https://doi.org/10.2352/ISSN.2470-1173.2020.16.AVM-203

A Study on Training Data Selection for Object Detection in Nighttime Traffic Scenes

Astrid Unger[1,2], *Margrit Gelautz*[1], *Florian Seitner*[2]
[1] *Institute of Visual Computing and Human-Centered Technology, TU Wien; Vienna, Austria*
[2] *emotion3D GmbH; Vienna, Austria*

Abstract

With the growing demand for robust object detection algorithms in self-driving systems, it is important to consider the varying lighting and weather conditions in which cars operate all year round. The goal of our work is to gain a deeper understanding of meaningful strategies for selecting and merging training data from currently available databases and self-annotated videos in the context of automotive night scenes. We retrain an existing Convolutional Neural Network (YOLOv3) to study the influence of different training dataset combinations on the final object detection results in nighttime and low-visibility traffic scenes. Our evaluation shows that a suitable selection of training data from the GTSRD, VIPER, and BDD databases in conjunction with self-recorded night scenes can achieve an mAP of 63,5% for ten object classes, which is an improvement of 16,7% when compared to the performance of the original YOLOv3 network on the same test set.

Introduction

As autonomous driving technology evolves from advanced driver assistance systems (ADAS) to true self-driving systems, the demand for robust object detection algorithms is continuously increasing. Cars operate under varying circumstances throughout the year, during different seasons and different times of day. However, publicly available databases of traffic scenes that provide training data for deep learning algorithms are traditionally biased towards daylight and environments with good visibility. Even though the incorporation of night scenes in training datasets has recently started receiving more attention [12], [10], the manual annotation of night scenes with demanding illumination conditions is tedious and time-consuming. The challenge is to find suitable combinations of previously existing and new datasets for training, while keeping the required amount of newly annotated data from the specific target application as low as possible. In our study, we explore this question in the context of object detection in nighttime and low-visibility traffic scenes. After a review of related literature, we describe the used datasets and our training and evaluation cycles. We then present experimental results obtained from gradual amplification of our training dataset in various test cycles.

Related Work

The application of deep learning techniques on nighttime traffic scenes is currently gaining attention. Several publications have been tailored to specific object classes in night scenes. For example, Kim et al. [6] study the detection of humans, and Lim et al. [5] concentrate on traffic sign recognition under illumination variations. Contrarily, our work seeks to detect a broad range of objects that usually appear in nighttime or low-visibility traffic scenes. A previous work by Anoosheh et al. [13] proposed to establish a relation between scene representations to overcome the lack of annotated nighttime data. With regards to the YOLO network, Tung et al. [9] examine YOLO's ability to detect objects in shifting illumination conditions. However, the authors do not retrain the network, which is the focus of our work.
Our goal is to retrain an existing object detector in order to create a stable object detector, which can detect objects in varying lighting situations. We decided to retrain YOLO (You Only Look Once) [7], which is considered one of the state-of-the-art object detectors. Redmon et al. [11] found that YOLO outperformed several object detection methods, in terms of both accuracy and speed. The comparison uses the COCO dataset [1], which we also use as a baseline in our study.

Datasets

Our work relies on retraining an existing Convolutional Neural Network (CNN) in order to study the influence of different data set selections and combinations on the recognition of objects in night scenes. We successively augment our training data set in view of our target application, which focuses on automotive night scenes. The following sections describe the data base and its usage in our training and evaluation cycles.

