https://doi.org/10.2352/EI.2025.37.15.AVM-108

Automotive Vision using Hybrid Event Sensors

Kamal Rana, Sean Fausz, Shijie Xiao, Zhongyang Huang and Bo Mu; OmniVision Technologies, Santa Clara, California 95054, USA

Abstract

Automotive vision plays a vital role in advanced driver assistance systems (ADAS), enabling key functionalities such as collision avoidance. The effectiveness of models designed for automotive vision is typically measured based on their ability to accurately detect objects in a scene. However, an often-overlooked factor for automotive vision is the speed of the detections that depends on the data collection rate of the sensors. With conventional image sensors (CIS), the object detection rate is limited by the no information region between two consecutive frames (hereafter we refer to it as blind time), which affects the response time of drivers and ADAS to external stimuli. While increasing the CIS frame rate decreases the blind time and enables faster decision-making, it comes at the cost of increased data rate and power consumption. In contrast, lower CIS frame rates reduce data rate and have lower power consumption, but result in longer blind intervals between frames, delaying response time, which could be critical in high-risk situations. This trade-off between data rate and decision-making speed can be addressed by utilizing hybrid sensors for automotive vision. Hybrid sensors integrate event pixels alongside with CIS pixels. Event pixels provide sparse yet high-temporal-resolution data, continuously capturing changes in scene contrast that complements dense low temporal information of CIS. In this work, we demonstrate that 7fps CIS frames combined with EVS data can achieve ~40% lower data rate compared to 20fps CIS, without compromising performance of object detections. Moreover, 7fps CIS combined with EVS maintains almost constant performance within the blind time and thus enables faster detection with low data rate and power.

Introduction

Advanced Driver Assistance Systems (ADAS) play an important role in assisting drivers with vehicle operation and significantly enhancing road safety by reducing the risk of accidents. Automotive vision is a key component of ADAS with various applications particularly in autonomous vehicles. Automotive vision using conventional image sensors (CIS) operate with limited frames for e.g. 20 frame rate per second and thus have a time period between two consecutive frames with no information of the scene (referred to as blind time). Blind time also affects a driver's response time to external stimuli, particularly in high-risk situations such as bad weather conditions. For an automotive sensor operating at 20–30 frames per second, the blind time ranges from 33 to 50 milliseconds, which is long enough to delay a driver's reaction. This delay is especially critical during high-speed driving, such as on highways where a vehicle traveling at 60 mph would cover approximately 26 meters in a second. In extreme weather conditions like snowfall, response times can be delayed by up to 1 second and thus limit the road safety [2]. Since human error is a leading cause of road accidents, faster reaction times to external stimuli are essential for improving road safety. Even advanced systems like Adaptive Cruise Control (ACC) have reaction times comparable to humans, typically ranging between 0.9 and 1.3 seconds [8]. In addition, Automatic Emergency Braking (AEB) takes more than a second to come to a complete stop if driving at a speed of 25 miles/hour [7]. The time AEB takes to make the vehicle come to a complete stop is higher with an increase in initial speed. One of the possible solutions to detect objects faster is by increasing the CIS frame rate, but it increases data rate and power consumption as well. Hence, solutions that can gather faster information without increasing data rate are highly desirable. A promising solution to gather faster information without increasing data rate is by using hybrid sensors. Hybrid sensors offer lower data rates while minimizing perceptual latency, making them an efficient and effective alternative for enhancing automotive vision systems.

Hybrid sensors integrate the capabilities of both event-based vision sensors (EVS) and conventional image sensors (CIS). EVS feature novel event pixels that detect relative changes in illumination when they exceed a predefined threshold [1]. These pixels output a stream of events encoding the location, polarity (sign of illumination change), and precise timestamp of each detected change. Since EVS operate asynchronously, they provide high temporal resolution with low latency. Additionally, they offer a high dynamic range, low power consumption, and are well-suited for various machine vision applications [1]. Recent studies on motion blur reduction and video frame interpolation have demonstrated the effectiveness of EVS in such tasks [3]. For automotive vision, combining EVS with CIS information will be an ideal solution for low data rate and high temporal information, because CIS sensors deliver high-resolution, dense spatial information, but suffer from low temporal resolution, leading to blind time between frames. In contrast, event pixels provide continuous, sparse information capturing brightness change information, effectively providing motion information between the blind time. By combining both modalities, hybrid sensors achieve an optimal balance, providing high-temporal-resolution information while maintaining a low data rate.

