https://doi.org/10.2352/ISSN.2470-1173.2019.5.MWSF-543

Statistical Sequential Analysis for Object-based Video Forgery Detection

[a]Mohammed Aloraini [1], [b]Mehdi Sharifzadeh [1], [c]Chirag Agarwal [1], [d]Dan Schonfeld [1]
[a]malora2@uic.edu, [b]mshari5@uic.edu,[c]cagarw2@uic.edu, [d]dans@uic.edu
[1] Department of Electrical and Computer Engineering, University of Illinois at Chicago, 851 S Morgan St, Chicago, IL 60607, USA

Abstract

Over the years, video surveillance systems have been used for indisputable evidence of a crime. Unfortunately, videos of the surveillance systems can be forged through adding (deleting) an object to (from) a video scene (i.e., object-based forgery) with invisible traces and little effort. In this paper, we propose a novel approach that uses spatial decomposition, temporal filtering, and sequential analysis to detect object-based video forgery and estimate a movement of removed objects. The results show that our approach not only outperforms a previous approach in detecting forged videos but it is also more robust against compressed and lower resolution videos. Also, our approach can effectively estimate a movement of different sizes of removed objects.

Introduction

For many years, surveillance videos have become essential for social security that monitors many organizations, and thus, it is important to ensure the reliability of these surveillance videos. If these recorded videos are abused, it could lead to many critical problems that are related to public security or legal evidence. That is, the fundamental challenge is to determine whether a recorded video is authentic or not especially when it is used as critical evidence for judgment [1]. Furthermore, with the advent of powerful and easy-to-use media editing tools, it enables an attacker to maliciously forge a video sequence through adding or deleting an object in a scene with invisible traces and little effort. This forged video is often eye-deceiving and appears in a way that is realistic, hence believable. That is, newspapers are sometimes tricked to use forged videos as if they are authentic. As a result, video contents should be carefully analyzed to ensure its originality and integrity, thus reducing digital crimes [2].

Creating an automatic approach for detecting forged video is a difficult problem because of the lack of truthful bases that can be used to verify the originality and integrity of video contents. Also, often forged video may not only run through inserting or removing an object in a scene but other complex processes are performed including compression, resizing, and rotation, which makes the detection more difficult [3]. Furthermore, if a forger tamper regions of a video by inpainting to remove an object (e.g., person) from the scene, it becomes more difficult to detect the tampered regions due to the high correlation between the tampered regions and the rest regions of the video. As a result, organizations that seek to validate video contents face a major challenge in proving the integrity of the video contents.

This issue leads to an increasing concern about the originality of video contents and the need to develop effective techniques to evaluate the originality, integrity, and authenticity of these video contents. Most of the existing video forensic approaches have been conducted to detect frame-based forgery that is created by inserting (deleting) frames into (from) a video scene. In [4],[5], inconsistent of optical flow vectors was used as evidence of unoriginality. The temporal correlations were used in [6], [7], [8] to detect tampered videos. In [9], [10], MPEG double compression artifacts were addressed on a macroblock-by-macroblock basis. In [11], motion compensation based approach was proposed to detect temporal interpolation in videos. In [12], Stamm et al. used Group Of Pictures (GOP) structure based approach to detect a frame deletion or addition fingerprint. Also, An anti-forensic approach was proposed in [12] to remove the frame deletion or addition fingerprint.

However, less attention has been paid to object-based video forgery that is created by adding objects to a video scene or removing objects from a video scene. Adding objects to a video scene (video splicing) is achieved through chroma key composition. A few approaches have been conducted to detect video splicing forgery. In [13], Su et al detected video splicing by examining changes of correlation patterns between the color signals on the edges of all objects. In [14], histograms of the DCT coefficients were used to classify a video as authentic or tamper. In [15], statistical correlation of blurring artifact was utilized to detect video splicing. Deep learning with autoencoders architecture was used in [16] to learn an intrinsic model of a given video.

