



(51) International Patent Classification:

A61B 5/024 (2006.01) G06T 5/50 (2006.01)

A61B 5/00 (2006.01) G06T 7/11 (2017.01)

G16H 30/40 (2018.01) G06V 40/16 (2022.01)

(21) International Application Number:

PCT/IB2024/055900

(22) International Filing Date:

17 June 2024 (17.06.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/521,433 16 June 2023 (16.06.2023) US

(71) Applicants: SECURIPOINT LLC [US/US]; 1900 Reston Metro Plaza, Suite 800, Reston, Virginia 20190 (US). UNIVERSITY OF NOTRE DAME DU LAC [US/US]; 1400 East Angela Blvd., South Bend, Indiana 46617 (US).

(72) Inventors: VANCE, Nathan; c/o University of Notre Dame du Lac, 1400 East Angela Blvd., South Bend, Indiana 46617 (US). SPETH, Jeremy; c/o University of Notre Dame du Lac, 1400 East Angela Blvd., South Bend, Indiana 46617 (US). SPORRER, Benjamin; c/o University of Notre Dame du Lac, 1400 East Angela Blvd., South Bend, Indiana 46617 (US). FLYNN, Patrick; c/o University of Notre Dame du Lac, 1400 East Angela Blvd., South Bend, Indiana 46617 (US). CARPENTER, Nathan; c/o Securiport LLC, 1900 Reston Metro Plaza, Suite 800, Reston, Virginia 20190 (US). SEGURA, Juan Manuel; c/o Securiport

LLC, 1900 Reston Metro Plaza, Suite 800, Reston, Virginia 20190 (US).

(74) Agent: WAHBA, Andrew; MILES & STOCKBRIDGE P.C., 1751 Pinnacle Drive, Suite 1500, Tysons Corner, Virginia 22102-3833 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: PROMOTING GENERALIZATION IN CROSS-DATASET REMOTE PHOTOPLETHYSMOGRAPHY

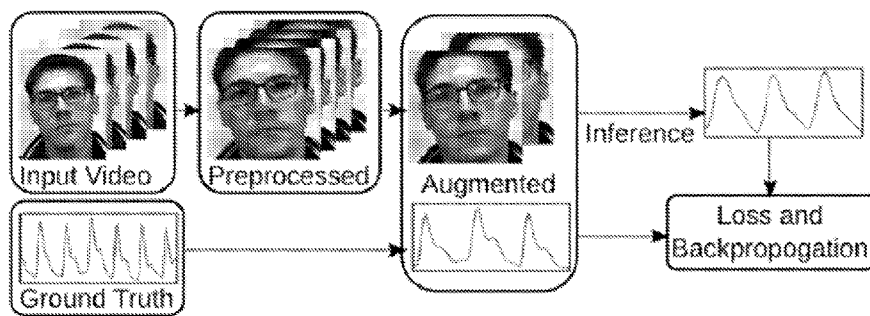


FIG. 3

(57) Abstract: Systems, devices, methods, and non-transitory computer-readable instructions for determining a physiological signal from a video stream, comprising capturing the video stream of a subject, the video stream including a sequence of frames, processing each frame of the video stream to identify a facial portion of the subject in each frame, determining a periodic physiological signal of the subject from the video stream using a plurality of datasets in which one or more augmentations were applied to the plurality of datasets.

Published:

- *with international search report (Art. 21(3))*
- *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

PROMOTING GENERALIZATION IN CROSS-DATASET
REMOTE PHOTOPLETHYSMOGRAPHY

PRIORITY INFORMATION

[0001] This application claims the benefit of the U.S. Provisional Patent Application No. 63/521,433 filed on June 16, 2023, which is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

[0002] The embodiments of the present invention generally relate to use of biometrics, and more particularly, to determination of physiological signals from video.

DISCUSSION OF THE RELATED ART

[0003] In general, biometrics may be used to track vital signs that provide indicators about a subject's physical state that may be used in a variety of ways. As an example, for border security or health monitoring, vital signs may be used to screen for health risks (e.g., temperature) or detect deception (e.g., change in pulse or pupil diameter). While sensing temperature is a well-developed technology, collecting other useful and accurate vital signs such as pulse rate (*i.e.*, heart rate or heart beats per minute) or pulse waveform has required physical devices to be attached to the subject. The desire to perform biometric measurement without physical contact has produced some video-based techniques.

[0004] Performing reliable pulse rate or pulse waveform estimation from a camera sensor is more difficult than contact plethysmography for several reasons. The change in reflected light from the skin's surface, because of light absorption of blood, is very minor compared to those caused by changes in illumination. Even in settings with ambient lighting, the subject's movements drastically change the reflected light and overpower the pulse signal.

[0005] Camera-based vitals estimation is a growing field enabling non-contact health monitoring in a variety of settings. While the number of successful approaches has increased, the size of benchmark video datasets with simultaneous vitals recordings has remained relatively stagnant. It is well-known across the machine learning community that increasing the quantity and diversity of training data is an effective strategy for improving performance.

[0006] Collecting remote physiological data is challenging for several reasons. First, recording many hours of high-quality videos results in an unwieldy volume of data. Second, recording a diverse population of subjects with associated medical data is difficult due to privacy concerns. Furthermore, synchronizing contact measurements with video recordings in diverse settings is highly dependent on the researcher's hardware infrastructure and lab setting. Even contact measurements used for ground truth contain noise, making data curation difficult. These difficulties contributing to data scarcity stifle model scaling and robustness.

[0007] Accordingly, the inventors have developed systems, devices, methods, and non-transitory computer-readable instructions that enable generalization in cross-dataset remote photoplethysmography.

SUMMARY OF THE INVENTION

[0008] Accordingly, the present invention is directed to generalization in cross-dataset remote photoplethysmography that substantially obviates one or more problems due to limitations and disadvantages of the related art.

[0009] Additional features and advantages of the invention will be set forth in the description which follows, and in part will be apparent from the description, or may be learned by practice of the invention. The objectives and other advantages of the invention will be realized and attained by the structure particularly pointed out in the written description and claims hereof as well as the appended drawings.

[0010] To achieve these and other advantages and in accordance with the purpose of the present invention, as embodied and broadly described, the generalization in cross-dataset remote photoplethysmography includes systems, devices, methods, and non-transitory computer-readable instructions for determining a physiological signal from a video stream, comprising capturing the video stream of a subject, the video stream including a sequence of frames, processing each frame of the video stream to identify a facial portion of the subject in each frame, determining a periodic physiological signal of the subject from the video stream using a plurality of datasets in which one or more augmentations were applied to the plurality of datasets.

[0011] In the various embodiments, the plurality of datasets includes ground truth data and previously captured video data. In the various embodiments, the ground truth data is captured using pulse oximeter. In the various embodiments, the one or more augmentations is applied to the ground truth data and the previously captured video data. In the various embodiments, the one or more augmentation include at least one of horizontal flip, illumination, and Gaussian noise. In the various embodiments, the periodic physiological signal is heart rate. In the various embodiments, the facial portion is cropped to 64 x 64 pixels. In the various embodiments, the video stream includes one or more of a visible-light video stream, a near-infrared video stream, a longwave-infrared video stream, a thermal video stream, and an audio stream of the subject. In the various embodiments, each frame of the media stream is cropped to encapsulate a region of interest that includes one or more of a face, cheek, forehead, or an eye. In the various embodiments, combining at least two of a visible-light video stream, a near-infrared video stream, and a thermal video stream are combined into a fused video stream.

[0012] It is to be understood that both the foregoing general description and the following detailed description are examples and explanatory and are intended to provide further explanation of the invention as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] The accompanying drawings, which are included to provide a further understanding of the invention and are incorporated in and constitute a part of this specification, illustrate embodiments of the invention and together with the description serve to explain the principles of the invention. In the drawings:

[0014] FIG. 1 illustrates a system for pulse waveform estimation.