In this work, we fuse conventional image sensor (CIS) data with event sensor data for automotive vision. We utilize the DSEC dataset that contains both CIS and EVS driving data for various scenarios with different light conditions. To transform EVS data to a frame based representation to be used as an input for convolutional neural networks**,** we accumulate events over a 20ms interval and transform it into a matrix representation. For evaluation, we use the YOLOv3 model and represent the fused CIS+EVS data using three channels: (1) empty channel, (2) CIS grayscale channel, and (3) event-based representation**.** We compared the data rate and performance of a 7 fps CIS+EVS system with a 20 fps CIS-only system and observed that 7 fps CIS+EVS consumes approximately 40% less data rate without compromising performance. As we use only CIS grayscale in 7 fps CIS+EVS, hereafter, we will refer to "7fps CIS+EVS" case as "7fps CIS (grayscale)+EVS".

Additionally, we evaluate the performance of 7 fps CIS (grayscale) +EVS during CIS blind time and observe consistent performance, demonstrating the CIS+ EVS effectiveness in providing continuous information. Our results show that combining CIS with EVS enables low-data-rate automotive vision while maintaining performance levels comparable to standard CIS systems. This approach offers a promising solution for reducing bandwidth and power consumption without sacrificing detection accuracy.

Figure 1: *Sample RGB Images from DSEC Dataset in various scenarios (different lightning conditions).*

Dataset

In this work, we utilize DSEC: A Stereo Event Camera Dataset for Driving Scenarios [6]. The dataset is publicly available from the DSEC dataset website. It contains driving dataset for various scenarios, including both conventional CIS frames (see Figure 1) and event data (see Figure 2).

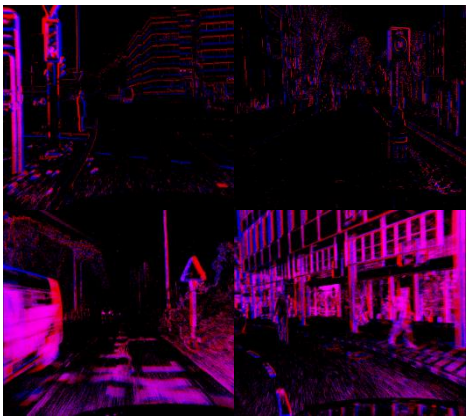


Figure 2: *Sample EVS images obtained from DSEC Dataset. These images are drawn by accumulating EVS data over a given time interval. The blue color shows the negative events and the red color shows the positive events.*

The event data has spatial resolution of 640 × 480 and the RGB images are downsampled to match the same resolution as EVS at 640 × 480. The dataset includes 20 frame rate per second with variable exposure time based on the scene and captured in a 12-bit raw format for CIS images. The dataset contains numerous sequences or clips of data captured but in our work, we use only a subset of the available dataset to limit the training time. The sequences we use for training are from "Zurich city" from "16-21a" and we use "thun_02_a" data for testing. The dataset is stereo; however, we only use the CIS images from the left camera.

Related Work

Previous studies on automotive vision using conventional image sensors (CIS) have primarily focused on object detection performance across various scenarios. Some studies have explored the impact of rain on detection accuracy, while others have examined the effects of lens blur or different lighting conditions [4]. However, these studies often overlook the influence of blind time between CIS frames and the detection rate for object detection.

Hybrid sensors that combine CIS and event-based vision sensors (EVS) have demonstrated significant potential in applications such as deblurring and video frame interpolation [3]. Research on automotive vision using hybrid CIS+EVS sensors has highlighted their advantages in detection performance. Notably, a recent study demonstrated the benefits of this hybrid approach in both performance and data rate [2]. However, existing studies fail to address the extensive data rate demands in real-world scenarios in the vehicles and the minimum CIS frames per second combined with events needed to maintain performance without degradation.

