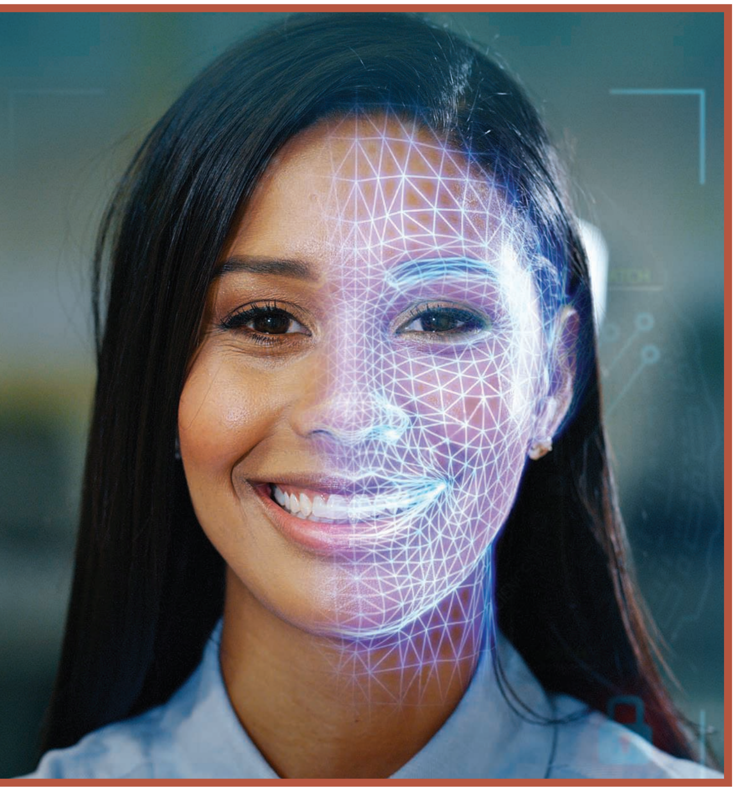
Zitong Yu, Xiaobai Li, and Guoying Zhao

Facial-Video-Based Physiological Signal Measurement

Recent advances and affective applications



©SHUTTERSTOCK.COM/HQUALITY

Monitoring physiological changes [e.g., heart rate (HR), respiration, and HR variability (HRV)] is important for measuring human emotions. Physiological responses are more reliable and harder to alter compared to explicit behaviors (such as facial expressions and speech), but they require special contact sensors to obtain. Research in the last decade has shown that photoplethysmography (PPG) signals can be remotely measured (rPPG) from facial videos under ambient light, from which physiological changes can be extracted. This promising finding has attracted much interest from researchers, and the field of rPPG measurement has been growing fast. In this article, we review current progress on intelligent signal processing approaches for rPPG measurement, including earlier works on unsupervised approaches and recently proposed supervised models, benchmark data sets, and performance evaluation. We also review studies on rPPG-based affective applications and compare them with other affective computing modalities. We conclude this article by emphasizing the current main challenges and highlighting future directions.

Introduction

In the past two decades, affective computing, the study of automatically processing, interpreting, and simulating human affects, has been attracting increasing attention and has been widely applied in everyday applications (e.g., remote education, autonomous driving, and psychotherapy).

Various emotion theories have been proposed in psychological studies. Researchers have diverse rather than unanimous opinions [1] about how to represent and measure emotions, and the debate is still ongoing. Two prevailing emotion models can be summarized, which describe emotions as either categorical or dimensional. Categorical models consider emotions as multiple discrete categories; e.g., one of the most typical models is Ekman's six basic emotions. On the other hand, dimensional models describe emotions as variants that change along two or more continuous dimensions, e.g., valence, arousal, and dominance. In affective computing, both categorical and dimensional models have been widely used. The selection of emotion

Digital Object Identifier 10.1109/MSP.2021.3106285
Date of current version: 27 October 2021

models is a key factor in affective data building (i.e., to achieve the data labels) and impacts the design of computational methods for different real-world applications.

Inspired by human affective perception and manners, machine intelligence is being used to explore and interpret emotions from various modalities, such as facial expressions, speech, and physiological responses. While explicit behaviors, including facial expressions and speech, can be faked, physiological responses (e.g., HR, respiration, and HRV) modulated by the autonomic nervous system are hard to voluntarily alter and are more reliable for affective computing under certain circumstances.