[0015] FIG. 2 illustrates an overview of the augmentations targeting the temporal domain.

[0016] FIG. 3 illustrates an overview of the temporal augmentation framework.

[0017] FIGs. 4(a)-(i) illustrate training and validation losses when training RpNet on the three datasets and applying augmentation settings.

[0018] FIG. 5 illustrates that speed augmentations reduce learned bias as reflected by a reduced mean error in cross dataset analysis between datasets with differing heart rate bands.

[0019] FIG. 6 illustrates that speed augmentations improve the accuracy of the model, reflected by an improved mean absolute error.

DETAILED DESCRIPTION OF THE EXEMPLARY EMBODIMENTS

[0020] Reference will now be made in detail to the embodiments of the present invention, examples of which are illustrated in the accompanying drawings.

[0021] Embodiments of user interfaces and associated methods for using a device are described. It should be understood, however, that the user interfaces and associated methods can be applied to numerous devices types, such as a portable communication device such as a tablet or mobile phone. The portable communication device can support a variety of applications, such as wired or wireless communications. The various applications that can be stored (in a non-transitory memory) and executed (by a processor) on the device can use at least one common physical user-interface device, such as a touchscreen. One or more functions of the touchscreen as well as corresponding information displayed on the device can be adjusted and/or varied from one application to another and/or within a respective application. In this way, a common physical architecture of the device can support a variety of applications with user interfaces that are intuitive and transparent.

[0022] The embodiments of the present invention provide systems, devices, methods, and non-transitory computer-readable instructions to measure one or more biometrics, including heart-rate and pulse waveform without physical contact with the subject. The embodiments can be used in combination with other biometrics that can include respiration, eye gaze, blinking, pupillometry, face temperature, oxygen level, blood pressure, audio, voice tone and/or frequency, micro-expressions, etc. In the various embodiments, the systems, devices, methods, and instructions collect, process, and analyze video taken in one or more modalities (e.g., visible light, near infrared, longwave infrared, thermal) to provide non-contrastive unsupervised learning of physiological signals from a video signal or video data (e.g., MP4).

[0023] As describe herein, the pulse or pulse waveform for the subject's heartbeat may be used as a biometric input to establish features of the physical state of the subject and how they change over a period of observation (e.g., during questioning or other activity).

Remote photoplethysmography (rPPG) is the monitoring of blood volume pulse from a camera at a distance. Using rPPG, blood volume pulse from video at a distance from the skin's surface may be detected. The disclosure of U.S. Application No. 17/591,929, entitled "VIDEO BASED DETECTION OF PULSE WAVEFORM", filed 3 February 2022, is hereby incorporated by reference, in its entirety.

[0024] FIG. 1 illustrates a system 100 for pulse waveform estimation. System 100 includes optical sensor system 1, video I/O system 6, and video processing system 101.

[0025] Optical sensor system 1 includes one or more camera sensors, each respective camera sensor configured to capture a video stream including a sequence of frames. For example, optical sensor system 1 may include a visible-light camera 2, a near-infrared camera 3, a thermal camera 4, or any combination thereof. In the event that multiple camera sensors are utilized (e.g., single modality or multiple modality), the resulting multiple video streams may be synchronized according to synchronization device 5. Alternatively, or additionally, one or more video analysis techniques may be utilized to synchronize the video streams. Although a visible-light camera 2, a near-infrared camera 3, a thermal camera 4 are enumerated, other media devices can be used, such as a speech recorder.

[0026] Video I/O system 6 receives the captured one or more video streams. For example, video I/O system 6 is configured to receive raw visible-light video stream 7, near-infrared video stream 8, and thermal video stream 9 from optical sensor system 1. Here, the received video streams may be stored according to known digital format(s). In the event that multiple video streams are received (e.g., single modality or multiple modality), fusion processor 10 is configured to combine the received video streams. For example, fusion processor 10 may combine visible-light video stream 7, near-infrared video stream 8, and/or thermal video stream 9 into a fused video stream 11. Here, the respective streams may be synchronized according to the output (e.g., a clock signal) from synchronization device 5.

[0027] At video processing system 101, region of interest detector 12 detects (i.e., spatially locate) one or more spatial regions of interest (ROI) within each video frame. The

ROI may be a face, another body part (e.g., a hand, an arm, a foot, a neck, etc.) or any combination of body parts. Example ROIs include a face, cheek, forehead, or an eye. Initially, region of interest detector 12 determines one or more coarse spatial ROIs within each video frame. Region of interest detector 12 is robust to strong facial occlusions from face masks and other head garments. Subsequently, frame preprocessor 13 crops the frame to encapsulate the one or more ROI. In some embodiments, the cropping includes each frame being downsized by bi-cubic interpolation to reduce the number of image pixels to be processed. Alternatively, or additionally, the cropped frame may be further resized to a smaller image.

[0028] Sequence preparation system 14 aggregates batches of ordered sequences or subsequences of frames from frame processor 13 to be processed. Next, 3-Dimensional Convolutional Neural Network (3DCNN) 15 receives the sequence or subsequence of frames from the sequence preparation system 14. 3DCNN 15 processes the sequence or subsequence of frames, by a 3-dimensional convolutional neural network, to determine the spatial and temporal dimensions of each frame of the sequence or subsequence of frames and to produce a pulse waveform point for each frame of the sequence of frames. 3DCNN 15 applies a series of 3-dimensional convolution, averaging, pooling, and nonlinearities to produce a 1-dimensional signal approximating the pulse waveform 16 for the input sequence or subsequences.

[0029] In some configurations, pulse aggregation system 17 combines any number of pulse waveforms 16 from the sequences or subsequences of frames into an aggregated pulse waveform 18 to represent the entire video stream. Diagnostic extractor 19 is configured to compute the heart rate and the heart rate variability from the aggregated pulse waveform 18. To identify heart rate variability, the calculated heart rate of various subsequences may be compared. Display unit 20 receives real-time or near real-time updates from diagnostic extractor 19 and displays aggregated pulse waveform 18, heart rate, and heart rate variability to an operator. Storage Unit 21 is configured to store aggregated pulse waveform 18, heart rate, and heart rate variability associated with the subject.

[0030] Additionally, or alternatively, the sequence of frames may be partitioned into a partially overlapping subsequences within the sequence preparation system 14, wherein a first subsequence of frames overlaps with a second subsequence of frames. The overlap in frames between subsequences prevents edge effects. Here, pulse aggregation system 17 may apply a Hann function to each subsequence, and the overlapping subsequences added to generate aggregated pulse waveform 18 with the same number of samples as frames in the original video stream. In some configurations, each subsequence is individually passed to the 3DCNN 15, which performs a series of operations to produce a pulse waveform for each subsequence 16. Each pulse waveform output from the 3DCNN 15 is a time series with a real value for each video frame. Since each subsequence is processed by the 3DCNN 15 individually, they are subsequently recombined.

[0031] In some embodiments, one or more filters may be applied to the region of interest. For example, one or more wavelengths of LED light may be filtered out. The LED may be shone across the entire region of interest and surrounding surfaces or portions thereof. Additionally, or alternatively, temporal signals in non-skin regions may be further processed. For example, analyzing the eyebrows or the eye's sclera may identify changes strongly correlated with motion, but not necessarily correlated with the photoplethysmogram. If the same periodic signal predicted as the pulse is found on non-skin surfaces, it may indicate a non-real subject or attempted security breach.