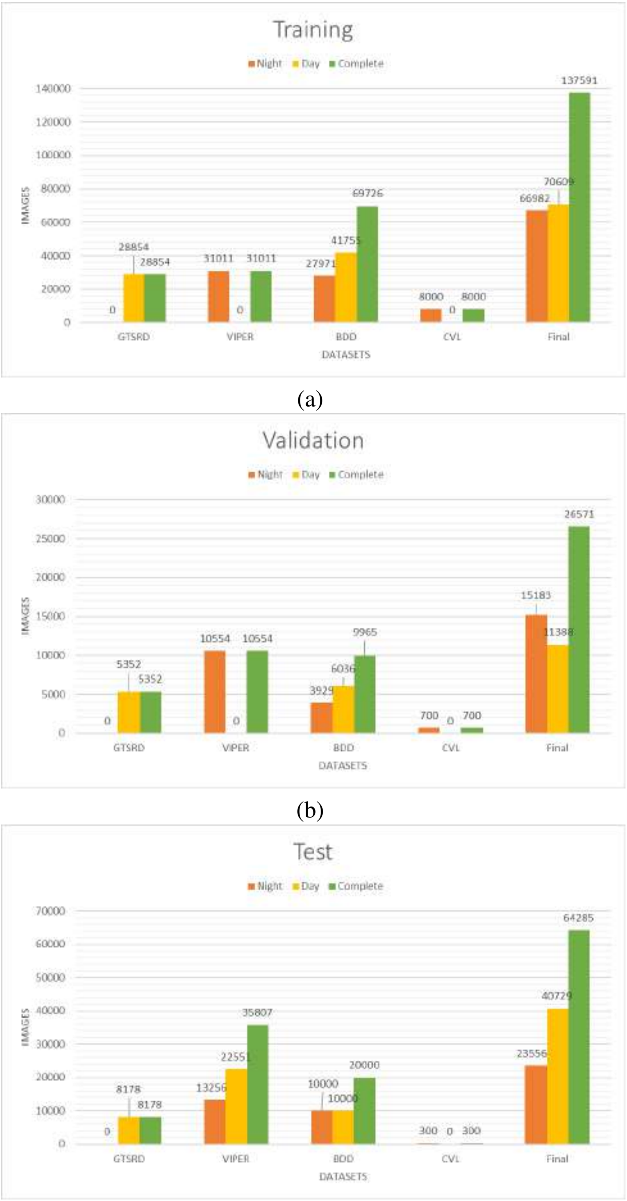
Four datasets were used in our study: The German Traffic Sign Recognition Database (GTSRD), VIsual PERception benchmark (VIPER), Berkeley Deep Drive 100k (BDD), as well as our own recordings, taken within the CarVisionLight (CVL) project. We review the main characteristics of these datasets and explain our motivation for incorporating them into our experiments.

GTSRD

The first dataset is the German Traffic Sign Recognition Database (GTSRD), which is a combination of the German Traffic Sign Recognition Benchmark (GTSRB) and German Traffic Sign Detection Benchmark (GTSDB) datasets. It was created by Stallkamp et al. [2] with the intention of providing a lifelike dataset of traffic signs for solving challenging computer vision and pattern recognition problems. It is a multi-class, single-image dataset, where each image contains one traffic sign. It comprises of over 40 classes of different street signs, as well as over 50k images and shows these signs in various lighting conditions. While the dataset is called “German Traffic Signs“, it should be pointed out that these road signs are subject to the Vienna Convention on Road Signs and Signals [3], which has been widely adopted in Europe and Russia and can therefore be used to solve traffic sign recognition problems in European and Russian street conditions. While the original dataset contains 40 classes, it should be noted that, for the final training set, they are all taken into account as a single "street sign" class, according to the requirements of our target application.

VIPER

The second dataset is the VIsual PERception benchmark (VIPER), which was assembled by Richter et al. [8]. It was created while driving, riding and walking 184 kilometers in diverse ambient conditions in the realistic virtual world of the Rockstar video game GTA5. It comprises over 250k high-resolution video frames, all annotated with ground-truth data for both low-level and high-level vision tasks. The dataset is split into training, validation, and test sets, containing 134K, 50K, and 70K frames, respectively. While each set contains a roughly balanced distribution of data acquired at different times of day and in different lighting conditions (day, sunset, rain, snow, night), we focused on the subset of night scenes for our study. The subset we used consists of over 15k labelled images showing exclusively night scenes with 30 classes, out of which 10 were deemed relevant for our case and therefore incorporated for training. The simulation model contains, for the most part, US American street conditions.



