Text Processing" by A. Ramisa, F. Yan, F. M. Noguer, and K. Mikolajczyk, proposes an adaptive CNN architecture that shares most of its structure for all the tasks. Addressing each problem then requires designing specific loss functions, and the authors introduce a metric based on the Great Circle Distance for geolocation and Deep Canonical Correlation Analysis for article illustration. All these technical contributions are exhaustively evaluated on a new dataset, BreakingNews, made of approximately 100K news articles (about 2 orders of magnitude more than similar existing datasets), and additionally including a diversity of metadata (like GPS coordinates and popularity metrics) that makes it possible to explore new problems.

- *T. Darrell is with Department of EECS, University of California, Berkeley, CA 94720. E-mail: trevor@eecs.berkeley.edu.*
- *C. Lampert is with Institute of Science and Technology, Klosterneuburg 3400, Austria. E-mail: chl@ist.ac.at.*
- *N. Sebe is with Department of Information Engineering and Computer Science, University of Trento, Trento, TN 38122, Italy. E-mail: sebe@disi.unitn.it.*
- *Y. Wu is with Department of EECS, Northwestern University, Evanston, IL 60208. E-mail: yingwu@eecs.northwestern.edu.*
- *Y. Yan is with Department of Computer Science, Texas State University, San Marcos, TX 78666. E-mail: tom_yan@txstate.edu.*

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2018.2804998

The paper "Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion" by I. D. Gebru, S. Ba, X. Li, and R. Horaud, proposes an audio-visual diarization method well suited for challenging scenarios consisting of participants that either interrupt each other, or speak simultaneously. In both cases, the speech-to-person association problem is a difficult one. The authors propose to combine multiple-person visual tracking with multiple speech-source localization in a principled spatiotemporal Bayesian fusion model. The diarization process was cast into a latent-variable dynamic graphical model.

2 Domain Adaptation Method

The second group of 3 papers are centered on domain adaptation methods for sharing information.

The paper "Webly-supervised Fine-grained Visual Categorization via Deep Domain Adaptation" by Z. Xu, S. Huang, Y. Zhang, and D. Tao, describes a simple but effective method for webly-supervised learning that performs knowledge transfer from existing datasets containing detailed annotations. The proposed method achieves two desired properties in a unified framework, namely scalability (by employing large-scale web images) and expertise (by introducing knowledge from sophisticated strongly supervised object recognition algorithms).

The paper "Domain Generalization and Adaptation using Low Rank Exemplar SVMs" by W. Li, Z. Xu, D. Xu, D. Dai, and L. Van Gool, proposes a new approach called Low-rank Exemplar SVMs (LRE-SVMs) for domain generalization by exploiting the low-rank structure of positive training samples from multiple latent source domains. Specifically, based on the recent work on exemplar SVMs, the authors propose to exploit the low-rank structure in the source domain by introducing a nuclear-norm based regularizer on the prediction matrix consisting of the predictions of all positive samples from all exemplar classifiers. The authors develop a new LRE-SVMs approach based on the least square SVMs (LRE-LSSVMs), which is much faster than the original LRE-SVMs method.

The paper "Deep Canonical Time Warping for Simultaneous Alignment and Representation Learning of Sequences" by G. Trigeorgis, M. Nicolaou, B. Schuller, and S. Zafeiriou, proposes a temporal alignment method based on deep architectures, which the authors dub Deep Canonical Time Warping (DCTW). DCTW discovers a hierarchical non-linear feature transformation for multiple sequences, where all transformed features are temporally aligned, and are maximally correlated. Furthermore, the authors consider the setting where temporal labels are provided for the data-at-hand. By modifying the objective function for the proposed method, the authors are able to provide discriminant feature mappings that may be more suitable for classification tasks.

3 Application

The last group of 4 papers is centered on interesting applications for sharing information.

The paper "Context-Aware Local Binary Feature Learning for Face Recognition" by Y. Duan, J. Lu, J. Feng, and J. Zhou, proposes a context-aware local binary feature learning (CA-LBFL) method for face recognition. In order to exploit more specific information from different scales, the authors have presented a context-aware local binary multi-scale feature learning (CA-LBMFL) method. Moreover, the authors have applied the above two methods to heterogeneous face matching by coupled learning methods (C-CA-LBFL and C-CA-LBMFL). The methods achieve better or very competitive recognition performance on four widely used benchmark face databases compared with the state-of-the-art face descriptors.

The paper "Collaborative Index Embedding for Image Retrieval" by W. Zhou, H. Li, J. Sun, and Q. Tian, explores the potential of unifying the index of CNN feature and SIFT feature for efficient and effective image retrieval. To adapt the deep CNN feature to the classic inverted index structure, the authors propose an energy ratio based method to improve the sparseness of the deep CNN feature. To integrate the index of SIFT and CNN feature, the authors propose a collaborative index embedding algorithm to alternatively upgrade the index files of CNN feature and SIFT feature.

