Combating the Impact of Video Compression on Non-Contact Vital Sign Measurement Using Supervised Learning

Ewa Magdalena Nowara
Rice University
Houston, TX
emn3@rice.edu

Daniel McDuff
Microsoft
Redmond, WA
damcduff@microsoft.com

Abstract

Imaging photoplethysmography (iPPG) and imaging ballistocardiography (iBCG) are popular approaches for unobtrusive camera-based measurement of vital signs. These involve recovering pulse signals from very subtle variations in video pixel intensities, which are easily corrupted by noise. Therefore, while the signal might be easy to obtain from high quality uncompressed videos, the signal-to-noise ratio drops linearly with video bit-rate. Uncompressed videos require large amounts of storage making them prohibitive to store, stream and transfer in large quantities. By learning compression specific models we show that supervised learning can be used to increase the signal-to-noise ratio (SNR) of pulse signals and reduce the mean absolute error (MAE) of heart rate estimates extracted from temporally compressed videos. We perform a systematic evaluation of the performance of our algorithm showing that the network trained on compressed videos consistently outperforms the model trained on the original less compressed compressed videos, both on videos with and without significant head motions. We found improvements in SNR of up to 8 dB and MAE of 6 BPM.

1. Introduction

Imaging photoplethysmography (iPPG) and imaging ballistocardiography (iBCG) leverage subtle changes in light reflected from the skin and motions of the body, respectively, to capture cardiac activity. These signals can be recovered from many types of commercially available and low-cost cameras (e.g., webcams, cellphones and DSLRs). Research over the past decade has shown that heart rate (HR) [17, 14, 1], heart rate variability (HRV) [13] and breathing rate (BR) [13] can be measured from video recordings of the human body. Measuring vital signs remotely with a camera offers several advantages over the contact devices traditionally used in pulse oximetry or electrocardiograms. 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, and reduce the risk of infection to, prematurely born babies [7, 5] and burn victims. 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 [19], driver monitoring [12] or cognitive engagement with computer applications [11].

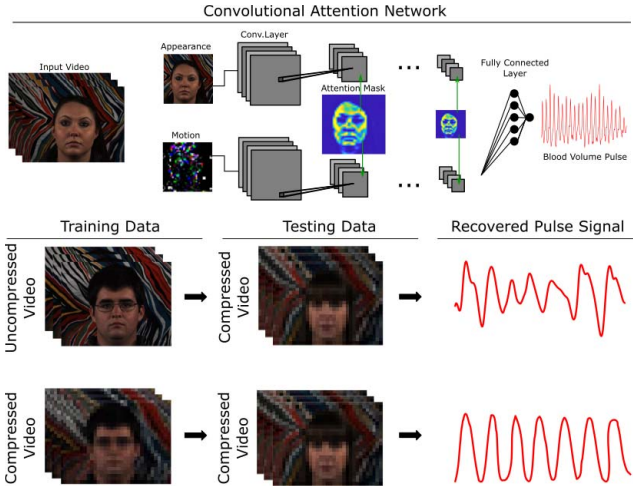

Figure 1. We show how a supervised deep-learning-based network can be used to combat the impact of video compression on non-contact physiological measurements. By training on compressed videos we achieve better performance when testing on compressed videos than when training on high-quality uncompressed videos.

A number of approaches have been proposed to achieve motion robustness in imaging-based measurement of physiology [14, 3, 16, 4, 18]. The existing methods achieve accuracy in heart rate estimation comparable to contact devices in many natural indoor applications provided that the video

was recorded with sufficiently high quality. Several video parameters, have been systematically studied and shown to impact the quality of iPPG signals, such as the camera quality [15], frame rate [2], the amount of spatial [2, 8] and temporal compression [9]. For a survey see [10].