Methods

Object detection is a fundamental task in computer vision that involves both identifying objects and determining their locations within an image. In this work, we use the YOLOv3 (You Only Look Once) model, widely recognized for its balance between speed and accuracy. YOLOv3 employs a multi-scale detection approach, enabling it to effectively detect objects of various sizes by predicting bounding boxes at three different scales: small, medium and large. The model's output consists of bounding boxes, confidence scores, and class probabilities (see Figure 3).

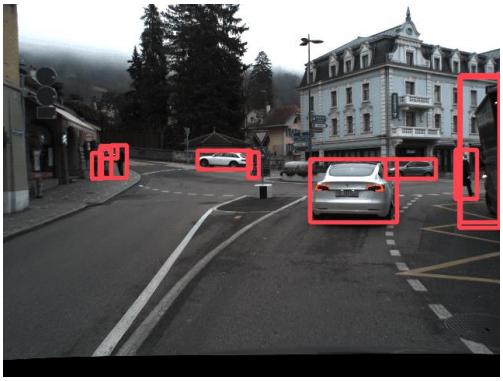


Figure 3: *A sample CIS image with the bounding boxes of objects.*

To eliminate redundant or overlapping detections, YOLOv3 applies non-maximum suppression (NMS), ensuring only the most relevant bounding boxes are retained.

Training YOLOv3 from scratch is both challenging and computationally expensive. Therefore, we use a pre-trained YOLOv3 model, originally trained on the COCO dataset, which includes 80 object categories. Notably, 8 of these classes are relevant to our study and are included in the DSEC dataset, including car, person, bicycle, truck, train, and bus. This approach allows us to leverage existing knowledge while adapting the model to our specific needs.

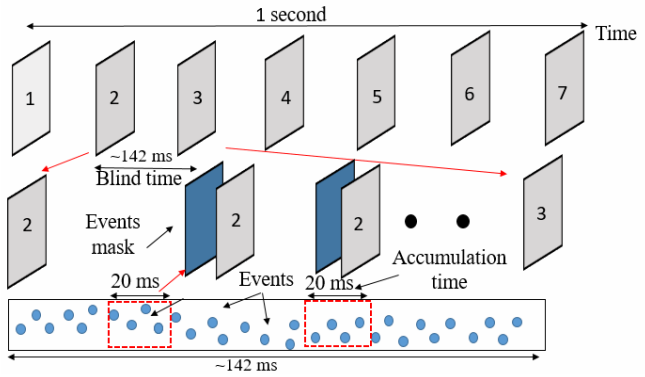


Figure 4: *The hybrid images from the 7fps CIS (grayscale)+EVS dataset were created by down sampling the 20fps CIS images to 7fps CIS images. Event data was grouped based on an accumulation time and added to the blue channel as an event mask for the hybrid frames over a consistent time interval. The previous CIS image is converted to grayscale and added to the green channel, while the red channel was left empty. This process is repeated for the next blind time between two CIS frames.*

Event data consists of the following information for each event: pixel location, timestamp, and polarity (positive or negative contrast change). To integrate this data with conventional CIS images and as an input to a neural network, we transform it into a 2D array format. Specifically, we accumulate event data over a 20ms period—a duration chosen after testing different time intervals. Longer accumulation periods resulted in blurry information, while shorter ones failed to capture sufficient events. Based on the accumulated event data, we generate a 2D event mask, where each pixel is assigned binary values either 0 or 1, depending on whether an event was triggered at that location during the accumulation time period. To maintain the original YOLOv3 architecture and to compare with CIS 3 channel images, we use a three-channel input for CIS+EVS. One channel contains the EVS mask information, another holds the grayscale CIS image and the third remains empty (see Figure 5). During the CIS blind time, we use the most recent CIS image available while continuously updating the EVS information at regular time intervals (see Figure 4). This ensures that even in the absence of new CIS frames, the model still receives updated event-based information, enhancing detection accuracy and temporal consistency.

