https://doi.org/10.2352/ISSN.2470-1173.2020.16.AVM-110

End-to-End Multitask Learning for Driver Gaze and Head Pose Estimation

Mahmoud Ewaisha*[*]; mahmoud.ewaisha.ext@valeo.com; Valeo Group; Cairo, Egypt
***Marwa El Shawarby*[*]; marwaelshawarby21@gmail.com**
Hazem Abbas; hazem.abbas@eng.asu.edu.eg; Ain Shams University; Cairo, Egypt
Ibrahim Sobh; ibrahim.sobh@valeo.com; Senior Expert, Valeo Group; Cairo, Egypt

Abstract

Modern automobiles accidents occur mostly due to inattentive behavior of drivers, which is why driver's gaze estimation is becoming a critical component in automotive industry. Gaze estimation has introduced many challenges due to the nature of the surrounding environment like changes in illumination, or driver's head motion, partial face occlusion, or wearing eye decorations. Previous work conducted in this field includes explicit extraction of hand-crafted features such as eye corners and pupil center to be used to estimate gaze, or appearance-based methods like Convolutional Neural Networks which implicitly extracts features from an image and directly map it to the corresponding gaze angle. In this work, a multitask Convolutional Neural Network architecture is proposed to predict subject's gaze yaw and pitch angles, along with the head pose as an auxiliary task, making the model robust to head pose variations, without needing any complex pre-processing or hand-crafted feature extraction.Then the network's output is clustered into nine gaze classes relevant in the driving scenario. The model achieves 95.8% accuracy on the test set and 78.2% accuracy in cross-subject testing, proving the model's generalization capability and robustness to head pose variation.

Introduction

One of the most important factors in accidents nowadays is driver's inattention which may involve the driver being distracted, asleep or fatigued or just lost in thought. According to [1], 80% of crashes and 65% of near crashes involve driver distraction. Moreover, in the context of self-driving cars, semi-autonomous vehicles requires the driver to be alert at all times so it can safely transfer the control of the car to the driver in case of a critical condition. It is therefore essential for driver assistance systems to include a component that specializes in driver monitoring.

A System that collects detectable information about the driver to determine their capabilities to drive safely is often referred to as Driver's Monitoring System. Such system comprises many components performing various tasks such as drowsiness detection, action recognition, blink rate detection, gaze estimation, etc.

A lot of progress has been made in this field, starting with the traditional methods that are considered intrusive [18], where the driver had to wear a form of gadget or electrodes to keep track of some biological measures like heart rate or brain activity. Other non-intrusive [7] camera-based methods have recently gained more popularity due to the rise of Deep learning specifically in computer vision tasks, where detecting the driver's body posture, head pose, blinking rate, eyelid closure or gaze direction can all be done using a single camera.
Among the different tasks of a driver monitoring system, gaze estimation is considered a very challenging task. One of the challenges is to make the solution robust to head pose variations, so the accuracy of the predicted gaze regions is not affected by the position of the head. Another challenge is the generalization capability, i.e. making the solution perform as good on subjects it has not seen before during training. While there has been a lot of research in driver gaze estimation systems, most of the existing solutions consists of complex architectures, or require heavy pre-processing or explicit feature extraction from images.

In this work, we propose a camera-based End-to-End person and head pose-independent solution to detecting the driver's gaze direction while maintaining simplicity of design. A single Convolutional Neural network is used that takes just an image of the driver as input, and predicts the driver's gaze yaw and pitch angles, with an additional auxiliary task to predict the driver's head pose which significantly enhances the predicted gaze angles.

Related Work

In this section difficulties facing gaze estimation task, and previous attempts to overcome them, a review of conventional gaze estimation techniques as well as deep learning techniques are discussed. Gaze estimation task faces a lot of difficulties like person independence, variation in head pose, subject wearing eye decorations, distance between the subject and the camera, and camera calibration.

