https://doi.org/10.2352/ISSN.2470-1173.2019.16.3DMP-004

Depth-map Estimation Using Combination of Global Deep Network and Local Deep Random Forest

SangJun Kim, Sangwon Kim, Deokwoo Lee and ByoungChul Ko; Dept. of Computer Engineering, Keimyung University, Daegu, S. KOREA

Abstract

This study propose a robust 3D depth-map generation algorithm using a single image. Unlike previous related works estimating global depth-map using deep neural networks, this study uses the global and local feature of image together to reflect local changes in the depth map instead of using only global feature. A coarse-scale network is designed to predict the global-coarse depth map structure using a global view of the scene and the finer-scale random forest (RF) is to be designed to refine the depth map based on combination of original image and coarse depth map. As the first step, we use a partial structure of the multi-scale deep network (MSDN) to predict the depth of the scene at a global level. As the second step, we propose local patch-based deep RF to estimate the local depth and smoothen noise of local depth map by combining MSDN global-coarse network. The proposed algorithm was successfully applied to various single images and yielded a more accurate depth-map estimation performance than other existing methods.

Introduction

Recently, as diverse applications are requiring 3-dimensional (3D) geometry, the importance of accurate depth estimation has been also increasing. From a few decades, there have been extensive research activities and significant improvement of performance in the fields of depth-map estimation. Depth-map estimation has been crucial and fundamental problem in the fields of computer vision and playing a key role in development of intelligence application systems including 3D modeling, computer graphics, virtual reality, augmented reality, and human-computer interaction.

To accomplish a highly accurate depth-map of a scene, all of the conditions should be sufficiently provided. To alleviate the limitations, the active method was introduced. One of the active methods, using structured light (SL) patterns, replaces one camera with a light source that generates a light pattern [1]. Once camera acquires an image of a 3D scene overlaid with the light patterns, and geometric relationship between the original and deformed light patterns, it provides sufficient information to recover depth of the target scene. Structured light patterns are able to provide highly accurate feature correspondence compared to the one using passive stereo vision method. Another widely employed active method is time-of-flight (ToF) that estimates depth of a target scene based on the known information of a speed of the light projected onto a target scene [2].

RGB-D camera has brought great attention with appearance of Microsoft Kinect leading to capability of low-cost and real-time acquisition of depth-map and color information for 3D modeling or reconstruction. Major components of RGB-D camera are RGB camera, infrared (IR) camera and projector, and calibration of or between all of the components is very important. RGB-D camera, usually employs a concept of SL or of ToF, and the accuracy of depth depends on calibration between RGB camera and depth sensor [3]. While passive and active methods have been widely applied to the field of computer vision, each of method is inherent with following limitations and disadvantages; Passive method usually suffers from texture-less region, occlusion or low-light condition, and active method suffers from external environment, e.g., ambient light, specular component of a target object, etc., and issues of low resolution [4]. Active methods with SL or ToF have difficulties in extracting depth-map in outdoor environment, and sometimes suffer from high power consumption.

To alleviate the problems and issues arisen in the methods all above, from practical perspectives, single image based depth-map estimation has been considered one of the most reliable alternatives to the conventional methods, e.g., stereo vision, SL, ToF, SFM, RGB-D sensors, or fusion (or hybrid). To extract depth-map using a single image, several approaches were proposed. Among many other methods, popularly employed methods were based on the properties of a lens in that quantification of a focus and defocus and based on the concept of learning based techniques such as neural network, deep learning, etc.

This paper proposes a new depth-map estimation method based on single image with combination of two learning based techniques. Contrast to the previous works on deep neural network (DNN) based depth-map estimation or other applications, the proposed approach combines global depth-map predicted form shallow structure of neural network with refined depth-map that is estimated from deep random forest (RF) to enhance depth-map accuracy.

Combination of global and local depth

Unlike related works, this study use the global and local feature of image together to reflect local changes in the depth map instead of using only global feature. Therefore, this study proposes a hybrid depth regression architecture which combines global-coarse network and local-patch based random forest (RF). As the first step, we use a partial structure of the multi-scale deep network (MSDN) [5] to predict the depth of the scene at a global level. Originally, MSDN consists of two component stacks such as a coarse-scale network and fine-scale network. A coarse-scale network is designed to predict the global-coarse depth map structure using a global view of the scene and the finer-scale network is to be designed to refine the depth map based on combination of original image and coarse depth map. However, because the finer-scale network is constructed with several convolutional layer using global image, this network also has limitations in predicting local depth. Therefore, in this paper, we use local patch-based deep RF

to estimate the local depth and smoothen noise of local depth map by combining MSDN global-coarse network as the second step.