Results

We compared the object detection performance of conventional CIS camera of 20 fps with the hybrid data consisting of both 7 fps CIS and EVS. For a fair comparison between these two cases, we create an extra 13 frames per second with the help of EVS (see method section for more details) for 7fps CIS(grayscale)+EVS case. We created this additional CIS(grayscale) +EVS frame in almost equal interval of time. As we used only grayscale CIS image in CIS(grayscale)+EVS image, we compared its performance with both 20fps CIS RGB and 20fps grayscale images. Using 20fps CIS RGB and 20fps CIS grayscale images, we achieved 0.54 Mean Average Precision (mAP) and 0.43 mAP, respectively. Whereas, 7fps CIS (grayscale) +EVS achieved 0.52 mAP (see Table 1). The slight drop in performance of 7fps CIS (grayscale) +EVS compared to 20fps CIS RGB images might be associated with only using the CIS grayscale images and creating labels itself in the intermediate time for 7fps CIS (grayscale) +EVS case.

Image Type	Performance (mAP)	Data Rate (bytes/s/pixel)	Blind Time (ms)
CIS Gray (20fps)	≈ 0.43	30	50ms
CIS RGB (20fps)	≈ 0.54	30	50ms
CIS Gray+EVS (7fps CIS +EVS)	≈ 0.52	10.5 + 8.14 = 18.64	≈ 0ms

Table 1: *The above displays the data rate, blind time and performance between 7fps CIS+EVS, 20fps CIS RGB, and 20fps CIS grayscale. The table shows that 7fps CIS+EVS achieves similar performance as 20fps RGB CIS while saving nearly 40% of the data rate compared to RGB CIS and having nearly no blind time. 7fps CIS+EVS has a similar data rate to a CIS camera capturing at 12fps. For EVS data rate calculation, we used 10 million events per second, the average from the data we used.*

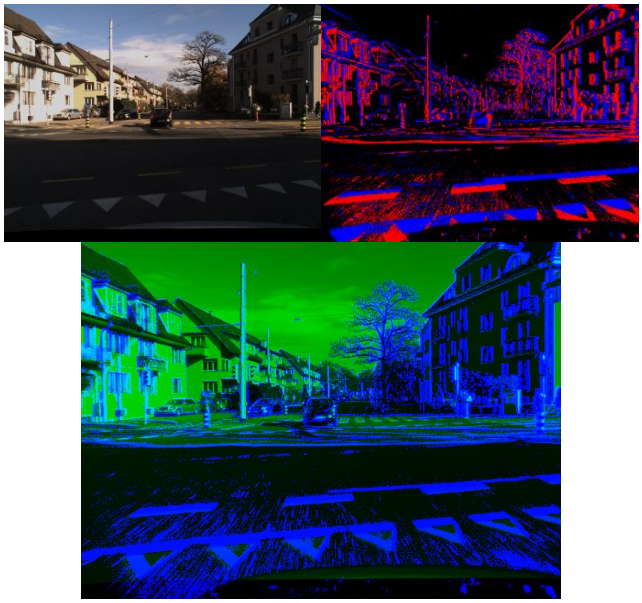


Figure 5*: Visualization of CIS image (top left), EVS image (top right) and hybrid CIS+EVS image. The EVS image contains both positive (red color) and negative (blue color) polarity events. For CIS+EVS, CIS info converted to grayscale and added to the green channel, event data (both polarities) added to the blue channel, and the red channel is left empty.*

To check the performance of CIS+EVS in the blind time between two consecutive CIS frames, we evaluated the method performance for various time interval within the blind time for 7fps CIS (grayscale)+EVS case. We observed almost constant mean average precision around the entire blind interval time.

The mean average performance varies from 0.54 to 0.5 within this blind time (see Figure 6). This slight drop in mAP with a large time difference between CIS and EVS channel (see method section) might be removed by more training or just using the entire DSEC dataset (please note we only used a fraction of the DSEC data for simplicity).

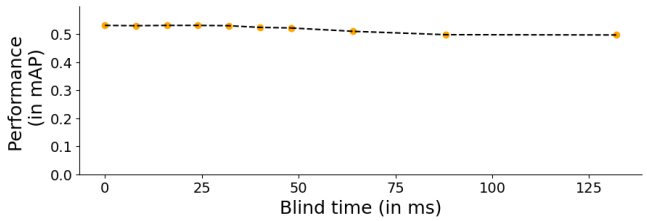


Figure 6: *The Mean Average Precision (mAP) of 7fps CIS (grayscale)+EVS case between the 7fps CIS blind time. The method achieved > 0.5 mAP between the blind time.*

