Eliminating Physiological Information from Facial Videos

Weixuan Chen and Rosalind W. Picard
Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA

***Abstract*—Vital signs, cognitive load, and stress can be remotely measured from human faces using video-capturing devices under ambient light, which raises both wide applications and privacy issues. To avoid immoral use of this technology, there is a need for methods to eliminate physiological information from facial videos without affecting their visual appearance. To meet the need, we develop a novel algorithm based on motion component magnification that inputs a video and outputs its replica with physiological signals removed. Facial video data has been collected from 18 participants in a study to assess the performance of our algorithm in thwarting heart rate measurement based on remote photoplethysmography. Our results show that the mean absolute error of heart rate measurement averaged among participants was increased from 0.254 beats per minute to above 17 beats per minute without causing visible artifact. This is the first demonstration of an algorithm that can achieve this kind of functionality.**

I. INTRODUCTION

We are living in a world where we are surrounded by so many intelligent video-capturing devices. These devices capture data about how we live and what we do. Recent research has also shown that they can be combined with computer vision algorithms to realize non-contact measurement of human physiology. For instance, as a new advance in the remote photoplethysmography (rPPG) technique [1], heart rate (HR), respiration rate (RR) and heart rate variability (HRV) can be extracted from facial videos in ambient light based on the subtle color changes of the skin caused by blood circulation [2], [3], [4]. Besides skin color changes, blood ejection into the vessels can also cause repetitive motions of the human body, measured as ballistocardiograph (BCG). It is also possible to estimate BCG from head motions in video [5] to get HR readings remotely. The remote measurement of these physiological responses has also been leveraged to build systems for remotely capturing cognitive load and stress during computer tasks [6], [7].

Meanwhile, these new technologies raise privacy issues. Now whenever you are in front of a camera, people can not only recognize your identity based on your appearance, but also monitor some aspects of your health and affective state. This information might be misused for manipulation in marketing, negotiation, and other situations. Moreover, there has been no way to eliminate this channel of information from facial videos without affecting the visual appearance.

Though there are many algorithms measuring physiological signals from facial videos, it is not trivial to develop a new method for eliminating these signals. First, most of the measurement algorithms are irreversible, which means the attenuation of their output signals can not be properly propagated to their input videos. For example, many methods take the average of multiple pixels in a region of interest to form raw physiological traces, but not every pixel within the region includes the needed physiological information. Therefore, to uniformly remove the estimated traces from the whole region would result in artifact on the non-relevant pixels. Second, the target of most of the measurement methods is to synthesize a single reading such as a heart rate from a video epoch instead of recovering the temporal shape of the physiological signal faithfully. Without the ability to estimate a signal faithfully, it will be also hard to eliminate it cleanly.

Thus we propose a novel method for eliminating physiological information from facial videos based on motion component magnification [8], an algorithm good at estimating and amplifying non-sinusoidal motions in videos. With any facial video as input, our method outputs a video visually the same but from which physiological signals such as rPPG can no longer be measured accurately. To our knowledge, it is for the first time this kind of functionality has been achieved.

In this paper, we review previous works on motion estimation and elimination for video, introduce the framework of our methodology, assess it on facial videos collected from an 18-participant user study, and discuss its parameters, limitations, and strengths.

II. PRIOR WORK

Motion estimation and elimination have long been used for video denoising. For example, videos captured from hand-held devices often have camera jitters. Several methods [9], [10] have been proposed to stabilize the camera motion by modeling the jitters using image-level transforms between consecutive frames. However, these methods deal with motions in a global manner, so they are unable to adaptively process motions only on one object such as physiology-related motions only on the human body.

To realize motion elimination at the object level, several previous works have introduced different conditions for selecting the motion of interest. Bai et al. have presented a semi-automated technique called video de-animation [11], in which motions are selectively processed based on their scales. The technique requires its users to draw a small set of strokes indicating the regions of the objects that should be immobilized. Then their algorithm warps the video to remove large-scale motion of these regions while leaving finer-scale, relative motions intact. This scale-based condition is not fit for selecting physiological information, as physiology-related motions can appear at different scales, and removing

This work was supported by the SDSC Global Foundation and MIT Media Lab Consortium.

the largest-scale motions of the human body could result in unnatural rigidity. Another way to select local motions for elimination is by looking at the frequency domain. For example, Rubinstein et al. [12] have successfully eliminated short-term motion jitters in time-lapse videos by assuming high-frequency changes are noise and low-frequency changes are signals. Based upon the assumption, they are able to recover a smoother version of the input video by reshuffling its pixels spatiotemporally. Nevertheless, the assumption is also not applicable to the estimation of physiological information, because physiological signals are not necessarily high in frequency relative to other motions.

Recently, a more sophisticated method called Eulerian motion magnification (EMM) [13] has been proposed, which adaptively extracts subtle motions using a linear band-pass filter. The method takes a standard video sequence as input, applies spatial decomposition using a Gaussian or Laplacian pyramid, and performs temporal linear filtering of the frames based on manually selected frequency parameters. The resulting signal is then amplified to reveal hidden temporal variations in videos that are difficult or impossible to see with the naked eye. Since the invention of Eulerian motion magnification, a series of improvements have also been made. To support larger magnification factors at high spatial frequencies, Wadhwa et al. [14] proposed a new Eulerian approach to motion processing, based on complex-valued steerable pyramids. Phase variations of the coefficients of a complex-valued steerable pyramid over time correspond to motion, and increasing the phase variations by a multiplicative factor can amplify subtle motions. Thus, the method computes the local phase variations, then a person picks a desired temporal frequency band to amplify using a linear filter, and finally reconstructs the modified video. In general, the linear EMM technique is better at magnifying small color changes, while the phase-based pipeline is better at magnifying subtle motions [15]. On top of the two core methods, a new image pyramid representation, the Riesz pyramid, was developed to accelerate the phase-based pipeline[16], and a layer-based approach was proposed to amplify small motions combined with larger ones using matting [17]. The EMM methods can naturally lend themselves to attenuation of motions in videos by setting their amplification factors to negative values. Based on this idea, success has been achieved in motion attenuation for turbulence removal and color amplification [14].

