https://doi.org/10.2352/ISSN.2470-1173.2020.8.IMAWM-185

A Local-Global Aggregate Network for Facial Landmark Localization*

Ruiyi Mao [a]***, Qian Lin*** [b] ***and Jan P. Allebach*** [a]

[a] ***School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, U.S.A***
[b] ***HP Labs, Palo Alto, CA, U.S.A***

Abstract

Facial landmark localization plays a critical role in many face analysis tasks. In this paper, we present a novel local-global aggregate network (LGA-Net) for robust facial landmark localization of faces in the wild. The network consists of two convolutional neural network levels which aggregate local and global information for better prediction accuracy and robustness. Experimental results show our method overcomes typical problems of cascaded networks and outperforms state-of-the-art methods on the 300-W [1] benchmark.

Introduction

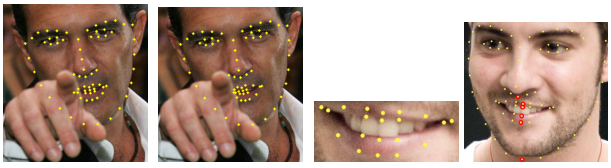
Facial landmark localization aims to automatically detect and localize distinctive human facial features, which is regarded as the basis for many face analysis tasks, such as face recognition, face model reconstruction and so on. Even though progress has been made in recent years, facial landmark localization in the unconstrained environment is still a very challenging problem in computer vision. The challenges come from the large variations of face appearance caused by different illuminations, facial expressions, angles of heads, and face image qualities.

Works done on this topic can be divided into two categories: generative methods and discriminative methods. For generative methods, a prior generative model for both the face shape and appearance is generated and used. Works such as Active Appearance Models (AAMs) [2, 3] and Active Shape Models (ASMs) [4, 5] use generative methods. For discriminative methods, the target location is directly inferred from facial appearances. Works such as Constrained Local Models (CLMs) [6, 7], Deformable Part Models (DPMs) [8, 9], and regression-based method [10, 11] are discriminative methods.

With the breakthrough and development of deep learning, Sun et al. [12] first applied a convolutional neural network in a cascaded regression framework. Since then, more works have been done using deep neural networks [13, 14], and some works further exploited a cascaded convolutional neural networks framework and achieved state-of-the-art performance [15, 16, 17]. All these cascaded frameworks share the same strategy of level-wise coarse-to-fine refinement, and have been demonstrated to have superior robustness and accuracy compared with previous methods.

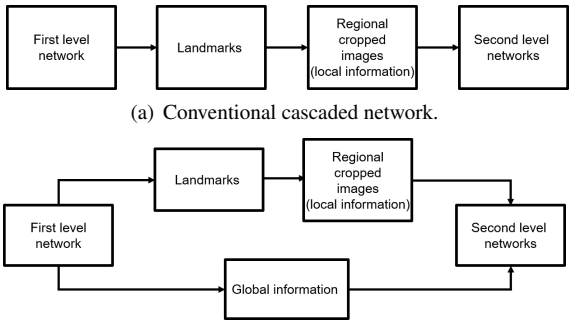
For all the works using a convolutional neural network cascade framework so far, a first level network is trained to obtain a rough prediction of all facial landmarks. The second or higher level subnetworks take regional cropped images as input and do refinements within the region. However, through experiments, we discovered that this cascade architecture does not work well for occluded faces and deformable facial components, for example mouths, which cause a major part of the error. Representative examples are shown in Fig. 1. In Figs. 1(a) and 1(b), part of the mouth is occluded by a hand, the mouth landmark prediction from the conventional second level network is even worse than that from the first level network. In another example, if the mouth landmarks generated by conventional second level networks (Fig. 1(c)) are shown in whole face scale in Fig. 1(d), it is obvious that the center facial landmarks of the nose, mouth, and chin marked by red circles are not lined up and are shifted from the mid perpendicular of the face, where they are supposed to be.

The problem for the conventional cascaded network is that with the regional cropped image as the only input, the second level subnetworks only have local information to make the prediction without any global information. Since facial landmarks



(a) (b) (c) (d)

Figure 1. *Mouth landmark prediction failure cases with conventional cascaded network. (a) One-level network prediction; (b) Conventional cascaded network prediction; (c) Conventional cascaded network prediction; (d) Mouth landmark prediction in the scale of whole face.*



(a) Conventional cascaded network.

