https://doi.org/10.2352/ISSN.2470-1173.2021.8.IMAWM-267

Remote estimation of respiration rate by optical flow using convolutional neural networks

Tianqi Guo[1]***, Qian Lin***[2]***, and Jan Allebach***[1]
[1] ***School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN***
[2] ***HP Labs, HP Inc., Palo Alto, CA***

Abstract

In this paper, we propose a novel system for remotely estimating the respiration rate of people. Periodic inhalation and exhalation during respiration cycles induce subtle upper body movements, which are reflected by the local image deformation over time when recorded by a digital camera. This local image deformation can be recovered by estimating the optical flow between consecutive frames. We propose the usage of convolutional neural networks designed for general image registration to estimate the induced optical flow, the periodicity of which is then leveraged to obtain the respiration rate by frequency analysis. The proposed system is robust to lighting condition, camera type (RGB, infrared), clothing, and posture (sitting in chair/lying in bed); and it could be used by individuals with a webcam, or by healthcare centers to monitor the patients at night.

Introduction

Vision-based estimation and monitoring of vital signs, including heart rate (HR), blood pressure (BP), respiration rate (RR), and body temperature, has gained increasing attention in the past few years, thanks to the advances in digital cameras and computing power [1].

In particular, the remote measurement of RR has proven critical in clinic settings including the prevention of respiratory impairment in the post-anesthesia care unit [2], early detection of abnormal respiratory rhythm in the neonatal intensive care unit [3, 4], and rapid and reliable assessment of patient in the emergency triage room for lowering the workload of nurses [5]. Moreover, systems and mobile applications have also been developed for RR monitoring during stationary bike telerehabilitation sessions in home settings [6], and for outdoor usage in mobile situations [7].

Traditional RR estimation methods, for example respiration belts or nasal probes, usually require the subject to stay relatively stationary during measurement, and wear the equipment with wires, which may cause discomfort and disturb the natural breathing pattern. And often times, special trainings are necessary to faithfully perform the measurement and obtain reliable readings. On the contrary, vision-based techniques are unobtrusive, contact-free, and are able to deliver real-time RR reading simply by recording via a regular webcam or thermal imager without extra hardware or operations.

In this paper, we propose a novel vision-based system for remote RR estimation by individuals with a webcam, or by healthcare centers to monitor the patients. Our method has the following contributions:

1. Our method is based on general image registration, which could be applied to any type of camera videos (RGB, near infrared, thermal), as long as there are visually noticeable changes in consecutive frames.
2. Our method can be used with customer level RGB cameras, which are more affordable than thermal imagers.
3. Our method does not depend on a remote photoplethysmography (rPPG) signal and directly estimates the motion of body, which is easy to implement and does not require a lot of tuned parameters.
4. Our method uses convolutional neural networks (CNN) for estimating the optical flow, which is robust to clothing, posture, and lighting conditions.

This paper is structured as follows. First, related works using vision-based techniques for remote RR estimation will be reviewed. Next, we give a detailed description of the proposed processing pipeline. Error analysis and ablation study will be performed on public datasets. Finally, we demonstrate that our C++ implementation can achieve real-time RR estimation with a decent GPU.

Related works

RR estimation using a thermal imager

Traditional RGB cameras collect photons reflected from objects for imaging in the visible range, while thermal imagers visualize emissive radiation from objects, which do not require additional lighting sources. The alternating cold and warm air flows through the nasal passages during inhalation and exhalation cycles have distinct thermal signatures, and this periodic temperature change can be leveraged for RR estimation.

Algorithms developed for RR estimation using a thermal imager usually start with manual initialization of the region of interest (ROI) [2, 5, 6, 7, 8, 9], or automatic detection of the nostril region [10]. Next, the subject is either asked to remain still [9], or the ROI is tracked by image registration [11] based on minimum eigenvalue features [12], off-the-shelf trackers [13, 14], gradient-based normalized cross-correlation [7], template matching [5], or a pan-tilt mechanism [6].

Next, the respiration signal can be recovered from the thermal footprints by either simply looking at a single point of interest [5], or taking the averaged pixel brightness within the tracked ROI [2, 6, 8, 10]. More sophisticated methods for respiration signal extraction include thermal voxel integration [7] over the nostril cross section, or reconstructing the minimum-temperature envelope for all brightness-varying pixels inside the ROI [9].

Finally, from the constructed respiration signal, the RR is

calculated by peak-counting in the temporal domain [5, 8, 9], or by peak-detection in the Fourier spectrum after a band-pass filtering [6]. Alternatively, methods including short-time autocorrelation function estimation [7, 15] and sliding short-window analysis [2, 10] can deliver time-dependent RR over time after proper windowing.

Although RR estimation using a thermal imager has the advantage of direct temperature measurement, and can be deployed in total darkness without any additional illumination source, it also suffers from several drawbacks. One critical aspect for the faithful construction of respiration signal is the proper detection, tracking, and alignment of the nostril region. Most of the current methods still rely on manual selection of the ROI and tracking based on hand-picked features, and they can only deal with the subject looking directly into the thermal imager in laboratory settings. However, in real world scenarios, motions like head turning, sleeping on stomach or side, switching between nasal breathing and mouth breathing may cause misalignment or loss of the ROI. Moreover, while typical costs for a spirometer for breath monitoring at home are below $100 and respiration belts for clinical usage are below $500, even an entry-level thermal imager with limited pixel resolution costs more than $500, preventing its wider usage for RR estimation.

RR estimation using an RGB or NIR camera

RGB and near-infrared (NIR) cameras of customer level quality have both been used by previous researchers to estimate RR, as a cheaper alternative to the thermal imagers.

Methods based on brightness changes

One way to extract the respiration signal from RGB videos is to directly look at the values in the 3 color channels. During respiration cycles, reflected light from the moving chest wall is collected by the RGB camera, and the brightness in each color channel varies periodically with the chest movement. Massaroni et al. [16, 17] proposed an algorithm to extract the image brightness variation around the neck region, and constructed a respiration signal as the total brightness sum of the pixels with top 5% brightness variations over time. The RR was then calculated from a simple zero-crossing counting of the signal, and compared with sound waves captured by a headphone. Similarly, Jorge et al. [3] looked at the color variation on the back skin of newborn babies in clinical settings. Frame-wise brightness differences were accumulated within segmented skin regions over every 10-second window as the respiration signal. The RR was then estimated by a 5-th order auto-regressive model [18], and compared with impedance pneumography. Although methods based on brightness changes of RGB cameras are usually straight-forward, they rely heavily on accurately determining the skin region, and may suffer when the ambient lighting is not optimal.