Gaze estimation challenges

Head pose variation is one of the major challenges for gaze estimation. The pose and orientation of the eye ball collectively depict the eye gaze [11]. Several hardware based approaches such as head-mounted cameras were introduced [8] to eliminate the effect of the changing of the eye appearance based on the orientation of the head. Others aimed to incorporate the head pose information in their dataset. Columbia Gaze dataset [15] was collected with proper calibration setup under constrained lab environment where subjects were asked to sit at a particular location and rest their chin on a stand. This setup made the dataset free from scaling and ensured the proper head pose labels. While [20], [3] collected the dataset in the wild using laptop and a tablet. While this setup allows for collection of large datasets, there is nearly no variation in the subjects' head pose due to the nature of the setup.

[*] Both authors contributed equally

Person Independence can also contribute to the difficulties in detecting eye gaze. The orientation of the face of every individual is different [13] which could lead to the model performing well on subjects present in the training data and failing to generalize on subjects it has not seen before.

Eye Decorations such as prescribed glasses can lead to some noise as the glare and the reflection on the glasses could affect the information that is needed for the problem of gaze estimation [3]. People with different glasses of different glares is an interesting challenge for an eye gaze estimation algorithm.

Techniques for gaze estimation can generally be categorized into feature-based and appearance-based methods.

Feature-based Methods

Feature-based methods are the methods that use extracted features from the eye such as eye corners, contours, orientation and ratio of the major and minor axes of the pupil ellipse or pupil-glint displacement determined by reflections of an external light source on the cornea.

Some of the previous works, [13], [3] used the hand-crafted features such as multi-level Histograms of Oriented Gradients (mHoG) to obtain important information for the eye gaze estimation. In order to accurately extract the relevant features, feature-based methods usually require high resolution images of the eye which is not easily acquired in the wild. Generally, two types of feature-based approaches exist, regression based and model based (geometric).

Regression-based Approach

The regression-based methods feed the features extracted from the image to a function which maps the features to gaze coordinates. Such a function can vary in complexity from a simple polynomial to a multi-layer neural network. Neural networks are often considered a good choice for regression tasks. A generalized regression neural network–based method was suggested in [21] where eye features such as glint coordinates were fed to a network to be mapped to screen coordinates. The type of extracted features allowed for moderate head movement. [23] suggested using Support Vector Regressor where an SVR is used to map the pupil and glint features to screen coordinates. However, these two-dimensional methods suffers a major drawback, which is failing in handling head pose variations well.

Model-based Approach

Three-dimensional model-based approaches uses the common physical structures of the human eye to calculate a 3D gaze direction vector by modelling this structure geometrically. This gaze direction vector can then be integrated it with information about the objects in the scene and used to calculate the point of regard.

Generally, these methods assume eye ball structure as spherical, which is not very accurate. To permit for free head motion, a large field of view is required, but a limited field of view is essential to capture sufficing high resolution eye images to provide reliable gaze estimates. To achieve this, multiple cameras are utilized where one camera is used for observing the head orientation and another camera for eye images, then the information from both cameras are combined and processed [19], [22], [16].

The use of multiple cameras seems to produce robust results but requires stereo calibration, complex fusion algorithms, as well as greater processing time when compared to methods that utilize a single camera. Also, due to the nature of the setup, this approach has very limited application to many settings of interest such as driver monitoring.

Appearance-based Methods

Detecting pupils and glints is essential when using feature-based methods, however, these extracted features are susceptible to error. Besides, there may be unrealized features that conveys information about gaze but is not modeled by the chosen features.

Appearance based models for gaze estimation do not explicitly extract features, but rather use the image contents as input with the intention of mapping these directly to screen coordinates. Thereby, these methods aims for implicitly extracting the underlying function for estimating the point of regard, relevant features, and personal variation, without the need for prior knowledge about the scene geometry or camera calibration. And since the mapping is made directly on the image contents, these methods do not need any calibration of cameras or geometry data [2].