(a)

(b)

(c)

Figure 1. *Distribution of day and night images in the original datasets (GTSRD, VIPER, BDD, CVL) and contribution to the final dataset (Final), used for training (a), data validation (b) and testing (c).*

BDD

The Berkeley Deep Drive (BDD) [4] dataset can be described as a large-scale diverse driving video database. It contains annotated images of different scenarios, day, night and dusk/dawn of 100k driving videos from more than 50k rides. These videos were filmed in New York, San Francisco, SF Bay Area and Berkeley, in diverse weather conditions (sunny, rainy, snowy). The recordings comprise six weather conditions, six scene types, and three distinct times of day for each image, with an approximate equal contribution between night and day. This dataset provides bounding box annotations for 10 different categories, all of which were chosen for our experiments.

CVL

The CarVisionLight (CVL) dataset was recorded at dusk and during night-time in the countryside of rural Austria. The recordings contain over 50 videos, of which over 9k images were annotated with a recently developed annotation tool by Groh et al. [14]. The bounding box annotations contain three classes: person, car and street sign.

Final Dataset Composition

The distribution of day and night images in the previously discussed datasets (GTSRD, VIPER, BDD, CVL) and their contributions to the final dataset (Final) can be seen in Figure 1. The subfigure (a) shows the training dataset, with over 130k images in the final set. The largest contributor of data is BDD, with a greater contribution of day images. The other datasets consist exclusively of day (GTSRD) or night (VIPER, CVL) images. The validation dataset (b) resulted in over 26k images in the final set. These images were exclusively selected for validation and not used in the training dataset. Finally, the test dataset (c) contains over 64k images, neither used for training nor validation. Images from any of these test sets were used for qualitative evaluation.

Training and Evaluation

Our study comprises 14 distinct training and evaluation cycles, each following the same pattern, which can be seen in Figure 2. After the initial evaluation of the original YOLOv3 CNN, trained on the Microsoft COCO dataset [1], the strengths and weaknesses were evaluated and the training was adapted accordingly. This triggered the start of another training session, now with the adapted training dataset. After training, another round of quantitative and qualitative evaluation started, with the quantitative evaluation done on the validation dataset and the qualitative evaluation on the test dataset, as well as distinct evaluation videos from the CVL dataset. This cycle was completed 14 times, resulting from a combination of all 4 datasets used.

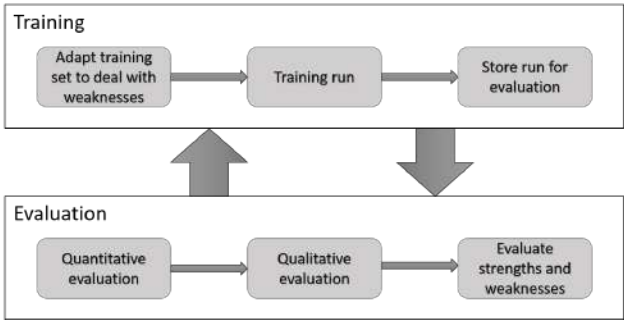


Figure 2. *Training and evaluation cycle, completed 14 times.*

The training and evaluation cycles can be characterized as follows:

- We started with an initial implementation of the YOLOv3 [7] network, which was originally trained on the Microsoft COCO dataset [1]. The COCO dataset contains over 200k labeled images with 80 common object classes, but neither traffic nor night scenes are specifically represented. An example can be seen in Figure 3 (a), where no objects could be detected.
- In a second step, we retrained the network with each dataset individually (VIPER, CVL, BDD, GTRSD), resulting in 4 individual networks.
- For the third step, we utilized combinations of two datasets each, resulting in 6 more networks.
- In a fourth step, we retrained the network with combinations of three datasets each, resulting in another 4 networks.
- Our final network was trained with over 150k images, using a combination of the GTSRD, VIPER, BDD and CVL datasets, and evaluated with over 20k images, which can be seen in more detail in Figure 1.

