https://doi.org/10.2352/ISSN.2470-1173.2019.7.IRIACV-454

Foreground-Aware Statistical Models for Background Estimation*[1]

Edgar A. Bernal; University of Rochester; Rochester, NY 14627
Qun Li; Microsoft Corporation; Bellevue, WA 98004

Abstract

Video-based detection of moving and foreground objects is a key computer vision task. Temporal differencing of video frames is often used to detect objects in motion, but fails to detect slow-moving (relative to the video frame rate) or stationary objects. Adaptive background estimation is an alternative to temporal frame differencing that relies on building and maintaining statistical models describing background pixel behavior; however, it requires careful tuning of a learning rate parameter that controls the rate at which the model is updated. We propose an algorithm for statistical background modeling that selectively updates the model based on the previously detected foreground. We demonstrate empirically that the proposed approach is less sensitive to the choice of learning rate, thus enabling support for an extended range of object motion speeds, and at the same time being able to quickly adapt to fast changes in the appearance of the scene.

Introduction

Foreground detection refers to a set of techniques that aim at distinguishing foreground objects from background or stationary areas in video streams. Motion detection techniques exploit the assumption that objects of interest are often in motion and rely on finding largely dissimilar regions across temporally adjacent frames. Consequently, such techniques may fail to detect slow-moving or stationary objects. Background estimation and subtraction techniques construct statistical models describing the pixel behavior of the stationary background, and perform detection by finding differences between the current video frame and the constructed background model. According to this approach, a historical statistical model describing the behavior of each pixel is constructed. Once constructed, the background model may be updated (*e.g.*, in adaptive techniques) or left unchanged (*e.g.*, in non-adaptive techniques). Non-adaptive methods have been largely abandoned due to the fact that the accuracy of the static background model usually decreases over time. Adaptive methods update the background model continuously with each incoming frame at a rate controlled by a predetermined learning rate factor. Foreground detection is performed by determining a measure of fit of each pixel value in the incoming frame relative to its constructed statistical model: pixels that do not fit their corresponding background model are considered foreground pixels. Adaptive background estimation models are thus able to detect slow-moving and stationary objects. However, selecting the learning rate involves a tradeoff between how fast the model is updated and the range of motion speeds that can be supported by the model: too small a learning rate results in background estimates which do not adapt quickly enough to fast changes in the appearance of the scene; conversely, too large a learning rate causes objects that stay stationary for extended periods (relative to the frame and learning rates) to be absorbed into the background estimate.

Related Work

Applications that perform analytics from video captured using stationary cameras are amenable to foreground and motion detection algorithms. The two most commonly used methods for motion detection are frame differencing (FD) [1, 2] and background estimation and subtraction [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. FD methods compromise detection performance in favor of computational efficiency, given that the rate at which the frames are processed determines the range of object speeds that can be reliably supported. On the other hand, background estimation and subtraction methods tend to be more agnostic to the time scale and parameter values.

Over the past few decades, various background estimation and subtraction algorithms have been proposed, each with its own pros and cons. These methods construct a background appearance model from either a single frame or a temporal sequence of frames. Popular techniques used to obtain and maintain the background model for adaptive background subtraction methods include: 1) temporal median filtering [3], where a sequence of images are median-filtered to obtain a background model; 2) running Gaussian averaging [4], where the past behavior of each pixel is modeled with a Gaussian distribution; 3) modeling via GMMs, which describe the probability of observing a certain pixel value at a given instant in time by means of a mixture of Gaussians with a fixed [5] or an adaptive number of components [6]; 4) kernel density estimation [7, 14], where the background distribution is given as a sum of Gaussian kernels. Other methods have been proposed to model background, including sequential kernel density approximation [15], co-occurrence of image variations [16], Eigen backgrounds [17], MinMax [18], etc. More recently, Local Binary Pattern (LBP)[19]-like features have been explored for background subtraction [10, 11, 12, 13]. Once the background model is available, foreground detection can be performed by determining a measure of fit of each pixel value in the incoming frame relative to its constructed model. For an in-depth discussion of various background modeling methods, we refer the reader to [8, 9].