Methods based on remote photoplethysmography (rPPG)

Another popular approach to construct the respiration signal is to leverage algorithms already developed for rPPG. Although rPPG was principally developed for measuring the heart rate [1], RR can be calculated from the rPPG signal by relatively sophisticated methods. van Gastel et al. [4] argued that as respiration causes blood pressure variations, the volume in the veins is also affected such that the rPPG signal is modulated in frequency, amplitude, and overall shape. By applying color transforming weights for RGB or NIR images [19, 20], the rPPG signal is constructed, and the RR can be detected in the lower range (10-40 breaths/min) of the Fourier power spectrum after proper scaling. More recently, Wei et al. [21] proposed the usage of blind source separation [22] to detect both RR and HR from the decomposition of a 6-channel input signal. The concatenated input signal consists of the spatially averaged RGB values from the mouth and neck regions over time. Finally, the RR was estimated from the frequency domain by a peak detector. Systems with NIR or night-vision cameras are of particular interest for RR estimation in dark settings, where lighting is usually limited or undesired. A Eulerian video magnification (EVM) framework [23] was utilized by He et al. [24] to magnify the illumination changes caused by blood flow and breathing in a small area around the neck region in NIR videos. The resulting RR from frequency analysis was compared with a respiration belt. Algorithms in this category typically deliver HR and RR in one shot, and can be used with NIR cameras at night. However, they usually involve extensive hand tailoring of parameters and carefully-tuned thresholds along the processing pipeline.

Methods based on optical flow

Alternatively, respiration signal can also be extracted from body motions induced by respiration cycles. Lin et al. [25] estimated the vertical chest motion over respiration cycles, and the median of the optical flow at each time step was extracted as the respiration signal. The number of completed respiration cycles was identified by counting zero-crossings. The proposed method was demonstrated for both sitting-up and lying-in-bed settings, and was compared with a respiration belt. Similarly, Chatterjee et al. [26] proposed an iterative algorithm to estimate the principle flow field around the thoraco-abdominal region in online mode. The respiration signal was constructed from the phase-synchronized principle flow field for every 12-second window. The dominant frequency in the Fourier spectrum of the signal was taken as the RR and compared with impedance pneumography. However, both algorithms were based on the vanilla optical flow method [27], which could only estimate the motion along the local brightness gradient. If the subject wears clothes of less visible texture, or if the video is contaminated with noise due to poor ambient lighting, the estimated optical flow might not describe the body motion faithfully.

Method

Benchmark datasets

For proof-of-concept demonstration and error analysis, we chose two public datasets. Please note that, in those two datasets RR and HR are both provided as the ground truth, but the latter was not used in the current study. Sample frames from the datasets are shown in Figure 1.

COHFACE dataset

In the COHFACE dataset [28], subjects were asked to look into a RGB webcam connected to a laptop for approximately 60 seconds. The dataset consists of 40 individual subjects, and each subject recorded 4 videos under different conditions (lighting, respiration pattern, etc.). There are 160 video clips in total with simultaneous thoracic stretch measured with a respiration belt (Fig-

ure 1-b). The ground truth RR for each video is obtained from the PSD of those belt measurement over the entire 60 s, assuming a constant RR. The videos in this dataset suffer from severe compression artifacts, and show structured noise in the background.

Sleep dataset

As a complementary imaging modality, the Sleep dataset [29] consists of 28 60-second video clips where subjects were lying in a bed in laboratory setting. A dual-camera imaging system was used to simultaneously capture thermal and NIR videos. The ground truth of RR was obtained by a qualified human observer. The thermal videos in this dataset were not in the raw single-channel format (temperature) from the thermal imager, but rendered in false RGB color (Figure 1-d). Therefore, only the NIR videos are used for current study.

Working principle

Periodic inhalation and exhalation during respiration cycles are associated with volume changes of lungs, as well as the expansion and contraction of the anteroposterior diameters of the rib cage and abdomen [30]. Among traditional respiration measurement techniques, impedance pneumography directly captures the volume changes, while other respiration belts with accelerometers, force sensors, or pressure sensors are designed to sense the motions of the chest wall.

When imaged by a digital camera over time, the periodic motions of the chest wall and upper body are reflected by the local image deformation between consecutive frames. In principle, any image registration techniques designed to restore local deformations should capture periodic changes in the deformation matrix when applied to such videos. The overall processing steps of the proposed algorithm are shown in Figures 2 (a)-(f), which will be introduced in the following sections.

FlowNet for optical flow estimation

Optical flow

When first introduced in the 1980s, optical flow was used to describe brightness variation in an image by analogy with a flow field [27]. Let $E(x,y,t)$ be the brightness at the point (x,y) in the image at time t. By analogy to the Navier-Stokes equations, the conservation equation of image brightness during motion of the pattern is

$$\begin{aligned}\frac{DE}{Dt} &= \nabla\cdot(E\vec{u}) + \frac{\partial E}{\partial t} \\ &= \frac{\partial E}{\partial x}u + \frac{\partial E}{\partial y}v + \frac{\partial E}{\partial t} = E_x u + E_y v + E_t = 0\end{aligned} \tag{1}$$

under the assumption of local coherent and rigid motion. Here $\frac{D\cdot}{D\cdot}$ denotes the total (material) derivative, and $\vec{u} = (u,v)$ is the local velocity (displacement) vector. At each pixel, the local brightness gradient (E_x, E_y) and temporal derivative E_t can be determined from the image sequence, which further gives the projected velocity along the local brightness gradient (E_x, E_y) as

$$U_{projected} = \frac{-E_t}{|(E_x,E_y)|} = \frac{-E_t}{\sqrt{E_x^2+E_y^2}} \tag{2}$$

Sample results are shown in Figures 3 (a)-(d). Please note how vectors are only visible where local gradient is large, and the vectors are always along the local brightness gradient regardless of actual motion directions.

To fully solve the under-determined system and obtain the two unknowns $\vec{u} = (u,v)$, Lucas et al. [11, 31] proposed to form an over-determined system by assuming all pixels in a 3×3 neighborhood S share the same local displacement. The over-determined system can be solved by a simple least-square fitting

$$\vec{u} = \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i (E_x^2)_i & \sum_i (E_x E_y)_i \\ \sum_i (E_y E_x)_i & \sum_i (E_y^2)_i \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i (E_x E_t)_i \\ -\sum_i (E_y E_t)_i \end{bmatrix} \tag{3}$$

