https://doi.org/10.2352/ISSN.2470-1173.2021.8.IMAWM-268

Robust Real-Time Heart Rate Measurement from Face Videos

Yang Cheng [a], **Qian Lin** [b], **Jan Allebach** [a]
[a] **School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, U.S.A.**
[b] **HP Labs, Palo Alto, CA 94304, U.S.A.**

Abstract

Heart rate, the speed of the heartbeat, has been regarded as one of the most important measurements to evaluate one's health. It can be used to measure one's anxiety, stress and illness; abnormalities of heart rate usually indicate potential disease one may have. Recent studies have shown that it is possible to directly measure the heart rate from a sequence of images that contain a person's face. Requiring only a webcam, this method largely simplifies the process of traditional methods, which require the use of a pulse oximeter attached to the fingertip to measure the PPG signal, or electrodes placed on the skin to measure the ECG signal. However, this most recent method, though attracting a lot of interest, still suffers from sudden movement of the head, or turning away from the camera. In this paper, we propose a novel robust method of generating reliable PPG signals and measuring the heart rate from only face videos in real time, which is invariant to the movement of the head. We have also conducted studies on how different factors, light conditions, the angle of the head and the distance of the head away from the camera, could affect the predictions of the heart rate. After conducting a thorough analysis, we can conclude that our method succeeds in producing accurate, robust and promising results.

Introduction

Heart rate measurement is considered as one of the most important measurements, since it is capable of evaluating one's anxiety, stress and illness. A normal resting heart rate lies within the range of 60 beats per minute (bpm) to 100 bpm. A resting heart rate higher than 100 bpm, aka tachycardia, can lead to a possible heart failure or stroke if left untreated. On the other hand, a resting heart rate less than 60 bpm, aka bradycardia, could probably cause a person to experience fatigue or shortness of breath. Until now, a lot of methods can be used to measure the heart rate, including but not limited to placing fingers on the wrist to count the number of heart beats in one minute, using a blood pressure monitor that has the cuff wrapped around the arm, and using an electrocardiograph. However, these methods are either inaccurate or inconvenient. Professional devices are helpful for achieving a more accurate heart rate. But they are undoubtedly cumbersome; some people do not feel comfortable physically contacting the device. With the invention of the pulse oximeter, people nowadays can simply clip this small device on their fingertip to measure the heart rate. The main idea behind the pulse oximeter is photoplethysmography (PPG), which measures the blood volume changes with the light from an LED, making use of the fact that the light is more absorbed by the blood than surrounding tissues [10, 8]. This method is fairly accurate, low-cost, and requires a simple small device; however, physical contact with the device is still unavoidable.

Until today, many works have been introduced to avoid the physical contact [2, 11], and with the explosive growth of deep learning, many recently proposed methods [5, 13, 7] are now driven by the large-scale datasets. These methods usually focus on researching the idea of a remote Photoplethysmogram, or rPPG. Similarly to PPG, the rPPG signal can be used to detect the blood volume changes, but is measured by a webcam that captures the face region instead of a pulse oximeter.

Naturally, several datasets have been collected to facilitate the development of data-driven methods [1, 6]. These datasets consist of videos of faces, each of which is accompanied by a synchronized PPG signal recorded by a professional device, such as a pulse oximeter.

In this paper, we present our work in measuring the remote PPG using a deep-learning-based method. We also conduct experiments to analyze the effects of different conditions on the performance of existing methods to show that our method achieves the best accuracy in predicting the heart rate.