One of the main limitations of background estimation approaches is that slow-moving or quasi-stationary foreground objects may get absorbed into the background model. Inspired by the context-driven model that incorporates motion information obtained from FD [20], a context-aware background subtraction method [21] was proposed wherein motion information (from FD) is used to determine the confidence level of a pixel belonging to the foreground, and only low-confidence pixels are used to update the background model. This method works effectively only on the edge regions of the foreground object due to the detection holes that are inherent to FD techniques. A split GMM model was proposed where two GMMs are used to model background and foreground separately [22]. A similar scheme was also adopted in [23]. In [24], an additional moving object classification step is conducted so that the background model is only updated after the foreground object moves away. In [25], a statistical approach that fuses temporal and spatial information was proposed, where temporal occurrence analysis of foreground/background data is performed.

Contributions

The main contribution of this paper is an algorithm for adaptive background estimation that is robust to the value of the learning rate. This robustness is achieved by closing the loop on the background estimation and foreground detection pipeline. Traditional adaptive background estimation algorithms are open-loop since the outcome of decisions based on the model are not used to make decisions about the model itself. We propose a closed-loop alternative that feeds back the foreground detection mask and updates the models corresponding to different pixels at different rates based on the detected foreground. The proposed algorithm is robust to the choice of learning rate by slowing down the model updating process at locations where foreground objects are detected, and speeding it up at other locations. As a secondary contribution, empirical validation of the improved robustness of the proposed algorithm is presented.

Traditional Statistical Background Modeling

Let F_t denote the t-th video frame represented as an array of pixel values (single or multiple color channels), where t represents a temporal index. Let BG_t denote the t-th background model represented as an array of pixel-wise statistical models. Statistical background model BG_{t+1} is estimated by updating the current background model BG_t based on current and previous video frames F_1 through F_t. Foreground mask FG_t is a binary mask indicating the location of detected foreground (ON pixels) and background (OFF pixels) areas. FG_t is estimated by performing a pixel-wise fit test between the values in F_t and the statistical models in BG_t; specifically, pixels that do not fit their corresponding background model are considered foreground pixels, and vice-versa. Fig. 1(a) illustrates the traditional process for adaptive background estimation and foreground detection.

Foreground-Aware Statistical Background Modeling

In this section we describe the proposed method for background modeling and foreground estimation. A high-level system schematic that illustrates inputs and output of each module is shown in Fig. 1(b).

Pixel Modeling

The temporal sequence of pixel values in a video stream can be interpreted as the instantiations of a random variable with a given distribution. Background estimation is achieved by estimating the parameters of the distributions that accurately describe the historical behavior of pixel values for every pixel in the scene. Specifically, at frame T, what is known about a particular pixel located at coordinates (i, j) is the history of its values $X_1, X_2, \dots, X_T = F(i, j, t), 1 \leq t \leq T$. F_t is the collection of pixel values $F(i, j, t)$ for all $1 \leq i \leq N_r$ and $1 \leq j \leq N_c$, where N_r and N_c are the number of rows and columns of the frames in the incoming video sequence.

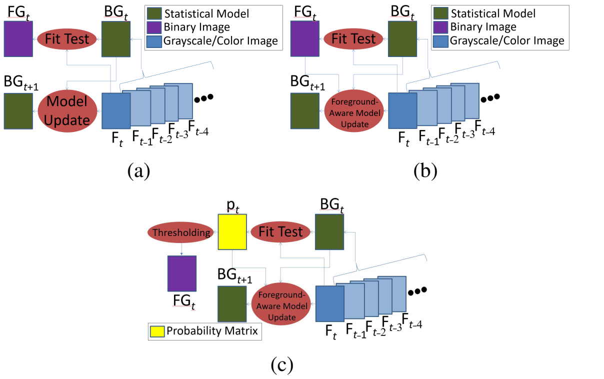


Figure 1. Flowchart of three background estimation/updating and foreground detection methods: (a) traditional, (b) proposed, (c) a more general algorithm.

