Research Article | Vol. 12, No. 1 / 1 January 2021 / Biomedical Optics Express 494
Biomedical Optics EXPRESS

Systematic analysis of video-based pulse measurement from compressed videos

EWA M. NOWARA,[1,*] DANIEL MCDUFF,[2] AND ASHOK VEERARAGHAVAN[1]

[1]Electrical and Computer Engineering Department, Rice University, 6100 Main St, Houston, TX 77005, USA
[2]Microsoft Research AI, 14820 NE 36th St, Redmond, WA 98052, USA
[*]emn3@rice.edu

Abstract: Camera-based physiological measurement enables vital signs to be captured unobtrusively without contact with the body. Remote, or imaging, photoplethysmography involves recovering peripheral blood flow from subtle variations in video pixel intensities. While the pulse signal might be easy to obtain from high quality uncompressed videos, the signal-to-noise ratio drops dramatically with video bitrate. Uncompressed videos incur large file storage and data transfer costs, making analysis, manipulation and sharing challenging. To help address these challenges, we use compression specific supervised models to mitigate the effect of temporal video compression on heart rate estimates. We perform a systematic evaluation of the performance of state-of-the-art algorithms across different levels, and formats, of compression. We demonstrate that networks trained on compressed videos consistently outperform other benchmark methods, both on stationary videos and videos with significant rigid head motions. By training on videos with the same, or higher compression factor than test videos, we achieve improvements in signal-to-noise ratio (SNR) of up to 3 dB and mean absolute error (MAE) of up to 6 beats per minute (BPM).

1. Introduction

Imaging photoplethysmography (iPPG) leverages subtle changes in light reflected from the skin to capture cardiac activity. These signals can be recovered from many types of commercially available and low-cost cameras (e.g., webcams and smartphones). Research over the past decade has shown that heart rate (HR) [1,2], heart rate variability (HRV) [3], and breathing rate (BR) [3] can be measured from video recordings of the human body via the photoplethysmogram or ballistocardiogram (BCG) [4]. Measuring vital signs remotely with a camera offers several advantages over the contact devices traditionally used in pulse oximetry or electrocardiograms (ECG). Imaging-based methods can be used in applications where direct contact with the skin should be minimized or where wearing contact devices hinders the tasks that need to be performed. It is particularly attractive for contexts that require unobtrusive, continuous, and long-term measurements. For example, non-contact sensors can avoid damaging the skin of prematurely born babies [5] and burn victims, reducing the risk of infections. Some applications may require long-duration measurements and using contact devices may be infeasible, cause discomfort or simply irritate or distract subjects, for example during sleep monitoring [6], driver monitoring [7] or cognitive engagement with computer applications [8]. Camera-based measurement of vital signs is also attractive in telehealth as there is no requirement for additional hardware, other than that typically used for a video conferencing call (i.e., webcam and computer) [9].

Existing imaging-based physiological measurement methods can provide heart rate estimates comparable to contact devices, provided that the video was recorded under good conditions (sufficiently high fidelity sensor, well-lit subject, small motions, etc.) [2,11–15]. Video

#408471 https://doi.org/10.1364/BOE.408471
Journal © 2021 Received 8 Sep 2020; revised 12 Nov 2020; accepted 7 Dec 2020; published 18 Dec 2020

Biomedical Optics EXPRESS

compression algorithms are designed to reduce the bitrate of videos while causing as little visual impact as possible. The algorithms typically remove subtle inter-frame intensity variations as these are deemed unlikely to impact the visual appearance, and it is true that these changes are often imperceptible to the naked eye. However, iPPG algorithms rely on these very subtle intensity and color variations to recover the pulse signal. Video compression partially removes this information through spatial and temporal compression. Consequently, as the compression level increases and the video bitrate decreases, the signal-to-noise ratio (SNR) of iPPG signals decreases linearly [16,17]. The low SNR of iPPG signals makes it very difficult to recover vital information from heavily compressed videos, especially in the presence of other sources of noise, such as head motions.

Most video datasets collected for iPPG measurement are captured as raw lossless images, or with very high bitrates, intentionally avoiding lossy video compression. Such datasets demand enormous amounts of storage [16]. For example, in AFRL dataset [10] a 5.5 minute video of one subject is on average 11.9 GB. Collecting, storing, streaming, and transferring such datasets becomes challenging, especially as the number of subjects, conditions, and duration of the recordings increases. This presents challenges for practical applications. For example, vital signs could be theoretically measured automatically over a video call with the physician from the patient's home. Unfortunately, streaming and processing an uncompressed video in real time during a video call presents challenges. Moreover, existing methods are unable to reliably extract iPPG signals from videos which are automatically compressed by videoconferencing applications. This large memory requirement also hinders sharing of large uncompressed video datasets. Therefore, most existing datasets suitable for iPPG measurements are very small.

Being able to use compressed videos would enable new applications, such as telemedicine. Reducing video bitrates would help facilitate sharing large datasets of video recordings, which in turn could help to advance the state-of-the-art. Furthermore, it would enable training algorithms on larger datasets, thus benefiting from the power of deep neural architectures [11]. Very few approaches have been proposed in the literature to overcome the problem of compression [18,19]. All of these approaches require enhancing the compressed videos instead of directly operating on them. The video enhancement step may be time consuming and require additional storage, making it unsuitable for many applications.

In this paper, we present a systematic analysis of how video compression impacts supervised iPPG algorithms. Unlike the previous methods, we do not rely on enhancing the videos before extracting the vital signs. Rather, we show that deep learning models trained on compressed videos can learn to recover iPPG signals directly. We demonstrate that training on videos compressed with the same or higher compression level than the compression level of videos in the test set achieves the best performance, as illustrated by Fig. 1. We use the current state-of-the-art attention-based deep learning approach [11] and evaluate performance at varying levels of compression. We also systematically analyze the generalizability of deep learning models across five different levels of compression (constant compression rate factor (CRF) = 12, 18, 24, 30, 36) and three different compression formats (H.264, H.265, MPEG-4). We report our results on a public dataset featuring stationary and motion conditions [10]. Finally, we analyze how the skin type affects the performance with increasing levels of compression. We show that our approach performs better than the baselines consistently across all skin types. The contributions of this work are the following: 1) We present an approach which can recover iPPG signals from compressed videos without enhancing them first. 2) We perform a systematic analysis of performance and generalizability across five compression levels and three compression formats. 3) We evaluate the effects of compression across different skin type. We hope that by demonstrating that compressed videos can be used to measure vital signs, this work will enable new applications, such as use in telemedicine, and enable collection and sharing of larger, more diverse datasets.