All of the EMM methods apply a temporal filter to the intensity or phase variations to select motions of interest within a specific frequency band, which is appropriate for sinusoidal motions. However, most physiological signals such as heart pulse and respiration are not pure sinusoidal motions, nor are they simple combinations of sinusoids from a narrow band of frequencies. If the filtering band of the EMM algorithms is selected to cover all the frequencies they contain, then a great deal of noise within the wide band will also remain to bury the motion of interest [8]. To better estimate non-sinusoidal motions in videos, a new Eulerian approach called "motion component magnification (MCM)" has been proposed [8]. Instead of applying linear filtering, the new approach uses principal component analysis (PCA) [18] to detect and amplify subtle motions from multiple pixels of a video, so that the estimated motions are more faithful to the original ones. The new method also yields fewer artifacts even for sinusoidal motions in some cases. In this paper we show a new capability: Replacing component magnification with component elimination, the MCM algorithm is promising for eliminating physiological information, such as pulse rate, from facial videos.

III. METHODS

The proposed methodology for eliminating physiological information from facial videos is depicted in Fig. 1, and introduced as follows.

A. Laplacian Pyramid

A subtle motion in facial videos related to physiology can appear at different scales. For example, for rPPG measurement, with different perspectives and different camera distances, skin regions on the face will have various sizes. Also, different spatial frequency bands might exhibit different signal-to-noise ratios, which should be treated separately. In order to estimate motions of interest across multiple scales, we compute an L-level Laplacian pyramid [19] for each channel of each video frame in the RGB color space. This process generates a set of band-pass filtered images at L levels, in which lower levels exhibit more fine-grain detail. If all the pixels within an image are processed together, a subtle motion taking up only a few pixels would be likely to be buried in noise. Thus every image is further split into 8 x 8 pixel blocks. We choose this size to be the same used in JPEG image compression. A block smaller than 8 x 8 does not include enough information about the motion we want to extract, while a larger size might incorporate too many objects and introduce noise. To make sure even the highest level of the pyramid covers an area larger than 8 x 8, L is selected as

$$L = \lfloor \log_2 \frac{\min(H,W)}{8} \rfloor + 1 \tag{1}$$

in which H and W are the height and width of the original video frame. When the image size is not exactly divisible by the block size, the image boundaries are padded with mirror reflections of themselves.

B. Principal Component Analysis

The intensity change of every pixel within a block forms a time series of length N, which can be considered as a channel of sensor recording. Therefore, all the 8 x 8 time series compose a data matrix $X_{N\times P}$, $P = 64$. To detect the greatest motion among the 64 channels, we apply PCA to X. As the first step of PCA, the column-wise means are subtracted from X and stored:

$$Y(n,p) = X(n,p) - \frac{1}{N}\sum_{n=1}^{N} X(n,p) \tag{2}$$

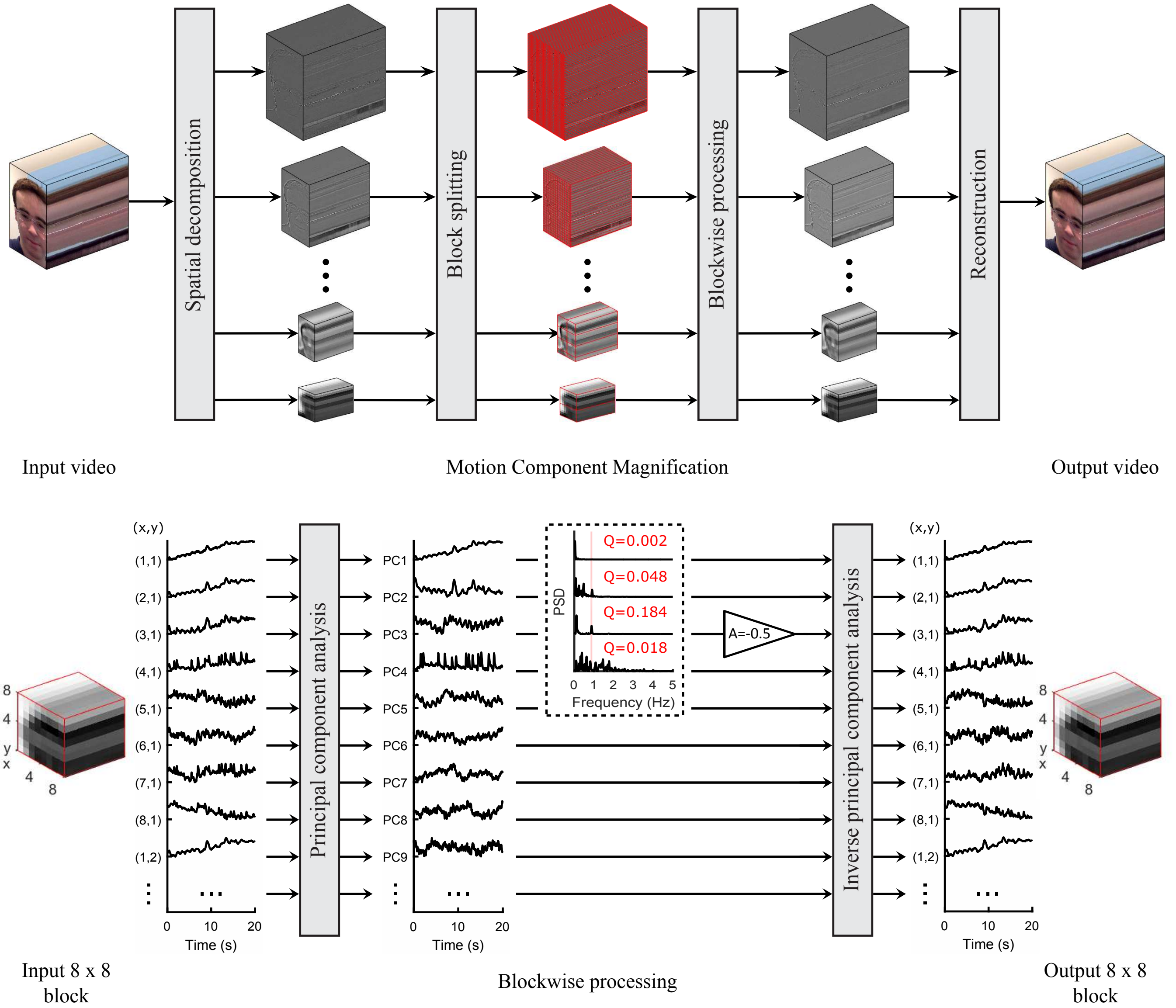

Fig. 1. Flow charts of our methodology. The red patch in the power spectral density (PSD) figures indicates the frequency band of interest: 0.8258 Hz - 0.9133 Hz. All the temporal signals have been normalized to the same amplitude for better visualization.