While the historical behavior of the values of a pixel can be described with different statistical models including parametric models that assume an underlying distribution and estimate the relevant parameters of the distribution, and non-parametric models such as kernel-based density estimation approaches, we discuss and implement the proposed algorithm with Gaussian Mixture Models (GMM), and note that our approach is equally applicable to other online modeling techniques. We model the recent history of behavior of values of each pixel as a mixture of K Gaussian distributions, so that the probability of observing the current value is

$$P(X_t) = \sum_{k=1}^{K} w_{kt} \Phi(X_t, \mu_{kt}, \Sigma_{kt}) \tag{1}$$

where w_{kt} is an estimate of the weight of the k-th Gaussian component in the mixture at time t, μ_{kt} is the mean value of the k-th Gaussian component in the mixture at time t, Σ_{kt} is the covariance matrix of the k-th Gaussian component in the mixture at time t, and $\Phi(\cdot)$ is the Gaussian probability density function. Once initialized, the model is updated according to the strategies described below.

Foreground Pixel Detection via Goodness-of-Fit Testing

Foreground detection is performed by determining a measure of fit of each pixel value in the incoming frame relative to its constructed statistical model. At time t, fit testing is performed by reading incoming frame F_t and the current background estimate BG_t and, for each pixel in the incoming frame, determining whether it belongs to the foreground or to the background according to its value and to its corresponding mixture model. The output of this stage is a binary mask FG_t with the same pixel dimensions as the incoming frame, with ON (OFF) values at foreground (background) pixel locations.

Foreground-Aware Background Model Updating

The foreground-aware background model update stage stores the current background model BG_t and updates it according to the foreground mask FG_t output by the foreground pixel detection stage, and the incoming frame F_t. The result is an updated background model BG_{t+1} to be stored and used in the processing of the new incoming frame F_{t+1}. One of the main limitations of traditional model-based background estimation algorithms is that the learning parameter α has to be carefully selected for the expected range of object velocity in the scene relative to the frame rate. In the proposed framework, the weights of the distribution are adjusted according to

$$w_{l(t+1)} = fg_t w_{lt} + (1 - fg_t)[(1-\alpha)w_{lt} + \alpha M_{(l,k)t}] \qquad (2)$$

where α is the learning or update rate and $M_{(l,k)t}$ is an indicator variable equaling 0 for $l \neq k$ and 1 for $l = k$, so that only the weight factor for the matching component in the mixture is updated; lastly, fg_t is the binary value of the foreground mask FG_t at the pixel whose model is being updated ($fg_t = 1$ and $fg_t = 0$ for foreground and background pixels, respectively.) Similarly, mean and covariance estimates for the matching components in the model distributions are updated according to

$$\mu_{t+1} = fg_t \mu_t + (1 - fg_t)[(1-\rho)\mu_t + \rho X_t], \qquad (3)$$

$$\sigma_{t+1}^2 = fg_t \sigma_t^2 + (1 - fg_t)[(1-\rho)\sigma_t^2 + \rho(X_t - \mu_{t+1})^T(X_t - \mu_{t+1})] \qquad (4)$$

where X_t is the value of the incoming pixel and $\rho = \alpha\Phi(X_t|\mu_k, \sigma_k^2)$ is the learning rate for the parameters of the matching component of the distribution, k. The mean and covariance estimates for the non-matching components are left unchanged. The effect of performing the updates in the manner described is that only models for background pixels get updated at each frame, which mitigates the risk for a foreground model being absorbed into the background, thus negatively affecting the background model for that pixel.

The model update approach described by Eqs. 2-4 uses the values in the foreground mask to make a hard decision as to whether a given model is to be updated or not. However, fit tests often yield probabilities that are indicative of the confidence of a pixel belonging to its respective background distribution. In some cases, it may be desirable to perform smoother model updates based on these probabilities, as illustrated in Fig. 1(c). In the figure, probability matrix P_t indicates the probability that each pixel value in frame F_t belongs to its respective distribution; a thresholding operation results in a foreground mask equivalent to that described above.