Biomedical Optics EXPRESS

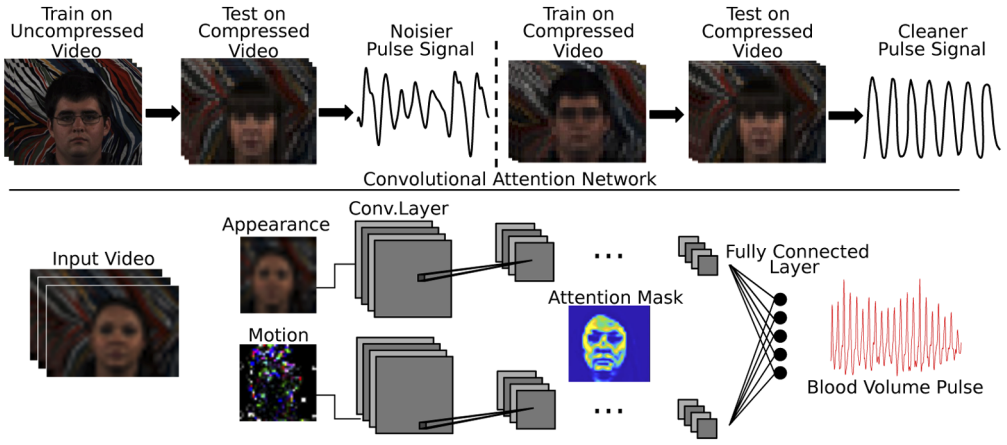

Fig. 1. A supervised deep learning network trained directly on compressed videos can combat the impact of video compression on camera-based vital signs measurements. (Images used with permission from AFRL dataset [10])

2. Background

2.1. Video compression

Raw video has a large memory footprint, therefore almost all video systems use compression algorithms. Applications that employ video compression include: video recording software, video over IP systems (e.g., Skype and Teams), video sharing sites (e.g., YouTube and Vimeo) and storage mediums (e.g., DVD and Blue-ray). The most commonly currently used compression formats are H.264, H.265, and MPEG-4. Most modern compression algorithms can dramatically reduce the bitrate of the video without significantly reducing the visual quality. Most forms of compression are lossy, making it difficult to recover subtle information after the video has been compressed.

Spatial information is reduced through *intra-frame compression* which is performed only within the current video frame. This type of compression is equivalent to image compression (e.g. JPEG coding). Intra-frame compression divides the frame into blocks of pixels. It uses the correlation between similar pixels located close to each other in an image to reduce the redundancy of similar information in the image. Instead of storing all pixel values directly, a small number of pixels is efficiently coded and the remaining pixel values are extrapolated from them. Each pixel block is spatially compressed by applying a discrete cosine transform (DCT) to the image and removing high frequency coefficients which are less crucial for the visual quality.

Color information is reduced through *chroma subsampling*. Human eyes are more sensitive to brightness information (luminance) and much less sensitive to color information (chrominance). Therefore, the chrominance information can be compressed more than luminance information. Similar to intra-frame compression, chroma subsampling is applied to individual frames in a video.

Temporal information is reduced through *inter-frame compression*. Unlike the intra-frame compression and chroma subsampling, inter-frame compression is performed for a group of consecutive video frames. Many regions in the video do not change much over time (e.g., static background regions). Therefore, less information needs to be stored to represent these regions. *Motion vectors* describe the difference between the current and the reference frame (I-frame). To compute the motion vectors, either the current and the previous frame is used to compute the predicted frames, referred to as the P-frames. Or, the current, and both the previous and the next frames are used to compute the bi-directionally predicted frames, referred to as B-frames. The

Biomedical Optics EXPRESS

P-frames and B-frames are placed in between the I-frames and similar to intra-frame compression are transformed and quantized to reduce the memory required. Videos with larger amounts of motion typically require higher bitrates to maintain the same visual quality.

Constant Compression Rate Factor (CRF) is an adaptive quantizer used to maintain a constant compression quality across videos with different amounts of motion and detail. Possible CRF values range between 0 and 51, where 0 is lossless and 51 is the most lossy. CRF values between 18 and 28 are most commonly used in applications where visual quality is important but memory savings are desirable.

Impact of Video Compression on iPPG Quality. Video compression methods are typically optimized for visual quality, not with physiological measurement in mind. It is often assumed that small color variations between frames or between spatial groups of pixels in an image are not important for the visual quality of the video and can be removed. While this compression may not affect the visual quality of the video, it has detrimental effects on iPPG signals. Imaging-based pulse measurement relies on those small variations and therefore compression algorithms significantly degrade the quality of the iPPG by removing that subtle spatial, color, and temporal information. At higher CRF values, the video is more compressed and the SNR of the pulse signals decreases more or less linearly with increasing CRF [16].

Figure 2 shows examples of iPPG waveforms computed using spatially averaged green channel pixel values in a video at different levels of compression with corresponding example frames. We used the green channel signals to demonstrate the impact of compression on iPPG signals independent of the post-processing algorithms because this channel has the strongest iPPG signal [1]. As the compression level increases, the iPPG waveforms become more noisy. Also, the more temporally compressed the video is, the less sharp the images are, showing the effects of intra-frame compression. The compression effects are particularly evident for CRF=36.

2.2. Imaging-based physiological measurement

Early imaging-based physiological measurement methods were focused on controlled recordings with limited noise to prove that reliable measurements could be captured from video [1,4]. As these methods have matured over the past two decades, the focus has moved towards more challenging scenarios, such as videos with large motion, different skin types, and uncontrolled ambient illumination [7,11,12,14,15]. However, less work has been done to overcome the detrimental effects of low video quality on iPPG signals. As a result, most existing methods are prone to noise artefacts with decreasing video quality.

Even though image intensities are usually spatially averaged to obtain iPPG signals, spatial compression does affect quality [18]. Rapczynski et al. [20] showed that temporal compression affects iPPG signals far more than spatial compression and chroma subsampling. Temporal compression is particularly problematic for iPPG signals because many of the compression algorithms remove small variations between frames imperceptible to the human eye in terms of the video quality, but containing information important for iPPG signals. The more temporally compressed the video, the lower the iPPG SNR and the more prone the iPPG signals to motion artifacts and other sources of noise.