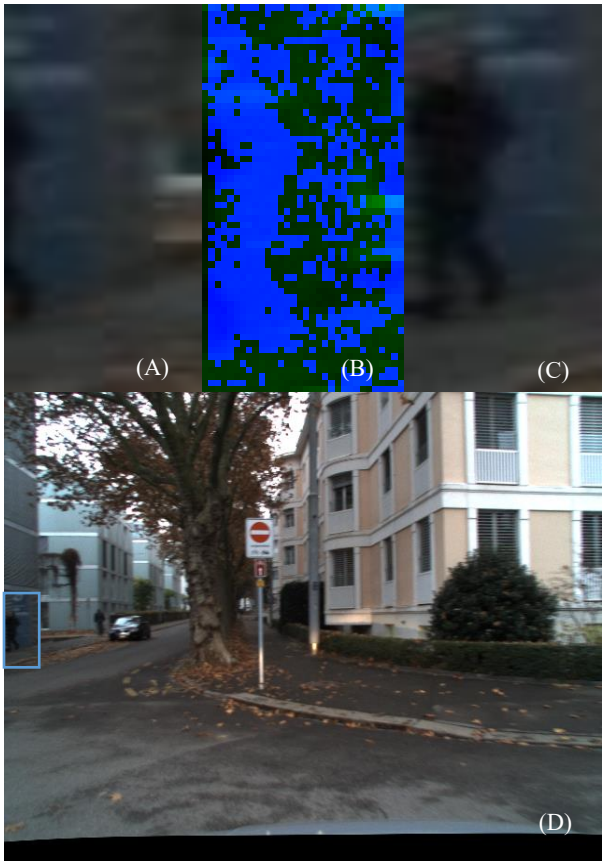


Figure 7: *Example of how EVS can help detect objects faster than CIS alone. Image (A) shows the faint outline of a person's shoulder and leg and will likely not be detected. Image (B) is a hybrid frame from the time between image (A) and image (C) and an EVS outline of a person can be seen. Image (C) shows the person is visible now in frame. Image (D) shows the entire image with a box to show the area zoomed in on for images A-C.*

Additionally, we calculated the data rate in bytes per second per pixel for both 20fps CIS and 7fps CIS(grayscale)+EVS case. A single CIS frame consume 1.5 bytes per second per pixel, while one second of EVS data consumes approximately 8.4 bytes per pixel. As a result, the 20 fps CIS setup has a data rate of 30 bytes per second per pixel, whereas the 7 fps CIS(grayscale)+EVS setup consumes only 18.64 bytes per second per pixel (see Table 1). This translates to approximately 40% lower data consumption for the 7 fps CIS(grayscale)+EVS setup while maintaining comparable detection performance.

Conclusion and Discussion

In this work, we demonstrate the advantages of hybrid sensors for automotive vision by evaluating their performance**,** data rate, and power consumption in comparison to conventional cameras. Traditional frame-based cameras inherently experience blind time between consecutive frames, leading to delays in object detection. These delays can significantly impact the response time of both human and machine drivers, particularly in high-speed scenarios or hazardous conditions where rapid detection is crucial for safety. Increasing CIS frame rate can solve this issue, but increases the data rate.

Hybrid sensors, which combine conventional image sensors (CIS) with event-based vision sensors (EVS), offer a promising solution for achieving faster object detection while maintaining a low data rate. EVS pixels provide high temporal resolution with sparse event-driven data, effectively capturing around 10,000 frames per second of data. Our results show that a 7 fps CIS (grayscale) combined with EVS achieves 20% higher mAP than the CIS grayscale and nearly the same object detection performance as a 20 fps CIS RGB, but with approximately 40% lower data rate. Furthermore, we analyzed the detection performance of the 7 fps CIS(grayscale)+EVS system within the blind intervals of consecutive frames and found that it maintained consistent accuracy throughout the entire blind time.

One of the key advantages of EVS is its ability to capture continuous, sparse information with almost no blind time. In certain cases, EVS can detect objects even when CIS fails to do so—such as when an object enters or exits the scene between the consecutive CIS frames (see Figure 7). We also generated additional CIS+EVS frames by combining 7fps CIS(grayscale) with EVS with comparable object detection performance, reinforcing the benefit of using EVS data. With this continuous event-driven information, objects can be detected at any moment while maintaining a lower data rate.

