https://doi.org/10.2352/ISSN.2470-1173.2021.6.IRIACV-302

ATTENTION-BASED LSTM NETWORK FOR ACTION RECOGNITION IN SPORTS

Mohib Ullah[1], *Muhammad Mudassar Yamin*[1], *Ahmed Mohammed*[1], *Sultan Daud Khan*[2], *Habib Ullah*[3] , *Faouzi Alaya Cheikh*[1]

[1] Norwegian University of Science and Technology, Norway.
[2] National University of Technology, Islamabad, Pakistan.
[3] University of Ha'il, Saudi Arabia.

ABSTRACT

Understanding human action from the visual data is an important computer vision application for video surveillance, sports player performance analysis, and many IoT applications. The traditional approaches for action recognition used hand-crafted visual and temporal features for classifying specific actions. In this paper, we followed the standard deep learning framework for action recognition but introduced channel and spatial attention module sequentially in the network. In a nutshell, our network consists of four main components. First, the input frames are given to a pre-trained CNN for extracting the visual features and the visual features are passed through the attention module. The transformed features maps are given to the bi-directional LSTM network that exploits the temporal dependency among the frames for the underlying action in the scene. The output of bi-direction LSTM is given to a fully connected layer with a softmax classifier that assigns the probabilities to the actions of the subject in the scene. In addition to cross-entropy loss, the marginal loss function is used that penalizes the network for the inter action classes and complimenting the network for the intra action variations. The network is trained and validated on a tennis dataset and in total six tennis players' actions are focused. The network is evaluated on standard performance metrics (precision, recall) promising results are achieved.

***Index Terms*—** Channel attention, Spatial attention, Bidirectional LSTM, Marginal loss.

1. INTRODUCTION

With the phenomenal development in computing technology, smartphones and other portable electronic devices are ubiquitous. This resulted in the generation of an astounding amount of visual data. According to a recent survey [1], 300 hours of videos are uploaded to youtube every single day. With more than 30 million visitors a day, around 5 billion videos are watched around the globe on youtube alone. Based on these statistics, automatic video labeling with the corresponding human actions brings in several online and offline applications. In the online setting, it helps in video retrieval for specific sports action like goals in soccer matches, kick service in tennis, etc. [2, 3], dance moves in the figure skating [4, 5], surveillance of public places [6, 7], and providing assistance to the elderly in smart homes [8, 9]. In a nutshell, human action recognition enables computers to infer human actions in a given video. In the last few decades, substantial progress has been made in the low-level vision tasks like image classification [10, 11], object detection [12, 13], segmentation [14, 15], tracking [16, 17], anomaly detection [18, 19], etc. However, high-level tasks like group behavior inference [20–22], cybersecurity [23, 24], individual action recognition [25–27], and pose estimation [28, 29] are still deemed as unsolved problems and there is room for improvement. In this context, researchers tackled the action recognition tasks in different ways. For example, Schindler et al. [30] researched the number of frames for the visual recognition of human actions. It is argued that short sequences of only 1-10 frames are enough to predict the underlying action with 90% accuracy. They used intuitive geometrical shape features for the optimal action prediction. Xia et al. [31] approximated the 3D skeleton joints locations from the data obtained through the Microsoft Kinect sensor in the spherical coordinate system. The 3D skeleton joints are used as the compact representation of the postures and the clustering mechanism is used to approximate the posture of humans in the pre-defined K postures. Yeffet et al. [32] used the local binary pattern for the appearance model and a linear classification for the prediction. The entire video is segmented into k slice and an accumulated histogram is computed. The histogram is used as the descriptor and the algorithm is trained and validated in similar settings. Other than visual data, different sensory data is also exploited for action recognition. For example, Ullah et al. [33] proposed a stacked LSTM network for action recognition and used 1D data obtained from the accelerometer and gyroscope of a smartphone. Similarly, Mimouna et al. [34] applied an entropy-based signal selection mechanism on the triaxial accelerometer data and trained a support vector machine for recognizing different human actions.