On the other hand, removing objects from a video scene is achieved by using inpainting algorithms such as [17] and [18]. A few approaches have been conducted to detect inpainted videos. In [19], Zhang et al. used ghost shadow artifacts, which were left during inpainting process, to detect inpainted videos. In [20], physical inconsistencies were used as evidence of inauthenticity. Statistical features of object contour was utilized in [21] to detect removed objects forgery. In [22], Chen et al. extracted steganalytic features from motion residual matrices and used these features to classify between three classes, which are pristine frames, forged frames, and double compressed frames.

In this paper, we study the problem of detecting object-based video forgery. It is difficult to add moving objects without leaving invisible traces due to possibly different motions and illuminations in videos. Hence, object-based video forgery often refers to removing objects from a video as illustrated in Fig.1, where the man in the red box has been removed from the scene. We propose an approach to detect removed moving objects from a video scene that is taken from a static camera and estimate a movement of removed objects. Hence, our approach can identify if a static

Figure 1: An example of object-based video forgery: Images on the top row indicate frames from the original video; Images on the bottom row indicate frames from the tampered video where the man in the red box has been removed from the scene.

scene in a video is naturally static or forged to be static.

The contributions of our approach can be summarized as follows:

- We have addressed a new and challenging object-based video forgery problem when compared to frame-based forgery problem.
- Our approach can estimate a movement of different sizes of removed objects using spatial decomposition.
- Our approach can detect temporal changes (i.e., pixels' changes) that are nearly invisible using sequential analysis.
- Our approach not only outperforms a previous approach in detecting forged videos, but it is also more robust against compressed and lower resolution videos.

The rest of the paper is organized as follows: Section 2 presents our proposed approach. Section 3 presents our experimental results, followed by conclusion in section 4.

Proposed approach

We briefly describe our approach in the following steps, as illustrated in Fig.3. First, we apply spatial decomposition to the video frames by means of Laplacian pyramid, followed by temporal high pass filter to detect edges spatially and highlight variations temporally. Then, sequential analysis is performed temporally to detect changes in pixels. These changes are candidates of tampered pixels and it needs to be verified. The forgery is confirmed if pixels changes form large spatial regions and last for short duration. Finally, The removed object's movement is estimated by summing all verified pixels' changes of the video frames.

Spatiotemporal Filter

We apply spatiotemporal filtering stage, which is presented in Fig.4, for two reasons. (a) Although the structure inpainting, texture inpainting or combined structural and textural inpainting are usually performed to remove the motion artifacts that is left from object removal, there are still some left traces for object-based video forgery, which is always exist near the object boundary and its boundary areas. These traces can be detected using spatiotemporal filter. (b) After applying spatiotemporal filter, pixels' values at static regions are close to zero that makes temporal (pixels') change detection more accurate as shown in Fig.2.

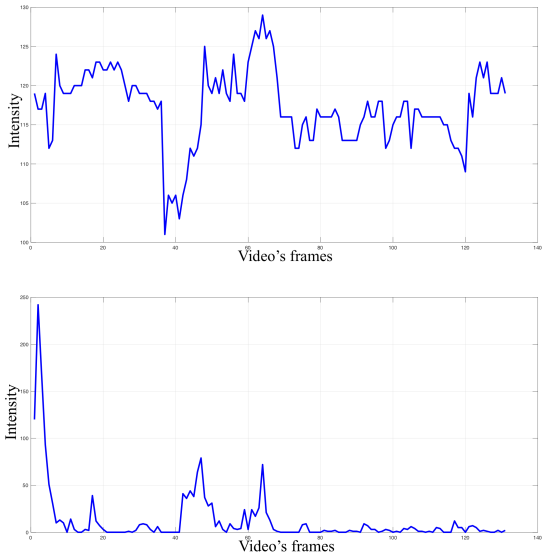


Figure 2: A pixel distribution before using the spatiotemporal filter is on the top and after using the spatiotemporal filter is on the bottom.

Since the size of the removed objects is unknown, a video is divided into frames, and spatial filtering (i.e., Laplacian pyramid [23]) is applied to each frame. The Laplacian pyramid subtracts each frame from its blurred version to form a video scale, down-samples each frame by half, and repeats this process until the minimum resolution of a frame is reached. This process constructs multi-scale videos that represent edges at different scales. Then, we perform temporal filtering in each scale by using the pixels values throughout time in a frequency band and apply a high-pass filter to remove static edges.