Then the principal component decomposition of Y is calculated as

$$T = YW \tag{3}$$

where W is a $P \times P$ matrix whose columns are the eigenvectors of $Y^{\top}Y$, and T consists of N row vectors $t_{(n)}$, $n = 1, \cdots, N$, each of which corresponds to a principal component score in the transformed co-ordinates in descending order of variance. For example, the first row of T with the highest variance is denoted as $t_{(1)}$.

C. Component Selection and Elimination

When there are multiple motions in the video and we are only interested in one of them, frequency-domain features can be exploited to separate them. For the four principal component scores with the highest variances $t_{(n)}$, $n = 1, \cdots, 4$, we compute their power spectral density (PSD) estimates $p_n(f)$, $n = 1, \cdots, 4$, $f \in (0, f_{Nyq})$ (f_{Nyq} is the Nyquist frequency) using the Lomb-Scargle method [20], [21], which is good at finding weak periodic signals in otherwise random unevenly sampled data. The method is chosen over other PSD estimation methods, mainly because many video-capturing devices have non-constant sampling rates.

Assume the fundamental frequency of our motion of interest lies in the band $[f_L, f_H]$. To evaluate its distinction in a PSD estimate, the normalized band power Q is calculated as the power within the interest frequency band divided by the full spectrum power:

$$Q(p_n) = \frac{\sum_{f=f_L}^{f_H} p_n(f)}{\sum_{f=0}^{f_{Nyq}} p_n(f)}, \quad n = 1, \cdots, 4 \tag{4}$$

The component among $t_{(n)}$ whose PSD estimate has the largest normalized band power Q_{max} is selected to represent the motion of interest, and denoted as $t_{(i)},\ i \in \{1, \cdots, 4\}$. To eliminate the motion, we can simply multiply $t_{(i)}$ by zero. However, this can not guarantee a thorough elimination of the motion, because other components $t_{(n)}$ might also contain minute power of the band. Thus a softer elimination factor A was applied to $t_{(i)}$ as formulated below

$$t_{(i)} = A \cdot t_{(i)},\ A \leq 0 \tag{5}$$

in which A can be set slightly negative to compensate for the residual power of the motion of interest in other components.

D. Reconstruction

After the targeted information is eliminated in (5), we rewrite the manipulated T as T'. In order to reconstruct the video, first, an inverse transform of (3) is applied to T':

$$Y' = T'W^{\top} \tag{6}$$

Then the subtracted column-wise means are added back:

$$X'(n,p) = Y'(n,p) + \frac{1}{N}\sum_{n=1}^{N} X(n,p) \tag{7}$$

After that, all $X'(n,p)$ are concatenated to recover the L levels of the Laplacian pyramid. Finally, the video is reconstructed from the pyramid levels in the usual way.

IV. USER STUDY

We run a user study to test the methodology on the elimination of rPPG. 18 participants (16 males, 2 females) between the ages of 23-50 years were recruited for this study, which was pre-approved by the Massachusetts Institute of Technology Committee On the Use of Humans as Experimental Subjects (COUHES). The participants have varying skin colors, with some of them wearing thick facial hair and/or glasses. Informed consent was obtained from all the participants prior to each study session.

In each experiment, a participant was seated still at a desk under ambient light for 30 s. An Intel RealSense Camera VF0800 was set up on the desk at a distance of approximately 0.5-1.0 m from the participant to capture facial videos. All videos were recorded in color (24-bit RGB with 3 channels 8 bits/channel) at a floating frame rate around 24 frames per second (fps) with pixel resolution of 1920 x 1080 and saved in MP4 format. Ground truth physiology was also measured with an FDA-cleared sensor (FlexComp Infiniti by Thought Technologies Ltd.) that recorded Blood Volume Pulse (BVP) from a finger probe at a constant sampling frequency of 256 Hz. The sensor data were synchronized with the video frames via timestamps. All the experiments were conducted in real work environments, in which many other people passed behind the participants very often.

To assess the extent to which the measurement accuracy of rPPG can be diminished by our methodology, an rPPG-based heart rate measurement algorithm needs to be implemented. Since major head motion is rare in our dataset, a simple but robust algorithm based on blind source separation (BSS) [3] was applied. The algorithm spatially averages all pixels in the facial region to form three color traces (RGB), decomposes the traces into three independent source components using Independent Component Analysis (ICA) [22], and always selects the second component (for the sake of simplicity) as the desired pulse signal. While implementing the algorithm on our dataset, we observed that sometimes the second component was not necessarily the best component corresponding to the pulse wave, so we slightly tweaked the algorithm by introducing a component selection step. In the step, a metric called pulse significance [23], which is the product of the normalized pulse band power and the power spectrum kurtosis, was calculated on each of the three independent source components, and the component with the highest pulse significance was chosen as the pulse representation for heart rate measurement.

As shown in Fig. 3, the videos we collected are in high resolution, but the facial regions of the participants only take a small area. Though our methodology is capable of processing the original full-resolution videos, it would consume a lot of computation power and time. Also, there are usually multiple human faces in our videos, as a result of which selective elimination of physiological information from a single face might be useful in some cases. Therefore, a face detection algorithm using OpenCV's Haar-like cascades [24] was applied to the first frame of each input video first. Among all the detected faces, we selected the largest one (always corresponding to the participants' faces), and defined the center 60% width and full height of its bounding box as the region of interest (ROI). Our rPPG elimination algorithm would be only employed within the ROIs of each video.