[0032] Although illustrated as a single system, the functionality of system 100 may be implemented as a distributed system. While system 100 determines heart rate, other distributed configurations track changes to the subject's eye gaze, eye blink rate, pupil diameter, speech, face temperature, and micro-expressions, for example. Further, the functionality disclosed herein may be implemented on separate servers or devices that may be coupled together over a network, such as a security kiosk coupled to a backend server. Further, one or more components of system 100 may not be included. For example, system 100 may be a smartphone or tablet device that includes a processor, memory, and a display, but may not include one or more of the other components shown in FIG. 1. The embodiments may be implemented using a variety of processing and memory storage devices. For example, a CPU and/or GPU may be used in the processing system to

decrease the runtime and calculate the pulse in near real-time. System 100 may be part of a larger system. Therefore, system 100 may include one or more additional functional modules.

[0033] Subtle quasi-periodic physiological signals such as blood volume pulse and respiration can be extracted from RGB video, enabling remote health monitoring and other applications. Advancements in remote pulse estimation – or remote photoplethysmography (rPPG) – are currently driven by supervised deep learning solutions. However, current approaches are trained and evaluated on limited benchmark datasets recorded with ground truth from contact-PPG sensors.

[0034] Remote Photoplethysmography (rPPG), or the remote monitoring of a subject's heart rate using a camera, has seen a shift from handcrafted techniques to deep learning models. While current solutions offer substantial performance gains, these models tend to learn a bias to pulse wave features inherent to the training dataset. Accordingly, the inventors develop augmentations to mitigate this learned bias by expanding both the range and variability of heart rates that the model uses while training, resulting in improved model convergence when training and cross-dataset generalization at test time. Through a 3-way cross dataset analysis, a reduction in mean absolute error from over 13 beats per minute to below 3 beats per minute is demonstrated.

[0035] Measuring a subject's heart rate is an important component of physiological monitoring. While methods such as photoplethysmography (PPG) exist for contact heart rate monitoring, a push has been made for non-contact remote photo-plethysmography (rPPG). rPPG is cheaper, requiring a commodity camera rather than a specialized pulse oximeter, and it is contact-free, allowing for applications in new contexts.

[0036] Initial techniques for rPPG employed algorithms involving a multi-stage pipeline. While these techniques may be highly accurate, their performance is adversely affected by dynamics common in videos such as motion and illumination changes. More recently, deep learning methods have been applied to rPPG, many of them outperforming handcrafted techniques. While deep learning techniques have benefits, they suffer drawbacks as well in terms of generalization. It has been shown that the learned priors in deep learning rPPG models are strong enough to predict a periodic signal in situations where a periodic signal is

not present in the input – a relevant attack scenario. It is demonstrated that a deep learning rPPG model may be biased toward predicting heart rate features such as the frequency bands and rates of change that appear in its training data, and therefore struggle to generalize to new situations.

[0037] Training of rPPG models incorporates various types of data augmentations in the spatial domain. In the embodiments, the data in the temporal domain is augmented — injecting synthetic data representing a wide spectrum of heart rates, thus enabling models to better respond to unknown heart rates. The embodiments are evaluated in a cross-dataset setup comprising significant differences between heart rates in the training and test subsets. For example, FIG. 2 illustrates an overview of the augmentations targeting the temporal domain according to an example embodiment.

[0038] There has been broad interest in rPPG, with applications including detection of heart arrhythmias such as atrial fibrillation, deepfake detection, and affective computing. However, early techniques were not robust to motion. The emergence of practical deep learning methods has enabled new methods for rPPG estimation, such as DeepPhys, a convolutional neural network (CNN) model which effectively predicts pulse waveform derivatives based on adjacent video frames. Additionally, a 3DCNN based approach has been developed for predicting the pulse waveform from video data.

[0039] Cross-dataset generalization is a common concern with deep learning techniques, specifically in that deep learning rPPG techniques tend to perform suboptimally when working outside of the heart rate range of the training set. In the various embodiments, speed and modulation augmentations for 3DCNN based models are provided, showing that this consideration mitigates much of the cross dataset performance loss experienced by current models.

[0040] A variety of computer systems, devices, methods, and non-transitory computer-readable instructions may be utilized for rPPG analysis. The example embodiments utilize the RPNet architecture, which is a 3DCNN-based approach. For example, the network architecture may be composed of 3D convolutions with max and global pooling layers for dimension reduction. The network may consume 64×64 pixel video over a 136-frame window, outputting an rPPG signal of 136 samples.

[0041] The preprocessing pipeline consists of the following steps. First, facial landmarks are obtained at each frame in the dataset using the MediaPipe Face Mesh tool. Second, the face is cropped at the extreme points of the landmarks, padded by 30% on the top and 5% on the sides and bottom, and the shortest dimension is extended to make the crop square. Third, the cropped portion is scaled to 64×64 pixels using cubic interpolation.

[0042] When a cross-dataset analysis is performed, the frame rate of all videos is reduced to the lowest common denominator, e.g., 30 FPS. This only affects the DDPM dataset, which is recorded at 90 FPS. The conversion is executed before the cropping step by taking the average pixel value over sets of three frames. The “averaging” technique is used rather than skipping frames in order to better emulate a slower camera shutter speed.

[0043] RpNet outputs rPPG waves in 136-frame chunks with a stride of 68 frames. These parameters can be used so that the model is small enough to fit on the used GPUs. To reduce edge effects, a Hann window may be applied to the overlapping segments and add them together, thus producing a single waveform.

[0044] To enable the determination of heart rates, a Short-Time Fourier Transform (STFT) of the output waveform is used with a window size of 10 seconds and a stride of 1 frame, thus enabling the use of the embodiments in application scenarios tolerant of a 10-second latency. The waveform may be padded with zeros such that the bin width in the frequency domain is 0.001 Hz (0.06 beats per minute (BPM)) to reduce quantization effects. Additionally, the highest peak is selected in the range of 0.66 and 3 Hz (i.e., 40 and 180 BPM) as the inferred heart rate.

[0045] The temporal aspect of the training data is augmented, affecting alternatively the heart rate or speed, and the change in heart rate or modulation. FIG. 3 illustrates an overview of the temporal augmentation framework according to an example embodiment. FIG. 3 shows how it fits into the training protocol.

[0046] To apply the speed augmentation, first randomly select a target heart rate between 40 and 180 BPM (i.e., the desired range of heart rates for which the model will be sensitive). This is set to be the same range as the peak selection used in the postprocessing step so that the model will be trained to predict the same heart rates that the rest of the system is designed to handle.

[0047] Second, the ground truth heart rate (i.e., obtained using the same STFT technique outlined above) is averaged over the 136 frame clip, as the source heart rate. Next, the length of data centered on the source clip is calculated to be:

$[136 \times \text{HR}_{\text{target}}/\text{HR}_{\text{source}}]$.

[0048] Third, the data in the source interval is interpolated such that it becomes 136 frames long. This process is applied to both the video clip and the ground truth waveform.

[0049] To apply the modulation augmentation, randomly select a modulation factor f based on the ground truth heart rate such that when the clip speeds up or slows down by a factor of f , the change in heart rate is no more than 7 BPM per second. This parameter was selected based on the maximum observed change in heart rate in the DDPM dataset. Further, constrain the modulation such that the video clip is modulated linearly by the selected factor over its duration, i.e. for normalized heart rates s and e at the start and end of the clip respectively, the normalized heart rate at each frame x in the n frame clip (set to 136 as in Section 3.1) is:

$$n\text{HR}(x) = s + \frac{x(e - s)}{n} \quad (1)$$

where $s = 2/(1+f)$ and $e = sf$. Then, integrate $n\text{HR}$ to generate a function yielding the positions $P(x)$ along the original clip at which to interpolate:

$$P(x) = xs + \frac{x^2(e - s)}{2n} + c \quad (2)$$

where $c = 0$ due to indexing starting at 0. Finally, linearly interpolate the n frames from the original video clip at every position $P(x)$ for all x in the range $[0..n]$, thus yielding the modulated video clip. Optionally, horizontal flip, illumination, and Gaussian noise spatial augmentations may be used.