Sequential Analysis

We reconstruct the Laplacian pyramid that transfers multi-scale videos to one video scale (i.e., the input video scale) to minimize the computation time of the sequential analysis. We then introduce a null hypothesis H_0 that states there is no change in a pixel's mean value and an alternative hypothesis H_1 that states there are changes in a pixel's mean value. The mean before the

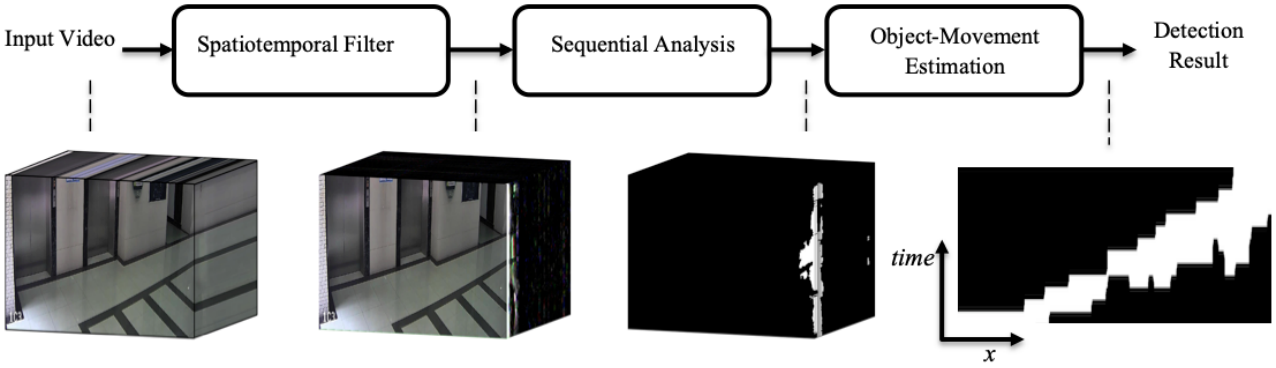


Figure 3: The flowchart of our approach.

change μ_0, is assumed to be known and the mean after the change is assumed completely unknown but different than μ_0. We begin by formulating the null and alternative hypothesis:

$$\begin{aligned} \mathbf{H_0} &= \{\mu : \mu = \mu_0, k < \tau\} \\ \mathbf{H_1} &= \{\mu : \mu \neq \mu_0, k \geq \tau\} \end{aligned} \tag{1}$$

We also assume pixels' values are drawn from a normal distribution ($x \sim \mathcal{N}(\mu_i, \sigma^2)$) and are independent and identically distributed and that the variance remains constant throughout frames while the mean is dependent on the scene. Using these hypotheses, we form the null and alternative likelihoods

$$L(H_0) = P(x/H_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{i=k}^{N} e^{\frac{-(x_i-\mu_0)^2}{2\sigma^2}} \tag{2}$$

$$L(H_1) = \sup_{\mu_v} P(x/H_1) = \sup_{\mu_v} \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{i=k}^{N} e^{\frac{-(x_i-\mu_v)^2}{2\sigma^2}} \tag{3}$$

Where x represents values of a pixel throughout video frames; N is the number of frames; μ , σ^2 donate the mean and variance of the pixel respectively. Using Equations 2 and 3, we form log likelihood

$$R_k^N = \ln \frac{\sup_{\mu_v} P(x/H_1)}{P(x/H_0)} = \ln \frac{\sup_{\mu_v} \prod_{i=k}^{N} e^{\frac{-(x_i-\mu_v)^2}{2\sigma^2}}}{\prod_{i=k}^{N} e^{\frac{-(x_i-\mu_0)^2}{2\sigma^2}}} \tag{4}$$

$$\hat{\mu_v} = \frac{1}{N-k+1} \sum_{i=k}^{N} x_i \tag{5}$$

The unknown mean (μ_v) is replaced by its maximum likelihood (ML) estimate (equation 5). Then log likelihood and generalized log likelihood becomes