In the implementation of our methodology, there are two parameters to be selected: the frequency band of the physiological signal of interest $[f_L, f_H]$ and the elimination factor A. For the frequency band $[f_L, f_H]$, we tested a fine estimate, a 0.05-Hz-wide band around each participant's heart rate detected by the BSS-based method, and a coarse estimate, [0.75 Hz, 2.5 Hz] for every participant (corresponding to 45 and 150 beats per minute). As for the elimination factor A, we observed visible artifact in most of the processed videos when $A < -1.5$, so we tested a span of values between 0 and -1.5. Using these different combinations of parameters, all the 18 videos were processed by our algorithm. The tweaked BSS-based rPPG measurement method was then applied to both the input and the output videos to generate a heart rate reading as well as a power spectrum of the estimated pulse signal for each video.

The heart rate readings and the power spectra were compared with the ground truth using three metrics: mean absolute error (MAE) of the heart rate, spectrum power at the heart rate normalized by the full spectrum power (normalized pulse power 1, shortened as NPP1), and spectrum power at the heart rate normalized by the band power within [0.75 Hz, 2.5 Hz] (normalized pulse power 2, shortened as NPP2). The metric NPP1 reflects the extent to which the pulse power is diminished among all frequencies, and the metric

NPP2 indicates the significance of the ground truth heart rate within the common heart rate band, which is more related to measurement accuracy. Among 18 input videos, 3 gave an MAE higher than 2 beats per minute (BPM). This suggests that their measurement accuracies for heart pulse were already low without any processing, so we will not report their results in the following sections. To further show the superiority of our methods over previous motion magnification methods in eliminating physiological information, we also implemented the linear Eulerian motion magnification algorithm [13], previously shown to be good at magnifying small color changes [15], using the fine estimate of frequency bands and equivalent elimination factors.

V. Results

Table I lists the three metrics MAE, NPP1, and NPP2 corresponding to each input or output video of each subject using the fine estimate of the heart rate frequency band. The same metrics for videos processed by our algorithm using the coarse frequency band and the EMM algorithm are attached in Appendix I. As shown in Table I, the average MAE of heart rate measurement is 0.254 BPM for the original videos, which is very accurate. After the videos were processed by our algorithm, the average errors increased nicely to 5.499 BPM ($A = 0$), 17.321 BPM ($A = -0.5$), 17.787 BPM ($A = -1.0$), and 17.977 BPM ($A = -1.5$). A heart rate reading with an error as high as 17 BPM is not meaningful; thus, our algorithm effectively eliminated heart beat information from the facial videos. For 13 of the videos, the MAE showed an increase after being processed by our algorithm using the fine frequency band. Meanwhile, the NPP1 and NPP2 metrics of all the participants decreased. To assess whether the changes are significant, a paired t-test was conducted between each output metric and its corresponding input metric. The resulting t-values and p-values are shown in Table I, Table III and Table IV. For the fine frequency band results in Table I, the MAEs were increased significantly at a 0.05 significance level when $A \leq -0.5$, and the normalized pulse powers were decreased significantly for all choices of A.

To more intuitively compare our method versus EMM, and the different selections of the elimination factor and the frequency band, we summarize Table I, Table III and Table IV in Fig. 2. There is a clear trend in all the three subplots that a larger A leads to bigger MAEs and lower NPP1s and NPP2s using our algorithm, while the EMM method can barely change any of the metrics. Furthermore, the changes of all the three metrics are always more significant when using the fine frequency band compared with using the coarse one.

Our algorithm is able to eliminate physiological information while preserving the visual appearance of videos. Fig. 3 shows the same frames from an input video and its processed output videos using different elimination factors. There is nearly no visual difference between the input and output videos. Only when the elimination factor A is as small as -1.5 in Fig. 3 (e) does a slight artifact become observable on the participants' eyes after zooming in. The power spectra of the pulse signals estimated from each video are also drawn in Fig. 3, in which attenuation of the pulse power along with the increase of A is apparent. To quantitatively assess how much distortion our algorithm caused to the videos, we computed two metrics: the root mean square error (RMSE) and the structural similarity (SSIM) index [25] between each input and output video within the ROI. The SSIM index is a method for measuring the perceived similarity between two images or videos. Its value is always between -1 and 1, and value 1 is only reachable in the case of two identical sets of data. Table II lists the RMSEs and SSIMs of our output videos using the fine frequency band. The same metrics for the coarse frequency band videos are attached in Appendix I. As shown in Table II, for every participant, RMSE increased and SSIM decreased as the elimination factor A was set to be smaller. Even when $A = -1.5$, all the videos' SSIMs remain much higher than 0.9, a common threshold for indicating high visual similarity.

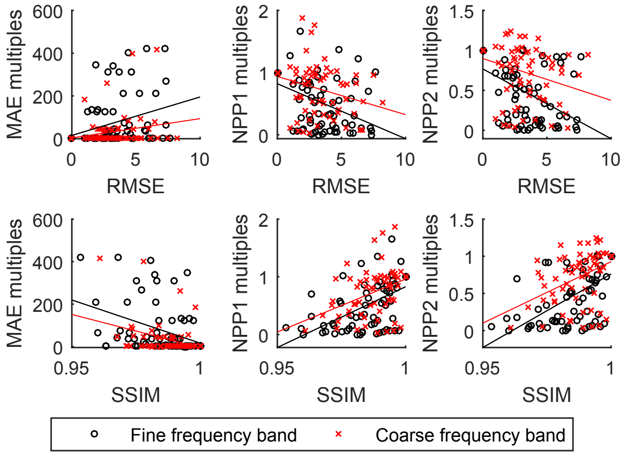

Fig. 4. Scatter plots show the relationships between the physiological elimination indicators (MAE = mean absolute error of heart rate measurement, NPP1 = normalized pulse power 1, NPP2 = normalized pulse power 2) and the video appearance distortion indicators (RMSE = root mean square error, SSIM = structural similarity index). MAE, NPP1 and NPP2 were converted into multiples by dividing them by their corresponding input values. Least-squares lines were also fit and superimposed in each subplot.

To sum up, by decreasing the elimination factor A within the range [-1.5, 0], both physiological attenuation and video distortion intensify. To show the relationship between them directly, we drew scatter plots of MAE/NPP1/NPP2 versus RMSE/SSIM for all participants and all elimination factor choices with least-squares lines superimposed in Fig. 4. The slopes of the least-squares lines suggest that with the same level of sacrifice in RMSE and SSIM, our method using the fine frequency band always achieves a greater improvement in physiological elimination compared with using the coarse frequency band.