With the availability of large scale data, deep learning-based approaches have sustained improvement in almost all

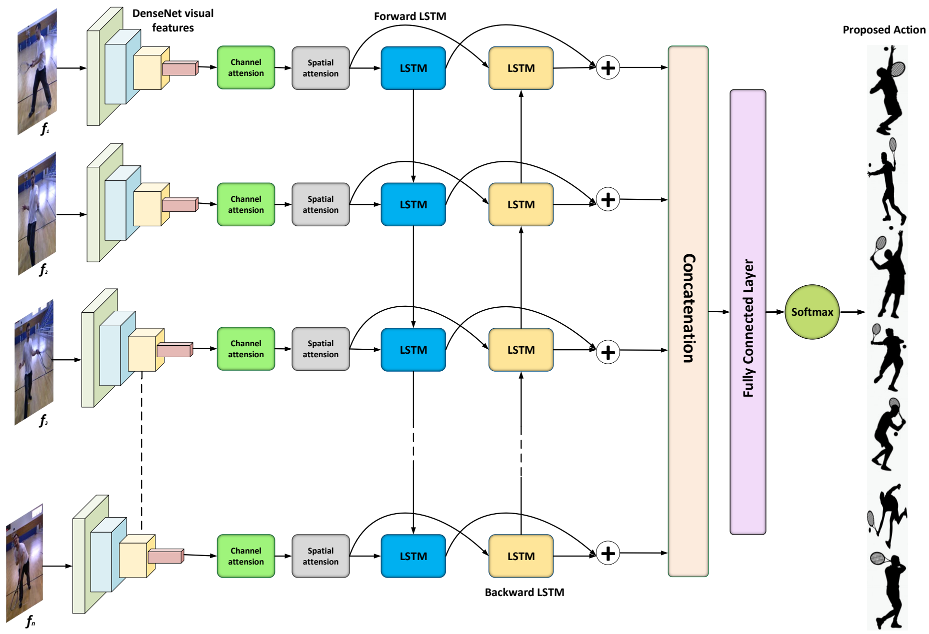


Fig. 1: The input frames are given to Densenet that output feature maps. The feature maps are passed through the channel and spatial attention module for refinement. The refine feature maps are inserted to the bidirectional LSTM network and consequently, a fully connected layer with softmax classifier output the action probabilities of the sports player.

the vision problems. In the context of deep learning, Han et al. [35] used the classical two-stream convolutional neural network architecture for action recognition but exploited transfer learning for training the network on a smaller amount of data. For the backbone architecture, a pre-trained Resnet is used. Ma et al. [36] exploited the human pose estimation model to find the position of human joints in the first stage and then used that information for generating a descriptor for the action classification. Li et al. [37] proposed a deep learning framework where the visual features are extracted through a pre-trained convolutional neural network (CNN) and used as the video descriptor. A part selection mechanism is applied to the extracted features that give useful video cubes to the bi-directional LSTM for final actional predication. Our work is based on the standard deep learning-based action recognition framework where we extract the visual features through a CNN and bi-directional LSTM to incorporate the temporal information for the action classification. However, we introduced spatial and temporal attention mechanism in the end-to-end network for emphasizing the most useful information in the video segment. In a nutshell, the key contributions of the proposed framework are the following:

- The design of an end-to-end deep network for sports player action recognition.
- The introduction of the channel and spatial attention module for emphasizing the useful features for action recognition.
- Exploiting marginal loss in addition to cross-entry for better separability between action classes.

The rest of the paper is organized in the following order. In section 2, the proposed method is briefly explained. The visual features are elaborated in section 3. The channel and spatial attention mechanisms are discussed in section 4. The overall cost function and hyperparameters are discussed in section **??**. A brief description of the dataset and the experiments are discussed in section 6 and section 7 concludes the paper with final remarks and the future directions.

2. PROPOSED METHOD

The block diagram of the proposed method is given in Figure 1. In a nutshell, the method is based on the classical deep architecture of the action recognition framework. The novel attributes of the method come from the use of the channel and spatial attention. The channel attention essentially helps the network in finding the most meaningful information in the input frame. Similarly, spatial attention enables the network to localize the most important information. Such a mechanism not only improves the performance of the network but also help the network in processing the frames more efficiently. Once the feature maps generated by Densenet is refined by the attention module, a bidirectional LSTM followed by the fully connected layer is used to classify the actions of the sport played in the input video clip. In the following section, each module of the method is explained.

3. DEEP FEATURES

To extract the visual features from the input frames, we used a pre-trained Densenet [38] and fine-tuned its parameter with the Tennish dataset. DenseNet is a special type of a CNN wherein the hierarchical structure, each layer gets inputs from all the preceding layers and in a similar fashion, forward it's own feature maps to all the subsequent layers. At each layer, feature map concatenation is used. One of the novel aspects of densenet is the layers concatenation at each layer which helps the network to be thinner and compact with comparable or even better performance. In our experiment, we used the densenet121. Technically, we are only interested to extract the deep features from each input frame. Therefore, the fully connected layer of the network is truncated and only the feature extraction module is kept for obtaining the visual description of each input frame. As the deeper layer of the network learns only the class-specific features Fig. 2, we consider features maps $F \in \mathbb{R}^{C \times H \times W}$ from the mid-layers of the network. In the succeeding step of the network, the extracted feature maps are processed by the channel and spatial module as given in section 4.