Only a few approaches have been proposed so far to alleviate the compression noise. McDuff et al. [18] used deep-learning-based super-resolution to enhance heavily spatially compressed video frames to improve iPPG estimation. The super-resolution method was able to recover high frequency spatial information in the facial images and result in more reliable iPPG signals. Yu et al. [19] presented the first method to recover iPPG signals from temporally compressed videos by using a deep-learning-based video-to-video generator to enhance compressed videos, followed by computation of iPPG signals from cleaner videos with an attention-based network. However, both of these methods require enhancing the images prior to iPPG computation making it time-consuming and requiring large memory. In this work, we use deep convolutional

Biomedical Optics EXPRESS

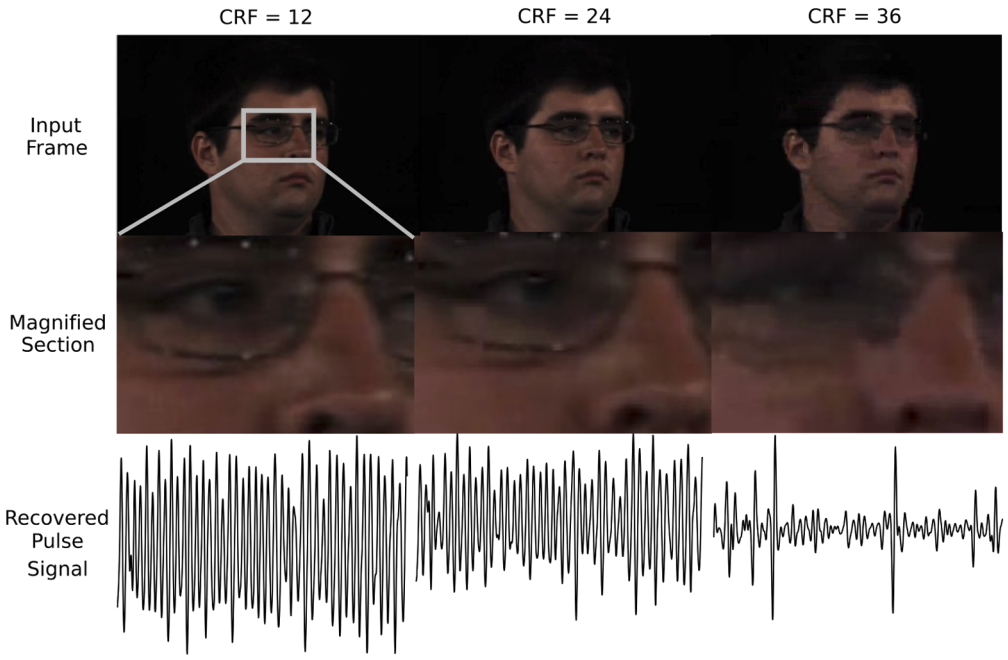

Fig. 2. An illustration of the detrimental effect of compression on the pulse signal within a video. Examples of face images from these videos (images used with permission from the AFRL dataset [10]), and zoomed in (grey) regions, respectively. Followed by examples of the processed green channel signals at different compression constant rate factors (CRF = 12, 24, 36). While the images from less and more compressed videos look similar and only slightly more blurry at high compression, at greater CRFs the signal-to-noise ratio of the pulse signal significantly decreases.

models trained directly on temporally compressed videos to recover the iPPG signals without the requirement to first enhance the video [11]. We show that training on compressed videos outperforms all compared baseline methods.

3. Methods

In this section, we present the details of the architecture of our approach used to recover the pulse signals from compressed videos, and how we evaluated its performance.

3.1. Deep learning architecture

To analyze the effects of video compression during training of a supervised network, we used a state-of-the-art convolutional attention network architecture (DeepPhys) [11]. We trained the network following the training parameters of Chen et al. [11] using subject-independent five-fold cross-validation where we used videos of 20 subjects for training and left out five subjects for testing. We averaged the results over the five validation splits, separately for all motion tasks. As illustrated in Fig. 1, the DeepPhys network uses two separate models trained on the motion representation and the appearance representation. The motion representation is computed from a normalized frame difference based on a skin reflection model [14]. The appearance representation is computed from the color and texture information from input RGB image frames. The appearance representation guides the motion representation to recover the pulse information from the skin regions and to differentiate it from other sources of variations, such as head motion or non-uniform illumination. The appearance and motion representations are learned jointly

Biomedical Optics EXPRESS

through an attention mechanism. We spatially averaged the input images to 36×36 pixels, using a bicubic interpolation, to reduce the camera quantization noise. Non-uniform variations which are caused by varying ambient light, motion, or skin types may vary across subjects and datasets, and would hinder the supervised learning model. Therefore, the input to the network is normalized with AC/DC normalization applied once for the entire video duration by subtracting the temporal mean and dividing by the standard deviation. The DeepPhys network outputs the pulse signal along with the attention mask which illustrates which regions in the video were used to compute the signal. The output signal was bandpass filtered in the physiological range ([0.7 Hz, 2.5 Hz]) and heart rate was estimated as the frequency with the highest power spectrum energy.

3.2. Dataset

To evaluate our approach we used the AFRL dataset containing videos of 25 participants, aged 18 to 28 years, recorded with a Basler Scout scA640-120gc GigE-standard color camera with a 16 mm fixed focal length lens [10]. The images were recorded as 8-bit, 658x492 pixel resolution at 120 frames per second (FPS). 17 of the participants were male, nine wore glasses, eight had facial hair and four had makeup. The dataset features participants with diverse skin types estimated with the following Fitzpatrick Sun-Reactivity Skin Types [21]: I=1, II=13, III=10, IV=1, V=0. ECG signals were recorded as ground truth physiological signals simultaneously with each video recording using a BioSemi ActiveTwo research-grade biopotential acquisition unit. The participants were recorded during 5.5 minute tasks with varying amount of head motion, resulting in 13 hours and 45 minutes of video data for the three motion tasks we used. Each task was recorded with a black uniform background and repeated with a textured background.

Stationary Task: The participants were asked to sit still and look at the camera, allowing for small natural head motions.

Medium Motion Task: The participants moved their heads horizontally at a speed of 20 degrees/second.

Large Motion Task: The participants were asked to randomly reorient their heads once every second towards one of nine positions evenly spaced in an arc around them. This was the most challenging motion task in this dataset because it simulated random head motion and introduced noise at frequencies close to the average resting heart rate (~ 1 Hz).

3.3. Evaluation metrics

We used two evaluation metrics for capturing the performance of the pulse signal recovery, mean absolute error (MAE) and pulse signal-to-noise ratio (SNR).

Mean absolute error (MAE):

$$\text{MAE} = \frac{\sum_{i=1}^{N} |R_i - \widehat{R}_i|}{N} \tag{1}$$

where N is the total number of time windows, R_i is the ground truth heart rate measured with a contact ECG sensor and $\widehat{R}_i$ is the estimated HR from the video recording.