Figure 3 shows the visual results of different training cycles on the same test image. Subfigure 3 (a) shows the result after training on the COCO dataset, where no objects could be found. Subfigure 3 (b) was trained on the VIPER dataset and shows that in this case only the car could be detected. Subfigure 3 (c) results from the combination of two datasets, namely BDD and VIPER. Here, several false positives were identified over the whole frame. Only the car was identified correctly. These problems continued to occur throughout all different combinations of two datasets. Subfigure 3 (d) shows the combination of three different datasets - CVL, BDD and VIPER. Now the network appeared to be steered into the right direction, as indicated by the correct identification of the car and one traffic sign. Finally, we combined all four datasets (BDD, VIPER, CVL and GTRSD) to achieve the best overall results, as illustrated by the correct detection of the car and two traffic signs in Figure 4.

Results

The evaluation metric we used for our study is Intersection over Union (IoU), with a threshold of 0.6. In other words, objects found with an IoU equal or greater than 0.6 resulted in a True Positive (TP) and everything else in a False Positive (FP). From this, we calculated the mean average precision (mAP) by dividing the true positives by the sum of all detected objects. All results were calculated using the validation and test datasets, which used a combination of 4 datasets (GTSRD, CVL, VIPER, and BDD), a more detailed breakdown was shown before in Figure 1.

Table 1 shows the classification results, split for each class. Results from the original network (IoU, initial), pretrained on the COCO dataset, are compared to the interim training cycle results (interim), as well as the final results (IoU, final). The interim training cycle results from the combination of two datasets - VIPER and BDD, where the final result is obtained from a combination of GTSRD, CVL, VIPER and BDD training data. It can be seen that, while an improvement when training with a combination of the VIPER and BDD (interim) training set was achieved, the result still has the possibility of improvement,

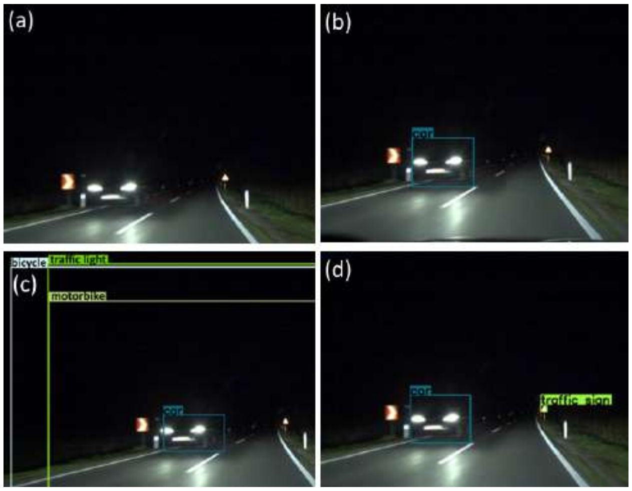


Figure 3. *Visual results of different training cycles on the same test image (not included in the training dataset), showing a car and two traffic signs on a street at night. The subfigures show the results after training on: (a) COCO dataset, (b) VIPER dataset, (c) combination of VIPER and BDD datasets, (d) combination of BDD, VIPER and CVL datasets.*