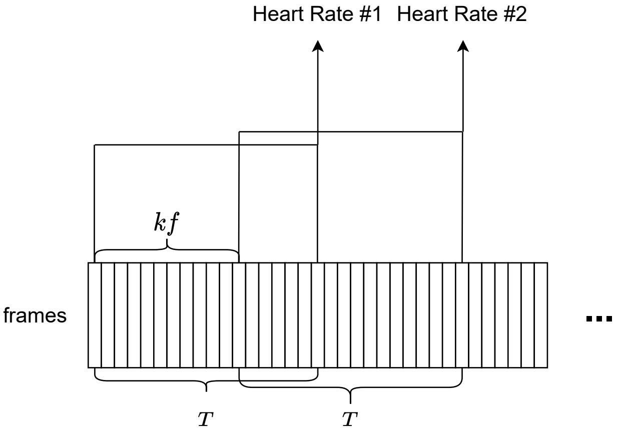


Figure 1. *Illustration of our heart rate estimation system. T denotes the number of frames that are used to calculate one heart rate reading, f denotes the frame rate of the webcam, and k denotes the interval of heart rate readings in seconds, e.g., $k = 1$ indicates that one heart rate reading is produced every second. Because the model needs T frames to produce a heart rate reading, the subjects need to wait for T frames or $\frac{T}{f}$ seconds to get the first heart rate reading.*

Related Work

To solve the problem of remotely measuring the heart rate, numerous different methods were proposed, while the majority of them follows these two steps: (1) extract an 1D signal that contains a strong pulsatile component, and (2) calculate the heart rate in beats per minute by either counting the peaks existing in the sig-

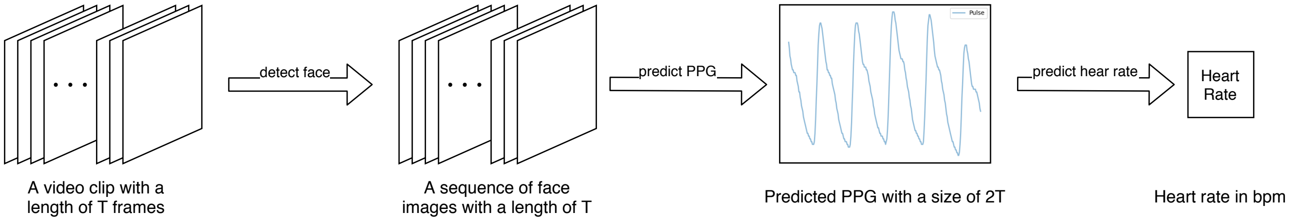

detect face
predict PPG
Pulse
predict hear rate
Heart Rate
A video clip with a length of T frames
A sequence of face images with a length of T
Predicted PPG with a size of 2T
Heart rate in bpm

***Figure 2.** Overview of the system for predicting the heart rate from a video clip.*

nal or locating the dominant frequency by conducting frequency analysis on the signal. Traditional methods include chrominance-based rPPG [2] and plane-orthogonal-to-skin (POS) [11]. Recently, many deep-learning-based methods have emerged. Niu et al. [5] proposed a spatial-temporal map to measure the heart rate. Yu et al. [13] proposed a 3D-CNN network to predict the heart rate and used the negative Pearson correlation as the loss function to train the network.

Methodology

Suppose that there is a webcam that is capable of capturing still images of the face at f frames per second (FPS), our method then estimates one measurement of heart rate from the video clip of T frames. If we denote k as the interval of heart rate readings in seconds, then one heart rate reading is produced every k seconds. The overall system is illustrated in Figure 1.

Our method of estimating the heart rate from a video clip of T frames can be summarized into a three-step process: (1) detecting the face that appears in every frame of the video and assigning a bounding box that encloses the face region, (2) using our trained model to predict the plethysmograph (PPG) from the sequence of face images, and (3) conducting frequency analysis on the signal to calculate the heart rate. Figure 2 illustrates the system for predicting the heart rate from a video clip of T frames.

Face Detection

The very first step of detecting the heart rate is to locate the face region where the color variation of the skin can be extracted. We adopted MTCNN [14] with pre-trained weights to detect the face that appears in the video. An illustration of the face region detection is shown in Figure 3. To further improve the efficiency of our method and reduce detected face bounding box's jittering, we chose not to detect the face in every frame; instead, we applied MTCNN on the n^{th} frame once and use the detected bounding box to crop the following $n+1^{\text{th}}$, $n+2^{\text{th}}$, $\ldots$, $n+X^{\text{th}}$ frames, where X depends on the frame rate of the camera. If we assume that a person's head should stay relatively still within 0.1 seconds, then

$$X = \lfloor 0.1 \times \text{FPS} \rceil \tag{1}$$

where $\lfloor \cdot \rceil$ denotes rounding to nearest integer. After the face regions of all frames in the video are determined, they should be cropped out from the frames to be used by the PPG prediction model.