Signal-to-noise ratio (SNR) was calculated as the ratio of the area under the curve of the power spectrum around the first and second harmonic of ground truth HR frequency to the area under the curve of the rest of the spectrum between 42 to 240 bpm [13]:

$$\text{SNR} = 10 \log_{10} \left(\frac{\sum_{42}^{240} (U_t(f) S(f))^2}{\sum_{42}^{240} ((1 - U_t(f)) S(f))^2} \right) \tag{2}$$

Biomedical Optics EXPRESS

where S is the power spectrum of the estimated iPPG signal, f is the frequency in BPM and $U_t(f)$ is equal to one for frequencies around the first and second harmonic of the ground truth heart rate (HR-6 bpm to HR+6 bpm and 2*HR-6 bpm to 2*HR+6 bpm), and 0 everywhere else.

For each test video we calculated these metrics on a set of 30 second time windows, with one second overlap, from each video. We then averaged each metric for all time windows to get a MAE and SNR for each subject video in the test set. We removed the first and last 15 seconds of each 5.5 minute recording. The error bars in Table 1, and in Fig. 3 were computed as the standard error defined as the standard deviation across all subjects' results and divided by the square root of the number of subjects.

Table 1. MAE and SNR results at varying compression (CRF = 12 - 36), computed from each video on a set of 30 second time windows, with one second overlap. Training on compressed videos with matching CRF leads to lower MAE and higher SNR.

	↓ Mean absolute error (MAE)					↑ Signal-to-noise ratio (SNR)				
Method	12	18	24	30	36	12	18	24	30	36
					Still					
Train on Matching CRF (Ours)	1.7 ± 0.5	**1.6±0.5**	**1.4±0.3**	**3.5±0.9**	**8.0±1.7**	**6.4±0.5**	**5.8±0.5**	**4.5±0.3**	**-2.4±0.9**	-7.4 ± 1.7
Train on CRF=12 [11]	1.7 ± 0.5	4.3 ± 1.4	4.5 ± 1.3	9.5 ± 1.8	14.4 ± 2.0	**6.4±0.5**	-0.5 ± 1.4	-3.9 ± 1.3	-8.6 ± 1.8	-11.3 ± 2.0
CHROM [13]	1.6 ± 0.3	2.6 ± 0.6	2.5 ± 0.5	11.9 ± 0.9	16.2 ± 1.0	2.4 ± 0.4	0.6 ± 0.4	0.3 ± 0.4	-4.2 ± 0.3	-5.5 ± 0.4
ICA [2]	2.3 ± 0.7	3.2 ± 1.1	3.3 ± 1.0	12.6 ± 1.3	15.4 ± 0.9	3.7 ± 0.8	2.3 ± 0.7	1.1 ± 0.7	-4.4 ± 0.5	**-5.2**± 0.4
POS [14]	**1.5±0.2**	1.9 ± 0.3	2.4 ± 0.5	12.8 ± 1.7	17.0 ± 1.8	3.6 ± 0.7	2.1 ± 0.6	1.2 ± 0.6	-4.4 ± 0.5	-6.0 ± 0.5
GREEN [1]	10.4 ± 1.1	11.4 ± 1.1	13.2 ± 1.2	15.5 ± 1.1	16.9 ± 1.2	-3.9 ± 0.5	-4.3 ± 0.5	-4.9 ± 0.5	-5.6 ± 0.4	-5.9 ± 0.5
BCG [4]	13.3 ± 1.9	13.4 ± 2	13.7 ± 1.9	15.5 ± 1.9	16.6 ± 2.1	-3.5 ± 0.7	-3.6 ± 0.7	-4.2 ± 0.6	-5.1 ± 0.6	-5.5 ± 0.6
					Medium Motion					
Train on Matching CRF (Ours)	**2.4±0.9**	**2.6±0.9**	**3.3±1.1**	**4.8±1.3**	12.2 ± 1.9	**2.5±0.9**	**1.6±0.9**	**-0.6±1.1**	**-4.1±1.3**	-9.1 ± 1.9
Train on CRF=12 [11]	**2.4±0.9**	7.1 ± 1.4	8.3 ± 1.5	10.5 ± 1.6	**11.6±1.5**	**2.5±0.9**	-6.2 ± 1.4	-7.3 ± 1.5	-9.6 ± 1.6	-11.0 ± 1.5
CHROM [13]	3.0 ± 0.7	5.9 ± 1.2	8.8 ± 1.3	16.4 ± 1.2	16.0 ± 1.4	0.1 ± 0.5	-0.9 ± 0.4	-2.3 ± 0.4	-4.3 ± 0.5	**-5.1±0.7**
ICA [2]	5.0 ± 1.1	6.0 ± 1.3	11.6 ± 1.8	18.2 ± 0.9	15.6 ± 0.8	-0.1 ± 0.6	-0.7 ± 0.6	-2.7 ± 0.5	-4.7 ± 0.6	-5.3 ± 0.8
POS [14]	3.8 ± 0.9	4.0 ± 0.9	6.8 ± 1.3	14.0 ± 1.4	15.8 ± 1.8	0.0 ± 0.7	0.0 ± 0.6	-1.9 ± 0.6	-4.2 ± 0.6	-5.4 ± 0.8
GREEN [1]	16.7 ± 1.2	16.2 ± 1.2	17.4 ± 1.2	17.0 ± 1.2	16.8 ± 1.2	-5.3 ± 0.5	-5.4 ± 0.5	-5.5 ± 0.5	-5.5 ± 0.5	-5.7 ± 0.5
BCG [4]	–	–	–	–	–	–	–	–	–	–
					Large Motion					
Train on Matching CRF (Ours)	**5.3±1.3**	**6.6±1.5**	**6.8±1.4**	**12.2±1.6**	**14.1±1.5**	**-3.1±1.3**	-4.7 ± 1.5	-5.3 ± 1.4	-9.5 ± 1.6	-11.7 ± 1.5
Train on CRF=12 [11]	**5.3**± 1.3	11.3 ± 1.5	12.5 ± 1.5	13.3 ± 1.4	14.2 ± 1.5	**-3.1±1.3**	-9.2 ± 1.5	-9.7 ± 1.5	-10.4 ± 1.4	-11.0 ± 1.5
CHROM [13]	11.0 ± 1.2	10.2 ± 1.2	17.3 ± 1.3	17.5 ± 1.4	17.2 ± 1.0	-3.7 ± 0.4	-3.7 ± 0.4	-4.9 ± 0.5	**-5.4±0.6**	**-5.3±0.5**
ICA [2]	13.1 ± 1.5	13.3 ± 1.9	16.6 ± 1.8	16.8 ± 1.7	16.7 ± 0.9	-4.3 ± 0.6	-4.6 ± 0.6	-5.4 ± 0.7	-5.7 ± 0.8	-5.4 ± 0.7
POS [14]	9.9 ± 1.5	9.3 ± 1.5	15.1 ± 1.7	15.7 ± 1.7	16.1 ± 1.4	-3.5 ± 0.6	**-3.4±0.6**	**-4.8±0.7**	-5.5 ± 0.8	-5.6 ± 0.6
GREEN [1]	17.3 ± 1.6	17.4 ± 1.6	17.2 ± 1.6	17.5 ± 1.6	17.1 ± 1.6	-6.3 ± 0.6	-6.4 ± 0.6	-6.4 ± 0.7	-6.4 ± 0.6	-6.4 ± 0.6
BCG [4]	–	–	–	–	–	–	–	–	–	

3.4. Compared benchmark methods

We compared the performance of the deep learning approach on each compression level to five signal processing methods described below.