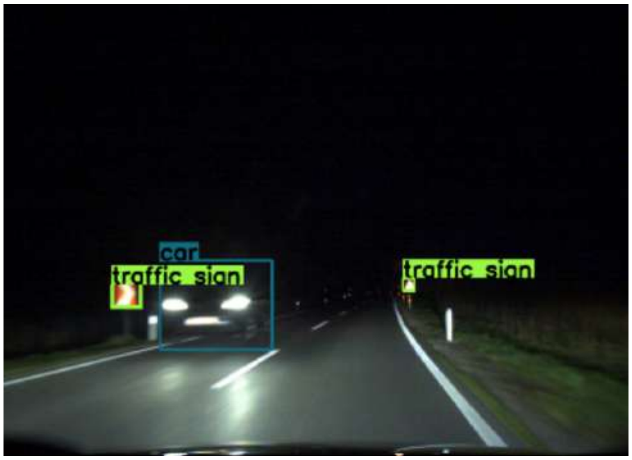


Figure 4. *Visual result of a combination of all four datasets (BDD, VIPER, CVL and GTSRD).*

which was achieved in training with the final dataset (final).
For the class “Traffic sign“, Table 1 shows a value of 0.00 for the column IoU, initial. This can be explained since the COCO dataset (initial) is missing the class “Traffic sign“ and therefore unable to detect these objects.

Overall, compared to the initial training, an improvement of 9% was achieved for the interim training cycle. The addition of the CVL and GTSRD dataset achieved an additional improvement of 7,7%, leading to an overall improvement of 16,7% compared to the initial training. An improvement of the computed mAP can be found in all classes - with the exception of the class “Rider“, which shows a decrease of 10% mAP. A possible explanation is that our night database had significantly less data regarding riders (on motorbikes or bicycles) at night, while parked motorbikes and bicycles were still found and could therefore be labeled successfully.

Table 1: Classification results obtained from different training cycles.

Class	*IoU, initial*	*IoU, interim*	*IoU, final*
Person	0.61	0.64	0.70
Car	0.68	0.77	0.83
Bus	0.34	0.41	0.44
Bicycle	0.59	0.61	0.62
Truck	0.43	0.59	0.59
Train	0.31	0.41	0.43
Rider	0.61	0.51	0.51
Motorbike	0.52	0.59	0.63
Traffic light	0.59	0.66	0.79
Traffic sign	0.00	0.35	0.81
All classes	0.46	0.55	0.63

Figure 5 shows object detection results of the final network under different lighting conditions, as well as displaying the confidence value of each detected object. Subfigure (a) shows a country road at dusk: one traffic sign is detected (with a confidence of 97.03%.). (b) shows a highway at night: two trucks (confidences of 98.63% and 37.51%) and a car (56.97%) are detected. (c) shows a village at day: several cars (confidences of 96.36%, 99.30%, 85.19% and 96.64%) and a person (93.59%) are detected. Finally, (d), a country road at night: one traffic sign is detected (with a confidence of 96.85%).

Figure 6 illustrates object detection results of the final network on different test images from various test sets. Subfigure (a) displays a scene from the CVL dataset showing a country road at night: a traffic sign and a car are detected. (b) shows a scene from the GTSRD test dataset: four traffic signs are detected. Next, subfigure (c) shows a simulated image of a city at night from the VIPER test dataset: two cars, a truck and a traffic light are detected. (d) shows a daytime city scene from the BDD test dataset: several cars, traffic lights and persons are detected.

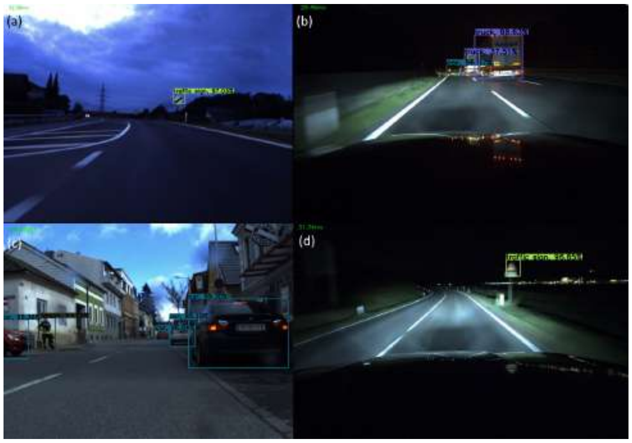


Figure 5. *Examples of object detection results of the final network under different lighting conditions, with computed confidence values. The images shown were taken from the CVL test dataset.*