[0050] A variety of metrics may be used for evaluation. These metrics utilize either the pulse waveform (provided as ground truth or inferred by RPNNet) or the heart rate. If the lengths of the ground truth and predicted waves differ (as is the case if the ground truth wave is not a multiple of 68 frames, i.e. the stride used for RPNNet), then data points from the end of the ground truth wave are removed such that they have the same length.

[0051] Each evaluation metric is calculated over each video in the dataset independently, the results of which are averaged. Example evaluation metrics will now be described.

[0052] The Mean Error (ME) captures the bias of the method in BPM, and is defined as follows:

$$ME = \frac{1}{N} \sum_{i=1}^N (HR'_i - HR_i) \quad (3)$$

where HR and HR' are the ground truth and predicted heart rates, respectively, where each contained index is the heart rate obtained from the STFT window, and N is the number of STFT windows present. Many rPPG methods omit an analysis based on ME since it is often close to zero due to positive and negative errors canceling each other out. However, ME is valuable for gauging the bias of a model in a cross-dataset analysis by explaining how the model is failing, i.e. whether the predictions are simply noisy or if they are shifted relative to the ground truth.

[0053] The Mean Absolute Error (MAE) captures an aspect of the precision of the method in BPM, and is defined as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^N |HR'_i - HR_i| \quad (4)$$

[0054] The Root Mean Squared Error (RMSE) is similar to MAE, but penalizes outlier heart rates more strongly:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (HR'_i - HR_i)^2} \quad (5)$$

[0055] The waveform correlation, *rwave*, is the Pearson's r correlation coefficient between the ground truth and predicted waves. When performing an inter-dataset analysis, the *rwave* value is further maximized by varying the correlation lag between ground truth and predicted waves by up to 1 second (30 data points) in order to compensate for differing synchronization techniques between datasets.

[0056] For cross dataset analysis, three rPPG datasets were used as an example, chosen to contain a wide range of heart rates: PURE, UBFC-rPPG, and DDPM. Key statistics for these three datasets are summarized in Table 1.

Dataset	Duration (s)	HR Avg	HR SD
PURE	68.307 $\pm$ 1.502	69.200 $\pm$ 6.026	1.638 $\pm$ 0.2682
UBFC	64.964 $\pm$ 1.516	100.801 $\pm$ 5.056	3.016 $\pm$ 0.525
DDPM	656.464 $\pm$ 22.310	96.982 $\pm$ 4.186	4.000 $\pm$ 0.286

Table 1. Average duration, heart rate (HR) in BPM calculated using the STFT, and average within-session standard deviation in HR within a 60 second window and a stride of 1 frame, for PURE, UBFC-rPPG, and DDPM. The 95% confidence intervals are calculated across sessions in the dataset.

[0057] The PURE dataset is useful for cross-dataset analysis for two key reasons. First, it has the lowest average heart rate of the three datasets, being about 30 BPM lower than the other two. Second, it has the lowest within-subject heart rate standard deviation.

[0058] The UBFC-rPPG dataset (shortened to UBFC) features subjects playing a time-sensitive mathematical game which caused a heightened physiological response. UBFC has the highest average heart rate of the three datasets and more heart rate variability than PURE, but less variability than DDPM.

[0059] The DDPM dataset is the largest of the compared datasets, with recorded sessions lasting nearly 11 minutes on average. It also features the most heart rate variability of the three, with a heart rate standard deviation of about 4 BPM. This is due to stress-inducing aspects (mock interrogation with forced deceptive answers) in the collection protocol of DDPM. Due to noise in the ground truth oximeter waveforms, 10 second segments in DDPM where the heart rate changes by more than 7 BPM per second may be masked out.

[0060] Training. For each of the three datasets, randomly partition the videos into five subject-disjoint sets, three of which are merged to generate splits for training, validation, and testing at 3/1/1 ratios, for example. Then, rotate the splits to generate five folds for cross-validation. For example, train for 40 epochs using the negative Pearson loss function

and the Adam optimizer configured with a 0.0001 learning rate. Models are selected based on minimum validation loss.

[0061] FIGs. 4(a)-(i) illustrate training and validation losses when training RpNet on the three datasets and applying three augmentation settings: none, speed, and speed+mod. Utilizing any sort of temporal augmentation causes the validation loss to converge with tighter confidence intervals. This is especially evident when training on the PURE dataset where the median validation loss confidence interval without temporal augmentations (FIG. 4a) drops from ± 0.174 to ± 0.081 and ± 0.078 with speed and speed+mod augmentations, respectively (FIGs. 4d and 4g). Furthermore, while it is apparent from FIG. 4c that training over DDPM without temporal augmentations may lead to overfitting, both temporal augmentation settings appear to avoid this problem (FIGs. 4f and 4i).

[0062] Across all combinations of augmentations and datasets, the validation loss converges to a lower value when temporal augmentations are used than when they are not. This is likely because the models are forced to generalize when the range and variability of heart rates they are exposed to is increased, limiting the effectiveness of simply memorizing a signal which looks like a heart rate and replaying it at a frequency common to the dataset.

[0063] The various embodiments trained and tested RpNet on each of the three datasets, both in a within dataset analysis (3 training-testing configurations with PURE-PURE, UBFC-UBFC, and DDPM-DDPM), and with a cross-dataset analysis (6 training-testing configurations with PURE-UBFC, PURE-DDPM, UBFC-PURE, UBFC-DDPM, DDPM-PURE, and DDPM-UBFC). Furthermore, three temporal augmentation settings were used, namely no temporal augmentation (none), speed augmentation (speed), and speed plus modulation augmentation (speed+mod). The results for the within-dataset analysis are shown in Table 2 and for the cross-dataset analysis are shown in Table 3.

Dataset	Augmentations	ME	MAE	RMSE	<i>P_{wave}</i>
PURE	none	-0.516 ± 1.814	1.176 ± 1.891	1.872 ± 3.067	0.694 ± 0.253
PURE	speed	-0.012 ± 0.461	0.694 ± 0.566	1.222 ± 1.456	0.753 ± 0.087
PURE	speed+mod	0.006 ± 0.389	0.639 ± 0.482	1.130 ± 1.347	0.752 ± 0.089
UBFC	none	0.922 ± 2.215	1.432 ± 2.301	2.238 ± 2.630	0.803 ± 0.024
UBFC	speed	0.016 ± 0.384	0.616 ± 0.201	1.346 ± 0.746	0.793 ± 0.020
UBFC	speed+mod	0.091 ± 0.139	0.502 ± 0.121	0.993 ± 0.335	0.798 ± 0.024
DDPM	none	-1.443 ± 5.725	4.167 ± 4.680	6.907 ± 6.504	0.569 ± 0.070
DDPM	speed	-0.773 ± 2.036	3.230 ± 2.267	5.897 ± 4.671	0.584 ± 0.052
DDPM	speed+mod	-1.048 ± 1.434	2.981 ± 1.738	5.485 ± 3.412	0.587 ± 0.057

Table 2. Results for the 9 within-dataset combinations of dataset and the temporal augmentations used. Heart rate metrics (ME, MAE, and RMSE) have units of BPM, and rwave is Pearson's r correlation over pulse waveforms.

[0064] While the temporal augmentations were intended to improve cross-dataset performance, a slight performance boost occurred in the within-dataset case. As shown in Table 2, all metrics except rwave on UBFC exhibited better performance when temporal augmentations were employed. However, in these cases the performance boost is slight, often falling within the 95% confidence intervals of the results without augmentation.