When intermediate probabilities are available, the updating rules implemented by the foreground-aware background model update module are as follows:

$$w_{l(t+1)} = (1 - p_t)w_{lt} + p_t[(1-\alpha)w_{lt} + \alpha M_{(l,k)t}] \qquad (5)$$

$$\mu_{t+1} = (1 - p_t)\mu_t + p_t[(1-\rho)\mu_t + \rho X_t] \qquad (6)$$

$$\rho_{t+1}^2 = (1 - p_t)\rho_t^2 + p_t[(1-\rho)\sigma_t^2 + \rho(X_t - \mu_{t+1})^T(X_t - \mu_{t+1})] \qquad (7)$$

where p_t is the output of the fit test for the pixel whose model is being updated. These modified update rules reflect the estimated confidence of a pixel belonging to its respective background distribution.

Experimental Results

The described algorithm was applied to a one-hour long video acquired in a challenging transportation scenario where multiple lanes merge into a single lane; this scenario is representative, for example, of a situation that is prevalent in multilane tolls. Foreground/motion detection is first performed to detect vehicles that are subsequently tracked. In this scenario, vehicles drive to and stop at a check point and then pull forward. Consequently, the length of time vehicles stay stationary varies from short (a few seconds) to relatively long (several minutes), and vehicle speeds span a range from 0 mph to roughly 15 mph. Such stop-and-go patterns of motion poses challenges to motion detection for the reasons elucidated above. The performance of our algorithm was compared to that of the foreground detector from [6]. We tested both algorithms at four different learning rates, namely $\alpha = 1 \times 10^{-k}$, for $k = 2, 3, 4, 5$.

Fig. 2 compares the estimated background images of both methods with different learning rates by the end of the video. As can be seen, the proposed foreground-aware background modeling method manages to maintain a clean (*i.e.*, no foreground objects having been absorbed) background across the video regardless of learning rate α (between 1×10^{-2} and 1×10^{-5}) chosen (top row images). In contrast to the results obtained with the proposed method, the quality of the background model using the traditional foreground detection method is highly sensitive to the chosen learning rate α. Foreground objects are clearly present in background estimates for the three largest values for α (Figs. 2(e)-2(g)), and slight smearing is visible for the smallest value of α (Fig. 2(h)). Wrongly estimated backgrounds will lead to erroneous foreground object detection, as illustrated in Fig. 3, which depicts the detection results in the form of red contours outlining the perimeter of foreground blobs. Since the vehicle at the service station was absorbed into the background model in the traditional approach (Figs. 2(e)-2(g)), it is not detected as a foreground object (Figs. 3(e)-3(g)). This is not an issue with the proposed method, regardless of the value of the learning rate (Figs. 3(a)-3(d)).

A quantitative performance analysis was also performed. The performance of the algorithms was measured as suggested by [26]: TP, the number of true positives (*i.e.*, number of foreground blobs detected that correspond to an actual vehicle), FP, the number of false positives (*i.e.*, number of foreground blobs detected that correspond to no vehicles), and FN, the number of false negatives (*i.e.*, number of vehicles without a corresponding foreground blob) were counted on a frame-by-frame basis across five thousand frames. The results are summarized in Table 1, where P stands for precision and R for recall.

Although the performance of the method in [6] (in terms of blob-level precision and recall) seems to approach that of the proposed method when the learning rate is set to 1×10^{-5}, the numbers from Table 1 do not reflect the pixel-level magnitude of its detection inaccuracy. Fig. 4 shows the detection of foreground objects by two methods. The proposed method outperforms the

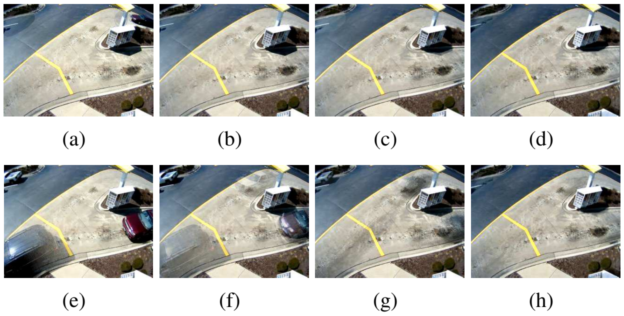

(a) (b) (c) (d)
(e) (f) (g) (h)