(b) The proposed local-global aggregate network (LGA-Net).

Figure 2. *Comparison of conventional and proposed cascaded network.*

*Research supported by HP Inc., Palo Alto, CA.

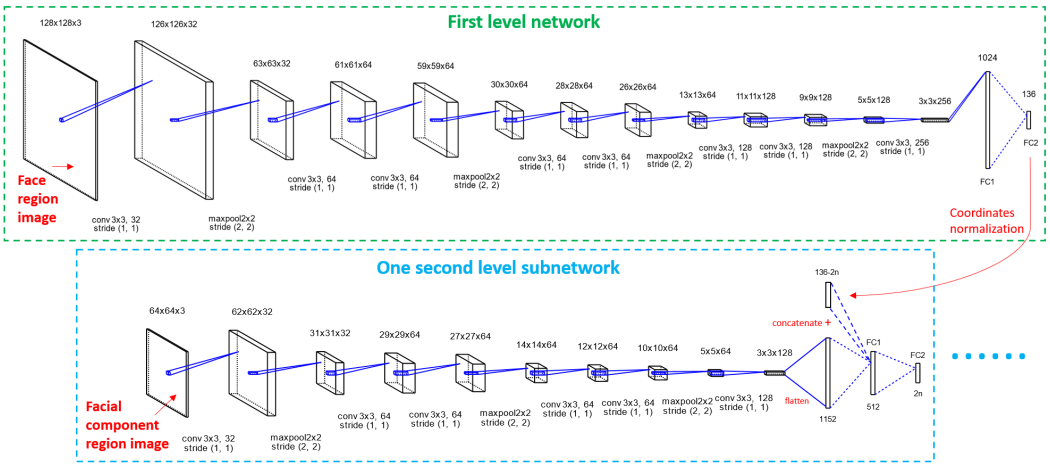


Figure 3. *Local-global aggregate network structure.*

are mutually correlated, regional landmark predictions from the subnetworks with no global facial information may sacrifice accuracy, especially for deformable facial components or faces regionally occluded or corrupted. To tackle this problem, a novel local-global aggregate network (LGA-Net) whose subnetworks are able to accept and aggregate both local and global information is presented. The network consists of two level of networks, and its second level subnetworks take both local and global information from first level network, which is illustrated in Fig. 2. The details of the network are discussed in Section 2.

Local-global Aggregate Network

In the proposed network, the first level network takes face images as input which is obtained by independent face detectors. Since the first level network is trained with full faces, which contain global information, it is capable of giving rough, but robust predictions of all the facial landmarks. However, its accuracy is very limited, especially for some deformable facial components such as mouth and eyes. In order to get both high accuracy and small overall model size, strategies of coarse-to-fine are involved, which introduce the cascaded neural network architecture. At each following cascaded level, the networks locally refine a subset of facial landmarks within the corresponding regional images obtained from the previous level.

For more robust landmark predictions in unconstrained environments, the local-global aggregate network's second level subnetworks take both regional images and global facial information from its first level as input. In this work, landmark coordinates obtained from the previous level network are chosen as the high-level global features. The benefit of this choice is that during training, the global features can be obtained directly from training data ground truth instead of from previous networks, so that each subnetwork can be trained and cascaded independently.

The detailed structure of the novel network cascade is shown in Fig. 3. The first level network takes a 3-channel 128×128 face image as input and outputs an array with 136 elements which are the 68-point landmark coordinates for the whole face. Facial components bounding boxes are generated from these coordinates. The original face image is then cropped by the components bounding boxes, re-sized to 64×64 and fed to the second level subnetworks as local information for landmark refinement. It should be noted that a maximum aspect ratio of 2:1 is set for the cropped component regional images. If the maximum ratio is exceeded in both training and testing, the shorter side of the regional image is forced to extend to keep a reasonable aspect ratio for better performance.