VI. Discussion and Future Work

One of the key parameters of our algorithm is the frequency band of the physiological signal of interest. We compared two different estimates of it: a fine version and a coarse version.

TABLE I

Mean absolute error of heart rate measurement (MAE, in the unit of beats per minute), normalized pulse power 1 (NPP1), and normalized pulse power 2 (NPP2) for each video. The parameter A represents the elimination factor. All the output videos were processed by our algorithm using the fine estimate of the heart rate frequency band. The last two rows show the t-test results between each output metric and its corresponding input metric.

#	Input video			Output video ($A = 0$)			Output video ($A = -0.5$)			Output video ($A = -1.0$)			Output video ($A = -1.5$)		
	MAE	NPP1	NPP2	MAE	NPP1	NPP2	MAE	NPP1	NPP2	MAE	NPP1	NPP2	MAE	NPP1	NPP2
1	0.084	0.166	0.013	0.084	0.154	0.013	0.084	0.151	0.012	0.084	0.142	0.011	22.566	0.010	0.002
2	0.379	0.121	0.029	0.113	0.087	0.019	2.577	0.021	0.004	7.769	0.012	0.004	15.878	0.003	0.001
3	0.416	0.034	0.002	14.595	0.018	0.001	14.595	0.018	0.001	14.595	0.020	0.002	0.917	0.024	0.001
4	0.104	0.031	0.005	0.104	0.032	0.005	0.104	0.042	0.004	0.104	0.038	0.005	0.104	0.031	0.005
5	0.681	0.039	0.005	25.176	0.002	0.000	45.317	0.002	0.000	45.317	0.001	0.000	45.317	0.000	0.000
6	0.083	0.108	0.019	0.083	0.087	0.009	0.083	0.081	0.013	0.434	0.021	0.004	28.343	0.000	0.000
9	0.084	0.119	0.023	0.401	0.105	0.020	0.401	0.095	0.012	0.401	0.076	0.006	0.886	0.040	0.003
10	0.600	0.036	0.006	0.100	0.060	0.005	22.570	0.000	0.000	22.570	0.000	0.000	22.071	0.006	0.001
12	0.088	0.055	0.010	0.088	0.030	0.003	35.466	0.017	0.001	36.920	0.009	0.001	36.920	0.006	0.000
13	0.491	0.060	0.009	0.491	0.034	0.006	21.793	0.018	0.004	21.793	0.009	0.002	35.801	0.003	0.001
14	0.110	0.097	0.006	0.110	0.058	0.004	23.216	0.031	0.003	23.216	0.018	0.002	23.216	0.011	0.001
15	0.094	0.040	0.006	32.516	0.011	0.001	29.021	0.001	0.000	29.021	0.002	0.000	29.021	0.002	0.000
16	0.065	0.064	0.008	8.095	0.004	0.001	8.095	0.005	0.001	8.095	0.016	0.004	8.095	0.022	0.004
17	0.422	0.096	0.015	0.422	0.122	0.010	56.388	0.006	0.001	56.388	0.009	0.001	0.422	0.081	0.011
18	0.103	0.098	0.005	0.103	0.082	0.003	0.103	0.071	0.002	0.103	0.054	0.002	0.103	0.030	0.001
Mean	0.254	0.078	0.011	5.499	0.059	0.007	17.321	0.037	0.004	17.787	0.028	0.003	17.977	0.018	0.002
t(15)				-1.960	3.184	4.550	-3.653	5.435	4.225	-3.789	6.338	4.396	-4.454	5.301	4.256
p				0.070	0.007	0.001	0.003	0.000	0.001	0.002	0.000	0.001	0.001	0.000	0.001

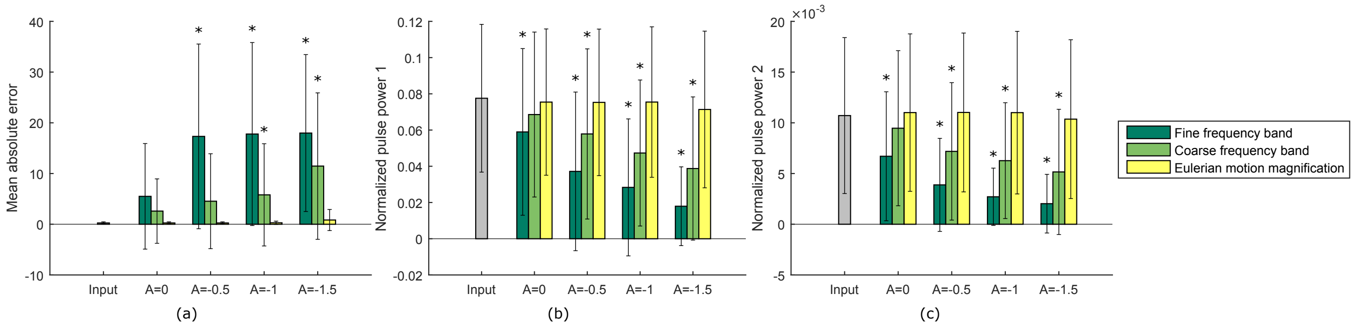

Fig. 2. (a) Mean absolute error of heart rate measurement, (b) normalized pulse power 1, and (c) normalized pulse power 2, averaged among all input videos and all output videos using different methods and different elimination factors. An asterisk indicates a significant difference from the input at a 0.05 significance level.