Figure 2. *Top: background estimates obtained via the proposed method with learning rates (a)* 1×10^{-2}*, (b)* 1×10^{-3}*, (c)* 1×10^{-4} *and (d)* 1×10^{-5}*; bottom: background estimates obtained via the method from [6] with learning rates (e)* 1×10^{-2}*, (f)* 1×10^{-3}*, (g)* 1×10^{-4} *and (h)* 1×10^{-5}*.*

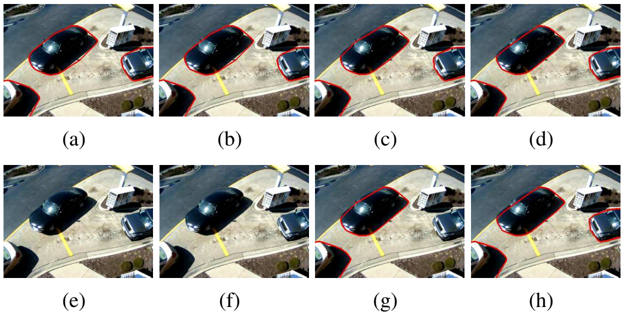

(a) (b) (c) (d)
(e) (f) (g) (h)

Figure 3. *Top: annotated video frame obtained via fit tests between displayed frame and estimated background with the proposed method and with learning rates (a)* 1×10^{-2}*, (b)* 1×10^{-3}*, (c)* 1×10^{-4} *and (d)* 1×10^{-5}*; bottom: estimated foreground with the method from [6] and with learning rates (e)* 1×10^{-2}*, (f)* 1×10^{-3}*, (g)* 1×10^{-4} *and (h)* 1×10^{-5}*.*

traditional one in terms of pixel-level precision and recall even though the blob-level performance numbers are similar.

Among the four different learning rates tested, the proposed method achieves its best performance when $\alpha = 1 \times 10^{-4}$ while the traditional method requires $\alpha = 1 \times 10^{-5}$, which could affect its capability to adapt to fast changes in the scene. The proposed method is less sensitive to the choice of learning rate α in that it provides almost equally satisfactory foreground detection performance for a wide range of learning rate values (from 1×10^{-2} to 1×10^{-5}) whereas the traditional method requires a carefully selected learning rate in order to perform well. This selection process would depend on camera geometry, frame rate and expected speed of motion of objects in the scene, and would only work robustly if the range of object motion speeds is somewhat narrow. The proposed method outperformed the traditional approach for every learning rate tested.

Comparison on foreground detection performance

Learning Rate	Method	TP	FP	FN	P	R
1×10^{-2}	Traditional	793	84	7608	90.42%	9.44%
	Proposed	8384	41	17	99.51%	99.80%
1×10^{-3}	Traditional	1839	87	6562	95.48%	21.89%
	Proposed	8385	4	16	99.95%	99.81%
1×10^{-4}	Traditional	5173	13	3228	99.75%	61.58%
	Proposed	8396	0	5	100.0%	99.94%
1×10^{-5}	Traditional	8352	16	49	99.81%	99.42%
	Proposed	8369	10	32	99.88%	99.62%

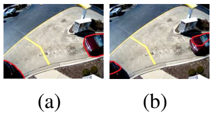

(a) (b)

Figure 4. *Sample annotated frame by (a) the proposed method, and (b) the traditional method, both with learning rate* 1×10^{-5}*.*

CONCLUSIONS

In this paper, we proposed a closed-loop approach to statistical background modeling for foreground estimation that leverages the additional information provided by foreground mask or fit test results. The proposed method robustly supports an increased range of motion speeds of objects for a given learning rate, and consequently is less sensitive to the choice of learning rate than existing methods. In particular, the proposed method greatly improves the detection performance with a relatively large learning rate value by preventing foreground object absorption, which in turn enables responsiveness of the background model to fast changes in the appearance of the scene.

References

[1] Jain, R. and Nagel, H.-H., "On the analysis of accumulative difference pictures from image sequences of real world scenes," *IEEE PAMI* **PAMI-1**, 206–214 (April 1979).

[2] Haritaoglu, I., Harwood, D., and Davis, L., "W4: real-time surveillance of people and their activities," *IEEE PAMI* **22**, 809–830 (Aug 2000).