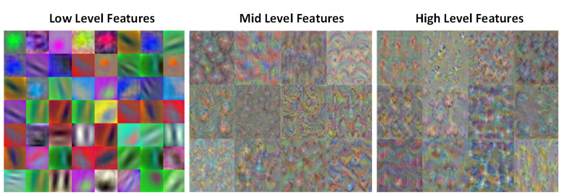


Fig. 2: Visualization of different layers of the CNN

4. ATTENTION MODULE

Different strategies have been explored for improving the performance of deep neural networks ranging from increasing the depth [39, 40], and width [41, 42] of the network to improving the cardinality of the network [43, 44]. Currently, the researcher has focused on incorporating attention mechanisms in the networks. In a nutshell, the attention mechanism is inspired by the human visual system. In this work, we followed a similar approach and used Convolutional Block Attention Module (CBAM) [45] for fusing the cross-channel and spatial information in a given frame. The details of CBAM is beyond the scope of this paper but can be read in [45].

5. BI-DIRECTIONAL LSTM

Recurrent neural networks generalize the feedforward neural network with a forward and a feedback connection. Long Short Term Memory (LSTM) is one of the implementations of the recurrent network and Bidirectional LSTM is an extended form of LSTM that help improve the model capacity and performance through a forward and backward propagation of information. Bidirectional LSTM is used for problems where all the information (future and past) about the video is available. Technically, the bidirectional LSTM network train two LSTMs on the input sequence of frames. The first from time t_1 to t_n while t_1 is the first frame and t_n is the last frame of the video clip. Similarly, the second LSTM takes the last frame as the input i.e. starting from t_n until t_1. Hence, the second LSTM gets the reverse copy of the same input video clip. Such a setting provides complete contextual information to the network and results in better performance than using only one LSTM. The output generated by the bi-direction is given to a fully connected layer that is followed by a six-way softmax classifier that assigns the probability score to each of the tennish action class.

6. EXPERIMENT

To evaluate the proposed method, we used a publicly available Tennis dataset [46]. The dataset consists of different imaging modalities like RGB, Depth, silhouette, 2D and, 3D skeleton video and keypoints of the skeleton joints. In our works, we used only the RGB data. In total, the dataset contains videos from 12 different tennis actions (Backhand, Backhand, Backhand volley, Backhand to hands, Flat service, Forehand flat, Forehand open stands, Forehand slice, forehand volley, kick service, slice service, smash). By large, there is considerable variation in the appearance of the player and the background. In our analysis, we used only 6 actions for training and testing. The videos are collected from 31 amateurs and 24 experienced players. For consistency, each action is performed several times which resulted in 8734 videos. Roughly, around 4 hours of videos are used for training and testing of the proposed network. The network is evaluated on standard performance metrics like the precision and recall and results are reported in table 1.

7. CONCLUSIONS

A deep learning framework for action recognition is introduced that exploits the channel and spatial attention module sequentially in the end-to-end network. The network consists of four main components. First, the input frames are given to a pre-trained CNN for extracting the visual features and the extracted features are refined through the attention mechanism. The refined feature maps are processed by the bi-directional LSTM network that exploits the temporal dependency among the frames for the underlying action in the scene. The bi-direction LSTM is succeeded by a fully connected layer with a softmax classifier that assigns the probabilities to the actions of the subject in the scene. In addition to cross-entropy loss, a marginal loss function is exploited that penalizes the network for the interaction class and complimenting the network for the intra action variations. The net-

Predicted Action

Actual Action		Forehand Volley	Backhand	Backhand Slice	Slice Service	Smash	Flat Service	Recall
	Forehand Volley	40	3	0	3	6	5	70.17%
	Backhand	3	38	4	0	0	0	88.37%
	Backhand Slice	5	3	45	0	0	0	84.90%
	Slice Service	0	0	2	44	2	1	89.79%
	Smash	0	1	0	2	37	0	92.50%
	Flat Service	1	0	0	0	0	42	97.67%
	Precision	81.63%	84.44%	88.23%	89.79%	82.22%	87.5%	

Table 1: Confusion Matrix of the test data. The off-diagonal elements correspond to the True positive. Other values in the column corresponds to the false positive while value along the row corresponds to the false negative.

work is trained and validated on the tennis dataset and in total six tennis actions are focused. The network is evaluated on standard performance metrics (precision, recall). The quantitative results show promising results on the validation test. In the future, we are aiming to incorporate temporal attention mechanisms and also exploit motion information through the dense optimal flow. Additionally, instead of using single LSTM cells, stacked LSTM will be explored for the forward and backward propagation.