Both iPPG and iBCG signals are very subtle and consequently most video datasets are captured high with bitrates to maximize the signal strength. To achieve the highest pulse signal-to-noise ratio videos need to be recorded as raw lossless images demanding enormous amounts of storage [9]. For example, a 5.5 minute video of one subject is on average 11.9 GB. Collecting, storing, streaming and transfering such datasets becomes challenging as the number of subjects, conditions, and durations of the recordings increase. This hinders the sharing of datasets and applications of imaging-based physiological measurement in applications such as telemedicine. Being able to use spatially and temporally compressed videos would help address these challenges, presenting new applications for the technology, making it easier to share video recordings and helping to advance the state-of-the-art in research. Not having to store raw images would also make collecting data easier, allowing researchers to record videos on any device with the default video settings. Furthermore, reducing video bit-rates would help facilitate training algorithms on datasets with a larger number of videos per batch, and thus benefit from the scalability of deep neural architectures [3].

The problem of very low spatial resolution has been addressed by using deep image super-resolution, making it possible to use heavily spatially compressed images and still recover the pulse signal [8]. 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. However, while it has been shown that increasing the amount of temporal compression linearly decreases the iPPG SNR [9], there has only recently been work specifically addressing the detrimental effects of temporal compression on iPPG signals [20]. Yu et al. propose a method that requires a two step process involving a network for video enhancement followed by a network for recovering the pulse signal. We choose to tackle the problem in an end-to-end fashion and train a signal network to recover the pulse signal from compressed videos.

We show that training the deep learning models on compressed videos performs significantly better than using deep learning models trained on clean, less compressed data, as illustrated by Figure 1. We use an attention-based deep learning approach [3], which outperformed all state-of-the-art iPPG methods on several publicly available datasets and a large dataset with different motion tasks [6], to evaluate our approach at five compression levels.

2. Background

2.1. Video Compression

Raw video has a large memory footprint, therefore compression algorithms are used in almost all video systems. Applications that use 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). There are several video compression methods which reduce the bit-rate of the video while retaining information important for visual quality. These algorithms may use intra-frame compression, inter-frame compression, or they may jointly use both intra- and inter-frame compression mechanisms.

Intra-frame compression uses the correlation between similar pixels located close to each other in an image. Predicted pixel values are computed by extrapolating from a small number of already coded pixels. Each frame is divided into blocks of pixels. Each block of pixels is then spatially compressed by applying a discrete cosine transform (DCT), dividing by a compression matrix (also called the quantization matrix) and rounding to reduce the number of coefficients required to represent the image. The same intra-frame compression methods are used for image compression, for example with JPEG coding.

Inter-frame compression is performed for a group of consecutive video frames. It uses reference frames (I-frames) which may be first compressed with intra-frame compression and motion vectors. The motion vectors may describe the difference between the current and the previous frame called predicted frames (P-frames), as well as the difference between the current frame and both the previous and the next frame called bi-directionally predicted frames (B-frames). This allows significant reductions in the amount of storage required for regions in the video where there are not large changes between the frames. The P- and B-frames are placed in between the I-frames and similarly to intra-frame compression are transformed and quantized to reduce the memory. Intra- and inter-frame compression algorithms are often used together for greater efficiency. In order to maintain a constant compression quality across videos with different information adaptive quantizers, such as constant compression rate factor (CRF) are used. 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 to reduce bit-rates when visual quality is important.

Video compression methods are typically optimized for visual quality, not with physiological measurement in mind. Compression algorithms often assume that small color vari-

ations 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. However, imaging-based physiological measurement relies on those small variations and therefore compression algorithms significantly degrade the quality of the iPPG and to a lesser extent iBCG signals by removing that subtle information. At higher CRF values, the video is more compressed and the SNR of the pulse signals decreases more or less linearly with CRF [8].

2.2. Imaging-Based Physiological Measurement

As imaging-based physiological measurement matures, the focus has moved towards increasing robustness to body motions [16, 18, 3]. Supervised deep learning methods have proven particularly successful [3]. However, significantly less work has attended to the effects of video quality on iPPG signals and most existing methods become less robust to motion with decreasing video quality. McDuff et al. [8] 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. [20] 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 models trained directly on temporally compressed videos to recover the iPPG signals without the requirement to first enhance the video. We use an end-to-end framework to measure pulse signals from input video without the necessity to first detect the face and segment the skin regions [3]. We show that training and testing on videos with the same compression level outperforms models trained on the original less compressed videos.