[0065] Turning to the cross-dataset case shown in Table 3, it was found that training on a dataset with higher heart rate variability and testing on a dataset with lower heart rate variability tends to produce better results than the reverse. This is especially evident in cross dataset cases involving DDPM, which has the highest heart rate variability as measured by heart rate standard deviation in Table 1.

Train	Test	Augmentations	ME	MAE	RMSE	Pearson
PURE	UBFC	none	-13.082 $\pm$ 12.972	13.690 $\pm$ 12.847	19.320 $\pm$ 13.359	0.532 $\pm$ 0.136
PURE	UBFC	speed	-3.340 $\pm$ 2.998	4.703 $\pm$ 3.083	9.219 $\pm$ 4.645	0.590 $\pm$ 0.102
PURE	UBFC	speed+mod	-1.491 $\pm$ 0.583	2.251 $\pm$ 0.671	5.191 $\pm$ 1.559	0.636 $\pm$ 0.053
PURE	DDPM	none	-27.633 $\pm$ 8.058	32.360 $\pm$ 3.934	38.397 $\pm$ 3.052	0.182 $\pm$ 0.015
PURE	DDPM	speed	-10.926 $\pm$ 11.184	24.343 $\pm$ 4.140	33.410 $\pm$ 3.694	0.221 $\pm$ 0.032
PURE	DDPM	speed+mod	6.436 $\pm$ 4.870	33.620 $\pm$ 2.018	42.494 $\pm$ 2.829	0.150 $\pm$ 0.015
UBFC	PURE	none	9.657 $\pm$ 3.971	11.532 $\pm$ 2.710	14.791 $\pm$ 2.751	0.619 $\pm$ 0.021
UBFC	PURE	speed	0.864 $\pm$ 1.074	2.196 $\pm$ 0.921	3.758 $\pm$ 1.289	0.671 $\pm$ 0.043
UBFC	PURE	speed+mod	0.938 $\pm$ 0.720	2.535 $\pm$ 0.920	4.246 $\pm$ 1.275	0.625 $\pm$ 0.025
UBFC	DDPM	none	-5.560 $\pm$ 4.479	14.947 $\pm$ 2.231	20.738 $\pm$ 2.366	0.264 $\pm$ 0.028
UBFC	DDPM	speed	-4.240 $\pm$ 6.961	18.574 $\pm$ 2.707	28.082 $\pm$ 3.056	0.251 $\pm$ 0.020
UBFC	DDPM	speed+mod	11.258 $\pm$ 4.904	32.914 $\pm$ 0.769	41.698 $\pm$ 0.834	0.174 $\pm$ 0.010
DDPM	PURE	none	26.092 $\pm$ 14.065	26.660 $\pm$ 13.435	30.915 $\pm$ 13.164	0.437 $\pm$ 0.099
DDPM	PURE	speed	1.256 $\pm$ 1.563	2.208 $\pm$ 1.824	3.905 $\pm$ 2.996	0.686 $\pm$ 0.061
DDPM	PURE	speed+mod	1.338 $\pm$ 1.477	2.509 $\pm$ 1.776	4.441 $\pm$ 2.991	0.673 $\pm$ 0.058
DDPM	UBFC	none	-0.358 $\pm$ 0.863	1.963 $\pm$ 1.135	3.745 $\pm$ 1.931	0.699 $\pm$ 0.050
DDPM	UBFC	speed	-0.431 $\pm$ 0.177	1.311 $\pm$ 0.282	3.140 $\pm$ 0.654	0.711 $\pm$ 0.028
DDPM	UBFC	speed+mod	-0.563 $\pm$ 0.383	1.160 $\pm$ 0.393	2.906 $\pm$ 1.112	0.734 $\pm$ 0.029

Table 3. Results for the 18 cross-dataset combinations of train dataset, test dataset, and temporal augmentations used. Heart rate metrics (ME, MAE, and RMSE) have units of BPM, while rwave is Pearson's r correlation over pulse waveforms.

[0066] As shown in the ME column of Table 3, it was observed that when training and testing between datasets of different heart rates without temporal augmentations, the bias as reflected by ME is strong, with UBFC-PURE yielding the ME closest to zero at over 9 BPM. Furthermore, these models are biased in the direction of the training dataset's mean heart rate, i.e. training on PURE which has relatively low heart rates results in a negative ME on UBFC and DDPM, while training on UBFC or DDPM results in a positive ME when testing on PURE. However, applying the speed augmentation causes ME to be much closer to zero than when no such augmentation is used. This is because the speed augmentation is intended to mitigate the heart rate bias inherent in the training dataset, thus causing it to generalize to any heart rates seen in the augmented training regime rather than simply those present in the dataset. With the mitigation of heart rate bias as reflected by improved ME scores, improvement in MAE and RMSE occurs in most cases. Furthermore, a boost in rwave was observed indicating that the models more faithfully reproduce the waveforms with low noise.

[0067] The modulation augmentation is intended to boost performance when training on a dataset with low heart rate variability such as PURE and testing on a dataset with high variability such as UBFC and DDPM. Modulation boosts performance for PURE-UBFC, though even with modulation PURE-DDPM fails to generalize. With the possible exception of DDPM-UBFC, the modulation augmentation does not positively impact cases when the training dataset already contains high heart rate variability, as is the case with UBFC and DDPM.

[0068] Poor results occurred in both cross dataset experiments where DDPM is the test dataset. Of those, the same trend was observed in PURE-DDPM as in other cases, i.e. that models trained with speed augmentations outperform those without, albeit in this case the performance is still quite poor. In UBFC-DDPM, models trained without speed augmentations achieve better results than with speed augmentations, which is a break from the trend observed in all other cases. Furthermore, whereas in other cases high MAE and RMSE errors are largely explained by bias as reflected in ME, this case has a relatively low ME relative to MAE and RMSE. In this case since the average heart rate between UBFC and DDPM is relatively close (differing by less than 4 BPM), overfitting to this band of heart rates is actually beneficial for the cross dataset analysis.

[0069] Furthermore, the “zero-effort” error rates achieved by a model which simply predicts the average heart rate for the dataset (97 BPM as in Table 1) has comparable error rates to UBFC-DDPM (MAE and RMSE are 17.804 and 22.113 respectively). These zero-effort results for the three datasets are reported in Table 4.

Dataset	MAE	RMSE
PURE	15.847	23.054
UBFC	14.085	17.256
DDPM	17.804	22.113

Table 4. Zero-effort errors obtained by predicting the average heart rate of the dataset for all subjects. In all cases ME is 0.

[0070] The cross dataset results are summarized in Table 5. The 95% confidence interval is calculated across 4 cross dataset combinations (omitting the cases when testing on DDPM as no models generalized) and 5 training folds. Combining both speed and

modulation losses yields optimal performance on all metrics. The box plots in FIG. 5 and FIG. 6 further demonstrate the reason why the temporal augmentations outperform the case without augmentations. FIG. 5 illustrates that speed augmentations reduce learned bias as reflected by a reduced ME in cross dataset analysis between datasets with differing heart rate bands according to an example embodiment. FIG. 6 illustrates that speed augmentations improve the accuracy of the model, reflected by an improved MAE. Accordingly, the bias of the model to predict heart rates similar to its training dataset has been significantly reduced, as is most clearly seen in the reduced absolute ME shown in FIG. 5, and the improved MAE shown in FIG. 6.