PPG Prediction

After T consecutive face images have been collected, their color space is firstly linearized, and then converted from RGB to $L^*a^*b^*$ (reference white: D65). A previous study by Yang et al. [12] showed that it is the image intensity that is primarily affected by the head movement instead of the image chromaticity. Therefore, by considering only the chromaticity channels a^* and b^*, we should be able to minimize the loss in accuracy caused by the head movement. The L^*, a^* and b^* channels of a face image are shown in Figure 4.

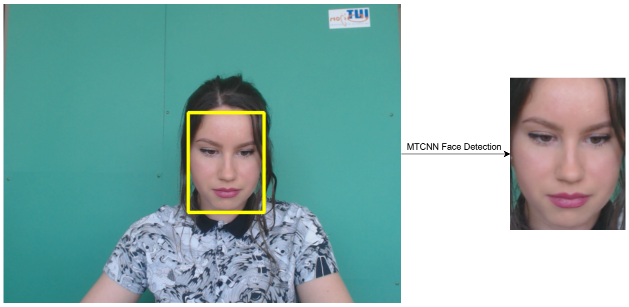


***Figure 3.** Face detection was applied to remove any irrelevant background.*

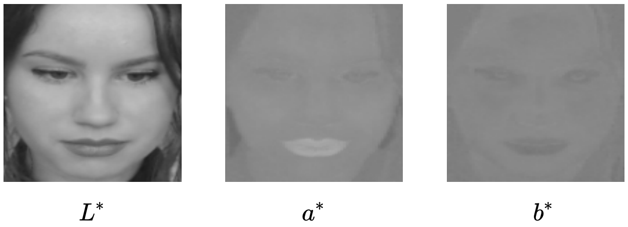


***Figure 4.** The L^*, a^* and b^* channels of a face image. A previous study showed that it is the image intensity that is primarily affected by the head movement instead of the image chromaticity.*

As illustrated in Figure 5, our PPG prediction model is capable of predicting a PPG signal of $2T$ samples from a video clip of T frames, i.e., if the frame rate of the video clip is f, then the sampling frequency of the output signal is $2f$. The details of our model will be discussed in PPG Prediction Model section. To determine the exact value of T, we should consider how it would affect the predicted heart rate. It should not be set too large, otherwise the sudden change of the heart rate might not be captured, aka poor time resolution. But it also should not be set too small, or the predicted heart rate might not be accurate, aka poor frequency resolution. Our setting of T differs in the training stage and the inference stage, which will be discussed in the Experiments section.

Heart Rate Calculation

To extract the heart rate from the PPG, we apply the Fourier Transform to the signal to represent it in the frequency domain.

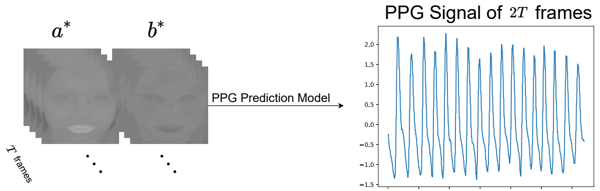


Figure 5. *Our model is capable of predicting a PPG signal of $2T$ samples from the video clip of T frames in the a^* and b^* channels.*

A band-pass filter is then applied to filter out any frequency that is below 0.9 Hz or above 3 Hz, since a normal person's heart rate is usually within the range from 54 bpm (beats per minute) to 180 bpm. Lastly, the dominant frequency, the frequency which corresponds to the highest peak, is chosen as the heart rate. The process is illustrated in Figure 6.