$$\begin{aligned} R_k^N &= \frac{1}{2\sigma^2} \left[\sum_{i=k}^{N} (x_i - \mu_0)^2 - \sum_{i=k}^{N} (x_i - \hat{\mu_v})^2 \right] \\ &= \frac{1}{2\sigma^2} \sum_{i=k}^{N} \left[(x_i - \mu_0)^2 - (x_i - \hat{\mu_v})^2 \right] \\ &= \sum_{i=k}^{N} \left[\frac{(\hat{\mu_v} - \mu_0) x_i}{\sigma^2} + \frac{\mu_0^2 - \hat{\mu_v}^2}{2\sigma^2} \right] \end{aligned} \tag{6}$$

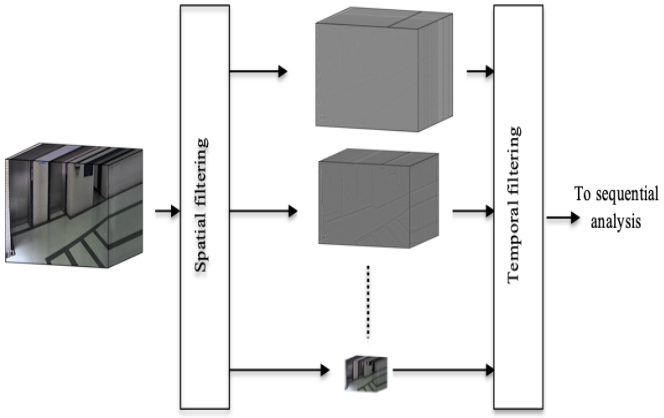


Figure 4: The overview of the spatiotemporal filter.

$$\begin{aligned} G &= \max_{1 \leq k \leq N} R_k^N \\ &= \max_{1 \leq k \leq N} \sum_{i=k}^{N} \left[\frac{(\hat{\mu_v} - \mu_0) x_i}{\sigma^2} + \frac{\mu_0^2 - \hat{\mu_v}^2}{2\sigma^2} \right] \end{aligned} \tag{7}$$

$$\tau = \min\{N \geq 1 : G \geq h\} \tag{8}$$

In equation 8, τ (alarm detection) donates frame number where the change occur; N is the discrete time index (frame index), and h is a threshold.

We use binary segmentation [24] to detect multiple changes. Binary segmentation starts by detecting single change point in the complete time series, if there is a change point, it splits the time series around this change point into two sub-series, and repeats this process until no change points are detected. By using binary segmentation, the time axis that represents a frame number will be divided into segments.

The null hypothesis is rejected if three conditions are met. First, G exceeds a certain threshold (h) to show there is a change in the pixel's mean. Second, the mean of any segment exceeds a certain threshold to identify whether this segment belongs to a background or a removed object. Third, the length of this segment is less than a certain threshold based on our definition that removed objects are moving as stated in introduction section.

A pixel's change detection is applied individually to the red, green, and blue channels of video frames. A change in a pixel is detected if this change happens on at least two channels in the

Table 1: The Comparison Results of Object-based Forgery Detection for Our Approach and Approach [21] Using Different Video Sets.

Video sets	Precision (%)		Recall (%)		F1 (%)	
	Our approach	Approach [21]	**Our approach**	Approach [21]	**Our approach**	Approach [21]
Uncompressed	**93.3**	85.7	**93.3**	80	**93.3**	82.7
Compressed (MPEG-4)	**92.8**	71.4	**92.8**	66.7	**92.8**	68.9
Low-resolution	**92.8**	78.5	**86.7**	73.3	**89.6**	75.8

same time interval to eliminate false alarms. Finally, we construct a binary video where a pixel equals one in frames that belong to changed segments and equals zero in frames that belong to unchanged segments. A video forgery is detected if at least 30 consecutive frames (i.e., one second) have a large area (i.e., 500 pixels) that contains only ones.

Object-Movement Estimation

The removed object's movement in a forged video is estimated by constructing another binary video where a pixel equals one on a frame where a change occurs until the last video frame and equals zero on the other frames. Once this binary video is constructed, the traces of a removed object can be visualized by plotting the last spatiotemporal XT slice, which is the top view of this video.