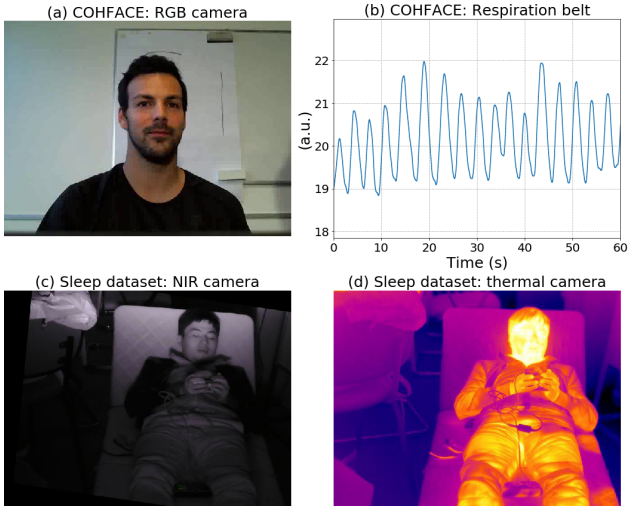


Figure 1. *Sample frames from the public datasets used in the current study.*

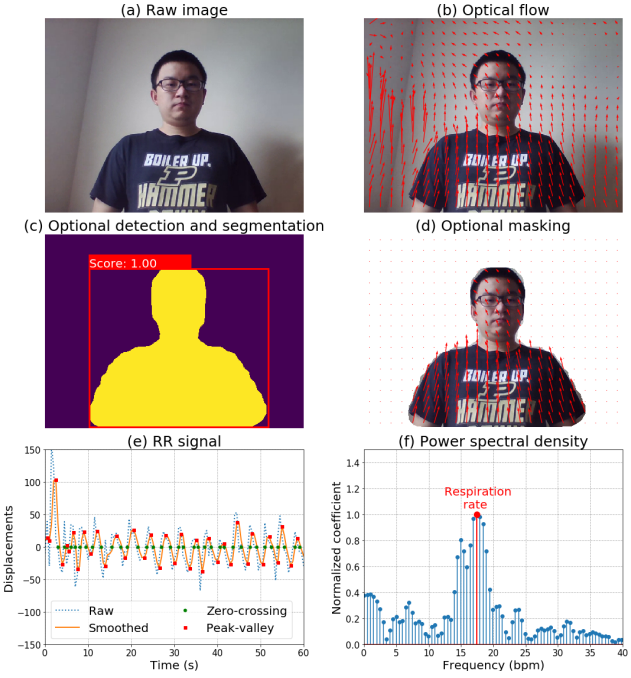


Figure 2. *Processing steps of the proposed RR estimation algorithm.*

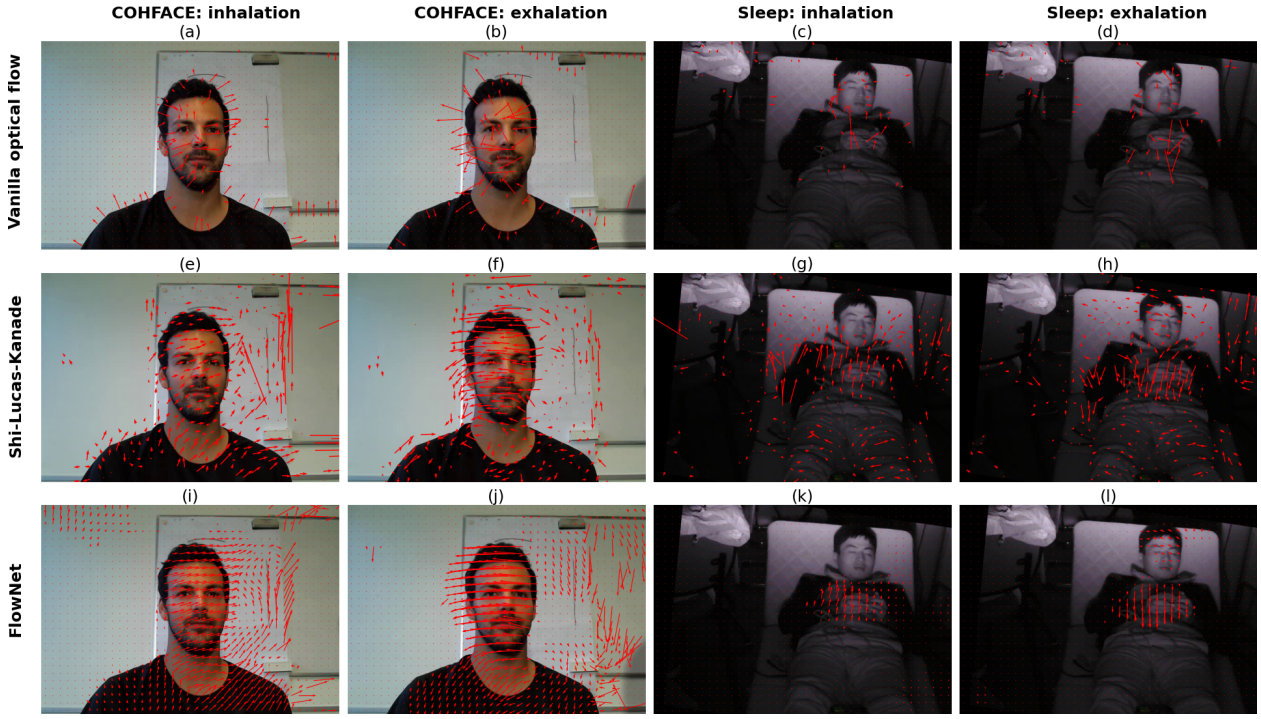


Figure 3. *Example outputs of different optical flow methods (first row: vanilla optical flow, second row: Shi-Lucas-Kanade, third row: FlowNet), all during representative inhalation and exhalation cycles. Please note that for vanilla optical flow and FlowNet the per-pixel results were interpolated onto* 16×16 *grids for better visual presentations.*

where $i \in S$ denotes all pixels in S centered at the pixel of interest. However, both the vanilla optical flow method and the improved Lucas-Kanade method heavily depend on the local brightness gradients to be prominent for reliable estimation. As a common practice, those methods are only applied to regions identified by some pre-processing methods[12], e.g. a Harris corner detector. Figure 3 (e-h) show example outputs from the Shi-Lucas-Kanade method in OpenCV [32]. The threshold for feature detection was set extremely low to track more feature points. Please note that only very few vectors are visible over the black T-shirt where less texture is available.

FlowNet

An alternative approach for reliably estimating the local image deformation is to utilize a CNN. In recent years, CNNs have proven successful in several aspects of computer vision, including image classification [33, 34], object detection [35, 36], pose estimation, and action recognition [37, 38], as well as dense prediction tasks like semantic and instance segmentation [39, 40], and optical flow estimation [41, 42].

Here, we propose the usage of pre-trained FlowNet2SD (small displacement) network [43] implemented in PyTorch [44] from NVIDIA for extracting respiration signals from the periodic upper body movement. FlowNet [41] takes one pair of input images, and predicts the local displacements at each pixel. It has an encoder-decoder architecture that consists of two distinct parts:

1. A contraction part that first extracts feature representations from two input images and reduces the spatial resolution through consecutive convolution, activation, and pooling layers. Then, the two feature maps are passed to a cross-correlation layer that recovers the spatial correspondence between two input images. The joint feature maps then go through deeper ConvLayers for higher levels of feature encoding.
2. An expansion part that takes the joint feature maps from the contraction part as input, and gradually predicts the local deformation vectors and recovers the spatial resolution by consecutive upconvolution and unpooling layers. Skip connections to the corresponding feature maps from the contraction part at each resolution level are also used for better preservation of fine local details.