The paper "Multi-Task Learning with Low Rank Attribute Embedding for Multi-Camera Person Re-identification" by C. Su, F. Yang, S. Zhang, Q. Tian, L. Davis, and W. Gao, proposes a multi-task learning (MTL) formulation with low rank attribute embedding for person re-identification. Multiple cameras are treated as related tasks, whose relationships are decomposed as a low rank structure shared by all tasks and task-specific sparse components for individual tasks by MTL. Both low level features and semantic/data-driven attributes are used. The authors have further proposed a low rank attribute embedding that learns attributes correlations to convert original binary attributes to continuous attributes, where incorrect and incomplete attributes are rectified and recovered. The objective function can be effectively solved by an alternating optimization under proper relaxation.

The paper "Unsupervised Transfer Learning via Multi-Scale Convolutional Sparse Coding for Biomedical Applications" by H. Chang, C. Zhong, J. Han, and J. Mao, proposes a Multi-Scale Convolutional Sparse Coding model (MSCSC) for unsupervised joint learning of filters at multi-scales with trainable collaboration among them, which, compared to CSC, leads to filters with improved scale-specificity and, subsequently, features with reduced redundancy across scales. Furthermore, such a joint learning strategy also provides an unsupervised solution for transfer learning, which is extremely helpful when the scale of labeled data is very limited.

4 Research Outlook

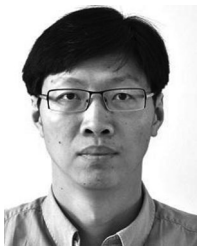
This special section focuses on new theory and algorithms for learning with shared information as well as applications putting these in practice. We believe this special section will offer a timely collection of information to benefit the researchers and practitioners working in the broad machine learning community and the research fields of computer vision and multimedia.

Trevor Darrell
Christoph Lampert
Nicu Sebe
Ying Wu
Yan Yan
Guest Editors

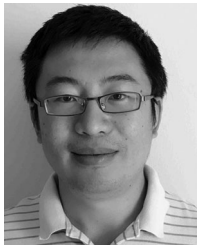
Trevor Darrell received the SM and PhD degrees from MIT, in 1992 and 1996, respectively. He is currently a full professor in the EECS Department, University of California, Berkeley and he is also appointed at the UC-affiliated International Computer Science Institute (ICSI). His group develops algorithms for large-scale perceptual learning, including object and activity recognition and detection, for a variety of applications including multimodal interaction with robots and mobile devices. His interests include computer vision, machine learning, computer graphics, and perception-based human computer interfaces. He was previously a professor in the MIT EECS Department from 1999-2008, where he directed the Vision Interface Group. He was a member of the research staff at Interval Research Corporation from 1996-1999.

Christoph Lampert received the PhD degree in mathematics from the University of Bonn, in 2003. In 2010 he joined the Institute of Science and Technology Austria (IST Austria) first as assistant professor and since 2015 as professor. His research on computer vision and machine learning won several international and national awards, including the best paper prize of CVPR 2008. In 2012 he was awarded an ERC Starting Grant by the European Research Council. He is an associate editor in chief of the *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, editor of the *International Journal of Computer Vision (IJCV)* and action editor of the *Journal for Machine Learning Research (JMLR)*.

Nicu Sebe received the PhD degree from Leiden University, The Netherlands, in 2001. He is currently a full professor in the Department of Information Engineering and Computer Science, University of Trento, Italy, where he is leading the research in the areas of multimedia information retrieval and human behavior understanding. He was a general co-chair of FG 2008 and ACM Multimedia 2013, and a program chair of CIVR 2007 and 2010, and ACM Multimedia 2007 and 2011. He was a program chair of ECCV 2016 and ICCV 2017. He is a senior member of the IEEE and ACM and a fellow of IAPR.



Ying Wu received the BS degree from the Huazhong University of Science and Technology, Wuhan, China, in 1994, the MS degree from Tsinghua University, Beijing, China, in 1997, and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 2001. He is Professor of electrical engineering and computer science with Northwestern University, Evanston, Illinois. He joined Northwestern University as assistant professor in 2001, and was promoted to associate professor in 2007 and to full professor in 2012. His research interests include computer vision, image and video analysis, pattern recognition, machine learning, multimedia data mining, and human-computer interaction. He serves as associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, the *IEEE Transactions on Image Processing (T-IP)*, the *IEEE Transactions on Circuit Systems and Video Technology (T-CSVT),* the *SPIE Journal of Electronic Imaging (JEI)*, and the *Journal of Machine Vision and Applications (MVA)*. He served many times as area chair of CVPR, ICCV, ICIP and ACM Multimedia. He is the program co-chair of CVPR'17. He received the Robert T. Chien Award at UIUC in 2001, and the National Science Foundation (NSF) CAREER award in 2003. He is a senior member of the IEEE.



Yan Yan received the PhD degree from the University of Trento, Italy, in 2014. He is currently an assistant professor in computer science with the Texas State University. He was research fellow with the University of Michigan (2016-2017) and University of Trento (2014-2016). He was a visiting scholar with Carnegie Mellon University, in 2013 and a visiting research fellow with Advanced Digital Sciences Center (ADSC), UIUC, Singapore in 2015. He is the recipient of Best Student Paper Award in ICPR 2014 and Best Paper Award in ACM Multimedia 2015. He served as a guest editor in the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *ACM Transactions on Multimedia Computing, Communications, and Applications* and the *Computer Vision and Image Understanding*.