Global-coarse depth map

This study use the global and local feature of image together to reflect local changes in the depth map instead of using only global feature. Therefore, this study proposes a hybrid depth regression architecture which combines global-coarse network and local-patch based RF. As the first step, we use a partial structure of the multi-scale deep network (MSDN) [5] to predict the depth of the scene at a global level. Originally, MSDN consists of two component stacks such as a coarse-scale network and fine-scale network. A coarse-scale network is designed to predict the global-coarse depth map structure using a global view of the scene and the finer-scale network is to be designed to refine the depth map based on combination of original image and coarse depth map. However, because the finer-scale network is constructed with several convolutional layer using global image, this network also has limitations in predicting local depth. Therefore, in this paper, we use local patch-based deep RF to estimate the local depth and smoothen noise of local depth map by combining MSDN global-coarse network as the second step.

Appearance and spatial location feature

From the training images, we first extract 32x32 patches randomly and group similar patches together using k-mean clustering. We measure similarity between patches using two features such as appearance and spatial location to ensure the accuracy of resulting clusters. As the first appearance feature, oriented center symmetric-local binary patterns (OCS-LBP) [6] is employed because it is invariant to changes in the image rotation and scaling. OCS-LBP is an eight dimensional histogram consisted of the magnitude of the closest gradient orientation bin ranging from 0° to 360° in 45° for every pixels included in a patch.
The second spatial location is the centroid of a patch to measure the spatial distance between a candidate patch and the centroid of comparison target cluster. If the spatial distance between a patch and all candidate clusters exceeds certain condition, then a new cluster with same appearance but different centroid is generated.

Local depth map estimation using deep RF

Although CNN is used to predict the depth map in many researches using a global view of the scene, deep convolution networks generally do not explicitly consider dependencies between local variables. Therefore, the size of the field of view is important for CNN's performance [7]. To derive depth considering various types of local views, we need to configure multiple CNNs optimized for each local view. However, it is inefficient to construct multiple deep networks in terms of computational resource and time.

Therefore, in this paper, instead of heavy CNN, a light random forest (RF) is connected to several layers to derive the depth of the local region. An RF is a decision tree ensemble classifier (regressor) and it is known to process very large amounts of data with high training speeds compared to conventional classifiers [8]. To connect light RF deeply, we propose a new algorithm that can extract a local depth map by improving gcForest algorithm [9] instead of heavy CNN.

The gcForest generates deep forest holding three characteristics behind the success of deep neural networks, i.e., layer-by-layer processing, in-model feature transformation and sufficient model complexity. This approach is a decision tree ensemble, with much less hyper-parameters than deep neural networks, and its model complexity can be automatically determined in a data-dependent way. From the experiments on even across different data from different domains, gcForest approach was able to get excellent performance compared to deep neural networks by using the same default setting and showed the possibility of constructing deep models without using backpropagation.

However because original gcForest is still deep and wide for fast processing, we modified the structure of original gcForest. The proposed deep RF architecture consists of two types of stages such as *multi-scale RF* and *cascade RF regressor*. The second stage is an ensemble of RF regressors, i.e., and *ensemble of ensemble*. Each stage of cascade receives feature information processed by its proceeding stage, and outputs its processing result to the next layer.

The first stage is the *multi-scale RF*. This stage includes different types of RF to encourage the diversity by randomly selecting samples for generating diverse decision trees data and randomly selecting a feature for split at each node of the tree. In this stage, the extracted patches are upsampled to 3 scales and the extracted OCS-LBP feature from each patch is used to construct the RF for each scale.

Because deep RF is designed to estimate local depth, we first extract patches randomly from training images with their depth values collected from depth sensor. After random patches are extracted, two features, OCS-LBP and spatial location (patch centroid) are also extracted from each patch. In terms of depth values, most of methods follow a regression task to estimate depth values. However, because it is difficult to regress the depth value to be exactly the ground-truth value, we discretize the depth values into several discrete bins (256) in log space inspired by [7].