3. Experiments

3.1. Deep Learning Architecture

Our goal is to systematically analyze the effect of video compression during training of a supervised network. For this purpose, we used the existing state-of-the-art convolutional attention network architecture [3]. It uses a motion representation and an attention mechanism using the appearance information to discriminate between the different motion sources. The motion representation is computed from a normalized frame difference based on a skin reflection model [18]. The appearance information is computed from the color and texture information from input image frames to guide the motion representation to recover physiological information from the skin region and differentiate it from other sources of variations, such as head motion or non-uniform illumination. The appearance and motion representations are learned jointly through an attention mechanism. As illustrated in Figure 1, DeepPhys uses two separate models trained on the motion representation from the difference of two consecutive frames and the appearance representation from the current input frame. The appearance model has the same architecture as the motion model but without the last three layers.

The frame difference used as input to the motion model is computed as follows:

$$D_l(t) = \frac{C_l(t + \Delta t) - C_l(t)}{C_l(t + \Delta t) + C_l(t)} \quad (1)$$

Where *C(t)* is the current raw RGB image frame and Δt is the time between frames, in this case 1/120 seconds. The illumination intensity is not spatially uniform on the face and changes with the changing distance of the skin to the light source causing uneven intensities and shadows on the face. These variations in videos used as training data would hinder the supervised learning model. Therefore, the input to the motion network is first normalized to reduce the dependency of the input signals on the light source and skin tone which vary across subjects and datasets. AC/DC normalization is applied once for the entire video duration by subtracting the temporal mean and dividing by the standard deviation.

The input to the appearance model are the raw RGB image frames, *C(t)*, normalized by centering to zero mean and scaled to unit standard deviation. We spatially averaged the input images to 36 x 36 pixels, using a bicubic interpolation,

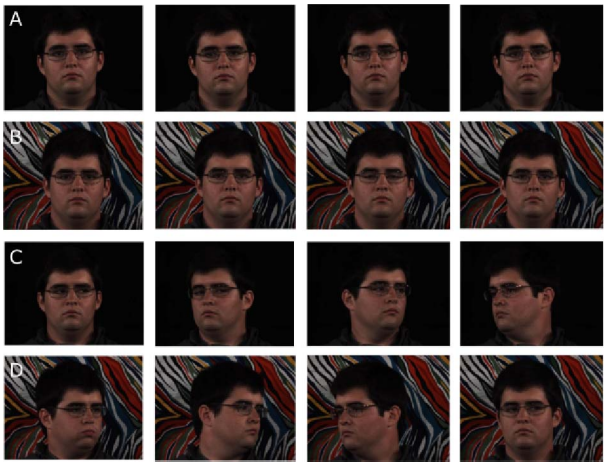

Figure 2. Examples of face images featuring stationary tasks (A and B) and random motion tasks (C and D). Each task was repeated with a solid background (A and C) and a background with texture (B and D).