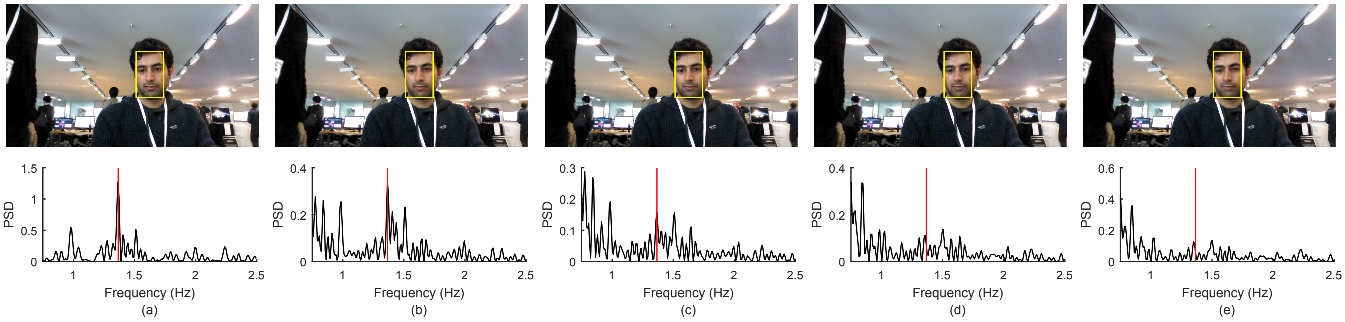

Fig. 3. Exemplary frames and power spectra of the estimated pulse signal from an input video (a) and output videos processed by different elimination factors: (b) $A = 0$, (c) $A = -0.5$, (d) $A = -1.0$, and (e) $A = -1.5$. The yellow bounding box indicates the region of interest we applied our algorithm to, and the red vertical lines mark the ground truth heart rate in each power spectrum.

TABLE II

Root mean square error (RMSE) and structural similarity (SSIM) between each output video and its corresponding input video within the regions of interest. The parameter A represents the elimination factor. All the output videos were processed by our algorithm using the fine estimate of the heart rate frequency band.

#	Input video		Output video ($A = 0$)		Output video ($A = -0.5$)		Output video ($A = -1.0$)		Output video ($A = -1.5$)	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
1	0.000	1.000	3.105	0.995	4.654	0.990	6.082	0.984	7.342	0.977
2	0.000	1.000	2.303	0.994	3.433	0.987	4.540	0.978	5.636	0.968
3	0.000	1.000	2.908	0.992	4.331	0.984	5.737	0.974	7.125	0.963
4	0.000	1.000	3.174	0.994	4.772	0.989	6.329	0.983	7.678	0.976
5	0.000	1.000	3.012	0.992	4.461	0.983	5.902	0.972	7.339	0.960
6	0.000	1.000	1.270	0.997	1.955	0.993	2.592	0.988	3.199	0.982
9	0.000	1.000	1.713	0.996	2.566	0.991	3.382	0.985	4.173	0.979
10	0.000	1.000	1.780	0.995	2.659	0.989	3.516	0.981	4.358	0.972
12	0.000	1.000	2.964	0.991	4.402	0.981	5.824	0.968	7.237	0.954
13	0.000	1.000	1.685	0.994	2.515	0.988	3.321	0.979	4.114	0.969
14	0.000	1.000	2.684	0.991	3.987	0.983	5.280	0.972	6.565	0.959
15	0.000	1.000	1.875	0.995	2.817	0.990	3.727	0.983	4.607	0.975
16	0.000	1.000	1.264	0.997	1.951	0.994	2.590	0.989	3.203	0.984
17	0.000	1.000	0.980	0.998	1.534	0.996	2.026	0.993	2.488	0.989
18	0.000	1.000	2.459	0.992	3.650	0.985	4.820	0.975	5.982	0.963
Mean	0.000	1.000	2.212	0.994	3.312	0.988	4.378	0.980	5.403	0.971

The fine version requires a non-contact physiological measurement method to be applied to the input video first, while the coarse version uses a general frequency band covering all common values of a physiological signal, which is much cheaper in terms of computation load. Our results showed that the fine version always outperforms the coarse version in physiological elimination using the same elimination factor. Thus the fine version is always recommended when computation power and speed are not under consideration.

In Table I and Table III, the decrease of the elimination factor A and the decrease of the heart rate measurement accuracy are generally correlated, but in theory their relationship should not be monotonic. As shown, A tends to work well when set slightly negative. However, if A was set to be too negative for a video, it would over-compensate the residual power and start to magnify the estimated physiological signal. This phenomenon can be observed in Table I on Participants 3 and 17, whose MAEs start to decrease when $A \leq -1.0$. Theoretically, there is an optimal A for each video, which can eliminate physiological information most cleanly. We have not yet found a way to estimate this optimal value, which is future work.

There are several other limitations in this study. First, the implementation of our algorithm is not yet fast enough to run in real time. As a result, it is not yet fit for online processing of live streams. We believe by optimizing the program it will be possible to achieve online analysis. Second, all the videos collected in our user study are 30s in length, which does not allow us to probe the feasibility of our method in the long run. In theory, all the vital signs of the human body change gradually under resting states. Consequentially, if the fine estimate of the frequency band is used in our algorithm, it needs to be updated once in a while. Finally, it would be also interesting to assess our algorithm on the elimination of facial BCG signals as well as physiological information measured from rPPG other than HR, such as RR and HRV. These limitations can be addressed in future work.

VII. Conclusions

We have shown that it is possible to eliminate physiological information from facial videos without affecting their visual appearance. We demonstrated a new algorithm for this elimination, which is based on motion component magnification. The new algorithm successfully increased the average error of HR measurement on 15 facial videos by 70 times. The success is attributed to the use of the new motion component representation, which was shown to work better than a previous EMM approach for representing the motion as sinusoids. Compared with using a coarse frequency band between 45 BPM and 150 BPM to extract the signal of interest, the method of using a fine frequency band estimated by a BSS-based rPPG measurement method was shown to achieve better performance at the cost of some extra computation. Finally, we showed that while tuning the elimination factor, there is a trade-off between physiological attenuation and video appearance distortion. Overall, a range of values of the main parameters was explored and shown to give significant attenuation of the physiological heart pulse signal, while maintaining suitably low root-mean-square error and significantly high visual similarity to the original.

References

[1] Y. Sun and N. Thakor, “Photoplethysmography revisited: from contact to noncontact, from point to imaging,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 463–477, mar 2016.

[2] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, “Remote plethysmographic imaging using ambient light,” *Optics express*, vol. 16, no. 26, pp. 21 434–21 445, 2008.