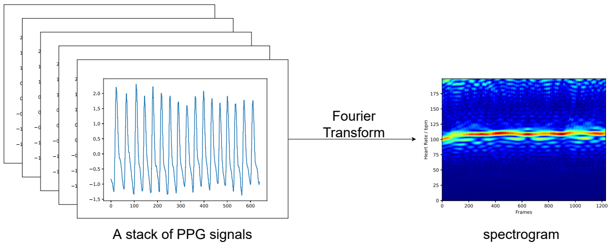


Figure 6. *The spectrogram is formed by taking Fourier Transform of each of the PPG signals. The dominant frequency is chosen as the heart rate.*

PPG Prediction Model

We designed a 3D-CNN [4] network structure that is suitable for predicting an 1D output (PPG signal) from a 3D input (video clip) and it performs better than existing methods in practice. Also, the network is lightweight enough for being used in real time. It takes only 0.2 seconds to predict one measurement of the heart rate from a video clip, thus making it satisfy the requirement of one heart rate reading per second. Figure 7 shows our network structure. Although not fully detailed, it illustrates the main ideas used in our network design. To simplify the plot, the number of feature channels is ignored. In the figure, T represents the size in the time dimension; N represents the size of the feature maps. For example, in the input video clip, T is the number of frames and N is the size of the frames; in the output, T is the number of samples in the PPG signal and $N = 1$.

The 3D-CNN network shown in Figure 7 resembles the U-Net network structure [9], but it is the size in the time dimension that is downsampled and upsampled, not the feature map size as in U-Net. The paths, represented by the green right arrows, follow the typical structure of a convolutional neural network (CNN) to extract the high-level features while keeping the size in the time dimension unchanged. Another path, represented by the blue arrows, is the contracting path where the size in the time dimension of the feature maps and the size of the feature maps are downsampled at the same time. The purpose of downsampling in the time dimension is to reduce any temporal noise and redundancy [13]. The last path, represented by the red up arrows, is the expansive path where the features are upsampled in the time dimension to reconstruct the PPG signal. The high-level features are combined with the upsampled output by the add operator in the expansive path, so that high-level features obtained from features at different temporal scales are all included. To downsample the size of the feature maps or the size in the time dimension, we apply the Max Pooling operator; to upsample the feature maps in the time dimension, we use Transposed Convolution [3].

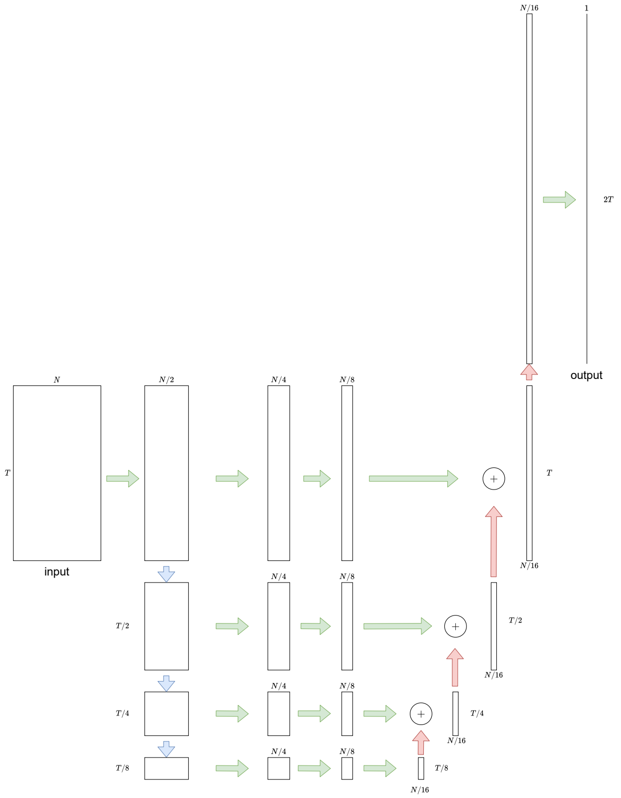


Figure 7. *The 3D-CNN network structure for predicting PPG from the video clip.*

Experiments

Dataset

To train, validate and test our PPG prediction model, we used the public dataset named UBFC-RPPG Dataset [1]. To further study how the head movement and the lighting condition affect the prediction accuracy, we collected our own dataset. An overview of these two datasets is shown in Table 1.