To ensure a fair comparison between the 20 fps CIS RGB and the 7 fps CIS (grayscale)+EVS setups, we used a three-channel input for YOLOv3. For the 7 fps CIS(grayscale)+EVS configuration, one channel contained the grayscale CIS frame, another contained EVS data, and the remaining channel was left empty. In future work, we plan to integrate all three CIS channels (RGB) with EVS data to improve performance, because using grayscale data of CIS images loses information present in the individual R, G and B channels.

Furthermore, we plan to utilize the entire DSEC dataset rather than a subset for training to enhance CIS+EVS performance. Additionally, the current EVS sensor data used in the work operates at a resolution of 640×480. To enhance EVS data quality as input to a neural network, we merged positive and negative event information into one channel. In the future work, we plan to use high-resolution EVS sensor data and use separate positive and negative event channels with CIS RGB channels as direct inputs to a neural network, potentially further improving detection performance.

References

[1] Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K. and Scaramuzza, D., 2020. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *44*(1), pp.154-180.s

[2] Gehrig, D. and Scaramuzza, D., 2024. Low-latency automotive vision with event cameras. *Nature*, *629*(8014), pp.1034-1040.

[3] Rana, K., Fausz, S., Yang, Z., Tu, F., Wang, Q., Fowler, B., Suess, A. and Mu, B., 2024. Benchmarking Motion Blur of Video Frame Interpolation Using Hybrid EVS+ CIS Against CIS. *Electronic Imaging*, *36*, pp.1-5.

[4] Juyal, A., Sharma, S. and Matta, P., 2021, June. Deep learning methods for object detection in autonomous vehicles. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 751-755). IEEE.

[5] Jiang, Z., Xia, P., Huang, K., Stechele, W., Chen, G., Bing, Z. and Knoll, A., 2019, May. Mixed frame-/event-driven fast pedestrian detection. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 8332-8338). IEEE.

[6] Gehrig, M., Aarents, W., Gehrig, D. and Scaramuzza, D., 2021. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, *6*(3), pp.4947-4954.

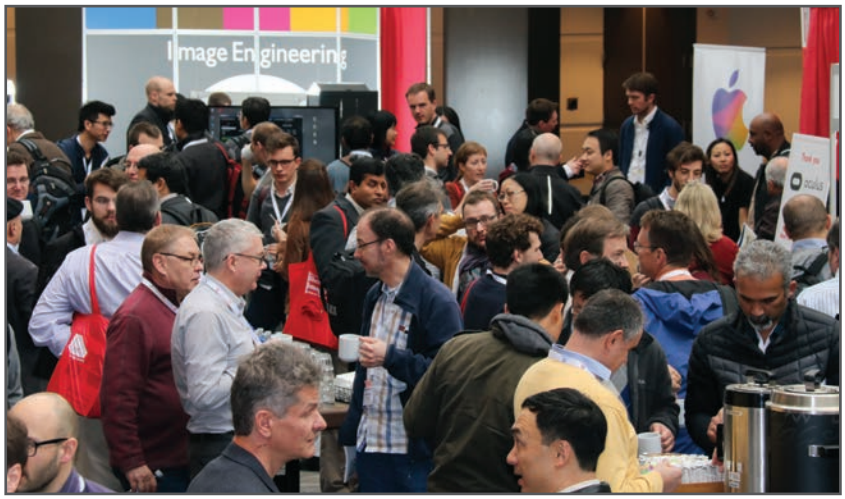
[7] Kidd, D.G., Riexinger, L.E., Perez-Rapela, D. and Jermakian, J.S., 2024. Pedestrian automatic emergency braking responses to a stationary or crossing adult mannequin during the day and night. *Traffic Injury Prevention*, *25*(sup1), pp.S116-S125.

[8] Makridis, M., Mattas, K., Borio, D., Giuliani, R. and Ciuffo, B., 2018, June. Estimating reaction time in adaptive cruise control system. In *2018 IEEE intelligent vehicles symposium (IV)* (pp. 1312-1317). IEEE.

JOIN US AT THE NEXT EI!

electronic IMAGING

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

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org