Appearance-based gaze estimation methods have a lot of advantages [17] but limited by training datasets because in most of the scenarios datasets allow constrained head poses and eye rotations. Training of the system requires a large amount of data that reflects the real-world variations [2]. To overcome this problem, a variety of datasets are available now. Earlier the accuracy of appearance-based methods depends on the head pose motion [6], which limits its applications. The present trend is shifted to allow head pose variation while collecting data.

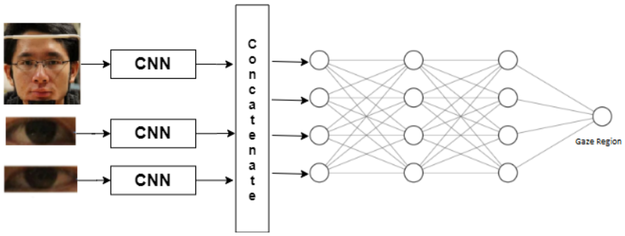


Figure 1: Images of face, right eye and left eye are each input to a CNN for feature extraction. Feature vectors obtained are then merged together then fed to a neural network to be mapped to gaze coordinates.

An example of appearance-based methods is Deep learning techniques, such as Convolutional Neural Networks (CNNs), which have been successfully used in challenging conditions such as those with variable illumination, unconstrained backgrounds and free head motion, and without the need for calibration, while achieving greater results than feature based methods.

Most of the previous work done using appearance-based methods try to incorporate information about the subject's head pose in the network.[4], [9] extract three patches from the original image, namely the left and right eyes and the whole face, then input each patch to a separate CNN, merge the feature vectors of each network together and use this vector as input to fully connected layers to finally estimate the gaze region as shown in Figure 1. [12] uses an image of a subject's eye as input to some shared convolutional layers, and uses the head pose information as another input to their architecture to switch between different sub-networks during training where each sub-network is respon-

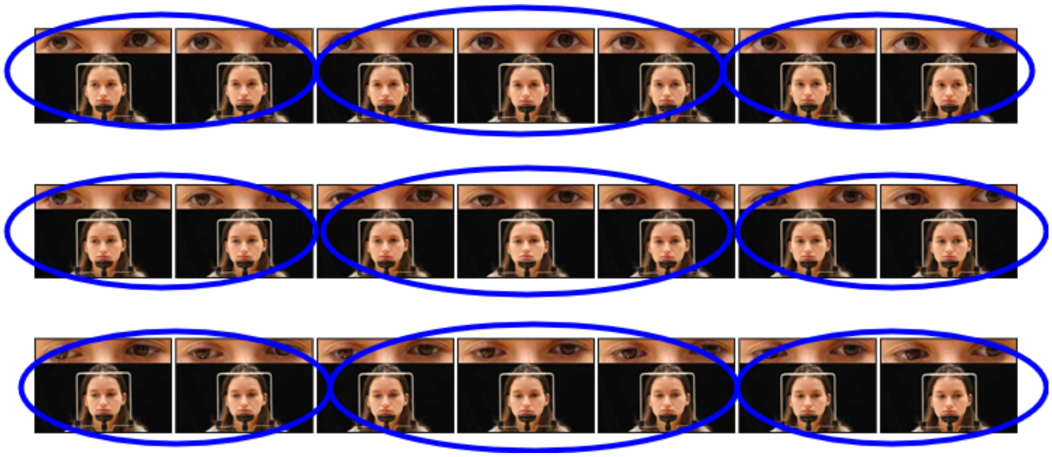


Figure 2: The original 21 gaze points and the clustering used to reduce them into a simplified and more practical 9 classes

sible for predicting the gaze region for a specific set of head pose angles. The main drawback of such methods is the complex setup and the use of multiple networks.

Proposed Method

An End-to-End solution to driver's gaze estimation is proposed using a single Convolutional Neural Network (CNN), without the need for any feature extraction or pipe-lined architectures used in other related work.