CHROM [13]. This method uses a linear combination of the chrominance signals obtained from the RGB video. Spatially averaged, bandpass filtered and normalized intensity signals [y_r,

Biomedical Optics EXPRESS

y_g, y_b] are used to calculate S_{win}:

$$S_{win} = 3(1 - \frac{\alpha}{2})y_r - 2(1 + \frac{\alpha}{2})y_g + \frac{3\alpha}{2}y_b \tag{3}$$

Where α is the ratio of the standard deviations of the filtered versions of A and B:

$$A = 3y_r - 2y_g \tag{4}$$

$$B = 1.5y_r + y_g - 1.5y_b \tag{5}$$

The resulting outputs are scaled using a Hanning Window and summed with the subsequent window (with 50% overlap) to construct the final pulse signal.

ICA [2]. The pixel intensities are spatially averaged for each camera channel to form time varying signals [x_R, x_G, x_B]. The signals are detrended. A Z-transform is applied to each of the detrended signals. The Independent Component Analysis (ICA) (JADE implementation) is applied to the normalized color signals.

POS [14]. A linear combination of the color channels is used:

$$X_s = \bar{x}_g - \bar{x}_b \tag{6}$$

$$Y_s = -2\bar{x}_r + \bar{x}_g + \bar{x}_b \tag{7}$$

Where $\bar{x}_r$, $\bar{x}_g$, $\bar{x}_b$ are spatially averaged intensity signals normalized by dividing by their temporal mean. X_s and Y_s are then used to calculate S_{win}, where:

$$S_{win} = X_s + \frac{\sigma(X_s)}{\sigma(Y_s)}Y_s \tag{8}$$

The resulting outputs of the window-based analysis are used to construct the final pulse signal in an overlap add fashion.

GREEN [1]. The GREEN method uses spatially averaged pixel intensities from the green camera channel only. To obtain the pulse signal, we normalized the green channel signal by subtracting the mean, dividing by the standard deviation, bandpass filtering, and detrending.

BCG [4]. The ballistocardiogram (BCG) is a motion-based method which measures the subtle head motions caused by the influx of blood with each heartbeat. BCG tracks features on the head and decomposes the tracked trajectories into a set of motion components using Principal Components Analysis (PCA). It chooses the motion component corresponding to the pulse signal based on the frequency spectrum.

For a fair comparison to the end-to-end deep learning approach, we detected and cropped the face region in each frame for all the benchmark methods using MATLAB's face detection (`vision.CascadeObjectDetector()`). Signals output by all methods were filtered with a Butterworth bandpass filter with pass-band frequencies of [0.7, 2.5] Hz.

3.5. Compared compression formats

We compared the performance with three commonly used compression formats: MPEG-4, H.264, H.265, described below.

MPEG-4 is an older and very simple video compression format. It usually uses fixed pixel macroblocks of 16 x 16, limiting its compression capability at different image resolutions.

H.264 is a more recent compression format than MPEG-4. The advantage of H.264 over MPEG-4 is that it can use variable pixel block-size segmentation for compression. This allows for smaller pixel macroblocks when needed and avoids encoding errors. Furthermore, H.264 is more efficient at motion prediction than MPEG-4 because it can encode repetitive patterns in the video more efficiently based on the surrounding pixel macroblocks.

Biomedical Optics EXPRESS

H.265 is the most recent and the most efficient of the three compression formats. Compared with H.264 and MPEG-4, H.265 has a higher compression ratio and can provide better image quality at the same bitrate. H.264 allows for at most 16 x 16 pixel macroblocks which are too small to be efficient in high resolution videos. H.265 can use larger 64 x 64 pixel macroblocks, making its encoding more efficient at all resolutions. Moreover, H.265 offers improved and more detailed motion prediction compared to H.264 and uses an additional filter that reduces artifacts at macroblock edges.

4. Experiments

We present results of HR measurements at different levels of compression. We analyze how the presented deep learning approach generalizes to videos with different compression levels and different compression algorithms. We also analyze whether darker skin types are more affected by higher compression.

4.1. Varying compression levels

We compressed the original videos, already moderately compressed with CRF 12, to obtain videos at five levels of compression using CRF of 12, 18, 24, 30 and 36. This results in approximate bitrates of 890, 534, 110, 87, 67 kb/s, respectively on videos with stationary individuals and 1500, 1077, 230, 221, 88 kb/s respectively on videos with large head motions. Compression rates larger than CRF = 36 destroyed most of the iPPG information [16], therefore we do not use more aggressive compression rates. We used an open-source codec producing H.264, H.265 and MPEG-4 compliant videos and the latest FFmpeg Windows 64-bit binary release (at the time of testing: N-94150-g231d0c819f).

For each compression level, we compared two deep learning approaches. First, we trained and tested the deep learning model on videos compressed with the same CRF as the test data. Second, we trained the deep learning model only on the original videos compressed with CRF = 12 and tested on videos compressed with different CRFs. The goal of these experiments was to test whether training on compressed videos, which are more noisy but are more similar to the test data, performs better than training on less compressed videos which have cleaner iPPG signals but are less similar to the test data. For all experiments we used a five-fold cross-validation where we used different subjects in the training and testing sets. For each fold, the training set contained 20 subjects and the test set contained five different subjects (i.e. subject independent validation). The presented results are the mean absolute heart rate error and BVP signal to noise ratio averaged over the five folds. We also compared the performance of these two deep learning approaches to non-machine-learning methods (CHROM, ICA, POS, GREEN, BCG). The results for these experiments at different compression levels are summarized in Table 1 and Fig. 3. We found that training on data with the same level of compression as the test set performs better than training on videos with lower compression (at CRF = 12), as expected. The highest compression level (CRF = 36) almost completely removed the pulse signal and the estimated HRs were close to random.