8. REFERENCES

[1] Danny, "37 mind blowing youtube facts, figures and statistics," Accessed 12 Mar, 2020. 2019.

[2] Brais Martinez, Davide Modolo, Yuanjun Xiong, and Joseph Tighe, "Action recognition with spatial-temporal discriminative filter banks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5482–5491.

[3] Md Nafee Al Islam, Tanzil Bin Hassan, and Siamul Karim Khan, "A cnn-based approach to classify cricket bowlers based on their bowling actions," *arXiv preprint arXiv:1909.01228*, 2019.

[4] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue, "Learning to score figure skating sport videos," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[5] Yongjun Li, Xiujuan Chai, and Xilin Chen, "End-to-end learning for action quality assessment," in *Pacific Rim Conference on Multimedia*. Springer, 2018, pp. 125–134.

[6] Habib Ullah, *Crowd Motion Analysis: Segmentation, Anomaly Detection, and Behavior Classification*, Ph.D. thesis, University of Trento, 2015.

[7] Lin Wu, Yang Wang, Ling Shao, and Meng Wang, "3-d personvlad: Learning deep global representations for video-based person reidentification," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3347–3359, 2019.

[8] Matteo Gabrielli, Pietro Leo, Fabrizio Renzi, and Sonia Bergamaschi, "Action recognition to estimate activities of daily living (adl) of elderly people," in *2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT)*. IEEE, 2019, pp. 261–264.

[9] Daniele Liciotti, Michele Bernardini, Luca Romeo, and Emanuele Frontoni, "A sequential deep learning application for recognising human activities in smart homes," *Neurocomputing*, 2019.

[10] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[11] Lars Schmarje, Monty Santarossa, Simon-Martin Schröder, and Reinhard Koch, "A survey on semi-, self-and unsupervised techniques in image classification," *arXiv preprint arXiv:2002.08721*, 2020.

[12] Sultan Daud Khan and et al., "Person head detection based deep model for people counting in sports videos," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[13] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[14] Yuhui Yuan, Xilin Chen, and Jingdong Wang, "Object-contextual representations for semantic segmentation," *arXiv preprint arXiv:1909.11065*, 2019.

[15] Fahad Lateef and Yassine Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, 2019.

[16] Mohib Ullah and Faouzi Alaya Cheikh, "A directed sparse graphical model for multi-target tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1816–1823.

[17] Gioele Ciaparrone and et al., "Deep learning in video multi-object tracking: A survey," *Neurocomputing*, vol. 381, pp. 61–88, 2020.

[18] Habib Ullah and Nicola Conci, "Crowd motion segmentation and anomaly detection via multi-label optimization," in *ICPR workshop on pattern recognition and crowd analysis*, 2012, vol. 75.

[19] Raghavendra Chalapathy and Sanjay Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[20] Habib Ullah and et al., "Multi-feature-based crowd video modeling for visual event detection," *Multimedia Systems*, pp. 1–9, 2020.

[21] Mark Marsden, Kevin McGuinness, Suzanne Little, and Noel E O'Connor, "Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–7.

[22] Mohib Ullah, Habib Ullah, Nicola Conci, and Francesco GB De Natale, "Crowd behavior identification," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1195–1199.

[23] Muhammad Mudassar Yamin, Mohib Ullah, Habib Ullah, and Basel Katt, "Weaponized ai for cyber attacks," *Journal of Information Security and Applications*, vol. 57, pp. 102722, 2021.

[24] Muhammad Mudassar Yamin, Basel Katt, Kashif Sattar, and Maaz Bin Ahmad, "Implementation of insider threat detection system using honeypot based sensors and threat analytics," in *Future of Information and Communication Conference*. Springer, 2019, pp. 801–829.

[25] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao, "Learning graph convolutional network for skeleton-based human action recognition by neural searching," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 2669–2676.

[26] Mohib Ullah, Habib Ullah, and Ibrahim M Alseadonn, "Human action recognition in videos using stable features," 2017.

[27] Jian Liu, Naveed Akhtar, and Ajmal Mian, "Adversarial attack on skeleton-based human action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[28] Bowen Cheng and et al., "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5386–5395.