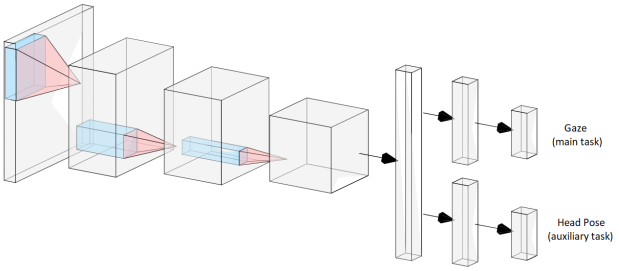


Figure 3: The proposed generic architecture with shared convolutional layers and 2 heads comprising of fully connected layers, one for gaze angle estimation, and the other for head pose estimation as an auxiliary task. (Figure generated by [5])

We tackled the problem as a regression problem where a Convolutional Neural Network is trained to predict the subject's gaze yaw and pitch angles as continuous values rather than using a classifier to directly predict the correct gaze region. These predicted gaze values are then clustered into 9 predefined classes. Results obtained using this approach were significantly better than the approach of using a classifier, since a pure classification approach to the problem fails to capture the underlying distance between gaze classes i.e. it can confuse two classes that are physically far away from each other, while a regression-based approach can understand and learn the distance between different gaze directions which is evident in our results. Furthermore, clustering the predicted continuous gaze values into classes is more convenient in the context of driving, as predicting the exact gaze angles is not necessary, but rather the region at which the driver is looking.

Additionally, along with the gaze angles, the model is trained to predict the subject's head pose angle as an auxiliary task in a multitask learning fashion, which has further enhanced the predicted gaze angles and made the model more robust to head pose variation without the need of explicitly training another model for head pose estimation like usually done in other related work.

Experimentation and Results

In this section, the dataset used as well as multiple experiments utilizing different approaches and architectures are described and compared.

Dataset

Columbia Gaze dataset [15] is used which contains a total of 5880 images of 56 different subjects (32 male, 24 female) with a resolution of 5184×3456 pixels. The data is diverse containing Asian, White, South Asian, Black, Hispanic and Latino subjects ranging from 18 to 36 years of age, and 21 of them wore prescription glasses. For each subject, there are images for each combination of five horizontal head poses $(0°, \pm 15°, \pm 30°)$, seven horizontal gaze directions $(0°, \pm 5°, \pm 10°, \pm 15°)$, and three vertical gaze directions $(0°, \pm 10°)$, which results in 21 gaze regions per head pose angle.

The limited size of the Columbia Gaze dataset makes it unsuitable for training CNNs from scratch. Pre-training our networks with large datasets seems convenient, however large datasets such as MPIIGaze [20] are comprised of cropped eye images only. Furthermore, they contain very limited head motion due to the nature of the recording environment in which subjects were gazing at a laptop or tablet screen. To overcome this, the pre-trained weights of VGG face descriptor [10] model is used to initialize the weights of the first 4 layers in the proposed architecture. Additionally, we obtain 42000 images using various augmentation techniques including adding Gaussian noise, changing brightness and contrast. Moreover, affine transformations are not applied as it will affect the gaze direction by changing the appearance of the eye. Finally, images are resized to 224×224.

For more practical and reliable results, two test sets are used; one comprising of 6 randomly selected subjects which are excluded from the training set to be used for cross-subject testing, the other is obtained by splitting the images of the remaining 50 subjects into 80% training and 20% testing sets.

Experiments

Four experiments investigating different approaches are conducted, namely classification, regression, feature fusion and lastly our proposed multitask learning approach. The configuration of the baseline model used in our experiments is illustrated in Table 1.