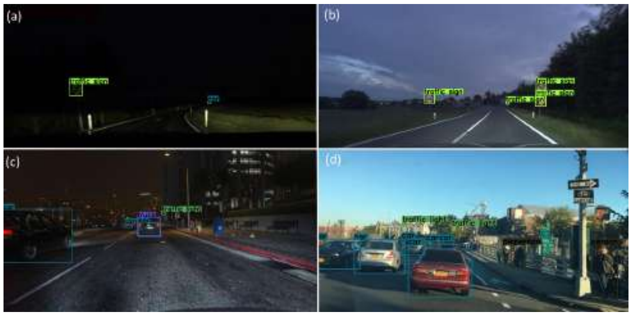


Figure 6. *Object detection result of the final network on different images: (a) CVL, (b) GTRSD, (c) VIPER, abd (d) BDD test datasets (not used for training).*

Conclusion

In our study, we performed several experiments with different combinations of training data sets for object detection in nighttime traffic scenes. Our training started with the YOLOv3 network, pre-trained on the Microsoft COCO dataset, and we successively added different combinations of four datasets - GTSRD, VIPER, BDD, and self-recorded CVL data - to our training data. We measured the detection rates in terms of mAP for ten object classes and observed an overall improvement of 16,7% between the original and final training results. In particular, the detection rates for the classes car, traffic light and traffic sign achieved values of about 80% and more. During the course of our research, we noticed a certain degradation of the performance on non-European traffic signs, which are underrepresented in our current training data and should be included in future work.

Acknowledgements

This work was supported by the CarVisionLight project, which is funded by the Austrian Research Promotion Agency (FFG) and the Austrian Federal Ministry BMVIT under the program “ICT of the Future” (project no. 861251).

References

[1] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. Zitnick. Microsoft COCO: Common objects in context. In European Conference on Computer Vision, ECCV 2014, pg. 740–755, (2014).

[2] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. The German traffic sign recognition benchmark: A multi-class classification competition. In IEEE International Joint Conference on Neural Networks, IJCNN 2011, pg. 1453–1460, (2011).

[3] Convention on Road Traffic, Vienna, 8 November 1968, United Nation Treaty Collection, Chapter XI, Section B, Number 19, available from https://treaties.un.org, last accessed 06.02.2020

[4] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. CoRR, abs/1805.04687, (2018).

[5] K. Lim, Y. Hong, Y. Choi, and H. Byun. Real-time traffic sign recognition based on a general purpose GPU and deep-learning. PLoS One, volume 12, pg. 1-22, (2017).

[6] J. H. Kim, H. Hong, and P. K. Ryoung. Convolutional neural network-based human detection in nighttime images using visible light camera sensors. Sensors (Switzerland), volume 17, pg. 1065–1091, (2017).

[7] J. Redmon and A. Farhadi. YOLOv3: An incremental improvement. CoRR, abs/1804.02767, (2018).

[8] S. R. Richter, Z. Hayder, and V. Koltun. Playing for benchmarks. In IEEE International Conference on Computer Vision, ICCV 2017, pg. 2232–2241, (2017).

[9] C. Tung, M. R. Kelleher, R. J. Schlueter, B. Xu, Y. H. Lu, G. K. Thiruvathukal, Y. K. Chen and Yang Lu. Large-scale object detection of images from network cameras in variable ambient lighting conditions. In 2nd International Conference on Multimedia Information Processing and Retrieval, MIPR 2019, pg. 393-398, (2019).

[10] M. B. Jensen, K. Nasrollahi, and T. B. Moeslund. Evaluating state-of-the-art object detector on challenging traffic light data. In IEEE Conference on Computer Vision and Pattern Recognition Workshops, pg. 882-888, (2017).