After all N patches included in training images have been extracted, a vector for patch $\mathbf{p}$ is constructed:

$$\mathbf{p}_i = \{(\mathbf{ocs_i}, \mathbf{sp_i}, \mathbf{dp_i})\}, i = 1,2, \dots N$$

where $\mathbf{ocs_i} = [ocs_{i1}, ocs_{i2}, \dots, ocs_{iM}]$ is composed of M dimensional CS-LBP feature vector according to the number of blocks, $\mathbf{sp_i} = [cx_i, cy_i]$ is composed of two-dimensional centroid vector, and $\mathbf{dp}$ is the discrete depth level of all pixels in a patch. with, and $\mathbf{dp}_i = [d_1, \dots, d_{1024}]$ is composed of a scalar depth label d (discretised depth level) for all pixels included in i-th patch marked by 3D depth sensor.

To invariant to image scaling, the scale of a patch gradually increases from scale 1 to scale 3. At scale 1, a patch is divided into 2 × 2 non-overlapping blocks, 3 × 3 non-overlapping blocks at scale 2, and 4 × 4 non-overlapping blocks at scale 3. The scale of a patch at stage 1 is 32 × 32 pixels, and the scale is increased by 20 × 20 pixels at successive scales.

As the first classifier, we train the RF regressor using OCS-LBP, which is extracted from a n × n sub-blocks and spatial location of a patch. Then, we train the RF using training dataset $A = \{\mathbf{p}_i = (\mathbf{ocs}_i, \mathbf{sp}_i, \mathbf{dp}_i) | i = 1,2, \dots N\}$ consisted of N patches.

In the training procedure of the RF, the individual decision tree first chooses a random subset A' from the training dataset, A. At node O, the sample A'_O is iteratively split into left and right subsets, A'_l and A'_r, by using the threshold, t, and split function, $f(\mathbf{p}_i)$, for the feature vector, $\mathbf{p}$. Then, several candidates are randomly created by the split function and threshold at the split node. From among these, the candidate that maximizes the information gain about the corresponding node is selected. The information gain, ΔE, is usually calculated by entropy estimation.

$$\Delta E = E(\mathrm{A}'_O) - \frac{|\mathrm{A}'_l|}{|\mathrm{A}'_O|} E(\mathrm{A}'_l) - \frac{|\mathrm{A}'_r|}{|\mathrm{A}'_O|} E(\mathrm{A}'_r) \quad (1)$$

where $E(\cdot)$ is the entropy computed for Gaussian kernel of the

classes in the set of training samples A'.

This study use the following Gaussian kernel inspired by [10] as the entropy instead of common Shannon entropy function to consider patch appearance and spatial location at the same time.

After training the regression tree, each leaf node predicts and stores a 2D depth vector using training samples in the leaf node. To predict the depth vector, we use the median of all the training samples in the leaf node instead of averaging because median is less sensitive to outliers (noises) than averaging. The final depth vector is generated by ensemble (arithmetic averaging) of each depth of all trees. Since RF regressor consists of three scales, we connect the depth vectors (1,024d, 32×32) created for each RF regressor into one dimension depth vector and the total 3,072 dimension depth vector is created per tree.

The second stage, *cascade RF regressor,* is to estimate the depth of the patch finally by applying the depth feature vector output from the first stage to cascade N layer RF sequentially. Unlike the study of [9], each RF layer consists of randomly generated regression trees instead of heterogeneous RFs to reduce the memory and computational time. The number of regression trees of each layer is same as 200.

Note that here we combine global depth map generated from global-coarse network with output of each layer to refine the local depth map and smoothen noise of local depth map. By combining two complement depth vector, we can expect in-model transformation as like a deep network model that is creating new features during the learning process. Moreover, it is expectable that more profit can be obtained if more augmented features are involved [9].