Augmentations	[ME]	MAE	RMSE	$r_{Pearson}$
none	12.349 $\pm$ 5.546	13.460 $\pm$ 5.335	17.192 $\pm$ 5.720	0.572 $\pm$ 0.056
speed	1.502 $\pm$ 0.803	2.604 $\pm$ 0.884	5.005 $\pm$ 1.536	0.664 $\pm$ 0.031
speed+mod	1.373 $\pm$ 0.570	2.501 $\pm$ 0.784	4.830 $\pm$ 1.174	0.677 $\pm$ 0.025

Table 5. Summaries of cross dataset performance under speed augmentation settings, omitting PURE-DDPM and UBFC-DDPM where no models succeed in generalizing. The absolute value of ME metrics before averaging is used.

[0071] The augmentations described herein are generally applicable to deep learning based rPPG as a whole, as these augmentation techniques may be implemented as a training framework for any model architecture that trains based on video inputs (or other data inputs) to produce waveform outputs.

[0072] The importance of temporal speedbased augmentations for the cross-dataset generalization of deep learning rPPG methods is demonstrated. In addition, a system for training deep learning rPPG models using two variants of this augmentation method is provided, i.e. speed augmentation affecting the heart rate, and modulation affecting the change in heart rate. These augmentations may be applied to any deep learning rPPG system which produces a pulse waveform from video inputs.

[0073] It will be apparent to those skilled in the art that various modifications and variations can be made in the generalization in cross-dataset remote photoplethysmography of the present invention without departing from the spirit or scope of the invention. Thus, it

is intended that the present invention cover the modifications and variations of this invention provided they come within the scope of the appended claims and their equivalents.

What is claimed is:

1. A computer-implemented method for determining a physiological signal from a video stream, the computer-implemented method comprising:

capturing the video stream of a subject, the video stream including a sequence of frames;

processing each frame of the video stream to identify a facial portion of the subject in each frame;

determining a periodic physiological signal of the subject from the video stream using a plurality of datasets in which one or more augmentations were applied to the plurality of datasets.

2. The computer-implemented method according to claim 1, wherein the plurality of datasets includes ground truth data and previously captured video data.

3. The computer-implemented method according to claim 2, wherein the ground truth data is captured using pulse oximeter.

4. The computer-implemented method according to claim 2, wherein the one or more augmentations is applied to the ground truth data and the previously captured video data.

5. The computer-implemented method according to claim 1, wherein the one or more augmentation include at least one of horizontal flip, illumination, and Gaussian noise.

6. The computer-implemented method according to claim 1, wherein the periodic physiological signal is heart rate.

7. The computer-implemented method according to claim 1, wherein the facial portion is cropped to 64 x 64 pixels.

8. The computer-implemented method according to claim 1, wherein the video stream includes one or more of a visible-light video stream, a near-infrared video stream, a

longwave-infrared video stream, a thermal video stream, and an audio stream of the subject.

9. The computer-implemented method according to claim 1, further comprising cropping each frame of the media stream to encapsulate a region of interest that includes one or more of a face, cheek, forehead, or an eye.

10. The computer-implemented method according to claim 1, further comprising instructions for:

combining at least two of a visible-light video stream, a near-infrared video stream, and a thermal video stream into a fused video stream.

11. A system for determining a physiological signal from a video stream, the system comprising:

a processor; and

a memory storing one or more programs for execution by the processor, the one or more programs including instructions for:

capturing the video stream of a subject, the video stream including a sequence of frames;

processing each frame of the video stream to identify a facial portion of the subject in each frame;

determining a periodic physiological signal of the subject from the video stream using a plurality of datasets in which one or more augmentations were applied to the plurality of datasets.

12. The system according to claim 11, wherein the plurality of datasets includes ground truth data and previously captured video data.

13. The system according to claim 12, wherein the ground truth data is captured using pulse oximeter.

14. The system according to claim 12, wherein the one or more augmentations is applied to the ground truth data and the previously captured video data.
15. The system according to claim 11, wherein the one or more augmentation include at least one of horizontal flip, illumination, and Gaussian noise.
16. The system according to claim 11, wherein the periodic physiological signal is heart rate.
17. The system according to claim 11, wherein the facial portion is cropped to 64 x 64 pixels.
18. The system according to claim 11, wherein the video stream includes one or more of a visible-light video stream, a near-infrared video stream, a longwave-infrared video stream, a thermal video stream, and an audio stream of the subject.
19. The system according to claim 11, further comprising cropping each frame of the media stream to encapsulate a region of interest that includes one or more of a face, cheek, forehead, or an eye.
20. The system according to claim 11, further comprising instructions for:
combining at least two of a visible-light video stream, a near-infrared video stream, and a thermal video stream into a fused video stream.