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In IEEE Conferenceon Computer Vision and Pattern Recognition, pg. 779–788, (2016).

[12] S. Sivaraman and M. M. Trivedi. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. IEEE Transactions on Intelligent Transportation Systems, volume 14, pg. 1773–1795, (2013).

[13] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. V. Gool. Night-to-day image translation for retrieval-based localization. CoRR, abs/1809.09767, (2018).

[14] F. Groh, D. Schörkhuber and M. Gelautz. A tool for semi-automatic ground truth annotation of traffic videos. Electronic Imaging 2020, (2020).

Author Biography

Astrid Unger *received her Bachelor of Engineering from the University of Salzburg, Austria and is currently finishing her Master in Visual Computing at Vienna University of Technology, Austria.*

Margrit Gelautz *received her PhD in Telematics from Graz University of Technology (1997). She is an associate professor at Vienna University of Technology, Austria, with a focus on image and video analysis & synthesis techniques. Her current research interests include 3D scene reconstruction, human-robot interaction and autonomous driving applications.*

Florian Seitner *received his Dipl.Ing and Dr. Techn. (PhD) degrees in Computer Science from the Vienna University of Technology. After working as a system designer on embedded video processing solutions at On Demand Microelectronics, he co-founded emotion3D, a company specialized in AI-based camera solutions for automotive in-cabin monitoring. Initially acting as CTO, Florian became CEO of emotion3D in 2014 and is responsible for the company's corporate and business strategy.*

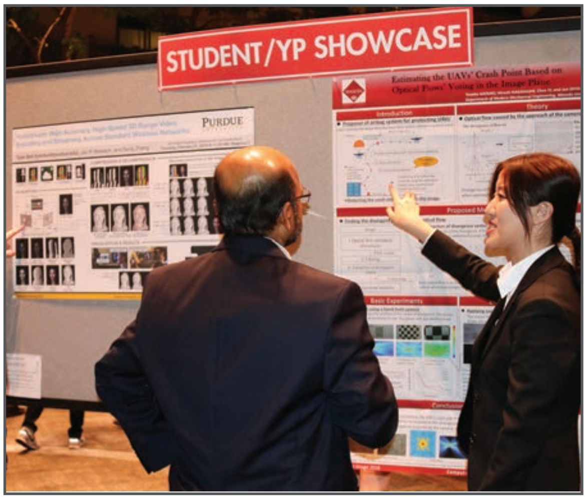
JOIN US AT THE NEXT EI!

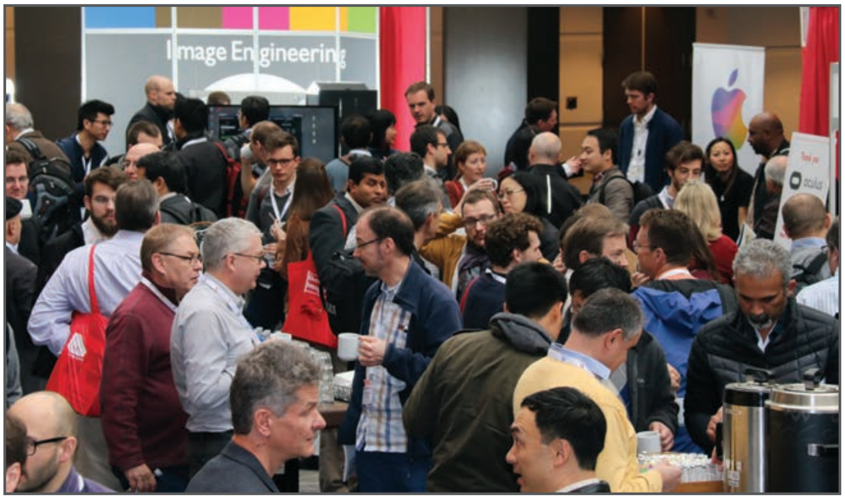
IS&T International Symposium on

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