Although the original FlowNet was trained on artificially generated unrealistic datasets (for example, flying chairs rendered on an arbitrary background), we shall see later it can recover optical flow from real-world videos, thanks to the strong feature extraction and representation power of CNNs. The dense prediction of optical flow from FlowNet is shown in Figures 3 (i)-(l) for visual comparison with the two other methods.

Detectron2 for segmentation mask

Optionally, we have integrated a person detection and segmentation module. We opted to use the `mask_rcnn_R_50_FPN_1x` pretrained model from the Detectron2 [45] model zoo. The PyTorch implementation of Mask R-CNN [39] has a multi-scale feature pyramid network [36] based on a 50-layer ResNet [46] backbone. When the segmentation mask is used, the optical flow is only gathered within the person region in each video frame (Figure 2-c).

The segmentation mask was introduced mainly to suppress the background noise (Figure 2-d) for error analysis on the COHFACE dataset due to its video compression issue. For direct RR estimation in the live mode, the segmentation mask is not necessary thanks to the compression-free and high-quality streaming video directly from a webcam.

Extract RR from the optical flow

The averaged value of the optical flow was calculated from each frame as the respiration signal (Figure 2-e). Depending on the posture of the subject and orientation of the camera, either the horizontal or vertical motion gives a stronger respiration signal with higher signal-to-noise ratio (SNR). For error analysis, we simply take the vertical component, as it is the primary movement for subjects in both benchmarking datasets.

In previous works, RR was estimated from the constructed respiration signal by simply counting peaks and zero-crossings in the temporal domain, or by frequency analysis via the power spectral density (PSD). In the current work, we follow the second route and obtain the PSD by fast Fourier transform after padding, Han-windowing, and mean-subtraction of the respiration signal. The dominant frequency in the PSD within the range of 2-40 bpm is identified as the RR (Figure 2-f).

Experiments

To quantitatively evaluate the accuracy of the proposed method, we perform an error analysis and ablation study on the two benchmark datasets. More specifically, we look for optimal design parameters including frame rate, video resolution, and video duration. We also test the effectiveness of optical flow method, ROI filtering, and RR estimation method. Performance metrics with representative parameters are shown in Figure 4.

Performance metrics

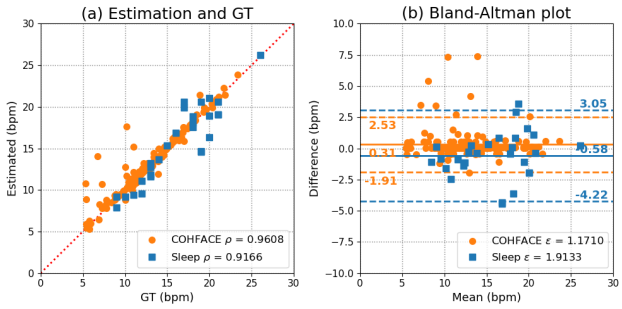


Figure 4. *Representative performance on benchmark datasets. (a) Estimated RR vs. ground truth (GT) values. (b) Bland-Altman plot, where the numbers from top to bottom are mean + 1.96 std., mean, and mean - 1.96 std., respectively.*

We report the following performance metrics as conventionally used by previous researchers in this field [2, 4, 5, 7, 10, 16, 17, 21, 25, 26, 28, 29]:

1. Pearson's correlation coefficient (ρ). We plot the estimated RR (R_{EST}) against the ground truth RR (R_{GT}) in Figure 4(a), and obtain their Pearson's ρ via

$$\rho = \frac{cov(R_{EST}, R_{GT})}{\sigma_{EST}\,\sigma_{GT}} \tag{4}$$

where $cov(R_{EST}, R_{GT})$ is the covariance, and $\sigma_{EST}, \sigma_{GT}$ are the standard deviations.

2. Root mean square error (ε). We plot the differences between R_{EST} and R_{GT} against the mean of the two as a Bland-Altman plot [47, 48] in Figure 4(b), and calculate the RMSE as

$$\varepsilon = \sqrt{\frac{\sum_{i=1}^{N}(R_{EST,i} - R_{GT,i})^2}{N}} \tag{5}$$

where N is the total number of video clips in each dataset.

Experiments on optimal parameters

In this section we experiment on several technical details of the proposed algorithm. Today's digital cameras of customer level typically have a video resolution from 480p to 1080p, and can usually achieve a frame rate of 30 Hz. However, it is not necessary to feed all the frames at full resolution to the CNN, as the computational cost might be unaffordable. Moreover, for online processing of live video, it is desirable to use shorter temporal window for RR estimation to achieve a faster response and lighter computational overhead.

Effect of frame rate

As the RR for a healthy adult is typically around 0.1-0.5 Hz (10-30 bpm), it is possible to skip camera frames for RR estimation to save computational cost. To determine the optimal time interval between consecutive frames to feed to CNN, we manually skip 0, 1, 4, 9, 19, and 29 frames to achieve equivalent sampling rate of the optical flow at 20, 10, 5, 2, and 1 Hz, respectively.

The results for the performance metrics are shown in Figures 5 (a) and (d). Both the correlation coefficient and RMSE have the best performance when the optical flow is sampled at around 4 Hz. When the sampling rate is too low, the respiration signal could not be constructed faithfully according to the Nyquist theorem. On the other end, optical flow requires the local displacement between consecutive frames to be adequate for reliable image registration. However, at a high sampling rate the displacement information and noise will be on the same order of magnitude, which explains the compromised performance.

When adequate computational resources are available, it is possible to construct the respiration signal at the live video frames per second (FPS, e.g. 30 Hz) while maintaining the optimal time interval (e.g. 0.25 s) by extracting optical flow between every frame (at 30 Hz) and its (0.25-s) delayed counterpart.

Effect of video resolution

As FlowNet delivers per-pixel local optical flow estimation, a pair of 480p frames results in more than 300k vectors at each time step, which are more than adequate to estimate the averaged flow velocity for the entire frame. Therefore, we resized the 480p videos from both datasets by bilinear interpolation, such that the shorter frame length is 480, 384, 288, 192, and 96. To make a fair comparison and maintain roughly the same total number of vectors for each frame, we resample the optical flow at 8×8, 6×6, 5×5, 3×3, 2×2 grid points for each resolution, respectively.