Experimental results

To the best of our knowledge, the only available video forgery datasets are SULFA [25] and SYSU-OBJFORG [22]. However, SULFA is a frame-based forgery, and SYSU-OBJFORG isn't realistic since a naked eye can identify its forged videos. Therefore, we collected a video set that is extracted from a static surveillance camera [22] where videos are uncompressed with a resolution of 1280x520pixels. Using this video set, we built our video set that contains 15 original videos that have only a static scene and 15 forged videos that are generated using recent inpainting algorithm [17]. Using our video set, we generated compressed and low-resolution video sets by compressing these videos using H.264/MPEG-4 with 1 Mbps and reducing the original resolution by half, respectively. Hence, we ended up with 90 videos to evaluate the effectiveness of our approach.

Evaluation Metric

Our approach focuses on detecting pixel changes that often occur near a removed object boundary, not on the whole area of this object. Hence, our approach is measured at the video level, which identifies whether a video is forged or not, rather than pixel level. By defining T_P as the correctly detected forged videos, F_P as original videos that have been incorrectly detected as forged and F_N as falsely missed forged videos, we compute *Precision*, *Recall*, and $F1$ as follows:

$$Precision = \frac{T_P}{T_P + F_P} \tag{9}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{10}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Presision + Recall} \tag{11}$$

Precision shows the probability that a detected forgery is truly a forgery, *Recall* indicates the probability that a forged video is detected, and $F1$ score combines precision and recall in a single value.

Comparison of Detection Results

To the best of our knowledge, there are only two recent approaches [21] and [22] that detect object-based video forgery. We can only compare our approach with [21] since the experiments in [22] are not convincing due to the conflicts of their classification measurements and the overfitting of their classifier as discussed in Section III.A and IV.B in their paper, respectively. The selected approach along with our approach are implemented on a machine with an *Intel Core i7 with* 8-*GB RAM*.

Results on the Data Set

The experimental results are shown in Table 1 that summarizes the *Precision*, *Recall*, and $F1$ score using our approach and approach [21] for three different video sets. We observe that our approach achieves the best performance on uncompressed video sets. Then its performance slightly decreases throughout compressed and low-resolution video sets, but it outperforms approach [21] throughout all the video sets. We conclude that our approach is more robust against compressed and lower resolution videos

An example of an estimation of a removed object movement for a video is shown in Fig.5. Our approach can estimate the movement even though a video is compressed or has lower resolution since our approach applies the spatiotemporal filter that can detect different sizes of removed objects.

Conclusion

In this paper, object-based video forgery is investigated and we have proposed an approach based on sequential analysis. Furthermore, we have shown that the proposed approach can estimate a movement of different sizes of removed objects using spatial decomposition and it can detect temporal changes that are nearly invisible using sequential analysis. Results show that our approach not only outperforms the other approach in terms of *Precision*, *Recall*, and $F1$ score but it is also more robust against compressed and lower resolution videos. Our further research will focus on improving the detection speed of the proposed approach since sequential analysis stage is computationally expensive.

References

[1] M. C. Stamm, M. Wu, and K. R. Liu, "Information forensics: An overview of the first decade," *IEEE Access*, vol. 1, pp. 167–200, 2013.

[2] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," *APSIPA Transactions on Signal and Information Processing*, vol. 1, 2012.

[3] K. Sitara and B. M. Mehtre, "Digital video tampering detection: an overview of passive techniques," *Digital Investigation*, vol. 18, pp. 8–22, 2016.

[4] J. Chao, X. Jiang, and T. Sun, "A novel video inter-frame forgery model detection scheme based on optical flow con-

Figure 5: An estimation result of a removed object movement for a video. Images on the top row indicate frames from the original video; Images on the middle row indicate frames from the tampered video where the girl in the red box has been removed from the scene; Images on the bottom row (from left to right) indicate the movement estimation for uncompressed, compressed, and low-resolution videos, respectively.

sistency," in *The International Workshop on Digital Forensics and Watermarking 2012*. Springer, 2013, pp. 267–281.