UBFC-RPPG Dataset [1] is one of the public datasets that is specifically for rPPG analysis. The videos were recorded using a Logitech C920H HD Pro webcam at 30 FPS with a resolution of 640 × 480 in uncompressed 8-bit RGB format. Corresponding PPG signals were recorded using a CMS50E pulse oximeter with a sampling frequency of 60 Hz. It consists of two sub-datasets that differ from how subjects were asked to behave during recording. The first sub-dataset contains 8 videos and each video has a length of approximately 80 seconds. Subjects were asked to sit still and be relaxed. The second sub-dataset includes 42 1-minute-videos, and subjects were asked to play a time sensitive mathematical game.

Table 1: An overview of two datasets used.

	UBFC-RPPG Dataset	Our Own Dataset
PPG Recording Device	CMS50E Pulse Oximeter / 60 Hz	CMS50D+ Pulse Oximeter / 60 Hz
Video Recording Device	Logitech C920H HD Pro / 30 FPS / 640 × 480 / Uncompressed	Logitech C930E / 30 FPS / 640 × 360 / JPEG
Lighting Conditions	Natural Light	Ranging from 2700 K to 5000 K
Contains Head Movement?	No	Yes
Subjects	Relaxed / Playing Math Games	Relaxed

Our Own Dataset: To evaluate our model and compare it further with existing methods in more complex settings, we collect our own dataset which contains some different settings: varying lighting conditions, with/without head movement. The videos were recorded using a Logitech C930E at 30 FPS with a resolution of 640×360 in JPEG format, and the length of each video is about 1 minute. The PPG signals were recorded using a CMS50D+ pulse oximeter with a sampling frequency of 60 Hz.

Training the Model

To train the model, the dataset was pre-processed so that each video clip consists of 96 face images, i.e., $T = 96$. Then the color temperature of the video clip was randomly changed to simulate different lighting conditions. After that, the color space of the video clip was converted from RGB to $L^*a^*b^*$ and only the a^* and b^* channels were kept. Finally, each frame of video clips was resized to 128×128, i.e., the input dimension is $96 \times 2 \times 128 \times 128$. The output dimension is then 192×1. In the training stage, the Adam optimizer was used, the learning rate was set to 0.001 and the number of epochs was set to 40.

Loss Function

We adopted the Negative Pearson Correlation used in [13] as the loss function to train the network. If x is the ground-truth PPG signal, $\hat{x}$ is the predicted PPG signal, and N is the length of the signal, then the loss function is given by

$$\mathscr{L} = 1 - \frac{\sum_{t=1}^{N}\left(\hat{x}_t - \bar{\hat{x}}\right)\left(x_t - \bar{x}\right)}{\sqrt{\sum_{t=1}^{N}\left(\hat{x}_t - \bar{\hat{x}}\right)^2}\sqrt{\sum_{t=1}^{N}\left(x_t - \bar{x}\right)^2}} \qquad (2)$$

Testing the Model

To test the model, we pre-processed the input video so that each video clip consists of 320 face images, i.e., $T = 320$. Under this setting, the subjects only need to wait for $\frac{320}{30} \approx 10.67$ seconds to get their first heart rate reading on a webcam running at 30 FPS. Then the color space of video clips was converted from RGB to $L^*a^*b^*$ and only the a^* and b^* channels are used. Lastly, each frame of video clips was resized to 128×128. Therefore, the dimension of the input, or the video clip, is $320 \times 2 \times 128 \times 128$, and the dimension of the output, or the PPG signal, is 640×1. Finally, to retrieve the heart rate from the PPG signal, we compute the Fourier Transform to locate the dominant frequency. However, to ensure that the frequency resolution is 1 bpm or $\frac{1}{60}$ Hz, we zero padded the predicted PPG signal so that it has 3600 samples, then the frequency resolution is given by $\Delta f = \frac{60}{3600} = \frac{1}{60}$ Hz, assuming that the frame rate of webcam is 30 Hz, so the sampling rate of the predicted PPG signal is 60 Hz, since our model is capable of predicting a PPG signal of $2T$ samples from a video clip of T frames.

Metrics

To make the performance results more intuitive, we applied frequency analysis on both the ground truth and the predicted PPG to extract the heart rate measurements.