The results for the correlation coefficient and RMSE are shown in Figures 5 (b) and (e). Both metrics are at their best

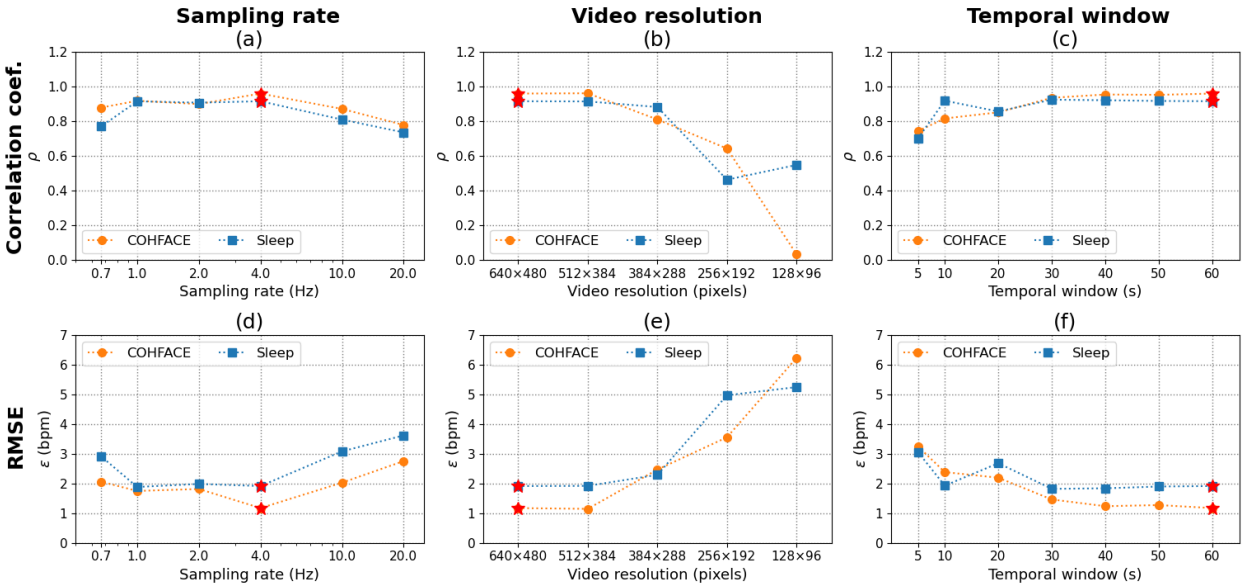


Figure 5. *Correlation coefficient (first row) and RMSE (second row) as functions of (first column) sampling rate, (second column) video resolution, and (third column) temporal window. The red stars correspond to the parameters used for Figure 4.*

when the videos are processed at the original resolution, as expected. As the video resolution shrinks, the metrics first remain relatively stable, and then degradate quickly only after the resolution is smaller than about 1/4 of the original resolution. It is therefore possible to obtain reasonable RR estimation at reduced resolution (e.g. 320p) for lighter computational cost.

One could use all 300k vectors to construct the respiration signal. However, we found that it only marginally improves the performance over the 8×8 resampled flow field, while adding 64 times undesired memory cost.

Effect of temporal window duration

The video clips in both datasets have a length of 60 seconds. However, it is not necessary to use the full video for a reliable RR measurement. We demonstrate this idea by windowing and clipping the complete respiration signal to 5 s, 10 s, 20 s, 30 s, 40 s, 50 s, and 60 s shorter durations aligned at the centers with the original signal. The windowed signal is then padded to the original length for comparable resolution in PSD.

The RR is then obtained from those shorter waveforms and compared with the ground truth, as shown in Figures 5 (c) and (f). When the entire respiration signal is used for RR estimation, the correlation coefficients are at the highest and RMSEs are at the lowest, since more information is used to obtain the PSD. Comparable performances can be achieved using as short as a 30-second-windowed respiration signal. Even with a 10-s signal, the correlation coefficients are still higher than 0.8, with the RMSEs staying below 2.5 bpm.

For constant RR, it is always more accurate to use a longer respiration signal for estimation, while in real-world applications, it is desirable to use a short temporal window for faster response in live estimation mode, especially when the RR varies with time.

Ablation study

In this section we experiment with alternative components along the processing pipeline. To be more specific, we evaluate the performance of traditional optical flow methods, we demonstrate the effect of ROI filtering, and finally we compare frequency analysis with temporal analysis for RR extraction. The results are shown in Figure 6.

Performance of traditional optical flow methods

We obtain the optical flow with the optimal sampling rate, video resolution, and temporal window using the vanilla optical flow method, and the Shi-Lucas-Kanade method implemented in OpenCV. We then mask the optical flows with a segmentation mask and extract the RR from the PSD, as in the proposed work flow. The comparison with FlowNet is shown in Figures 6 (a) and (d).

The vanilla optical flow method has poor performance due to its incapability of capturing the actual flow directions, while the FlowNet method out-performs the Shi-Lucas-Kanade method by a small margin. With careful parameter tuning for the corner detector and feature tracker, the Shi-Lucas-Kanade method still proves to be a powerful alternative, considering its lighter computational cost. But here, the Shi-Lucas-Kanade method also benefits from the segmentation masks predicted by the CNN. In a separate experiment, the correlation coefficient and RMSE drop to 0.85 and 2.80 bpm, respectively, on the Sleep dataset when there are no pre-defined ROIs.

Effect of ROI detection

The construction of the respiration signal can benefit from pre-defined ROIs identified by CNN. More specifically, one can mask out the regions where motions are not introduced by respiration to suppress noise. We evaluate its effectiveness by exploring two variants: 1) using only the regressed bounding box and taking all vectors within the box for respiration signal construction;