[5] Y. Wu, X. Jiang, T. Sun, and W. Wang, "Exposing video inter-frame forgery based on velocity field consistency," in *Acoustics, speech and signal processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2674–2678.

[6] W. Wang and H. Farid, "Exposing digital forgeries in video by detecting duplication," in *Proceedings of the 9th workshop on Multimedia & security*. ACM, 2007, pp. 35–42.

[7] V. K. Singh, P. Pant, and R. C. Tripathi, "Detection of frame duplication type of forgery in digital video using sub-block based features," in *International Conference on Digital Forensics and Cyber Crime*. Springer, 2015, pp. 29–38.

[8] J. Yang, T. Huang, and L. Su, "Using similarity analysis to detect frame duplication forgery in videos," *Multimedia Tools and Applications*, vol. 75, no. 4, pp. 1793–1811, 2016.

[9] W. Wang and H. Farid, "Exposing digital forgeries in video by detecting double quantization," in *Proceedings of the 11th ACM workshop on Multimedia and security*. ACM, 2009, pp. 39–48.

[10] D. Liao, R. Yang, H. Liu, J. Li, and J. Huang, "Double h. 264/avc compression detection using quantized nonzero ac coefficients," in *Media Watermarking, Security, and Forensics III*, vol. 7880. International Society for Optics and Photonics, 2011, p. 78800Q.

[11] P. Bestagini, S. Battaglia, S. Milani, M. Tagliasacchi, and S. Tubaro, "Detection of temporal interpolation in video sequences," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3033–3037.

[12] M. C. Stamm, W. S. Lin, and K. R. Liu, "Temporal forensics and anti-forensics for motion compensated video," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 4, pp. 1315–1329, 2012.

[13] Y. Su, Y. Han, and C. Zhang, "Detection of blue screen based on edge features," in *Information Technology and Artificial Intelligence Conference (ITAIC), 2011 6th IEEE Joint International*, vol. 2. IEEE, 2011, pp. 469–472.

[14] J. Xu, Y. Yu, Y. Su, B. Dong, and X. You, "Detection of blue screen special effects in videos," *Physics Procedia*, vol. 33, pp. 1316–1322, 2012.

[15] M. A. Bagiwa, A. W. A. Wahab, M. Y. I. Idris, S. Khan, and K.-K. R. Choo, "Chroma key background detection for digital video using statistical correlation of blurring artifact," *Digital Investigation*, vol. 19, pp. 29–43, 2016.

[16] D. D'Avino, D. Cozzolino, G. Poggi, and L. Verdoliva, "Autoencoder with recurrent neural networks for video forgery detection," *Electronic Imaging*, vol. 2017, no. 7, pp. 92–99, 2017.

[17] A. Newson, A. Almansa, M. Fradet, Y. Gousseau, and P. Pérez, "Video inpainting of complex scenes," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 1993–2019, 2014.

[18] M. Ebdelli, O. Le Meur, and C. Guillemot, "Video inpaint-

ing with short-term windows: application to object removal and error concealment," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3034–3047, 2015.
[19] J. Zhang, Y. Su, and M. Zhang, "Exposing digital video forgery by ghost shadow artifact," in *Proceedings of the First ACM workshop on Multimedia in forensics*. ACM, 2009, pp. 49–54.
[20] V. Conotter, J. F. O'Brien, and H. Farid, "Exposing digital forgeries in ballistic motion," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 283–296, 2012.
[21] C. Richao, Y. Gaobo, and Z. Ningbo, "Detection of object-based manipulation by the statistical features of object contour," *Forensic science international*, vol. 236, pp. 164–169, 2014.
[22] S. Chen, S. Tan, B. Li, and J. Huang, "Automatic detection of object-based forgery in advanced video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 11, pp. 2138–2151, 2016.
[23] P. J. BURT and E. H. ADELSON, "The laplacian pyramid as a compact image code," *IEEE TRANSACTIONS ON COMMUNICATIONS*, vol. 3, no. 4, 1983.
[24] J. Bai, "Estimating multiple breaks one at a time," *Econometric theory*, vol. 13, no. 3, pp. 315–352, 1997.
[25] G. Qadir, S. Yahaya, and A. T. Ho, "Surrey university library for forensic analysis (sulfa) of video content," 2012.