Table 1: Model Configuration

Type	Configuration
Input	224x224x3 image
Convolution Convolution Maxpooling	#maps:64, k:3 x 3, s:1, p:1 #maps:64, k:3 x 3, s:1, p:1 Window:2 x 2, s:2
Convolution Convolution Maxpooling	#maps:128, k:3 x 3, s:1, p:1 #maps:128, k:3 x 3, s:1, p:1 Window:2 x 2, s:2
Convolution Convolution Maxpooling	#maps:256, k:3 x 3, s:1, p:1 #maps:256, k:3 x 3, s:1, p:1 Window:2 x 2, s:2
Convolution Convolution Maxpooling	#maps:256, k:3 x 3, s:1, p:1 #maps:256, k:3 x 3, s:1, p:1 Window:2 x 2, s:2
Flatten	-
Dense Dropout	#hidden units:256 0.5
Dense Dropout	#hidden units:128 0.5
Output	Classification: 9 neurons Regression: 2 neurons

Baseline experiments

Experiment 1. First, a simple classification approach is investigated where the input image is fed to our baseline model illustrated in Table1. The output layer is a softmax layer with 9 output neurons representing the 9 gaze classes. This approach yielded 78.7% on the test set and 64.8% in cross-subject testing. However, it was evident in the confusion matrix that the model confused classes that were physically far from each other, which is expected from a classification approach as the softmax layer has no way of understanding the correlation between classes.

Experiment 2. A regression approach is studied where the same CNN and configuration were used but the output layer was now comprised of 2 neurons representing the gaze yaw and pitch angles as continuous values. While the loss and metric used for evaluating the model is the Mean Squared Error (MSE) for each of gaze angles, we have also used the classification accuracy obtained by clustering the predicted continuous values into the same 9 classes (Fig2) used in the classification experiment to be able to compare the results. The results obtained using this approach are 88.2% on the test and 70.1% in cross-subjects testing which is better than the previous experiment.

Incorporating Head Pose Information

In the following experiments, our aim was to include head pose information in the network. Two approaches are discussed namely feature fusion and multitask learning.

Experiment 3. For feature fusion, a separate network is pre-trained to detect subjects' head pose. We then train the regression-based gaze network from experiment 2 and use the pre-trained head pose network as a feature extractor and concatenate its resulting features with features obtained from the gaze network. Results obtained using this approach did not enhance our previously obtained results.

Experiment 4. In this experiment, a multitask learning approach is investigated, where we train a single network to perform two tasks, predicting gaze yaw and pith angles, and an auxiliary task of predicting subject's head pose as shown in Figure 2. We use the same network configuration in 1 but use two heads i.e. sets of fully connected layers after the flattening layer instead of one; one head predicts yaw and pitch gaze angles and the other predicts the head pose angle as an auxiliary task.
Results from the four experiments are shown in Table 2. As expected, it is clear that, the multitask learning approach yielded the best results of 95.8% accuracy on the test set and 78.2% in cross-subject testing. Furthermore, to better understand what the network has learned, saliency maps[14] with respect to both head pose and gaze are shown in figure 4.

Table 2: Results

Experiment	Accuracy (Test Set)	Accuracy (Cross-subject)
Exp1. Classification	78.7%	64.8%
Exp2. Regression	88.2%	70.1%
Exp3. Feature Fusion	81.1%	66.3%
Exp4. Multitask Learning	**95.8%**	**78.2%**