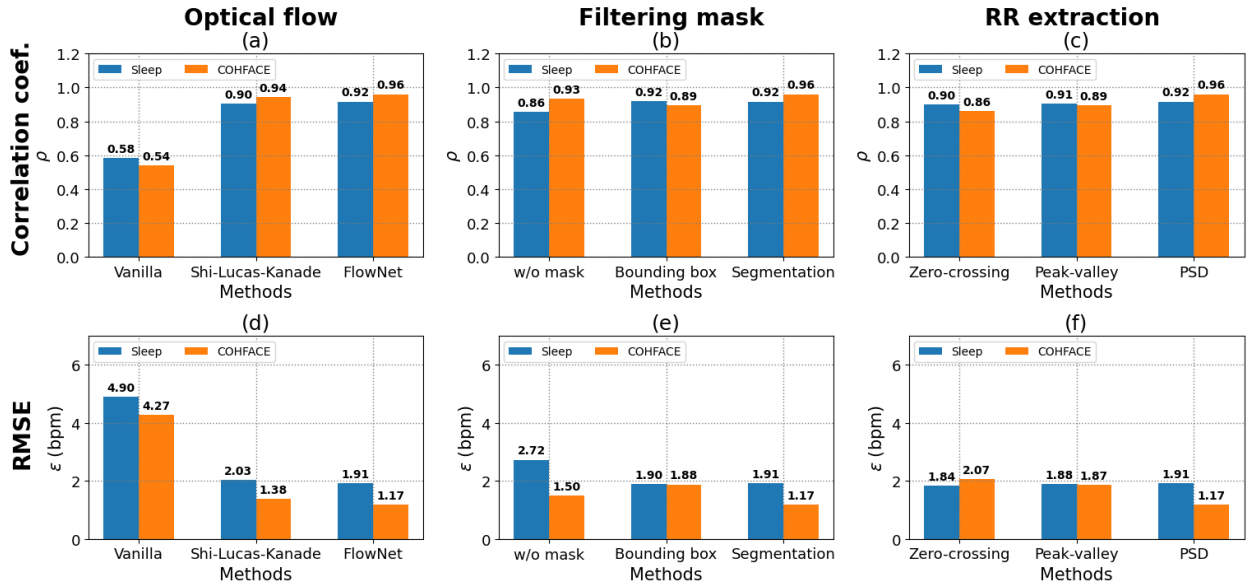


Figure 6. *The effects of (first column) optical flow method, (second column) filtering mask, and (third column) RR extraction method on the correlation coefficient (first row) and RMSE (second row).*

2) using the densely predicted segmentation mask at each pixel to determine if a vector is kept for signal construction. In general, person detection and bounding box regression require less computational cost than the segmentation task. As a baseline, we also report the performance when no ROI is defined, and the optical flow over the entire frame is harvested for respiration signal construction. The results are shown in Figures 6 (b) and (e).

Compared with the baseline method (w/o mask), using a bounding box for filtering doesn't always improve the performance, while the segmentation mask improves the correlation coefficient and reduces the RMSE on both datasets. When there are multiple persons or background motions in the frame, it is beneficial to include ROI localization for noise suppression, if moderate computational cost is affordable.

RR estimation from temporal analysis

As direct temporal analysis on the respiration signal for RR extraction is surprisingly popular in the literature, we also report its comparison with frequency analysis from PSD as shown in Figures 6 (c) and (f). For the zero-crossing method, the signal is first detrended by simply subtracting the mean, and the RR is counted as half of the total number of zero-crossings. For the peak-valley method, the peaks and valleys are detected based on a topographic prominence of 5 to neglect local fluctuations, and the RR is taken as the mean of the peak and valley counts. A representative detection of those points of interest is shown in Figure 2 (e).

It is clear that all three methods have similar correlation coefficient and RMSE. However, the methods in the temporal domain require more careful detrending and peak prominence selection as manual inputs, while for PSD one can simply take the RR that corresponds to the highest coefficient within a natural band (e.g. 10-30 bpm).

Implementation details

We prototyped the proposed algorithm, and carried out all the numerical experiments in Python 3.7. The official implementations of FlowNet 2 [43] and Detectron 2 [45] were running in PyTorch 1.5 [44] with the CUDA toolkit 10.1 [49].

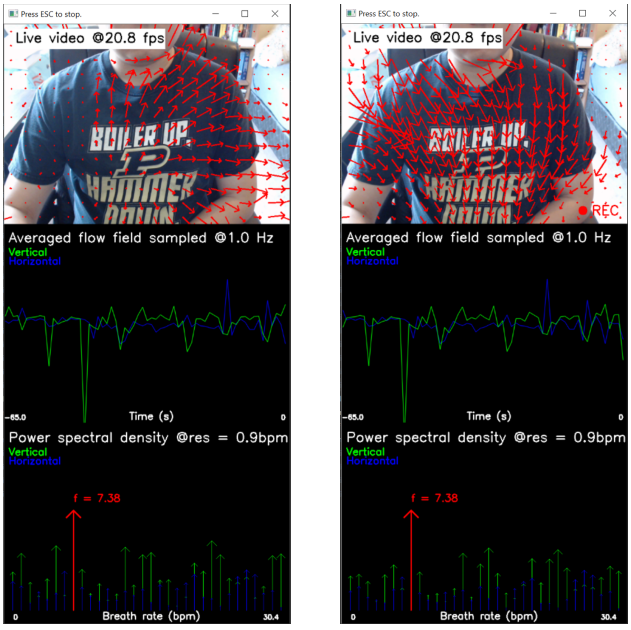


Figure 7. *Sample screenshots of the executable running on Windows 10 during (left) inhalation and (right) exhalation.*

The proposed working pipeline (without the segmentation mask) was then implemented in C++ for cross-platform deployment. The FlowNet with loaded pre-trained weights was traced and serialized in Python, and then imported into C++ as Torch-

Script modules.

We tested the live performance of the C++ implementation built on two customer-level PCs with LibTorch 1.6 for respective platforms. Sample screenshots of the executable running on Windows are shown in Figure 7. Based on the previous analysis in Figures 5 (a) and (d), the optical flow sampling rate was limited to 1 Hz by feeding image pairs to FlowNet once every 1 second. When grabbing frames from a 640p webcam on a single CPU thread and inferencing on GPU, the live FPS is shown in Table 1 below.

Table 1: Live FPS of the C++ implementation

OS	Ubuntu 18.04.5	Windows 10
CPU	i7-4710HQ	i7-6700K
GPU	GeForce GTX 860M	GeForce GTX 1070
FPS	29.4	20.8

Conclusion

In this paper, we proposed a novel system for remotely estimating the respiration rate of people. We leverage convolutional neural networks to extract optical flow induced by subtle upper body movements during periodic inhalation and exhalation cycles.

We tested the performance of our implementation on public datasets on RR estimation, and showed improvements over traditional optical flow methods in terms of Pearson's correlation coefficient and root mean square error. We also performed numerical experiments, and demonstrated the robustness and reliability of the proposed system. The optical flow can be sampled at reduced rate from 1 Hz to 10 Hz to lower computational cost. We also found that videos as low as 320p resolution can deliver reliable RR measurements; and a video clip as short as 10 s is adequate for fast measurement response.

We further implemented the proposed working pipeline in C++ by converting the pre-trained CNN to serialized TorchScripts using a sample tracing technique provided by PyTorch. We demonstrated the feasibility of running live RR estimation using CNN by cross-platform deployment on Ubuntu and Windows machines with GPU inference.

Future work may include further performance evaluation of our algorithm on additional datasets where clothing and lighting conditions are properly controlled. The design of light-weight CNNs for deployment on mobile devices with limited computational powers is also a promising direction. Moreover, potential integration with vision-based remote heart rate and blood pressure estimation algorithms could prove useful for fast vital sign screening in clinics.

One major drawback of current method is that we still require the subject to remain relatively stationary for about 10 s. This is because the FlowNetSD variant we use is ideal for inter-frame displacements on the order of several pixels. To overcome this, one could use CNNs designed for long-range image registration, coupled with a proper person tracking algorithm.