Author Biography

Mohammed Aloraini *received his BS in electrical engineering from Qassim University in 2011 and the M.S. degree in electrical and computer engineering from University of Illinois at Chicago in 2014. He is now pursuing his PhD in electrical and computer engineering at University of Illinois at Chicago. His current research interests include multimedia forensics and information security.*

Mehdi Sharifzadeh *received his BS in electrical engineering from Sharif University of Technology in 2012. Currently, he is a PhD student and researcher in electrical and computer engineering at University of Illinois at Chicago. His current researches are in machine learning, and problems in image processing and computer Vision.*

Chirag Agarwal *received his B.Tech in Electronics and Communication engineering from Future Institute of Engineering and Management, Kolkata, India in 2012. Currently, he is a PhD student in the department of Electrical and Computer engineering at Univeristy of Illinois at Chicago. His current research interests are in adversarial machine learning, deep neural networks and computer vision. Along with his current research he works on the application of deep learning in the field of medical Imaging.*

Dan Schonfeld *received the B.S. degree in electrical engineering and computer science from the University of California at Berkeley in 1986 and the M.S. and Ph.D. degrees in electrical and computer engineering from the Johns Hopkins University, Baltimore, MD, in 1988 and 1990, respectively. In 1990, he joined the University of Illinois at Chicago, where he is currently a Professor in the Department of Electrical and Computer Engineering. He has authored over 120 technical papers in various journals and conferences. His current research interests are in multi-dimensional signal processing, image and video analysis, computer vision, and genomic signal processing.*

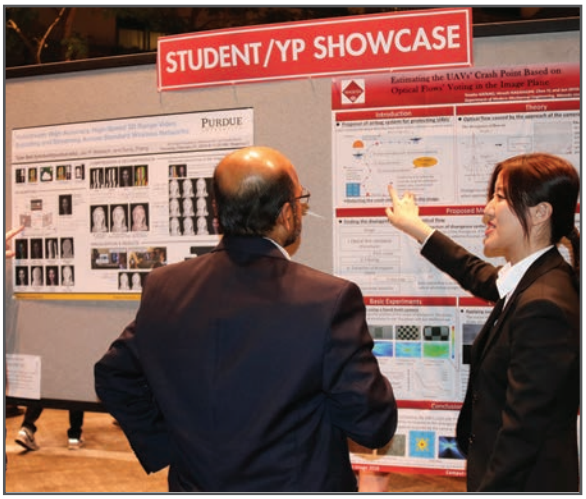
JOIN US AT THE NEXT EI!

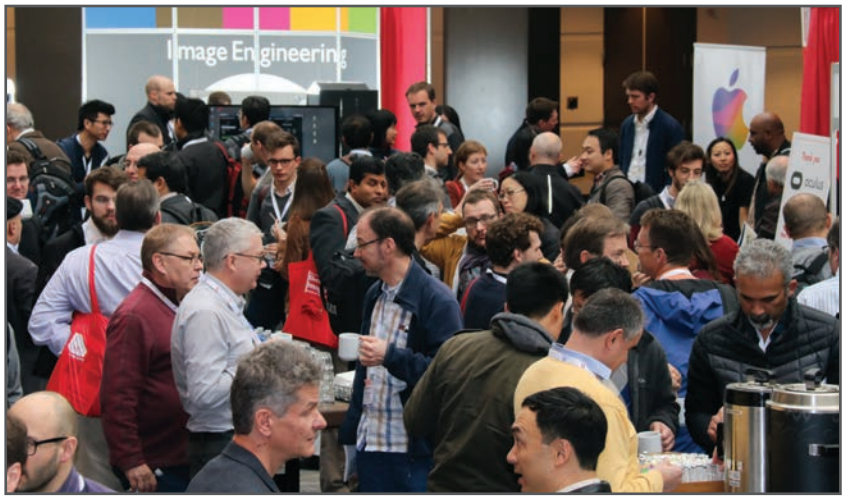
IS&T International Symposium on

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