Traditionally, special contact sensors are needed to measure physiological signals. For example, electrocardiography (ECG) is used for measuring electrical cardiac activity, in PPG, an oximeter is used for measuring the blood volume pulse (BVP), and a breathing belt is used to measure respiration. Contact-based measurements suffer from two drawbacks: 1) inconvenience and discomfort, especially for long-term monitoring and for human–human interactions/human–computer interactions (HCIs); and 2) the constraint-based status impedes the expression of spontaneous emotions. One study [2] showed that it is possible to remotely measure PPG signals under ambient light, and the topic of rPPG measurement has attracted great attention in recent years. Figure 1(a) shows the trend of rPPG-related publications during the last 10 years.

The fundamental mechanism of remote PPG measurement is illustrated in Figure 1(b). Facial skin contains rich blood vessels. When ambient light shines on the skin, part of the light is absorbed (mainly by the hemoglobin in the blood), and the rest is reflected and captured by a camera. The heart pumps blood through the body, and for a local skin region, the count or density of hemoglobin fluctuates with the pulsation, which then changes the amount of light absorbed. The rPPG technology uses a camera to capture the periodic color change of a local skin region, which is dependent on the amount of absorbed light. A general framework for rPPG signal measurement approaches is illustrated in Figure 2. Physiological features (e.g., HR, respiration, and HRV) can be extracted from the recovered rPPG signals and used for various affective applications.

One major challenge is that the recorded skin-color changes indicating rPPG signals are very subtle and can be easily affected by noises such as environmental light variations and the subjects' head movements. In previous rPPG studies, efforts

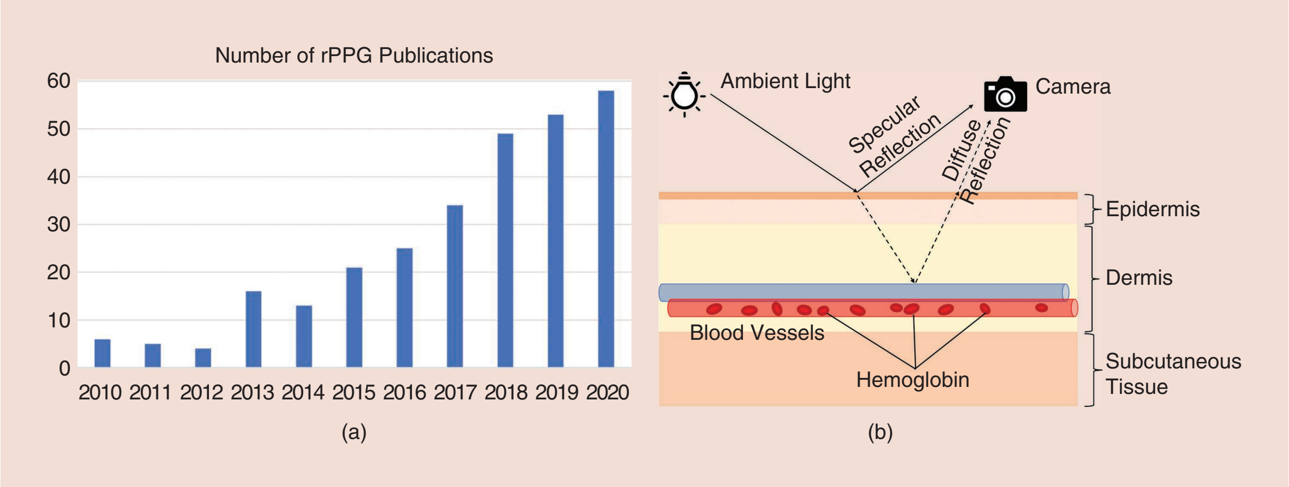


FIGURE 1. (a) The number of rPPG publications over the past decade (obtained through a Google Scholar search with the keywords: allintitle: "remote photoplethysmography," "remote heart rate," "remote physiological," "rPPG" and "iPPG"). (b) A reflection model of rPPG. iPPG: imaging photoplethysmography.