The developed system can be used by individuals with a customer-level RGB webcam, or by healthcare centers with near-infrared cameras or thermal imagers to monitor the vital signs of patients in total dark settings.

Acknowledgments

This research project was sponsored by HP Labs, HP Inc., Palo Alto, CA. FlowNet 2 and Detectron 2 are distributed in accordance with the Apache License 2.0 by the Apache Software Foundation.

References

[1] C. H. Antink, S. Lyra, M. Paul, X. Yu, and S. Leonhardt, "A Broader Look: Camera-Based Vital Sign Estimation across the Spectrum," *Yearbook of Medical Informatics*, vol. 28, no. 1, pp. 102–114, 8 2019.

[2] N. Hochhausen, C. B. Pereira, S. Leonhardt, R. Rossaint, and M. Czaplik, "Estimating respiratory rate in post-anesthesia care unit patients using infrared thermography: An observational study," *Sensors (Switzerland)*, vol. 18, no. 5, 5 2018.

[3] J. Jorge, M. Villarroel, S. Chaichulee, A. Guazzi, S. Davis, G. Green, K. McCormick, and L. Tarassenko, "Non-Contact Monitoring of Respiration in the Neonatal Intensive Care Unit," in *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017 - 1st International Workshop on Adaptive Shot Learning for Gesture Understanding and Production, ASL4GUP 2017, Biometrics in the Wild, Bwild 2017, Heteroge*, 2017.

[4] M. van Gastel, S. Stuijk, and G. de Haan, "Robust respiration detection from remote photoplethysmography," *Biomedical Optics Express*, vol. 7, no. 12, p. 4941, 12 2016.

[5] H. E. Elphick, A. H. Alkali, R. K. Kingshott, D. Burke, and R. Saatchi, "Exploratory Study to Evaluate Respiratory Rate Using a Thermal Imaging Camera," *Respiration*, vol. 97, no. 3, pp. 205–212, 3 2019.

[6] R. Chauvin, M. Hamel, S. Briere, F. Ferland, F. Grondin, D. Letourneau, M. Tousignant, and F. Michaud, "Contact-Free Respiration Rate Monitoring Using a Pan-Tilt Thermal Camera for Stationary Bike Telerehabilitation Sessions," *IEEE Systems Journal*, vol. 10, no. 3, pp. 1046–1055, 9 2016.

[7] Y. Cho, S. J. Julier, N. Marquardt, and N. Bianchi-Berthouze, "Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging," *Biomedical Optics Express*, vol. 8, no. 10, p. 4480, 10 2017.

[8] A. Basu, A. Routray, R. Mukherjee, and S. Shit, "Infrared imaging based hyperventilation monitoring through respiration rate estimation," *Infrared Physics and Technology*, vol. 77, pp. 382–390, 7 2016.

[9] K. Mutlu, J. E. Rabell, P. Martin Del Olmo, and S. Haesler, "IR thermography-based monitoring of respiration phase without image segmentation," *Journal of Neuroscience Methods*, vol. 301, pp. 1–8, 2018.

[10] C. B. Pereira, M. Czaplik, V. Blazek, S. Leonhardt, and D. Teichmann, "Monitoring of cardiorespiratory signals using thermal imaging: A pilot study on healthy human subjects," *Sensors (Switzerland)*, vol. 18, no. 5, 5 2018.

[11] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, vol. 2. Morgan Kaufmann Publishers Inc., 1981, pp. 674–679.

[12] J. Shi and C. Tomasi, "Good features to track," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern*

Recognition. Publ by IEEE, 1994, pp. 593–600.

[13] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2259–2272, 2011.

[14] Z. Kalal, K. Mikolajczyk, and J. Matas, "Face-TLD: Tracking-learning-detection applied to faces," in *Proceedings - International Conference on Image Processing, ICIP*, 2010, pp. 3789–3792.

[15] C. B. Pereira, X. Yu, M. Czaplik, R. Rossaint, V. Blazek, and S. Leonhardt, "Remote monitoring of breathing dynamics using infrared thermography," *Biomedical Optics Express*, vol. 6, no. 11, p. 4378, 11 2015.

[16] C. Massaroni, E. Schena, S. Silvestri, F. Taffoni, and M. Merone, "Measurement system based on RBG camera signal for contactless breathing pattern and respiratory rate monitoring," in *MeMeA 2018 - 2018 IEEE International Symposium on Medical Measurements and Applications, Proceedings*. Institute of Electrical and Electronics Engineers Inc., 8 2018.

[17] C. Massaroni, D. Lo Presti, D. Formica, S. Silvestri, and E. Schena, "Non-contact monitoring of breathing pattern and respiratory rate via rgb signal measurement," *Sensors (Switzerland)*, vol. 19, no. 12, 6 2019.

[18] L. Tarassenko, M. Villarroel, A. Guazzi, J. Jorge, D. A. Clifton, and C. Pugh, "Non-contact video-based vital sign monitoring using ambient light and auto-regressive models," *Physiological Measurement*, vol. 35, no. 5, pp. 807–831, 5 2014.

[19] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, 2013.

[20] G. De Haan and A. Van Leest, "Improved motion robustness of remote-PPG by using the blood volume pulse signature," *Physiological Measurement*, vol. 35, no. 9, pp. 1913–1926, 9 2014.

[21] B. Wei, X. He, C. Zhang, and X. Wu, "Non-contact, synchronous dynamic measurement of respiratory rate and heart rate based on dual sensitive regions," *BioMedical Engineering Online*, vol. 16, no. 1, 1 2017.

[22] M. Pal, R. Roy, J. Basu, and M. S. Bepari, "Blind source separation: A review and analysis," in *2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE 2013*, 2013.

[23] H. Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics*, vol. 31, no. 4, 7 2012.

[24] X. He, R. Goubran, and F. Knoefel, "IR night vision video-based estimation of heart and respiration rates," in *SAS 2017 - 2017 IEEE Sensors Applications Symposium, Proceedings*, 2017.

[25] K. Y. Lin, D. Y. Chen, and W. J. Tsai, "Image-Based Motion-Tolerant Remote Respiratory Rate Evaluation," *IEEE Sensors Journal*, vol. 16, no. 9, pp. 3263–3271, 5 2016.

[26] A. Chatterjee, A. P. Prathosh, P. Praveena, and V. Upadhya, "Real-Time Visual Respiration Rate Estimation with Dynamic Scene Adaptation," in *Proceedings - 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering, BIBE 2016*. Institute of Electrical and Electronics Engineers Inc., 12 2016, pp. 154–160.

[27] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185–203, 8 1981.

[28] G. Heusch, A. Anjos, and S. Marcel, "A Reproducible Study on Remote Heart Rate Measurement," *arXiv:1709.00962*, 9 2017.