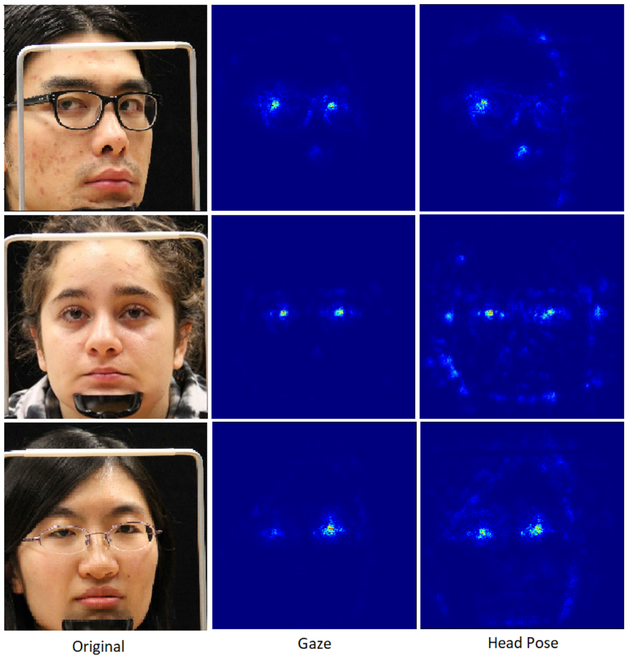


Figure 4: Saliency Maps with respect to both gaze and head pose outputs are shown. Gaze task pays more attention around eyes while head pose task pays more attention around head boundaries and eyes too

Conclusion

In this work an End-to-End Multitask learning solution to gaze estimation is proposed, that uses a single Convolutional Neural Network to predict the subject's gaze direction given an input image containing the subject's face, as opposed to other pipe-lined architectures that require a complex pre-processing and facial feature extraction. The proposed architecture has two main contributions to enhance the accuracy of the detected gaze region. **First**, it is regression-based, i.e. it predicts the subject's gaze yaw and pitch angles as continuous values. This comes from the understanding of the problem which implies that there is an underlying distance between gaze regions that pure classification may fail to capture. **Second**, is using Multitask Learning in which the network is trained to predict the subject's head pose angle as an auxiliary task along with its main task of predicting gaze. Since the appearance of the eye varies with the head pose, training one network on both gaze and head pose estimation tasks simultaneously has proven to enhance the results. Being End-to-End makes the design much simpler and significantly reduces the computational cost that arises when using multiple networks.

References

[1] Gregory M Fitch, Susan A Soccolich, Feng Guo, Julie McClafferty, Youjia Fang, Rebecca L Olson, Miguel A Perez, Richard J Hanowski, Jonathan M Hankey, and Thomas A Dingus. The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk. Technical report, 2013.

[2] Dan Witzner Hansen and Qiang Ji. In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):478–500, 2009.

[3] Qiong Huang, Ashok Veeraraghavan, and Ashutosh Sabharwal. Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets. *Machine Vision and Applications*, 28(5-6):445–461, 2017.

[4] Shreyank Jyoti and Abhinav Dhall. Automatic eye gaze estimation using geometric & texture-based networks. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2474–2479. IEEE, 2018.

[5] Alexander LeNail. Nn-svg: Publication-ready neural network architecture schematics. *Journal of Open Source Software*, 4(33):747, 2019.

[6] Carlos H Morimoto and Marcio RM Mimica. Eye gaze tracking techniques for interactive applications. *Computer vision and image understanding*, 98(1):4–24, 2005.

[7] Rizwan Ali Naqvi, Muhammad Arsalan, Ganbayar Batchuluun, Hyo Sik Yoon, and Kang Ryoung Park. Deep learning-based gaze detection system for automobile drivers using a nir camera sensor. *Sensors*, 18(2):456, 2018.

[8] Basilio Noris, Karim Benmachiche, and Aude Billard. Calibration-free eye gaze direction detection with gaussian processes. In *In Proceedings of the International Conference on Computer Vision Theory and Applications*, number CONF, 2008.

[9] Viral Parekh, Ramanathan Subramanian, and CV Jawahar. Eye contact detection via deep neural networks. In *International Conference on Human-Computer Interaction*, pages 366–374. Springer, 2017.

[10] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[11] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[12] Rajeev Ranjan, Shalini De Mello, and Jan Kautz. Light-weight head pose invariant gaze tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2156–2164, 2018.

[13] Timo Schneider, Boris Schauerte, and Rainer Stiefelhagen. Manifold alignment for person independent appearance-based gaze estimation. In *2014 22nd international conference on pattern recognition*, pages 1167–1172. IEEE, 2014.

[14] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[15] Brian A Smith, Qi Yin, Steven K Feiner, and Shree K Nayar. Gaze locking: passive eye contact detection for human-object interaction. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 271–280, 2013.