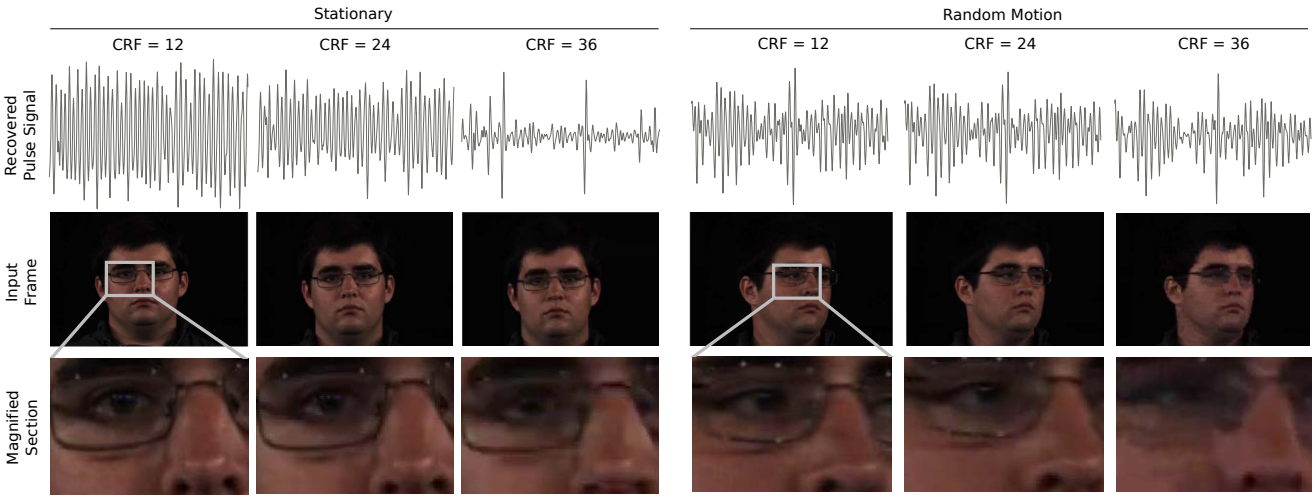

Figure 3. Examples of the detrended and filtered green channel signals for the stationary (left) and random motion (right) tasks at different compression rate factors (CRF = 12, 18, 24, 30, 36). Examples of face images from these videos are shown, as are zoomed in segments. At greater CRFs the images become more blurry, especially in the motion condition, and the signal-to-noise ratio of the iPPG signal in the green channel decreases.

to reduce the camera quantization noise.

We used average pooling instead of max pooling because in traditional iPPG applications the signal quality can be improved when several weaker and stronger features are combined instead of only keeping the strongest one. Two-dimensional facial attention masks were estimated using a 1×1 convolution filter right before every pooling layer to synthesize masks from different levels of appearance features. The spatial maps guide the motion model by determining which pixels belong to the skin regions and whether they contain iPPG signals. Hyperbolic tangent (tanh) was used as the activation functions because symmetry improved iPPG performance.

The last layer of the motion model has linear activation units and a mean squared error (MSE) loss to form a continuous signal as output. 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

We used the dataset of facial videos collected by Estepp et al. [6]. Examples of the images in the dataset are presented in Figure 2. Videos of 25 participants, aged 18 to 28 years, were recorded with a Basler Scout scA640-120gc GigE-standard color camera with a 16 mm fixed focal length lens. The images were recorded as 8-bit, 658x492 pixel resolution at 120 frames per second (FPS). Seventeen of the participants were male, nine wore glasses, eight had facial hair and four had makeup. The dataset features participants with diverse skin tones estimated with the following Fitzpatrick Sun-Reactivity Skin Types [9]: I-1, II-13, III-10, IV-2, V0. Ground truth contact physiological signals were measured simultaneously with each video recording using a BioSemi ActiveTwo research-grade biopotential acquisition unit. The participants were recorded during six five-minute tasks. Each task was recorded with a black uniform background and repeated with a textured background. We evaluated two representative tasks in this work.

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

Random Motion Task: The participants were asked to randomly reorient their head once every second towards one of nine positions evenly spaced in an arc around them. The random sequence of which point to look at was provided during the data collection as an audio recording. 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 ($\sim$ 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). 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. Mean absolute error (MAE):

$$\text{MAE} = \frac{\sum_{i=1}^{N} |R_i - \widehat{R}_i|}{N} \tag{2}$$

Table 1. Mean absolute errors (MAE) and signal-to-noise ratios (SNR) at different levels of compression for networks trained on a) less compressed videos and b) compressed videos. Training on compressed videos was performed at the same CRF as those videos in the testing set. Results are for participant independent experiments. Overall, training on compressed videos leads to lower MAE and higher SNR than training on less compressed videos. Bold numbers reflect significantly lower MAE/higher SNR.