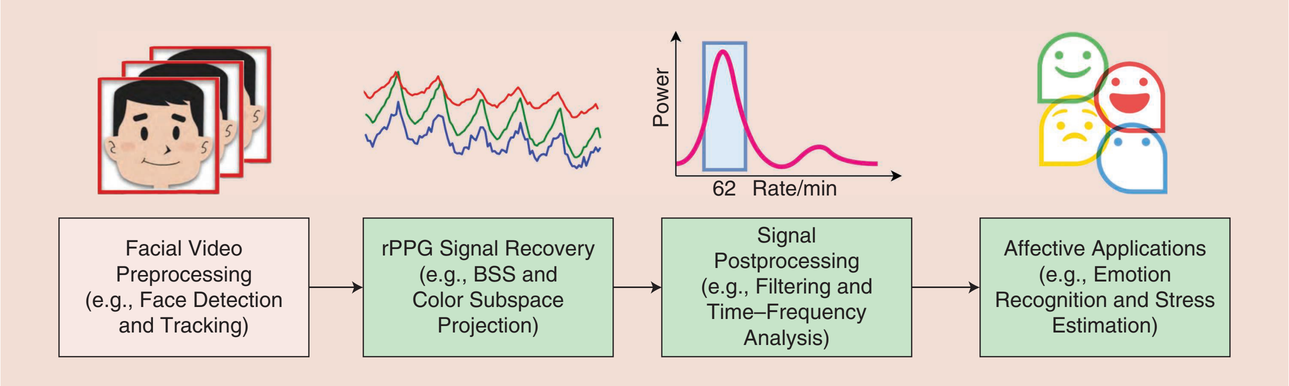

Power
62 Rate/min
Facial Video Preprocessing (e.g., Face Detection and Tracking)
rPPG Signal Recovery (e.g., BSS and Color Subspace Projection)
Signal Postprocessing (e.g., Filtering and Time–Frequency Analysis)
Affective Applications (e.g., Emotion Recognition and Stress Estimation)

FIGURE 2. A general framework for facial rPPG measurement and its affective applications. BSS: blind source separation.

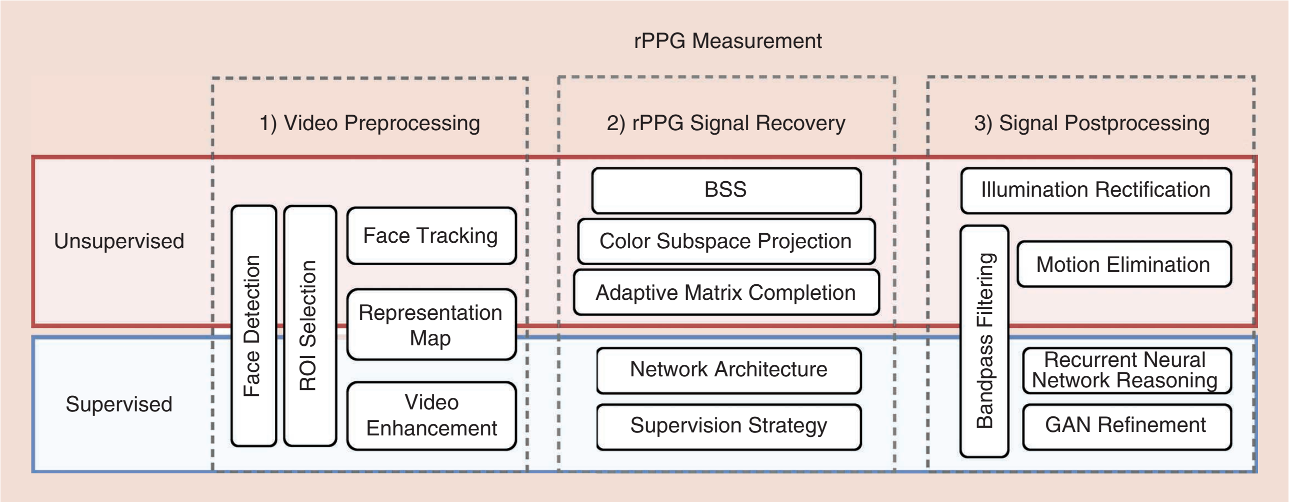


FIGURE 3. A general rPPG measurement framework: three steps (column) and two groups (row) of approaches. ROI: region of interest. BSS: blind source separation. GAN: generative adversarial network.

have been made to alleviate the impacts of noise for more accurate rPPG measurement. Besides developing approaches for robust rPPG measurement, other studies have also explored utilizing the remotely measured physiological signals for various affective applications. This research has demonstrated the great potential of the rPPG technology.