Unsupervised methods (CHROM, ICA, POS, and GREEN) are also affected by compression (see Table 1). POS performs better than the deep learning model trained on CRF = 12 at high compression levels, but not as well as the deep learning model trained on matching CRF (see Tables 1 and Fig. 3). We only compared the performance of the motion-based BCG method on the stationary videos, because this method is very sensitive to rigid head motions and performs poorly on videos with motion. BCG is not affected by the compression as much as methods using iPPG, suggesting that compression removes more intensity information than motion information (see Fig. 3). However, the results using the BCG method are poor overall.

Biomedical Optics EXPRESS

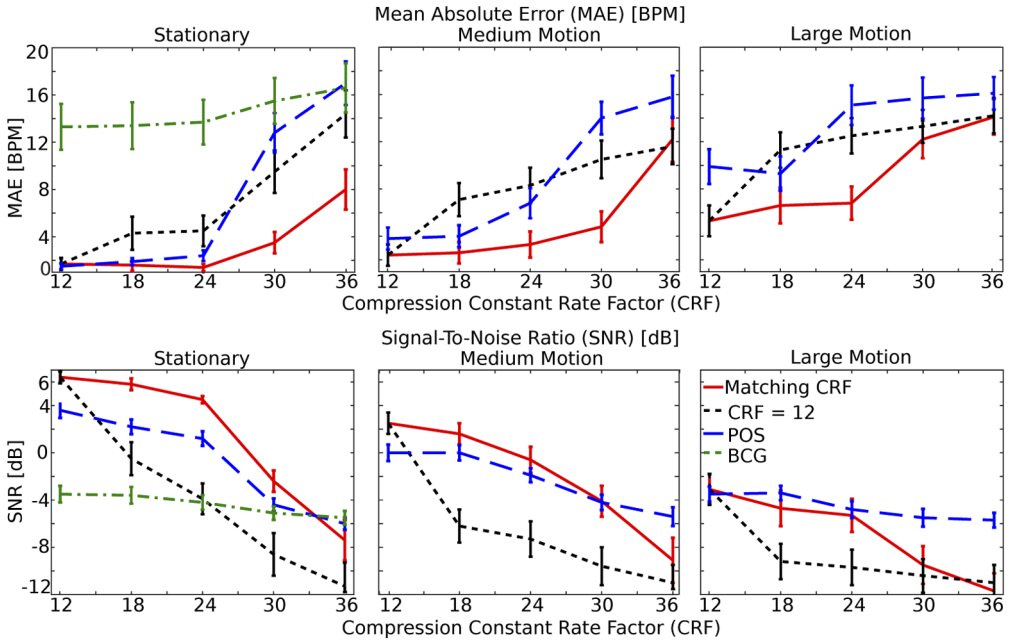

Fig. 3. Results at varying compression levels. Training on compressed videos with matching CRF leads to lower MAE and higher SNR. Unsupervised methods, such as POS are also affected by compression but they perform better than models trained on videos compressed with CRF=12 at higher compression levels.

4.2. *Generalizability across compression levels*

In realistic settings it may not always be possible to use a model trained on videos with exactly the same compression. Moreover, having to train and store a separate model for each possible compression may be expensive. We tested how well models trained on different compression levels generalize to more and less compressed videos. The results for these experiments are shown in Fig. 4. We found that while the performance drops when the model is trained on a different compression level, the model is able to generalize with sufficient accuracy to videos compressed with different CRF. On stationary videos the models can generalize to videos compressed differently by up to 12 CRF levels from the training set CRF. Moreover, we found that training on videos with higher compression and testing on less compressed data performs better than the opposite. The model is able to learn the compression noise better from more heavily corrupted videos and generalize to cleaner data. On videos with motion tasks, the model is not able to generalize as well to different compression levels.

4.3. *Generalizability across compression algorithms*

In addition to different compression levels, different compression algorithms may be used to compress videos. We compared the generalizability of the deep learning models across three commonly used compression algorithms: H.264, H.265 and MPEG-4. For these experiments we used moderately compressed videos with CRF = 18.

We found that the deep learning models do not generalize across different compression algorithms as well as they do across different compression levels (see Fig. 5). Each compression format removes intensity information differently, which affects the iPPG signals in different ways. Therefore, training on videos compressed with one compression algorithm makes it challenging for the model to remove compression artefacts introduced by another compression algorithm.

Biomedical Optics EXPRESS

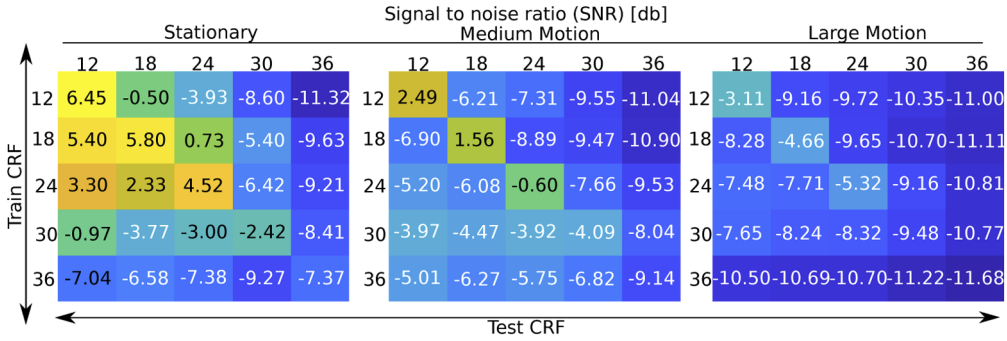

Fig. 4. Generalizability across compression levels. The closer the training set CRF is to the test set CRF, the better the performance. However, training on more compressed data and testing on less compressed performs better than the opposite.

Furthermore, we found that H.254 was the best and MPEG-4 the worst for pulse measurement. In the experiments across different levels of compression we show that training on more compressed data performs better on less compressed data, than the other way around. Similarly, we found that training on MPEG-4 and testing on others was better than the opposite. Training on H.265 and testing on H.264 also seemed slightly better than the opposite but there was no consistent trend. The reason why H.254 performs better than the more recent H.265 could be that H.265 might be removing and filtering more subtle information crucial for iPPG.