As the global information input to the second level networks, the first level landmark coordinates are normalized according to each of the corresponding facial component bounding boxes and fed to the second level subnetworks. To avoid landmark predictions from first level interfering with the second level subnetworks' predictions too much, the corresponding landmark coordinates that are going to be predicted by the second level subnetwork are excluded from the input of first level landmark coordinates. To effectively aggregate local and global features, in the second level networks, the high level features obtained from its last convolutional layer are flattened to a one-dimensional array, which is then concatenated with the landmark coordinates from the first level network to form a new feature array containing both local and global features. This array is then fed to the fully connected layers to make the refined landmark prediction. Finally, the landmark coordinates from all second level networks are combined with face contour landmarks from first level network to form the final prediction. There are in total 4 second level subnetworks which are left eyebrow, left eye, nose, and mouth. In order to make the whole network smaller, the right eyebrow and right eye are processed through the left eyebrow and left eye subnetworks by horizontally flipping the regional image and corresponding landmark coordinates from the first level network.

Network Training

Datasets

To train the proposed network, several public datasets with 68-point facial landmark configurations are used, including 300-W [1], AFW [8], HELEN [18], IBUG, LFPW [19], and MENPO [20]. In total, 10831 distinct raw training images with 68 point annotations are used to train and test the networks.

Mouth Landmark predictions	RMSE (%)
One-level network	3.935
Conventional cascaded network	3.941
LGA-Net (Ours)	3.479

Table 1. Mouth landmark prediction accuracy of different networks on 300-W test dataset

Data Augmentation

For each training image, bounding box random expansion [21], random rotation, horizontal flipping, and random Gaussian blurring are applied. In this work the maximum bounding box random expansion shift of all the sides is 0.3 of the width and height of the initial face bounding box.

In this work, we manually divide the whole training dataset into two subsets: normal augmentation set and strong augmentation set. The normal augmentation set contains common training samples and the strong augmentation set contains rarer and more challenging training samples such as rare and challenging head poses, facial expressions, and illuminations. In this work, 200 augmented training samples are generated from each data sample in the strong augmentation set; and 50 augmented training samples are generated from each data sample in the normal augmentation set.

Training Detail

In order to train a specific network, the landmark coordinates need to be normalized based on their own bounding box. For the first level and second level networks, it is the face bounding box and the corresponding facial component bounding box, respectively. For example, consider a bounding box whose top left corner coordinate is (x_{bbox}, y_{bbox}), with w and height h. (x, y) is one landmark coordinate before normalization. Then the normalized landmark coordinate (x_{norm}, y_{norm}) based on the bounding box is

$$(x_{norm}, y_{norm}) = (\frac{x - x_{bbox}}{w}, \frac{y - y_{bbox}}{h}) \quad (1)$$

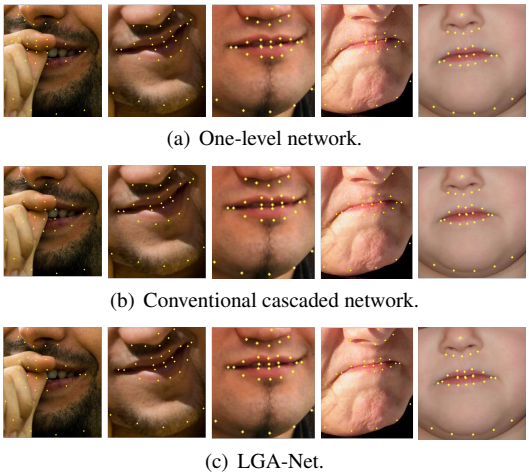
The same landmark normalization process is applied on both the first and second level networks using corresponding face or facial component bounding boxes.

For all the networks, the Euclidean loss is chosen to be the loss function for network training. Each network in the first and second level is trained separately. For all the networks, the dataset is divided into a training set which contains 80% of data samples and a validation set, which contains 20% of the images. Adam [22] is used for network parameters optimization.

Experimental Results

The accuracy of the landmark prediction was measured by the point-to-point RMS error between each predicted landmark and the ground truth annotations, normalized by the face's interocular distance, as proposed in [8].

Since the mouth is the most complicated and deformable component on the face and its failure cases have been shown before, the comparison between mouth landmark predictions by one-level network, a conventional cascaded network, and LGA-Net are shown in Table. 1 and Fig. 4. It can be seen that the proposed network has better performance in both robustness and accuracy compared with the one-level network and conventional cascaded network. It also does well on the facial mid-perpendicular check as discussed before, which the conventional cascaded network handles badly.