[3] M.-Z. Poh, D. J. McDuff, and R. W. Picard, “Non-contact, automated cardiac pulse measurements using video imaging and blind source separation,” *Optics express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[4] M. Z. Poh, D. J. McDuff, and R. W. Picard, “Advancements in non-contact, multiparameter physiological measurements using a webcam,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.

[5] G. Balakrishnan, F. Durand, and J. Guttag, “Detecting pulse from head motions in video,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3430–3437, 2013.

[6] D. McDuff, S. Gontarek, and R. Picard, “Remote measurement of cognitive stress via heart rate variability,” *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2957–2960, 2014.

[7] D. J. McDuff, J. Hernandez, S. Gontarek, and R. W. Picard, “COGCAM: Contact-free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera,” *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4000–4004, 2016.

[8] W. Chen and R. W. Picard, “Motion Component Magnification,” *Submitted to IEEE Transactions on Visualization and Computer Graphics*, 2017.

[9] F. Liu, M. Gleicher, H. Jin, and A. Agarwala, “Content-preserving warps for 3D video stabilization,” *ACM Transactions on Graphics*, vol. 28, no. 3, p. 1, 2009.

[10] Y. Matsushita, E. Ofek, W. Ge, X. Tang, and H.-Y. Shum, “Full-frame video stabilization with motion inpainting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 7, pp. 1150–1163, 2006.

[11] J. Bai, A. Agarwala, M. Agrawala, and R. Ramamoorthi, “Selectively De-Animating Video,” *ACM Trans. Graph. Article*, vol. 31, no. 10, pp. 1–10, 2012.

[12] M. Rubinstein, C. Liu, P. Sand, F. Durand, and W. T. Freeman, “Motion denoising with application to time-lapse photography,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011, pp. 313–320.

[13] 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 (Proc. SIGGRAPH)*, vol. 31, no. 4, pp. 1–8, 2012.

[14] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, “Phase-based video motion processing,” *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 32, no. 4, p. 1, 2013.

[15] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, “Eulerian video magnification for revealing subtle changes in the world,” 2012. [Online]. Available: http://people.csail.mit.edu/mrub/evm/

[16] N. Wadhwa, M. Rubinstein, F. Durand, and W. Freeman, “Riesz pyramid for fast phase-based video magnification,” in *2014 IEEE International Conference on Computational Photography (ICCP)*, 2014.

[17] M. A. Elgharib, M. Hefeeda, and W. T. Freeman, “Video Magnification in Presence of Large Motions,” *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4119–4127, 2015.

[18] I. Jolliffe, “Principal component analysis,” 2002.

[19] Burt P. and E. Adelson, “The laplacian pyramid as a compact image code,” *IEEE Trans. Commun.*, vol. 31(4), no. 4, pp. 532–540, 1983.

[20] N. R. Lomb, “Least-squares frequency analysis of unequally spaced data,” *Astrophysics and space science*, vol. 39, no. 2, pp. 447–462, 1976.

[21] J. D. Scargle, “Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data,” *The Astrophysical Journal*, vol. 263, pp. 835–853, 1982.

[22] P. Comon, “Independent component analysis, a new concept?” *Signal Process*, vol. 36, no. 94, pp. 287–314, 1994.

[23] W. Chen, J. Hernandez, and R. W. Picard, “Non-contact physiological measurements from near-infrared video of the neck,” *Submitted to Biomedical Optic Express*, 2016.

[24] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001, p. 511.

[25] Z. Wang, a. C. Bovik, H. R. Sheikh, and E. P. Simmoncelli, “Image quality assessment: form error visibility to structural similarity,” *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

APPENDIX I

TABLE III

Mean absolute error of heart rate measurement (MAE, in the unit of beats per minute), normalized pulse power 1 (NPP1), and normalized pulse power 2 (NPP2) for each video. The parameter A represents the elimination factor. All the output videos were processed by our algorithm using the coarse estimate of the heart rate frequency band. The last two rows show the t-test results between each output metric and its corresponding input metric.

#	Input video			Output video (A= 0)			Output video (A= −0.5)			Output video (A= −1.0)			Output video (A= −1.5)		
	MAE	NPP1	NPP2	MAE	NPP1	NPP2	MAE	NPP1	NPP2	MAE	NPP1	NPP2	MAE	NPP1	NPP2
1	0.084	0.166	0.013	0.084	0.167	0.013	0.084	0.167	0.013	0.084	0.134	0.011	6.425	0.014	0.003
2	0.379	0.121	0.029	0.379	0.128	0.026	0.379	0.116	0.021	0.113	0.099	0.016	0.379	0.116	0.021
3	0.416	0.034	0.002	0.416	0.030	0.002	0.416	0.027	0.002	0.917	0.018	0.002	0.917	0.018	0.002
4	0.104	0.031	0.005	0.104	0.034	0.005	0.104	0.031	0.005	0.367	0.026	0.004	0.104	0.030	0.004
5	0.681	0.039	0.005	1.226	0.033	0.005	31.164	0.020	0.002	31.164	0.016	0.002	31.164	0.013	0.002
6	0.083	0.108	0.019	0.083	0.091	0.016	0.434	0.064	0.010	3.018	0.026	0.005	3.535	0.015	0.002
9	0.084	0.119	0.023	0.084	0.117	0.024	0.084	0.118	0.023	0.401	0.118	0.021	0.401	0.109	0.019
10	0.600	0.036	0.006	0.600	0.038	0.005	0.600	0.048	0.005	0.100	0.034	0.004	0.600	0.044	0.003
12	0.088	0.055	0.010	22.866	0.029	0.003	8.636	0.016	0.001	8.151	0.006	0.001	36.435	0.004	0.000
13	0.491	0.060	0.009	0.491	0.035	0.007	21.793	0.024	0.005	21.793	0.021	0.006	25.614	0.019	0.006
14	0.110	0.097	0.006	0.110	0.079	0.007	0.110	0.053	0.007	0.110	0.051	0.007	0.110	0.056	0.005
15	0.094	0.040	0.006	0.094	0.075	0.007	0.094	0.070	0.007	0.094	0.043	0.005	37.437	0.014	0.002
16	0.065	0.064	0.008	11.842	0.007	0.001	3.683	0.004	0.001	0.471	0.059	0.007	0.065	0.106	0.008
17	0.422	0.096	0.015	0.422	0.105	0.015	0.422	0.018	0.003	20.066	0.005	0.001	20.066	0.003	0.001
18	0.103	0.098	0.005	0.103	0.061	0.006	0.103	0.092	0.004	0.103	0.054	0.005	8.755	0.022	0.002
Mean	0.254	0.078	0.011	2.594	0.069	0.009	4.540	0.058	0.007	5.797	0.047	0.006	11.467	0.039	0.005
t(15)				-1.413	1.590	1.735	-1.798	2.591	3.190	-2.160	4.079	3.210	-3.016	3.088	4.173
p				0.179	0.134	0.105	0.094	0.021	0.007	0.049	0.001	0.006	0.009	0.008	0.001