After the first stage 'multi-scale RF regressor', the transformed training set $B = \{\mathbf{p_i'} = (\mathbf{Crdp_i}, \mathbf{Ctdp_i}) | i = 1,2, \dots N\}$ is generated. This dataset B consists of N patches with 1,024 dimensional transformed feature vector. In detail, global depth map predicted from global-coarse network of *i*-th patch region, $\mathbf{Crdp_i} = [cd_1, \dots, cd_{1024}]$, and the concatenation of a scalar depth predicted from multi-scale RF regressor of *i*-th patch, $\mathbf{Ctdp}_i = [d_1, \dots, d_{1024}, \dots, d_{2048}, \dots, d_{3072}]$. The transformed feature, which is augmented with the coarse depth map predicted by the global-coarse network, will then is used to train from the first layer to L layer of cascade RF regressor, respectively. This procedure will be repeated till convergence of validation performance.

To automatically train the number of RF layers and RF parameters by avoiding overfitting, we split the training set B into two parts, such as growing set B^g and estimating set B^e as taking 80% of the training data for growing set and 20% for estimating set. To determine the optimal objective function of the regression tree, we use the least-squares error function (g) to find the split function that minimizes errors [11]:

$$g_j = \sum_{Q_j} (\mathrm{D}_j - \overline{\mathrm{D}}_j)^2 - \sum_{i \epsilon \{l,r\}} \left(\sum_{Q_j^i} \left(\mathrm{D}_j - \overline{\mathrm{D}}_j \right)^2 \right) \quad (2)$$

where Q_j indicates the set of training samples arriving at node *j*. Here, *l* and *r* are the left and right split nodes, respectively, and $\overline{\mathrm{D}}_j$ indicates the mean depth vector of individual depth D_j for all training samples reaching the *j*-th node.

After training the regression trees of *l*-th RF layer, each leaf node predicts and stores a depth vector using training samples in the leaf node. Then the loss function is estimated to decide whether or not the RF is extended to next layer. As the loss function, we use mean squared error (MSE) because it is easy to implement and generally works well. To calculate MSE of *l*-th layer, we take the difference between predictions (D_i^t) of T trees included in RF regressor and the ground truth depth ($\widehat{D}_i$) of i-th patch and average it out across the whole dataset. The general form of the loss function.

Given a test patch, it goes through the multi-scale RF regressor procedure to get its corresponding transformed feature representation, and then go through the cascade RF regressor till the last layer. Then the final predicted depth of each regression tree are averaged to produce the final depth map of *i*-th patch.

Experimental Results

This study use the NYU depth v2 dataset [Silberman] consist of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. It contains 464 indoor scenes taken from 3 cities and composed of 1,449 densely labelled pairs of aligned RGB and depth images and 407,024 new unlabelled frames. The image size is down-sampled by half from 640x480 to 320 x 240 pixels.

For training coarse CNN network, shuffle into a list of 220K after evening the scene distribution (1,200 images per scene) and performed the data augmentation with random online transformation, such as scaling, rotation, translation, color conversion, and flipping as the same method of [5]. Therefore, coarse network is trained using 2M samples with SGD and 32 batch size. Learning rate is 0.001 for layers 1-5 and 0.1 for coarse full layers 6 and 7.

To train the proposed deep RF, we used same augmented images from NYU depth v2 dataset. In this study, we set the maximum size (number) of trees as of RF as 50 trees, because the accuracy no longer improves as the tree number of trees increases over 50. After the multi-scale RF regressor was constructed using dataset A, dataset B was applied to the cascade RF regressor and produced 5 layers deep RF. The number of regression trees of one layer is 10 and tree depth is 15.

Evaluation of proposed approach

To verify the effectiveness of the proposed depth estimation method, we compared the performance of six state-of-the-art methods: (1) Liu et al. [12], which uses DNN, (2) Eigen et al. [5], which uses a multi-scale deep network, (3) Roy et al. [26], which estimates depth using neural regression forest, (4) Eigen and Fergus [13], which uses a common multi-scale convolutional architecture, (5) Chakrabarti et al. [14], which estimates depth-map by harmonizing over-complete local network, (6) Lee et al. [15], which is based on Fourier domain analysis, (7) global-coarse network, (8) proposed depth-map estimation method consisting of 30 regression trees of RF. The eight methods used for performance comparison commonly are based on DNN.

The color version of results are shown in http://cvpr.kmu.ac.kr/Depthmap_results. This result shows the depth-map estimation results of some examples NYU v2 2006[1].