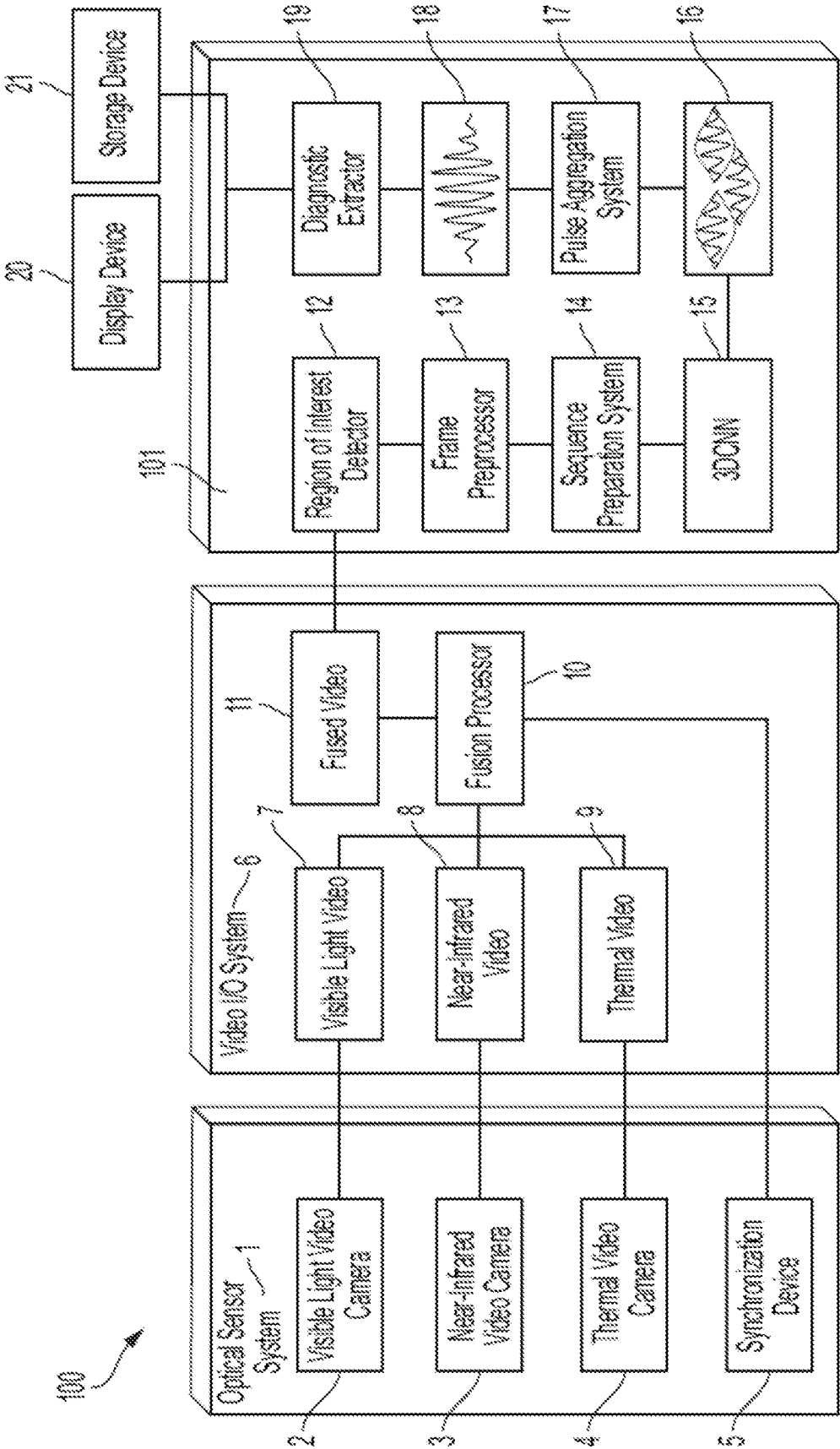


FIG. 1

2/6

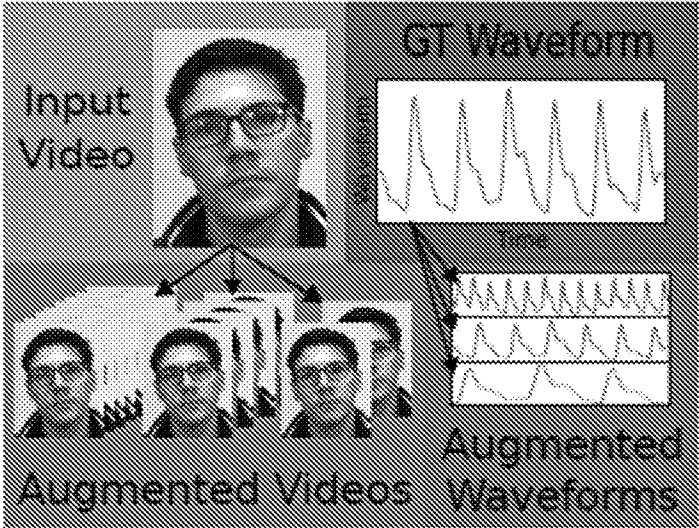


FIG. 2

3/6

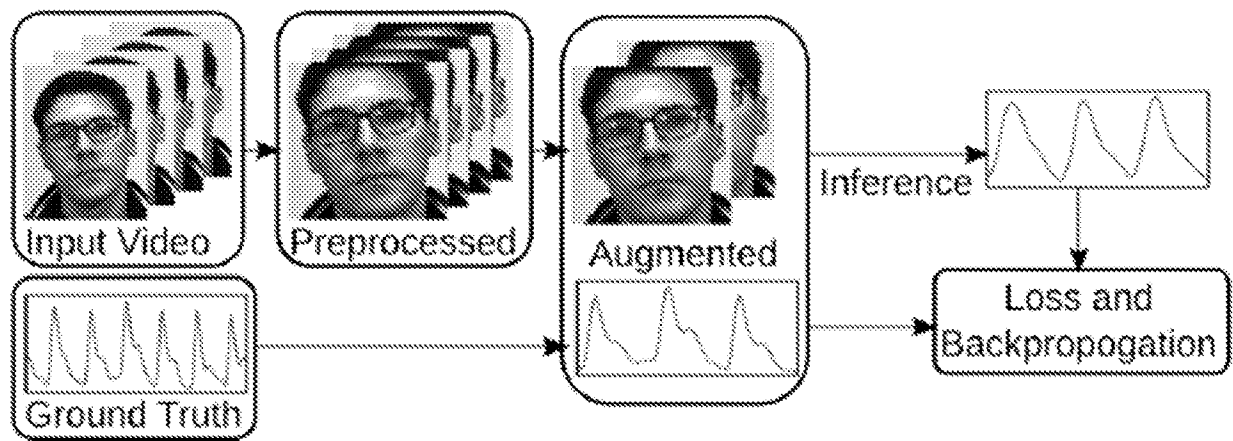


FIG. 3

4/6

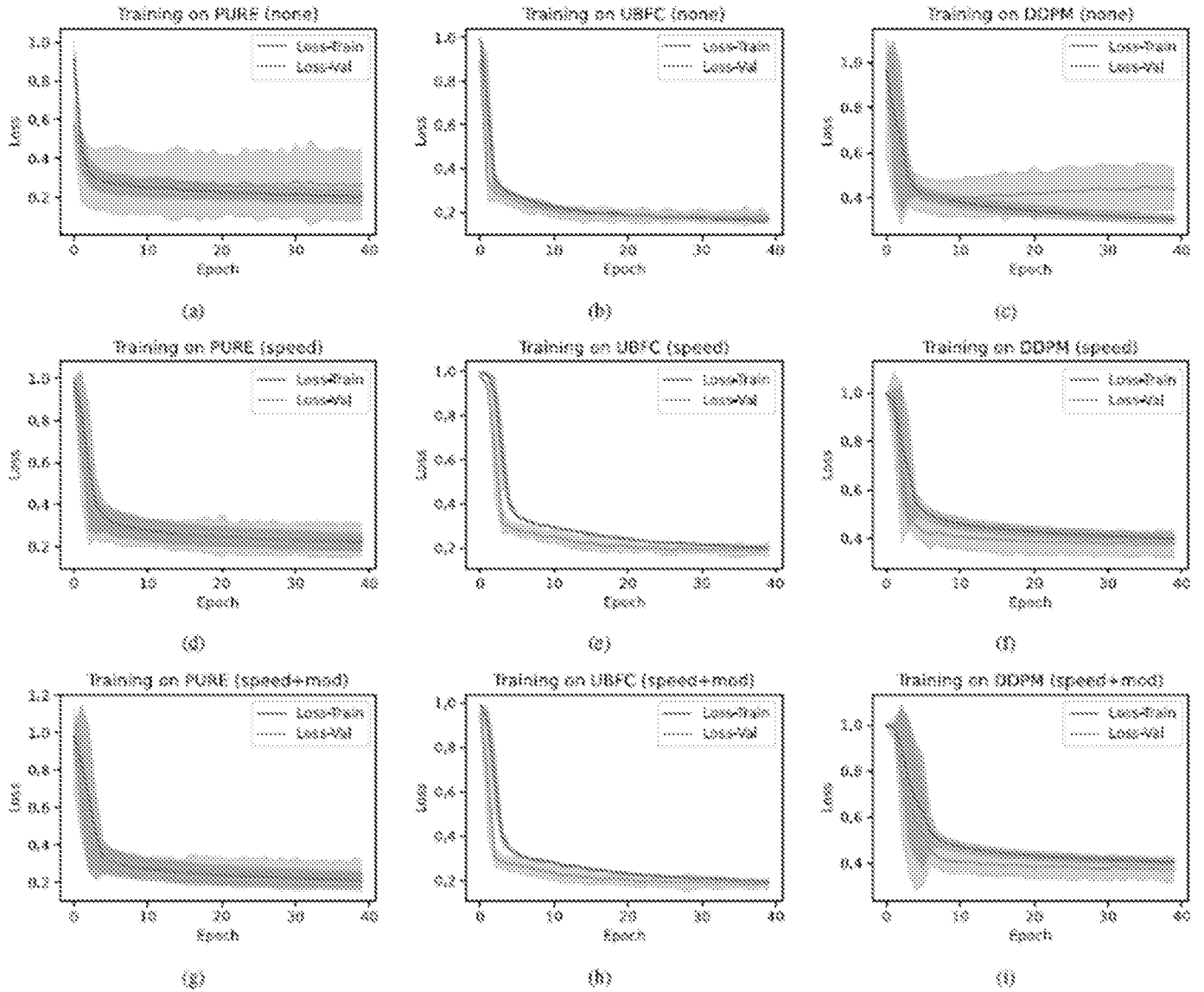


FIG. 4

5/6

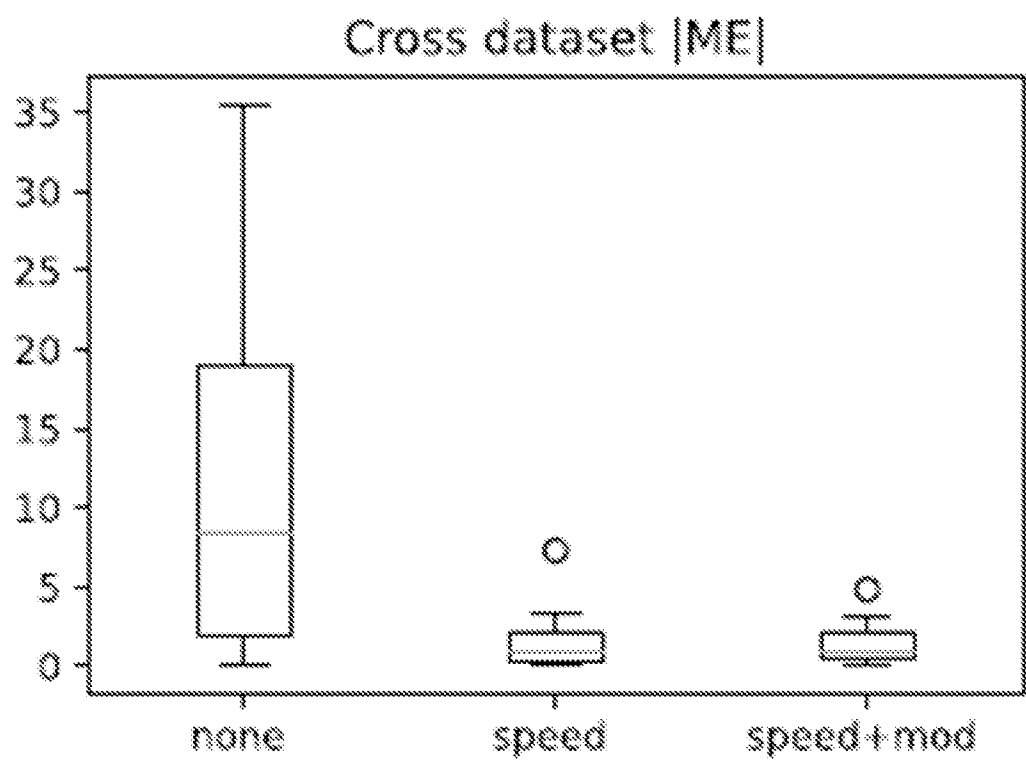


FIG. 5

6/6

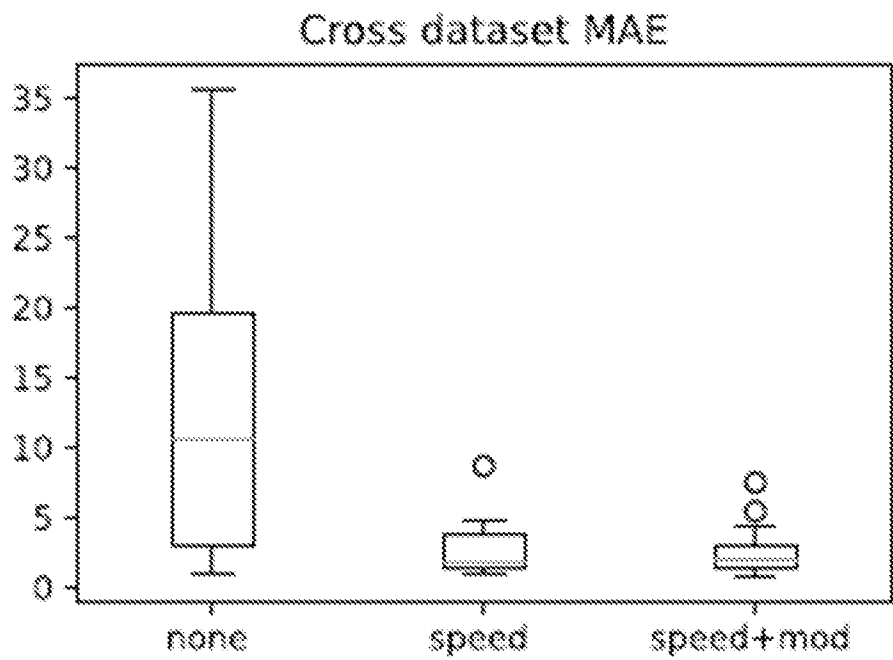


FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/IB2024/055900

A. CLASSIFICATION OF SUBJECT MATTER A61B 5/024 (2006.01)i; A61B 5/00 (2006.01)i; G16H 30/40 (2018.01)i; G06T 5/50 (2006.01)i; G06T 7/11 (2017.01)i; G06V 40/16 (2022.01)i According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) A61B 5/024(2006.01); A61B 5/00(2006.01); G06F 21/32(2013.01); G06K 9/00(2006.01); G06K 9/62(2006.01); G06T 7/90(2017.01) Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched Korean utility models and applications for utility models Japanese utility models and applications for utility models Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) eKOMPASS(KIPO internal) & Keywords: photoplethysmography, video-stream, face, bio-signal		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2017-0367590 A1 (UNIVERSITA` DEGLI STUDI DI TRENTO (UNIVERSITY OF TRENTO) et al.) 28 December 2017 (2017-12-28) paragraphs [0007]-[0108]; and claims 1-13	1-20
A	US 2021-0209388 A1 (ILKE DEMIR et al.) 08 July 2021 (2021-07-08) claims 1-20; and figures 1A-15	1-20
A	WO 2018-002275 A1 (KONINKLIJKE PHILIPS N.V.) 04 January 2018 (2018-01-04) claims 1-15; and figures 1-7	1-20
A	US 2019-0332757 A1 (AZ BOARD OF REGENTS ON BEHALF OF AZ STATE UNIV.) 31 October 2019 (2019-10-31) claims 1-20; and figures 1-17	1-20
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents: “A” document defining the general state of the art which is not considered to be of particular relevance “D” document cited by the applicant in the international application “E” earlier application or patent but published on or after the international filing date “L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) “O” document referring to an oral disclosure, use, exhibition or other means “P” document published prior to the international filing date but later than the priority date claimed “T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention “X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone “Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art “&” document member of the same patent family		
Date of the actual completion of the international search 02 October 2024		Date of mailing of the international search report 02 October 2024
Name and mailing address of the ISA/KR Korean Intellectual Property Office 189 Cheongsa-ro, Seo-gu, Daejeon 35208, Republic of Korea Facsimile No. +82-42-481-8578		Authorized officer LEE, Kang Ha Telephone No. +82-42-481-5687

INTERNATIONAL SEARCH REPORT

International application No.

PCT/IB2024/055900

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	NATHAN VANCE et al., Promoting Generalization in Cross-Dataset Remote Photoplethysmography , 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, pp. 5985-5993, 17 June 2023 [retrieved on 06-08(Aug)-2024] from <ieeexplore.ieee.org/abstract/document/10208448> pages 5985-5992	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/IB2024/055900

Patent document cited in search report				Publication date (day/month/year)	Patent family member(s)		Publication date (day/month/year)
US	2017-0367590	A1	28 December 2017	US	10335045	B2	02 July 2019
US	2021-0209388	A1	08 July 2021	US	11687778	B2	27 June 2023
				US	2023-0351178	A1	02 November 2023
WO	2018-002275	A1	04 January 2018	CN	109640821	A	16 April 2019
				CN	109640821	B	28 June 2022
				EP	3478176	A1	08 May 2019
				EP	3478176	B1	17 August 2022
				JP	2019-519319	A	11 July 2019
				JP	6932150	B2	08 September 2021
				US	10987030	B2	27 April 2021
				US	2019-0209052	A1	11 July 2019
US	2019-0332757	A1	31 October 2019	US	11580203	B2	14 February 2023