Let $y \in \mathbb{R}^N$ be the vector of ground truth heart rate measurements in bpm from a video, $\hat{y} \in \mathbb{R}^N$ be the vector of predicted heart rate measurements in bpm, and N be the number of heart rate measurements from a video. The following two metrics were used to evaluate the methods:

- Root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{N}\left(\hat{y}_t - y_t\right)^2}{N}} \qquad (3)$$

- Mean absolute error (MAE):

$$\text{MAE} = \frac{\sum_{t=1}^{N}\|\hat{y}_t - y_t\|}{N} \qquad (4)$$

Table 2: Performance comparison of existing methods and our method on the second sub-dataset in the UBFC-rPPG dataset.

	RMSE	MAE
POS [11]	6.92	3.77
Ours	**5.03**	**3.00**

Table 3: Performance comparison for the effect of head movement.

	w/o head movement		w. head movement	
	RMSE	MAE	RMSE	MAE
POS	**0.62**	**0.37**	3.91	2.15
Ours	0.76	0.56	**1.64**	**0.82**

Results

To compare the performance of an existing method and our method, we used them to predict the heart rate in videos from the second sub-dataset of UBFC-rPPG dataset, and compared it

Table 4: Performance comparison for the effect of lighting conditions

	2700 K		3300 K		3900 K		4500 K		5000 K	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
POS [11]	**1.03**	0.58	2.01	1.04	2.45	1.03	2.49	1.41	**0.62**	**0.37**
Ours	**1.03**	**0.57**	**1.32**	**0.88**	**1.38**	**0.81**	**1.89**	**1.02**	0.76	0.56

with the ground truth to calculate the root mean squared error (RMSE) and the mean absolute error (MAE). Table 2 shows that our method achieves better results for predicting the heart rate than POS [11] does.

To further analyze whether or not our method still achieves the best results under more complex situation, we designed experiments using our own dataset to show if the following conditions affect the performance of methods or not:

- with/without head movement
- varying lighting conditions

To study the effect of head movement, the subjects were asked to measure the heart rate with their head still, then measure again with their head moving. The color temperature was fixed at 5000 K; the subjects were close to the webcam and they were relaxed. Table 3 shows that our method is much less affected by the head movement.

To study the effect of varying lighting conditions, the subjects were asked to measure the heart rate 5 times at 2700 K, 3300 K, 3900 K, 4500 K and 5000 K. They were asked to keep their heads still, sit close to the webcam, and be relaxed. Table 4 shows that our method is less affected by the varying lighting conditions.

Conclusion

In this paper, we presented a method to predict the heart rate of a person robustly by simply using a webcam without physically contacting the person. We created our own dataset and conducted experiments to investigate the effects of head movement, varying lighting conditions and distance from the webcam on the performance of our method. The results show that our method achieves a better accuracy of measuring the heart rate, and is less affected by complex conditions.

References

[1] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82 – 90, 2019. Award Winning Papers from the 23rd International Conference on Pattern Recognition (ICPR).

[2] G. de Haan and V. Jeanne. Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

[3] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *ArXiv e-prints*, mar 2016.

[4] S. Ji, W. Xu, M. Yang, and K. Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[5] X. Niu, H. Han, S. Shan, and X. Chen. SynRhythm: Learning a Deep Heart Rate Estimator from General to Specific. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3580–3585, 2018.

[6] X. Niu, H. Han, S. Shan, and X. Chen. VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision*, pages 562–576. Springer, 2018.

[7] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, and X. Chen. Robust Remote Heart Rate Estimation from Face Utilizing Spatial-temporal Attention. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–8, 2019.

[8] E. A. Pelaez and E. R. Villegas. LED power reduction trade-offs for ambulatory pulse oximetry. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2296–2299, 2007.

[9] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[10] K. Shelley and S. Shelley. *Pulse Oximeter Waveform: Photoelectric Plethysmography*, pages 420–423. 01 2001.

[11] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan. Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017.

[12] Y. Yang, C. Liu, H. Yu, D. Shao, F. Tsow, and N. Tao. Motion robust remote photoplethysmography in CIELab color space. *Journal of Biomedical Optics*, 21(11):1 – 6, 2016.

[13] Z. Yu, X. Li, and G. Zhao. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks, 2019.