Conclusion and future works

This study proposed a new depth-map estimation method based on a single image with combination of two learning based techniques. Contrast to the previous works on deep neural network based

[1] We referred the results of six methods from the experimental evaluations of [15]

depth-map estimation or other applications, the proposed approach combines global depth-map predicted form shallow structure of neural network with refined depth-map that is estimated from deep random forest to enhance depth-map accuracy. In particular, the second stage, cascade RF regressor, is to estimate the depth of the patch finally by applying the depth feature vector output from the first stage to cascade N layer RF sequentially. By using deep RF for estimating load fine depth, this approach required much less hyper-parameters than deep neural networks, and its model complexity could be automatically determined in a data-dependent way. From the performance evaluation with a few state-of-the-arts algorithms, the proposed method showed a higher uniform performance in terms of Accuracy and Error.

For the future work, we concentrating on solving computational overload related to global-coarse network and local-patch based RF by reducing the layer structure and designing light version of FR for implementing in hardware resource.

Acknowledgement

Following are results of a study on the "Leaders in INdustry-university Cooperation +" Project, supported by the Ministry of Education and National Research Foundation of Korea and was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2016-0-00564, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding.

References

[1] A. Dipanda, S. Woo, Towards a real-time 3D shape reconstruction using a structured light system, Pattern Recognition, Vol. 38, no.10, pp. Oct. 2005.

[2] Yan Cui,Sebastian Schuon, Derek Chan,Sebastian Thrun,Christian Theobalt, 3D shape scanning with a time-of-flight camera, CVPR, pp. 1173-1180, 2010.

[3] Peter Henry,Michael Krainin, Evan Herbst,Xiaofeng Ren,Dieter Fox , RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments, Experimental Robotics, pp 477-491, 2014

[4] Q. Yang, K-H. Tan, B. Culbertson and J. Apostolopoulis, Fusion of Active and Passive Sensors for Fast 3D Capture, Proceedings of 2010 IEEE International Workshop on Multimedia Signal Processing, Oct. 2010.

[5] D. Eigen, C. Puhrsch and R. Fergus, Depth Map Prediction from a Single Image using a Multi-Scale Deep Network, Proceedings of Advances in Neural Information Processing Systems 27, 2014

[6] M. Jeong, B.C. Ko, J. Y. Nam, "Early detection of sudden pedestrian crossing for safe driving during summer nights," IEEE Trans. Cir. Sys. Vid. Tech., vol. 27, no. 6, pp.1368-1380, 2017

[7] Yuanzhouhan Cao, Zifeng Wu, Chunhua Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," IEEE Transactions on Circuits and Systems for Video Technology (Early Access),DOI: 10.1109/TCSVT.2017.2740321, pp.1-9, 15 August 2017.

[8] ByoungChul Ko, JuneHyeok Hong, Jae-Yeal Nam, "Human action recognition in still images using action poselets and a two-layer classification model", Journal of Visual Language and Computing, Volume 28, Pages 163–175, June 2015.

[9] Zhi-Hua Zhou, Ji Feng,"Deep forest: towards an alternative to deep neural networks," Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, pp. 3553-3559, 2017

[10] P. Kr¨ahenb¨uhl and Vladlen Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in Proc. Adv. Neural Inf. Process. syst., pp.1-9, 2011.

[11] Mira Jeong, Soo Young Kwak, ByoungChul Ko, Jae-Yeal Nam, "Driver Facial Landmark Detection in Real Driving Situation," IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Circuits and Systems for Video Technology, 10.1109/TCSVT.2017.2769096, vol. 99, pp.1-15, 2017

[12] F. Liu, C. Shen and G. Lin, Deep Convolutional Neural Fields for Depth Estimation from a Single Image, Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2015), pp. 5162-5270, June. 2015.

[13] A. Roy and S. Todorovic, Monocular Depth Estimation Using Neural Regression Forest, Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2016), pp. 5506 – 5514, June 2016.

[14] A. Chakrabarti, J. Shao, and G. Shakhnarovich. Depth from a single image by harmonizing overcomplete local network predictions. In NIPS, pages 2658–2666, Dec. 2016.