[3] Zhou, Q. and Aggarwal, J., "Tracking and classifying moving objects from video," *IEEE Workshop on Performance Evaluation of Tracking and Surveillance* (Jan 2001).

[4] Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A., "Pfinder: real-time tracking of the human body," *IEEE PAMI* **19** (Jul 1997).

[5] Stauffer, C. and Grimson, W. E. L., "Adaptive background mixture models for real-time tracking," in [*CVPR*], **2**, 246–252 (1999).

[6] Zivkovic, Z., "Improved adaptive gaussian mixture model for background subtraction," in [*ICPR*], **2**, 28–31 (Aug 2004).

[7] A. Elgammal, D. H. and Davis, L., "Non-parametric model for background subtraction," in [*ECCV*], 751–767 (Jun 2000).

[8] Piccardi, M., "Background subtraction techniques: a review," in [*Systems, Man and Cybernetics, IEEE International Conference on*], **4**, 3099–3104 (Oct 2004).

[9] Benezeth, Y., Jodoin, P.-M., Emile, B., Laurent, H., and Rosenberger, C., "Review and evaluation of commonly-implemented background subtraction algorithms," in [*ICPR*], 1–4 (Dec 2008).

[10] St-Charles, P.-L., Bilodeau, G.-A., and Bergevin, R., "Subsense: A universal change detection method with local adaptive sensitivity," *IEEE TIP* **24**, 359–373 (Jan 2015).

[11] Yao, J. and Odobez, J., "Multi-layer background subtraction based on color and texture," in [*CVPR*], 1–8 (June 2007).

[12] Nonaka, Y., Shimada, A., Nagahara, H., and Taniguchi, R., "Evaluation report of integrated background modeling based on spatio-temporal features," in [*CVPRW*], 9–14 (June 2012).

[13] St-Charles, P.-L., Bilodeau, G.-A., and Bergevin, R., "A self-

adjusting approach to change detection based on background word consensus," in [*WACV*], 990–997 (Jan 2015).

[14] Elgammal, A., Harwood, D., and Davis, L., [*ECCV*], ch. Non-parametric Model for Background Subtraction, 751–767, Springer, Berlin, Heidelberg (2000).

[15] Han, B., Comaniciu, D., Zhu, Y., and Davis, L. S., "Sequential kernel density approximation and its application to real-time visual tracking," *IEEE PAMI* **30**(7), 1186–1197 (2008).

[16] Seki, M., Wada, T., Fujiwara, H., and Sumi, K., "Background subtraction based on cooccurrence of image variations," in [*CVPR*], **2**, II65–II72 (June 2003).

[17] Oliver, N., Rosario, B., and Pentland, A., "A bayesian computer vision system for modeling human interactions," *IEEE PAMI* **22**, 831–843 (Aug 2000).

[18] Haritaoglu, I., Harwood, D., and Davis, L., "W4: A real time system for detecting and tracking people," in [*CVPR*], 962–962 (Jun 1998).

[19] Ojala, T., Pietikainen, M., and Harwood, D., "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in [*ICPR*], **1** (Oct 1994).

[20] Desa, S. M. and Salih, Q. A., "Image subtraction for real time moving object extraction," in [*Proceedings of the International Conference on Computer Graphics, Imaging and Visualization*], 41–45, IEEE Computer Society, Washington, DC, USA (2004).

[21] Garcia, A. and Bescós, J., "Real-time video foreground extraction based on context-aware background substraction," in [*Technical Report TR-GTI-UAM-2007-02 2007*],

[22] Wang, R., Bunyak, F., Seetharaman, G., and Palaniappan, K., "Static and moving object detection using flux tensor with split gaussian models," in [*CVPRW*], 420–424 (June 2014).

[23] Chen, Y., Wang, J., and Lu, H., "Learning sharable models for robust background subtraction," in [*ICME*], 1–6 (June 2015).

[24] Phuong, L. T. and Binh, N. T., [*Human Object Classification Based on Nonsubsampled Contourlet Transform Combined with Zernike Moment*], 212–222, Springer International Publishing, Cham (2016).