[14] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

Author Biography

***Yang Cheng** is a Ph.D. candidate at Purdue University. He received his B.S. in Electrical Engineering from Purdue University (2018) and is currently pursuing Ph.D. in Electrical Engineering from Purdue University. His research areas of interest includes image processing and computer vision.*

***Dr. Qian Lin** is a research scientist working on computer vision and deep learning research in HP Labs. She is also an Adjunct Full Professor of Electrical and Computer Engineering, Purdue University. She received her BS from Xian Jiaotong University in China, her MSEE from Purdue University, and her Ph.D. in EE from Stanford University. Dr. Lin is the inventor/co-inventor for 44 issued patents. She was awarded Fellowship by the Society of IS&T in 2012, and Outstanding Electrical Engineer by the School of ECE of Purdue University in 2013. She was promoted to the rank of HP Fellow in 2019.*

***Jan P. Allebach** received his B.S. from the University of Delaware in 1972, his M.S. from Princeton University in 1975* *and his Ph.D. from Princeton University in 1976. He is now the Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University. Allebach is a Fellow of the IEEE, the National Academy of Inventors, the Society for Imaging Science and Technology (IS&T), and SPIE. He was named Electronic Imaging Scientist of the Year by IS&T and SPIE, and was named Honorary Member of IS&T, the highest award that IS&T bestows. He has received the IEEE Daniel E. Noble Award, and is a member of the National Academy of Engineering. He recently received the OSA/IS&T Edwin Land Medal and the IS&T Johann Gutenberg Prize. He recently served as an IEEE Signal Processing Society Distinguished Lecturer (2016-2017). His current research interests include image rendering, image quality, color imaging and color measurement, printer and sensor forensics, and digital publishing.*

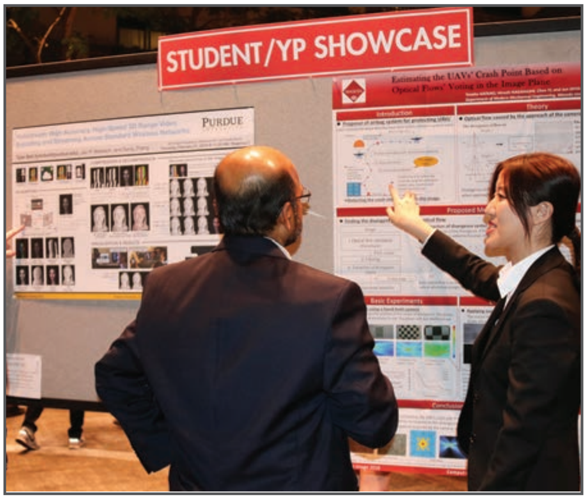
JOIN US AT THE NEXT EI!

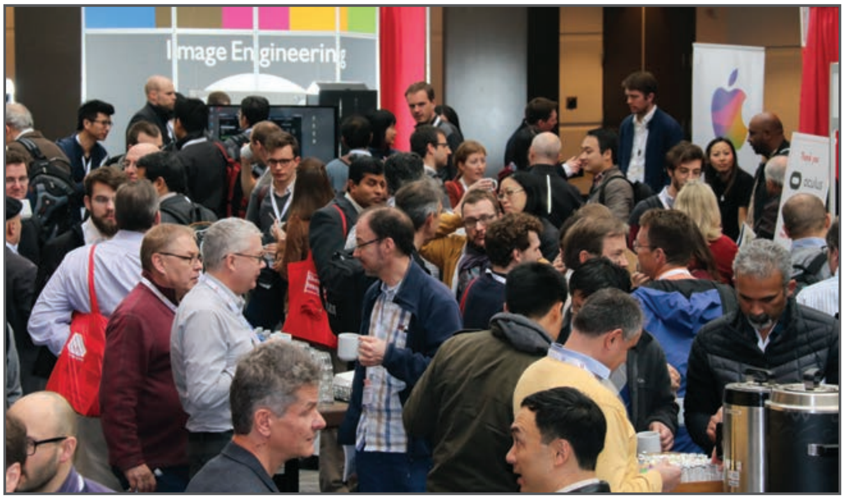
IS&T International Symposium on

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