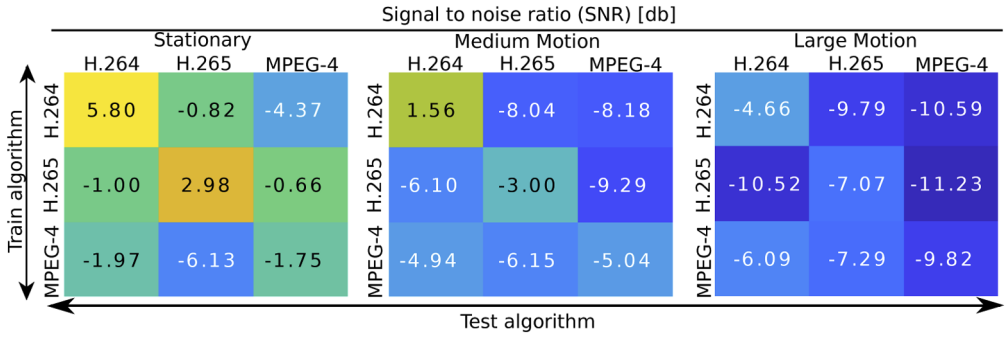

Fig. 5. Generalizability across compression algorithms. The models do not generalize well across different compression algorithms, especially in the presence of motion.

4.4. Effect of compression on different skin types

Darker skin types have a higher melanin concentration which absorbs more light and leads to lower iPPG SNRs [22]. We analyzed how compression affects iPPG signals for subjects across a range of skin types and whether iPPG signals extracted from participants with darker skin types tend to perform worse at higher compression because of the lower SNR. We only considered the stationary task for these experiments to separate the effects of compression from motion. The SNR results for different skin types and different levels of compression are shown in Fig. 6.

There is a benefit to training on videos with matching compression level for all skin types. At higher compression darker skin types are more affected by the compression artefacts than lighter skin types. However, at low compression the trend is not consistent across skin types. The results on skin types II and III are a bit worse than I, but skin type IV is better. The average improvements achieved by training on matching compression were: 3.2 dB, 6.2 dB, 6.5 dB, 8.3

Biomedical Optics EXPRESS

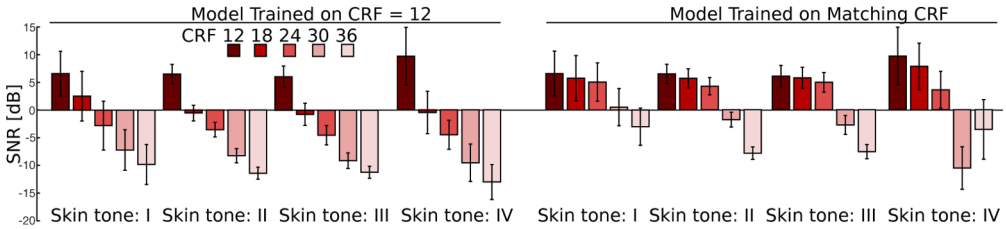

Fig. 6. Signal-to-noise ratio (SNR) at different levels of compression for different skin types. Darker skin types are more affected at large compression.

dB for skin types I, II, III, IV, respectively on CRF = 18, and 6.8 dB, 3.6 dB, 3.6 dB, 9.5 dB on CRF = 36. However, there was only one participant in skin type category I and IV in the dataset we used, making it difficult to draw definitive conclusions about whether darker skin types are more prone to compression artefacts.

5. Discussion

We have systematically compared the performance of pulse wave recovery and heart rate estimation using a supervised deep neural network. Our results show that training on compressed videos has a significant advantage when testing on compressed videos with a similar CRF. Unsupervised methods (e.g., POS) are also affected by video compression but they can perform better at high compression levels when compared to deep learning models trained only on high-quality less compressed videos.

5.1. Challenges with generalization

Models trained on higher levels of compression can generalize to videos with similar and lower compression levels. However, there is a limited generalization across different compression levels in presence of motion and little generalization across different compression formats. This limited generality may not necessarily be a hindrance for certain applications which may have access to information about how the testing videos are compressed. In these situations appropriate models can be selected in advance based on the compression level.

5.2. Effects of compression on pulse frequency

In order to understand how video compression affects the pulse signals, we plotted spectrograms of pulse signals obtained from videos with different compression levels. We compared spectrograms of pulse signals obtained with a model trained on CRF = 12, a model trained on matching CRF and the GREEN method (see Fig. 7). We compared the spectrograms of the deep learning models to the spectrograms of the GREEN method in order to analyze how the intensity signals are affected by compression without denoising with deep learning. As the compression increases, the pulse signal amplitude decreases and noise is introduced randomly at all frequencies. On videos with large rigid head motions, a strong frequency peak is visible at 1 Hz corresponding to the frequency of the head motion, especially at higher compression levels. However, videos with medium head motion also contain strong frequency peaks at 1 - 1.3 Hz, especially at high compression. We are uncertain what causes this effect but it could be caused by the way key frames are computed during compression, or the post-processing and filtering steps. It is unlikely that our model is “guessing" the most common HR frequency in the dataset because it is only trained on two input video frames, making it unlikely to overfit to the frequency information. Models trained on CRF = 12 are not able to recover the iPPG signal frequency as well as the models trained on matching CRF videos.

Biomedical Optics EXPRESS

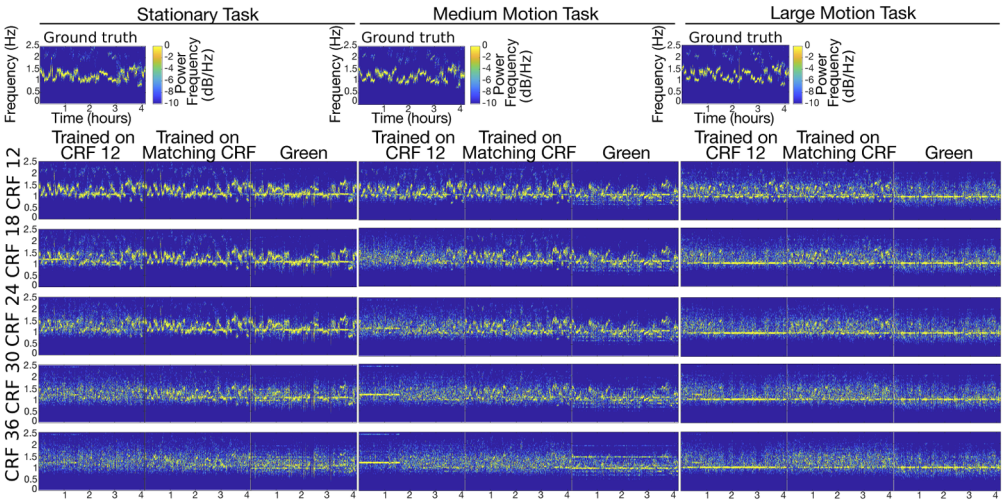

Fig. 7. Spectrograms of iPPG signals obtained at different compression levels with a model trained on CRF = 12, a model trained on matching CRF, and GREEN method.