[29] Akif Quddus Khan, Salman Khan, Mohib Ullah, and Faouzi Alaya Cheikh, "A bottom-up approach for pig skeleton extraction using rgb data," in *International Conference on Image and Signal Processing*. Springer, 2020, pp. 54–61.

[30] Konrad Schindler and Luc Van Gool, "Action snippets: How many frames does human action recognition require?," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[31] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 20–27.

[32] Lahav Yeffet and Lior Wolf, "Local trinary patterns for human action recognition," in *IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 492–497.

[33] Mohib Ullah, Habib Ullah, Sultan Daud Khan, and Faouzi Alaya Cheikh, "Stacked lstm network for human activity recognition using smartphone data," in *2019 8th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2019, pp. 175–180.

[34] Amira Mimouna, Anouar Ben Khalifa, and Najoua Essoukri Ben Amara, "Human action recognition using triaxial accelerometer data: selective approach," in *International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE, 2018, pp. 491–496.

[35] Yamin Han, Peng Zhang, Tao Zhuo, Wei Huang, and Yanning Zhang, "Going deeper with two-stream convnets for action recognition in video surveillance," *Pattern Recognition Letters*, vol. 107, pp. 83–90, 2018.

[36] Miao Ma and et al., "Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos," *Pattern Recognition*, vol. 76, pp. 506–521, 2018.

[37] Wenhui Li, Weizhi Nie, and Yuting Su, "Human action recognition based on selected spatio-temporal features via bidirectional lstm," *IEEE Access*, vol. 6, pp. 44211–44220, 2018.

[38] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[39] Christian Szegedy and et al., "Going deeper with convolutions," in *IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[40] Dongyoon Han, Jiwhan Kim, and Junmo Kim, "Deep pyramidal residual networks," in *IEEE conference on computer vision and pattern recognition*, 2017, pp. 5927–5935.

[41] Sergey Zagoruyko and Nikos Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[42] Christian Szegedy and et al., "Rethinking the inception architecture for computer vision," in *IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[43] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[44] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[46] Sofia Gourgari, Georgios Goudelis, Konstantinos Karpouzis, and Stefanos Kollias, "Thetis: Three dimensional tennis shots a human action dataset," June 2013.

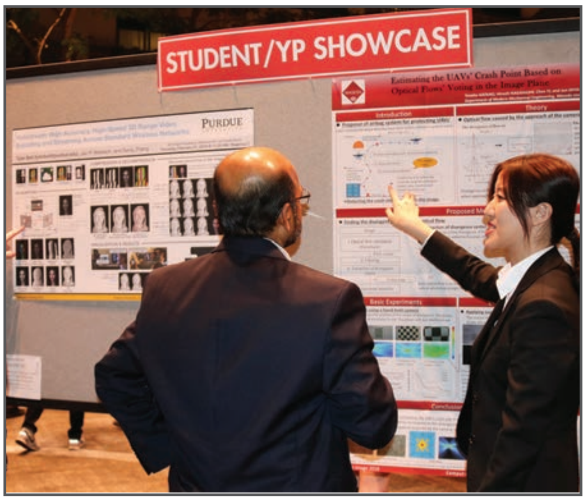
JOIN US AT THE NEXT EI!

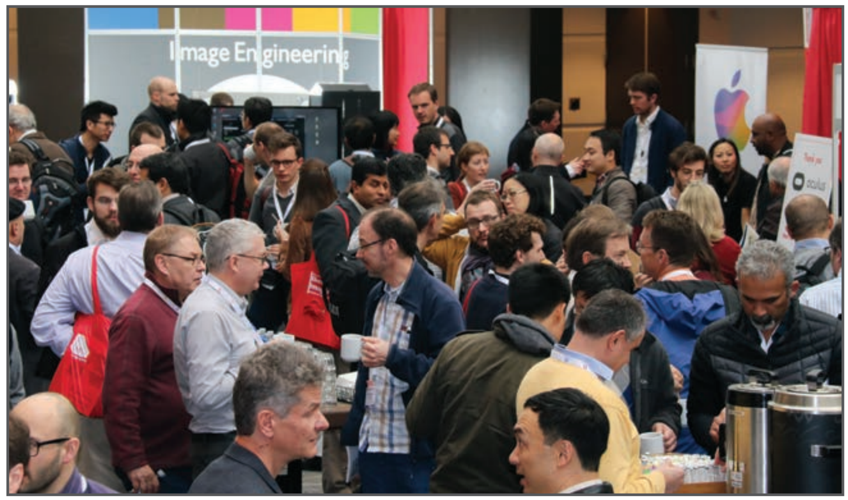
IS&T International Symposium on

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