TABLE IV

Mean absolute error of heart rate measurement (MAE, in the unit of beats per minute), normalized pulse power 1 (NPP1), and normalized pulse power 2 (NPP2) for each video. The parameter A represents the elimination factor. All the output videos were processed by the Eulerian motion magnification algorithm. The last two rows show the t-test results between each output metric and its corresponding input metric.

#	Input video			Output video (A= 0)			Output video (A= −0.5)			Output video (A= −1.0)			Output video (A= −1.5)		
	MAE	NPP1	NPP2	MAE	NPP1	NPP2	MAE	NPP1	NPP2	MAE	NPP1	NPP2	MAE	NPP1	NPP2
1	0.084	0.166	0.013	0.084	0.168	0.013	0.084	0.168	0.013	0.084	0.168	0.013	0.084	0.168	0.012
2	0.379	0.121	0.029	0.379	0.124	0.030	0.379	0.124	0.030	0.379	0.123	0.030	0.379	0.122	0.030
3	0.416	0.034	0.002	0.416	0.034	0.002	0.416	0.034	0.002	0.416	0.034	0.002	0.416	0.035	0.002
4	0.104	0.031	0.005	0.104	0.036	0.006	0.104	0.036	0.006	0.104	0.036	0.006	0.104	0.031	0.005
5	0.681	0.039	0.005	0.681	0.038	0.005	0.681	0.040	0.006	1.226	0.039	0.005	1.226	0.034	0.005
6	0.083	0.108	0.019	0.083	0.110	0.020	0.083	0.112	0.021	0.083	0.113	0.021	0.083	0.112	0.020
9	0.084	0.119	0.023	0.084	0.119	0.023	0.084	0.118	0.023	0.084	0.124	0.023	0.084	0.106	0.019
10	0.600	0.036	0.006	0.600	0.036	0.006	0.600	0.035	0.006	0.600	0.036	0.006	0.600	0.039	0.007
12	0.088	0.055	0.010	0.088	0.054	0.010	0.088	0.056	0.011	0.088	0.061	0.012	0.088	0.064	0.012
13	0.491	0.060	0.009	0.491	0.060	0.010	0.491	0.057	0.009	0.491	0.054	0.009	0.491	0.054	0.009
14	0.110	0.097	0.006	0.110	0.098	0.006	0.110	0.098	0.006	0.110	0.098	0.006	0.110	0.097	0.006
15	0.094	0.040	0.006	0.094	0.041	0.006	0.094	0.041	0.006	0.094	0.041	0.006	0.094	0.042	0.006
16	0.065	0.064	0.008	0.065	0.065	0.008	0.065	0.065	0.008	0.065	0.064	0.008	0.065	0.062	0.008
17	0.422	0.096	0.015	0.422	0.085	0.014	0.422	0.085	0.014	0.422	0.091	0.014	0.422	0.093	0.014
18	0.103	0.098	0.005	0.103	0.066	0.006	0.103	0.060	0.005	0.103	0.050	0.004	8.274	0.013	0.001
Mean	0.254	0.078	0.011	0.254	0.075	0.011	0.254	0.075	0.011	0.290	0.075	0.011	0.835	0.071	0.010
t(15)				NaN	0.926	-1.948	NaN	0.845	-2.034	-1.000	0.617	-1.501	-1.069	1.071	0.806
p				NaN	0.370	0.072	NaN	0.412	0.061	0.334	0.547	0.156	0.303	0.302	0.434

TABLE V

Root mean square error (RMSE) and structural similarity (SSIM) between each output video and its corresponding input video within the regions of interest. The parameter A represents the elimination factor. All the output videos were processed by our algorithm using the coarse estimate of the heart rate frequency band.

#	Input video		Output video (A= 0)		Output video (A= −0.5)		Output video (A= −1.0)		Output video (A= −1.5)	
	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM	RMSE	SSIM
1	0.000	1.000	2.772	0.996	4.172	0.992	5.435	0.987	6.500	0.981
2	0.000	1.000	2.499	0.995	3.743	0.989	4.960	0.981	6.122	0.973
3	0.000	1.000	3.396	0.994	5.084	0.987	6.736	0.980	8.242	0.971
4	0.000	1.000	3.087	0.995	4.645	0.990	6.167	0.985	7.485	0.978
5	0.000	1.000	2.451	0.993	3.645	0.987	4.827	0.978	5.999	0.968
6	0.000	1.000	1.229	0.997	1.895	0.994	2.514	0.989	3.101	0.984
9	0.000	1.000	1.467	0.997	2.229	0.993	2.945	0.988	3.626	0.982
10	0.000	1.000	1.473	0.996	2.238	0.991	2.975	0.985	3.696	0.977
12	0.000	1.000	2.717	0.992	4.048	0.984	5.361	0.973	6.662	0.961
13	0.000	1.000	1.460	0.995	2.211	0.990	2.934	0.983	3.639	0.974
14	0.000	1.000	2.204	0.994	3.290	0.988	4.357	0.980	5.396	0.972
15	0.000	1.000	1.913	0.996	2.895	0.991	3.829	0.985	4.690	0.978
16	0.000	1.000	1.014	0.998	1.608	0.995	2.138	0.992	2.640	0.988
17	0.000	1.000	0.974	0.998	1.528	0.996	2.019	0.993	2.481	0.989
18	0.000	1.000	1.787	0.995	2.683	0.989	3.546	0.982	4.385	0.974
Mean	0.000	1.000	2.029	0.995	3.061	0.990	4.050	0.984	4.978	0.977