(a) One-level network.

(b) Conventional cascaded network.

(c) LGA-Net.

Figure 4. *Mouth landmark prediction of different networks.*

To evaluate the performance of our facial landmark localization network, the most widely recognized 300-W benchmark testset is used. We compared our method with recent state-of-the-art methods on the 300-W Common Subset, Challenging Subset and Fullset in Table. 2. It shows that our method has the highest accuracy compared with other state-of-the-art methods. Fig. 5 gives some examples from the test dataset. It can be seen that the test dataset contains great variations in pose, expressions, and lighting conditions; and our network is very robust and able to give superior accurate facial landmark predictions.

Figure 5. *Landmark prediction examples using the LGA-Net.*

Conclusion

In this paper, we present a novel local-global aggregate network for facial landmark localization. In our method, two levels of CNN are carefully designed to aggregate both local and global information for more robust and accurate landmarks prediction compared with a conventional cascaded networks. The experimental results show the state-of-the-art performance of the proposed method, which demonstrates its superiority over other facial landmark localization algorithms.

Method	Common	Challenging	Full Set
SDM [10]	5.57	15.40	7.52
CFAN [23]	5.50	16.78	7.69
LBF [11]	4.95	11.98	6.32
CFSS [24]	4.73	9.98	5.76
TCDCN [13]	4.80	8.60	5.54
Fan et al. [16]	4.76	8.25	5.45
Honari et al. [25]	4.67	8.44	5.41
TSR [26]	4.36	7.56	4.99
RCN+ (L+ELT) [27]	4.20	7.78	4.90
RF-CHN [28]	4.03	6.84	4.58
DCFE [29]	3.83	7.54	4.55
Chen et al. [17]	3.73	7.12	4.47
PCD-CNN [30]	3.67	7.62	4.44
LGA-Net (Ours)	**3.42**	**6.23**	**4.18**

Table 2. Mean RMSE (%) on 300-W benchmark Common Subset, Challenging Subset and Fullset (68-point).

References

[1] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[2] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *European Conference on Computer Vision*. Springer, 2008, pp. 504–513.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J Graham, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[5] F. Kahraman, M. Gokmen, S. Darkner, and R. Larsen, "An active illumination and appearance (AIA) model for face alignment," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–7.

[6] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *European Conference on Computer Vision*. Springer, 2008, pp. 340–353.

[7] J. M. Saragih, S Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[8] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2879–2886.

[9] M. Uřičář and V. Franc, "Detector of facial landmarks learned by the structured output SVM," *VIsAPP*, vol. 12, pp. 547–556, 2012.

[10] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[11] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[12] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3476–3483.

[13] Z. Zhang, P. Luo, C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 94–108.

[14] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4177–4187.

[15] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *2013 IEEE International Conference on Computer Vision Workshops*, 2013, pp. 386–391.

[16] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *Image and Vision Computing*, vol. 47, pp. 27–35, 2016.

[17] X. Chen, E. Zhou, Y. Mo, J. Liu, and Z. Cao, "Delving deep into coarse-to-fine framework for facial landmark localization," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 2088–2095.

[18] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*. Springer, 2012, pp. 679–692.

[19] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.

[20] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 170–179.

[21] R. Mao, Q. Lin, and J. Allebach, "Robust convolutional neural network cascade for facial landmark localization exploiting training data augmentation," in *Imaging and Multimedia Analytics in a Web and Mobile World 2018, (Part of IST Electronic Imaging 2018)*, 2018.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *European Conference on Computer Vision*. Springer, 2014, pp. 1–16.

[24] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4998–5006.

[25] S. Honari, J. Yosinski, P. Vincent, and C. Pal, "Recombinator networks: Learning coarse-to-fine feature aggregation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5743–5752.

[26] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark

detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3317–3326.

[27] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz, "Improving landmark localization with semi-supervised learning," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1546–1555.

[28] Weiliang Chen, Qiang Zhou, and Roland Hu, "Face alignment by combining residual features in cascaded hourglass network," in *2018 25th IEEE International Conference on Image Processing*. IEEE, 2018, pp. 196–200.