[15] Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim and Chang-Su Kim,"Single-Image Depth Estimation Based on Fourier Domain Analysis", pp. 330-338, June.CVPR 2018

[16] David Eigen, Rob Fergus, Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture, Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2015), pp.2650-2658, 2015

Author Biography

Sang Jun Kim received his B.S. degrees in Computer Engineering from Keimyung University, Daegu, Korea, in Feb. 2017. He is currently a M.S. student in the Department of Computer Engineering, Keimyung University, Daegu, Korea. He received the best paper award in 2017 from Korea Computer Congress (KCC2017). His current research interests include advanced driver assistance systems using computer vision and deep learning (tmsor5@naver.com).

SangWon Kim received his B.S. degrees in Computer Engineering from Keimyung University, Daegu, Korea, in Aug. 2018. He is currently a M.S. student in the Department of Computer Engineering, Keimyung University, Daegu, Korea. His current research interests include advanced driver assistance systems using computer vision and deep learning (eddiesangwonkim@gmail.com).

Deokwoo Lee received the B.S. degree in electrical engineering from Kyungpook National University, Daegu, Korea, in 2007, and he received M.S and Ph.D. degree in electrical engineering from North Carolina State University, Raleigh, NC, USA in 2008 and 2012, respectively. His research interests are in the areas of image and signal processing, computer vision and pattern recognition. He is currently an assistant professor at Keimyung University, Daegu, Republic of Korea (dwoolee@kmu.ac.kr).

Byoung Chul Ko (Corresponding author) received his B.S. degree from Kyonggi University, Suwon, Korea, in 1998 and M.S. and Ph.D. degrees in Computer Science from Yonsei University, Seoul, Korea, in 2000 and 2004, respectively. He is currently a professor in the Department of Computer Engineering, Keimyung University, Daegu, Korea. His current research interests include content-based image retrieval, vision-based fire detection, advance driver assistance systems, and facial emotional recognition (niceko@kmu.ac.kr)

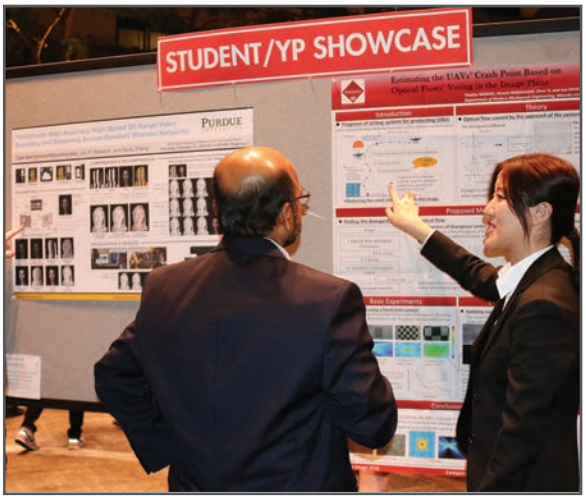
JOIN US AT THE NEXT EI!

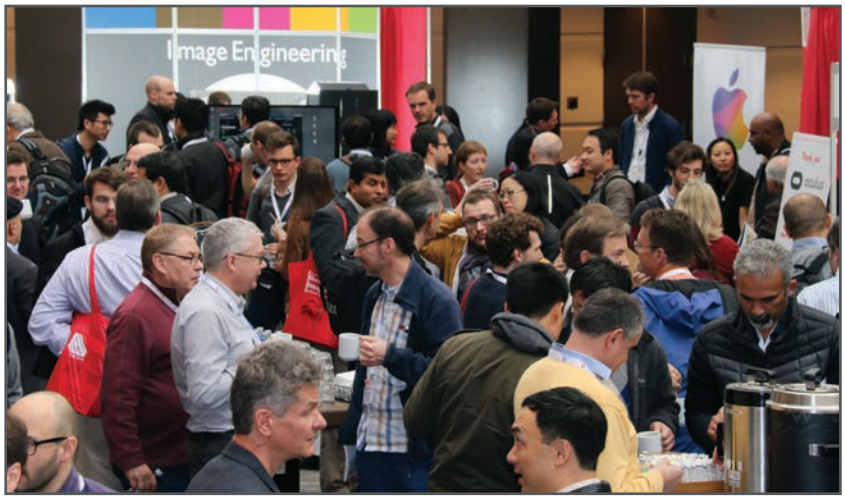
IS&T International Symposium on

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