[16] Kay Talmi and Jin Liu. Eye and gaze tracking for visually controlled interactive stereoscopic displays. *Signal Processing: Image Communication*, 14(10):799–810, 1999.

[17] Kar-Han Tan, David J Kriegman, and Narendra Ahuja. Appearance-based eye gaze estimation. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pages 191–195. IEEE, 2002.

[18] Laurent Vaissie and Jannick Rolland. Head mounted display with eyetracking capability, August 13 2002. US Patent 6,433,760.

[19] K Preston White, Thomas E Hutchinson, and Janine M Carley. Spatially dynamic calibration of an eye-tracking system. *IEEE Transactions on Systems, Man, and Cybernetics*, 23(4):1162–1168, 1993.

[20] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.

[21] Zhiwei Zhu and Qiang Ji. Eye and gaze tracking for interactive graphic display. *Machine Vision and Applications*, 15(3):139–148, 2004.

[22] Zhiwei Zhu and Qiang Ji. Novel eye gaze tracking techniques under natural head movement. *IEEE TRANSACTIONS on biomedical engineering*, 54(12):2246–2260, 2007.

[23] Zhiwei Zhu, Qiang Ji, and Kristin P Bennett. Nonlinear eye gaze mapping function estimation via support vector regression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1132–1135. IEEE, 2006.

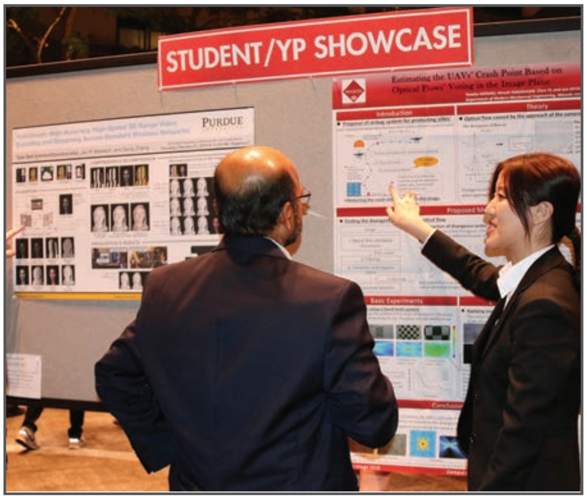
JOIN US AT THE NEXT EI!

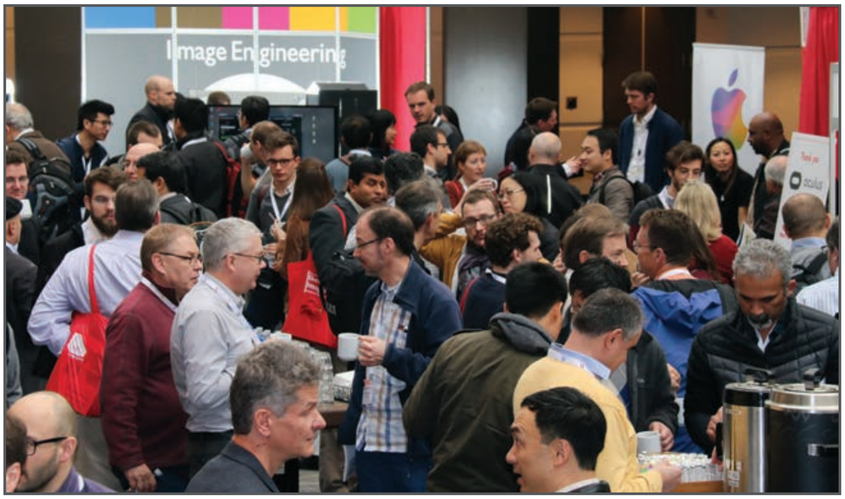
IS&T International Symposium on

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!

My first working AR prototype

Today's platforms

Optimis this is << 20

Ubiquitous consumer AR

• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org