[25] Boulmerka, A. and Allili, M. S., "Background modeling in videos revisited using finite mixtures of generalized gaussians and spatial information," in [*ICIP*], 3660–3664 (Sept 2015).

[26] Smith, K., Gatica-Perez, D., Odobez, J., and Ba, S., "Evaluating multi-object tracking," in [*CVPRW*], 36–36 (June 2005).

Author Biography

Edgar A. Bernal received the M.Sc. and Ph.D. degrees in Electrical Engineering from Purdue University, West Lafayette, IN, in 2002 and 2006, respectively. He is the Associate Director for the Rochester Data Science Consortium at the University of Rochester, in Rochester, NY. Prior to joining UofR, he was a Principal Scientist at the United Technologies Reseach Center in E. Hartford, CT, and a Senior Research Scientist with the Palo Alto Research Center, Webster, NY. His current research interests include computer vision, machine and deep learning, and multimodal data fusion.

Qun Li received the M.S. and Ph.D. degrees in Electrical Engineering from the University of Illinois at Chicago (UIC), IL, in 2012 and 2013, respectively. She joined Microsoft Researchm Redmond, WA, as a Data Scientist in 2016. Before that, she was with Palo Alto Research Center , a Xerox Company, Webster, NY, since 2013, as a Computer Vision Research Scientist. Her research interests include high-order data analysis, image and video analysis, computer vision, pattern recognition, compressive sensing, 3D imaging, and deep learning.

*[1] This work was performed while the authors were with PARC, A Xerox Company

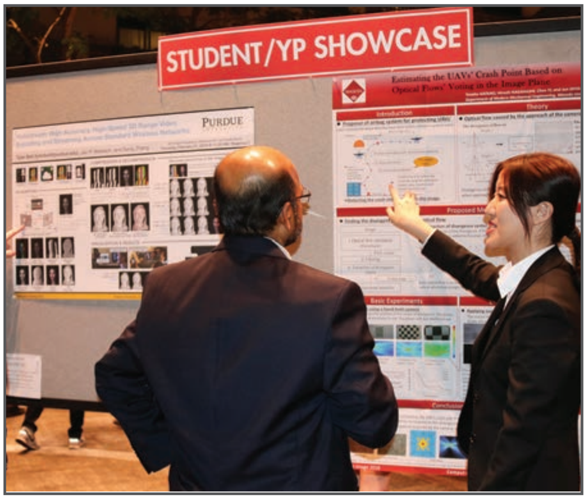
JOIN US AT THE NEXT EI!

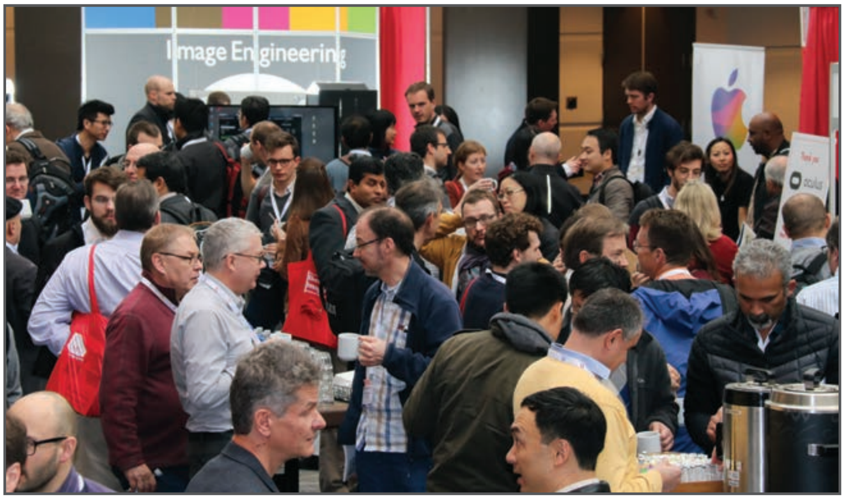
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

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





AR is a fundamental and transformative interface and platform.

20+ years from my initial working demo to today's AR displays

Optimistic that it won't take another 20 to reach ubiquitous consumer AR!

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •

• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org