5.3. Learned spatial features

In order to understand why training on videos with matching CRF performs well we visualized what the network is paying attention to at each compression level. The convolutional attention network that we used outputs, in addition to the pulse signal, the attention masks which show which regions in the video frame were used to compute the signals. Examples of attention masks output on test set images at different levels of compression are shown in Fig. 8. At low compression the attention masks mostly focus on facial skin regions. As the compression increases, models trained on matching compression tend to focus more around the edges (and thus motion). However, models trained on CRF = 12 consistently output masks lending more attention to skin regions (and thus color changes), regardless of the compression level. Even though the models trained on CRF = 12 are able to learn the spatial and color information more effectively, they are not able to recover iPPG signals as well as models trained on matching compression levels. The reason for this could be that the models trained on matching CRF are able to learn to place additional weight on motion features related to BCG signals rather than solely relying on color changes (that are distorted or removed in highly compressed videos). Compression might remove more color information than motion, therefore, learning to find skin regions well may not be as useful as being able to focus on the edges which might have more useful motion information. This claim is supported by the results shown in Table 1 and Fig. 3. BCG is not affected by compression as much as other methods relying on the intensity-based iPPG signals.

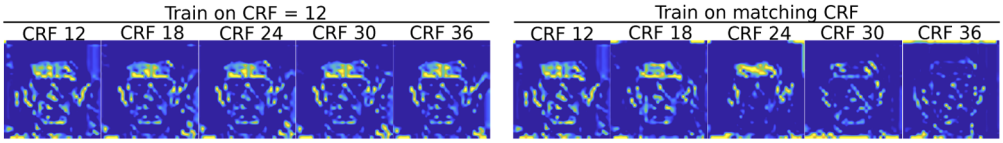

Fig. 8. Attention masks output by the deep learning model at different compression levels for 1) model trained on CRF = 12 and 2) model trained on matching CRF. As the compression increases, models trained on matching CRF focus more around the edges of the face, whereas models trained on CRF = 12 consistently focus on the skin regions.

Biomedical Optics EXPRESS

6. Conclusions

Video compression impacts the performance of imaging-based physiological measurement algorithms, including machine learning and traditional signal processing methods. We have presented a systematic analysis of the performance of these methods at varying compression levels and different compression formats. We demonstrated that it is possible to obtain reliable pulse measurements and heart rate estimates from compressed videos even in the presence of large rigid head motions, so long as the network is trained with examples of videos compressed with the same or higher compression level.

We have shown that models trained on videos with the same level of compression as the videos in the test set achieve the best performance, as expected. However, models trained on one compression level can generalize to videos with different compression levels between CRF 12 and CRF 24. The models generalize best when they are trained on higher levels of compression than the compression of the videos in the test set. When large motion is present it is more difficult to generalize to different compression levels. Generalizing across different compression formats is more challenging because each compression algorithm removes color and temporal information differently.

Different videos and different datasets may contain very different content and different amount of motion, resulting in variable compression parameters at the same CRF level. A limitation of our method is that a model trained on the same CRF of one dataset may generalize less to videos of a different dataset with the same CRF but different content and different motion. However, we expect a model trained on the same CRF will perform better than algorithms trained on different compression.

With supervised neural models becoming more and more popular for iPPG, we hope that this work will help advance the state-of-the-art in image-based physiological measurement by demonstrating that compressed videos may be used for these measurements, alleviating the cost of storing uncompressed videos.

Funding. National Science Foundation (CCF-1730574, CNS-1801372, EEC-1648451).

Disclosures. The authors declare no conflicts of interest.

References

1. W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," Opt. Express **16**(26), 21434–21445 (2008).
2. M.-Z. Poh, D. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," Opt. Express **18**(10), 10762–10774 (2010).
3. M.-Z. Poh, D. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," IEEE Trans. Biomed. Eng. **58**(1), 7–11 (2011).
4. G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2013), pp. 3430–3437.
5. Y. Ethawi, A. Al Zubaidi, G. Schmölzer, S. Sherif, M. Narvey, and M. Seshia, "Clinical applications of contactless imaging of neonates using visible, infrared light and others," Adv. Biomed. Sci. **3**, 39 (2018).
6. C. C. Yang, C.-W. Lai, H. Y. Lai, and T. B. Kuo, "Relationship between electroencephalogram slow-wave magnitude and heart rate variability during sleep in humans," Neurosci. Lett. **329**(2), 213–216 (2002).
7. E. M. Nowara, T. K. Marks, H. Mansour, and A. Veeraraghavan, "Sparseppg: Towards driver monitoring using camera-based vital signs estimation in near-infrared," in *Computer Vision and Pattern Recognition (CVPR), 1st International Workshop on Computer Vision for Physiological Measurement*, (2018).
8. D. McDuff, J. Hernandez, S. Gontarek, and R. W. Picard, "Cogcam: Contact-free measurement of cognitive stress during computer tasks with a digital camera," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, (ACM, 2016), pp. 4000–4004.
9. X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Advances in Neural Information Processing Systems* (2020).
10. J. R. Estepp, E. B. Blackford, and C. M. Meier, "Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, (IEEE, 2014), pp. 1462–1469.
11. W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), pp. 349–365.

Biomedical Optics EXPRESS

12. S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), pp. 2396–2404.
13. G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rppg," IEEE Trans. Biomed. Eng. **60**(10), 2878–2886 (2013).
14. W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," IEEE Trans. Biomed. Eng. **64**(7), 1479–1491 (2017).
15. M. Kumar, A. Veeraraghavan, and A. Sabharwal, "Distanceppg: Robust non-contact vital signs monitoring using a camera," Biomed. Opt. Express **6**(5), 1565–1588 (2015).
16. D. McDuff, E. B. Blackford, and J. R. Estepp, "The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, (IEEE, 2017), pp. 63–70.
17. E. Nowara and D. McDuff, "Combating the impact of video compression on non-contact vital sign measurement using supervised learning," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, (2019), p. 0.
18. D. McDuff, "Deep super resolution for recovering physiological information from videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (2018), pp. 1367–1374.
19. Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement," arXiv preprint, arXiv:1907.11921 (2019).
20. M. Rapczynski, P. Werner, and A. Al-Hamadi, "Effects of video encoding on camera-based heart rate estimation," IEEE Trans. Biomed. Eng. **66**(12), 3360–3370 (2019).
21. T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," Arch. Dermatol. **124**(6), 869–871 (1988).
22. E. M. Nowara, D. McDuff, and A. Veeraraghavan, "A meta-analysis of the impact of skin tone and gender on non-contact photoplethysmography measurements," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (2020), pp. 284–285.