	Mean Absolute Error (BPM)				**Signal-to-Noise Ratio**			
	Stationary		Random Motion		Stationary		Random Motion	
	Trained on CRF = 12	Trained on Matching CRF	Trained on CRF = 12	Trained on Matching CRF	Trained on CRF = 12	Trained on Matching CRF	Trained on CRF = 12	Trained on Matching CRF
CRF 12	1.76 ± 0.29	1.76 ± 0.29	4.94 ± 0.9	4.94 ± 0.9	8.09 ± 0.76	8.09 ± 0.76	-1.86 ± 0.71	-1.86 ± 0.71
CRF 18	4.24 ± 0.81	$\mathbf{1.54 \pm 0.17}$	11.3 ± 1.32	$\mathbf{6.46 \pm 1.05}$	-0.45 ± 0.67	$\mathbf{5.8 \pm 0.71}$	-9.03 ± 0.96	$\mathbf{-4.56 \pm 0.72}$
CRF 24	4.65 ± 0.61	$\mathbf{1.55 \pm 0.17}$	12.09 ± 1.3	$\mathbf{6.74 \pm 1.04}$	-3.6 ± 0.62	$\mathbf{4.84 \pm 0.67}$	-9.42 ± 1.06	$\mathbf{-4.9 \pm 0.61}$
CRF 30	9.55 ± 0.87	$\mathbf{3.4 \pm 0.41}$	13.27 ± 1.24	12.09 ± 1.05	-8.56 ± 0.45	$\mathbf{-2.53 \pm 0.61}$	-10.29 ± 1.09	-9.5 ± 0.53
CRF 36	10.69 ± 0.9	$\mathbf{6.84 \pm 0.73}$	14.3 ± 1.32	15.33 ± 0.98	-9.51 ± 0.4	$\mathbf{-6.78 \pm 0.41}$	-11.01 ± 1.14	-11.79 ± 0.46

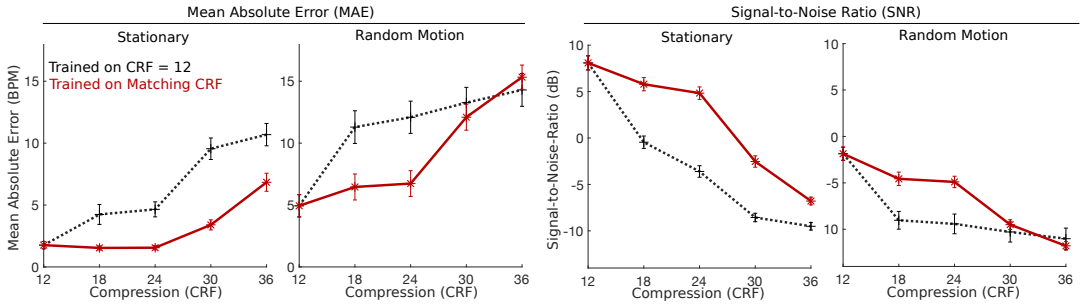

Figure 4. Mean absolute error in heart rate estimation in BPM and SNR of the recovery pulse wave in dB at different compression rate factors. We compare the network trained and tested on compressed videos and trained on less compressed videos and tested on compressed videos. As the compression level increases training on less compressed videos leads to worse performance (higher MAE and lower SNR) than when training and testing on videos with the same compression level. The error bars reflect the standard error in the measures.

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 [4]:

$$\text{SNR} = 10 \log_{10} \left(\frac{\sum_{42}^{240} ((U_t(f)S(f))^2}{\sum_{42}^{240} ((1 - U_t(f))S(f))^2} \right) \quad (3)$$

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.

The error bars 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.

3.4. Video Compression Experiments

McDuff et al. [9] showed that the SNR of iPPG signals decreases more or less linearly with increasing compression rate up to CRF = 36. Compression rates larger than CRF = 36 destroyed most of the iPPG information. The work also compared using x264 and x265 codecs and found that x264 performed worse on compressed videos in presence of motion. In this work we tested five levels of compression using CRF of 12, 18, 24, 30 and 36, corresponding to approximate bit rate of 890, 534, 110, 87, 67 kb/s on videos with stationary tasks and 1500, 1077, 230, 221, 88 kb/s on videos with random motion tasks.

We compressed the original less compressed videos using the x264 codec. We used an open-source codec producing H.264 compliant video and the latest FFmpeg Windows 64-bit binary release (at the time of testing: N-94150-g231d0c819f). Figure 3 shows examples of iPPG waveforms computed using spatially averaged green channels of the videos at different levels of compression with the corresponding images. We chose to use the green channel to demonstrate the impact of compression on iPPG signals independent of the post-processing algorithms because this channel has the strongest iPPG signal [17]. We spatially averaged the video frames, detrended and bandpass filtered the resulting signals using a passband of 0.7 Hz to 2.5 Hz. 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, especially on the random motion task.