[29] M. Hu, G. Zhai, D. Li, Y. Fan, H. Duan, W. Zhu, and X. Yang, "Combination of near-infrared and thermal imaging techniques for the remote and simultaneous measurements of breathing and heart rates under sleep situation," *PLoS ONE*, vol. 13, no. 1, 1 2018.

[30] K. Konno and J. Mead, "Measurement of the separate volume changes of rib cage and abdomen during breathing." *Journal of Applied Physiology*, vol. 22, no. 3, pp. 407–422, 1967.

[31] B. D. Lucas, "Generalized Image Matching by the Method of Differences," Ph.D. dissertation, Carnegie Mellon University, 1984.

[32] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 12 2015.

[34] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January. Institute of Electrical and Electronics Engineers Inc., 11 2017, pp. 5987–5995.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[36] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January. Institute of Electrical and Electronics Engineers Inc., 11 2017, pp. 936–944.

[37] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense Human Pose Estimation in the Wild," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 12 2018, pp. 7297–7306.

[38] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "R-CNNs for Pose Estimation and Action Detection," *arXiv:1406.5212*, 6 2014.

[39] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October. Institute of Electrical and Electronics Engineers Inc., 12 2017, pp. 2980–2988.

[40] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 833–851.

[41] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, H. Hazırbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "FlowNet: Learning Optical Flow with Convolutional Networks," in *Proceedings of the IEEE International Conference on Computer Vision*. Institute of Electrical and Electronics Engineers Inc., 2015, pp. 2758–2766.

[42] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January. Institute of Electrical and Electronics Engineers Inc., 11 2017, pp. 1647–1655.

[43] F. Reda, R. Pottorff, J. Barker, and B. Catanzaro, "flownet2-pytorch: Pytorch implementation of FlowNet 2.0: Evo-

lution of Optical Flow Estimation with Deep Networks," https://github.com/NVIDIA/flownet2-pytorch, 2017.

[44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. K. Xamla, E. Yang, Z. Devito, M. Raison Nabla, A. Tejani, S. Chilamkurthy, Q. Ai, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

[45] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem. IEEE Computer Society, 12 2016, pp. 770–778.

[47] J. Martin Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 2 1986.

[48] J. M. Bland and D. G. Altman, "Measuring agreement in method comparison studies," *Statistical Methods in Medical Research*, vol. 8, no. 2, pp. 135–160, 4 1999.

[49] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with CUDA," *Queue*, vol. 6, no. 2, pp. 40–53, 3 2008.

Author Biography

Tianqi Guo received his BS in mechanical engineering from Nanjing University of Science and Technology (2013) and his master in mechanical engineering from Purdue University (2016). He is now a Ph.D. candidate in Electrical and Computer Engineering at Purdue University, working on research projects in communications, networking, signal and image processing.

Dr. Qian Lin is a HP Fellow working on computer vision and deep learning research. She is also an adjunct professor at Purdue University, supervising Ph.D. students in their deep learning research. Dr. Lin joined Hewlett-Packard Company in 1992. She received her BS from Xi'an Jiaotong University in China, her MSEE from Purdue University, and her Ph.D. in Electrical Engineering from Stanford University. Dr. Lin is inventor/co-inventor for 45 issued patents. She was awarded Fellowship by the Society of Imaging Science and Technology (IS&T) in 2012, and Outstanding Electrical Engineer by the School of Electrical and Computer Engineering of Purdue University in 2013.

Jan P. Allebach is Hewlett-Packard Distinguished Professor of Electrical and Computer Engineering at Purdue University, West Lafayette, Indiana with courtesy appointments in Computer Science and Psychological Sciences. Imaging has been a central theme of his research; and he has made many contributions in the areas of printing and scanning, content repurposing, image quality, and image aesthetics. Professor Allebach has received numerous recognitions for his research and teaching. He is a Fellow of IEEE, IS&T (The Society for Imaging Science and Technology), and SPIE. He was named Electronic Imaging Scientist of the Year by SPIE and IS&T, and received Honorary Membership from IS&T, which is its highest award. He received the Daniel E. Noble Award for Emerging Technologies an IEEE Field Award. Allebach also received the Edwin Land Medal from the Optical Society of America and IS&T, and the Johann Gutenberg Prize from IS&T. He was elected to Membership in the National Academy of Engineering, and Fellowship in the National Academy of Inventors. From Purdue University, he has received ten different awards for teaching, research, and mentorship. He has served two separate terms as IEEE Signal Processing Society Distinguished Lecturer.

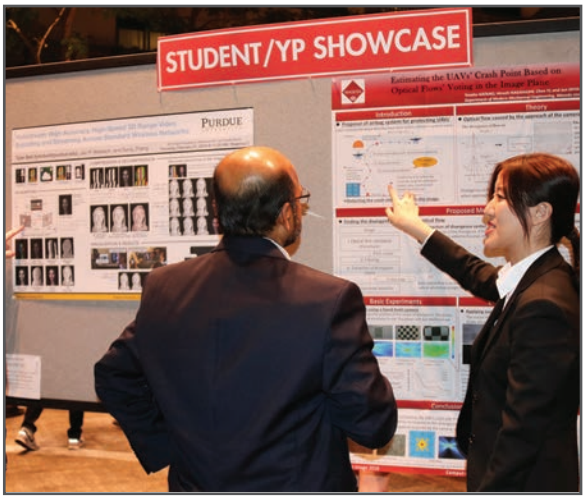
JOIN US AT THE NEXT EI!

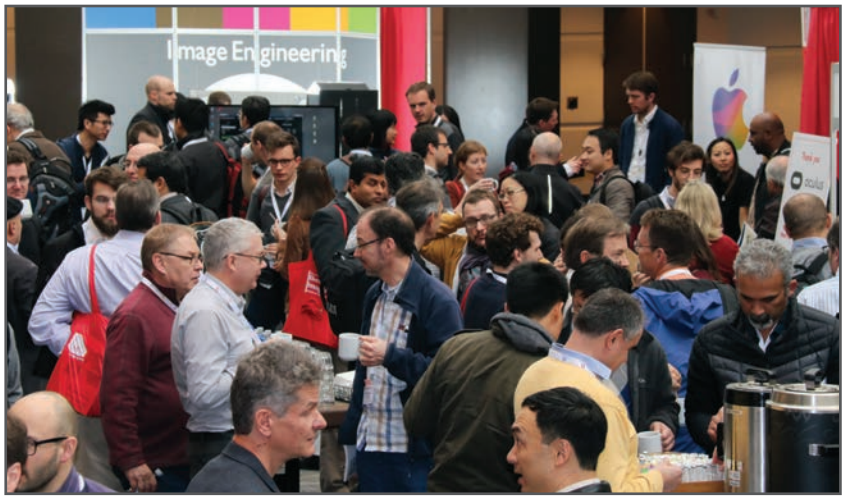
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





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

www.electronicimaging.org

IS&T
imaging.org