[29] Roberto Valle, Jose M Buenaposada, Antonio Valdes, and Luis Baumela, "A deeply-initialized coarse-to-fine ensemble of regression trees for face alignment," in *European Conference on Computer Vision*. Springer, 2018, pp. 585–601.

[30] Amit Kumar and Rama Chellappa, "Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 430–439.

Author Biography

Ruiyi Mao *received his B.Eng degree in Electrical and Computer Engineering from Huazhong University of Science and Technology, China, and University of Birmingham, UK, in 2012. He received his Ph.D. degree in Electrical and Computer Engineering at Purdue University in 2018 and is now an software engineer working on image/video processing, computer vision and machine learning at Apple.*

Qian Lin *is HP Fellow working on computer vision and deep learning research at HP Labs. Dr. Lin joined the Hewlett-Packard Company in 1992. She received her BS from Xi'an Jiaotong University in China, her MSEE from Purdue University, and her Ph.D. in Electrical Engineering from Stanford University. Dr. Lin is inventor/co-inventor for 44 issued patents. She was awarded Fellowship by the Society of Imaging Science and Technology (IS&T) in 2012, and Outstanding Electrical Engineer by the School of Electrical and Computer Engineering of Purdue University in 2013.*

Jan P. Allebach *is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. Allebach is a Fellow of the IEEE, the National Academy of Inventors, the Society for Imaging Science and Technology (IST), and SPIE. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IST, the highest award that IST bestows. He received the IEEE Daniel E. Noble Award, and the OSA/IST Edwin Land Medal. He is a member of the National Academy of Engineering. He served as a IST Visiting Lecturer (1991-1992), and twice as an IEEE Signal Processing Society Distinguished Lecturer (1994-1995, 2016-2017).*

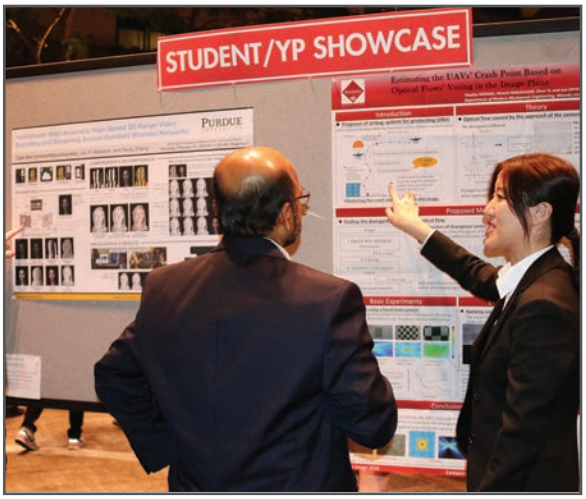
JOIN US AT THE NEXT EI!

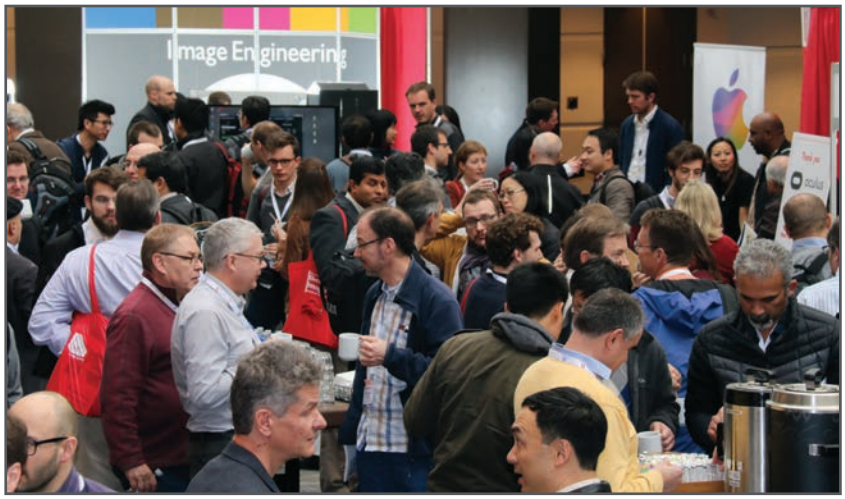
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org