For each compression level, we performed two experiments. First, we trained and tested the deep learning model on compressed videos. Second, we trained the deep learning model on the original less compressed videos and tested on compressed videos. The original videos we used as benchmark were already moderately compressed with CRF = 12. 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. The results for these experiments at different compression levels and compared to the performance on the original less compressed videos as benchmark are summarized in Table 1 and Figure 4. For all experiments we used a five-fold cross-validation where we used different subjects in the training and test set. For each validation fold, the training set contained 20 subjects and the test set contained five different subjects. The presented results are the means averaged over the five validation sets. We found that training on data with the same level of compression as the test set performs better than training on videos with less compression (at CRF = 12). The highest compression level (CRF = 36) almost completely removes the pulse signal and the estimated HR is close to random, making both DeepPhys networks trained on videos with both compression levels perform comparably poorly.

4. Discussion

We 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 compared to training on less compressed videos. The SNR of the recovered blood volume pulse is largest when training on videos of match compression level. As videos are compressed the performance of the neural network trained on CRF=12 videos decreases with the compression factor (see Figure 4 - black dotted line). However, when we retrain the network with videos at a similar compression factor as those in the testing set the performance is more robust until CRF is greater than 24 (see Figure 4 - red solid line). This suggests that the network architecture is able to learn compression specific information that helps in the recovery of the pulse signal when videos are compressed. Our results illustrate that it is not always beneficial to train on "cleaner" data if that data differs from the domain of application of the model. Many datasets are compressed and therefore it may be important to train models on compressed data, or at least include a proportion of compressed videos in the training set. Our results show that if this is not done then at higher compression a supervised deep neural network may perform significantly more poorly. One could imagine training several iPPG models and at test time selecting the model that best matches the compression level of the video being processed, thus being able to adapt the model to suit the data.

5. Conclusions

Video compression (both intra-frame and inter-frame) algorithms impact the performance of imaging-based physiological measurement algorithms, including iPPG and iBCG. We have presented a systematic analysis of the performance of training a supervised deep neural network on temporally compressed videos. We have shown that the performance improves when the model is trained on videos with the same level of compression as the videos in the test set, instead of training on less compressed videos with a higher SNR. Our proposed approach shows that it is possible to obtain reliable pulse measurements and heart rate estimates from compressed videos even in presence of large motion, so long as the network is trained with examples of compressed videos. All our experiments were conducted in a participant independent manner. We hope that this work helps advance the state-of-the-art in image-based physiological measurement by alleviating the time and memory requirements involved in the storage of uncompressed raw images or videos.

References

[1] 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*, pages 3430–3437, 2013.

[2] E. B. Blackford and J. R. Estepp. Effects of frame rate and image resolution on pulse rate measured using multiple camera imaging photoplethysmography. In *Medical Imaging 2015: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 9417, page 94172D. International Society for Optics and Photonics, 2015.

[3] W. Chen and D. McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365, 2018.

[4] G. De Haan and V. Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60(10):2878–2886, 2013.

[5] G.-F. Deng, Y.-S. Hung, W.-K. Ho, and H.-H. Lin. Remote measurement of infant emotion via heart rate variability.

[6] 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)*, pages 1462–1469. IEEE, 2014.

[7] 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. 2018.

[8] D. McDuff. Deep super resolution for recovering physiological information from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1367–1374, 2018.

[9] 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)*, pages 63–70. IEEE, 2017.

[10] D. McDuff, J. R. Estepp, A. M. Piasecki, and E. B. Blackford. A survey of remote optical photoplethysmographic imaging methods. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 6398–6404. IEEE, 2015.

[11] 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*, pages 4000–4004. ACM, 2016.

[12] 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.

[13] M.-Z. Poh, D. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE transactions on biomedical engineering*, 58(1):7–11, 2010.

[14] M.-Z. Poh, D. McDuff, and R. W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010.

[15] Y. Sun, V. Azorin-Peris, R. Kalawsky, S. Hu, C. Papin, and S. E. Greenwald. Use of ambient light in remote photoplethysmographic systems: comparison between a high-performance camera and a low-cost webcam. *Journal of biomedical optics*, 17(3):037005, 2012.

[16] 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*, pages 2396–2404, 2016.

[17] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. *Optics express*, 16(26):21434–21445, 2008.

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

[19] 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. *Neuroscience letters*, 329(2):213–216, 2002.

[20] 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.