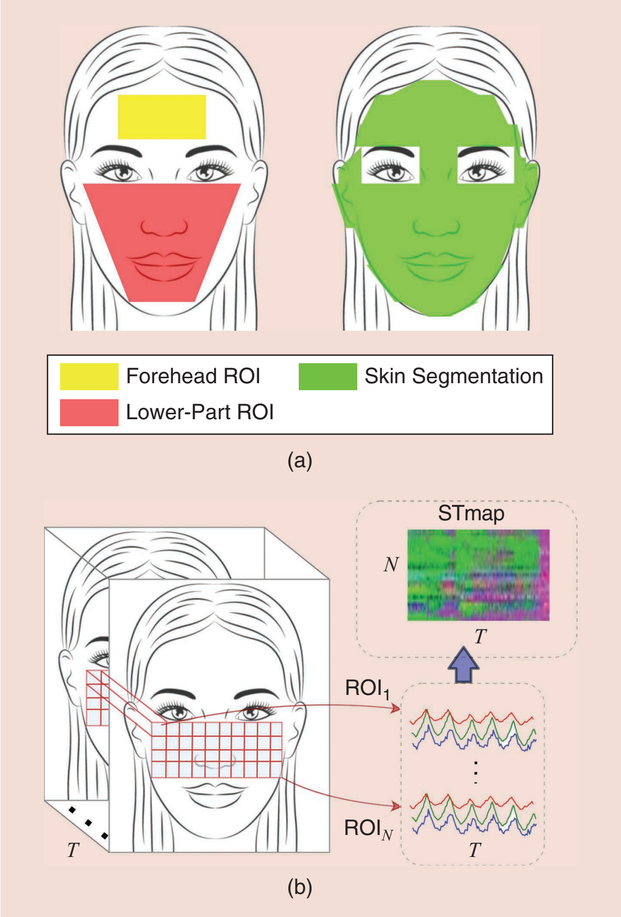


FIGURE 4. (a) Different facial ROIs for rPPG recovery. (b) Spatiotemporal representation of multiple facial ROIs, i.e., the spatiotemporal map (STmap).

Approaches to rPPG measurement

Despite each approach's specialties, a general framework of prevailing rPPG measurement approaches can be summarized in three steps: video preprocessing, rPPG signal recovery, and signal postprocessing. Each step may involve multiple processes; the most frequently employed ones are shown in Figure 3. The prevailing unsupervised (red) and supervised (blue) rPPG approaches will be introduced subsequently.

Early-stage unsupervised rPPG approaches

Early-stage ([2]–[6], 2007–2016) rPPG approaches usually involve straightforward signal processing steps and do not rely on supervision from the contact-measured physiological signals.

Video preprocessing and region-of-interest selection

The density distribution of blood vessels varies in different facial regions; thus, it is important to select effective regions of interest (ROIs) with rich PPG clues. Face detection is usually first applied to localize the face region and remove the background. Then ROI selection is performed in both the spatial and temporal domains. Researchers have explored different intraframe facial ROIs in the spatial domain. One study [2] found that the forehead region [see Figure 4(a), yellow] works better than other facial regions as the recovered rPPG signals have a higher signal-to-noise ratio. However, the forehead might be occluded by hair or a hat. Other studies [3] preferred to use the lower facial regions [see Figure 4(a), red] of the cheek and nose areas. An alternative solution is to include as many skin pixels as possible as larger ROIs [4] can produce more stable rPPG signals that are less affected by random noise. Skin segmentation is employed to segment all skin pixels within the face region for rPPG measurement [see Figure 4(a), green] according to the color contrast between skin and nonskin parts. Besides defining one single ROI, one approach [5] employed multiple small ROIs divided from a large ROI [see Figure 4(b)] for the rPPG measurement. Such local ROI banks could provide more synergic rPPG clues from various facial regions and mitigate the impact of occluded facial regions.

To eliminate the noisy fluctuations of the recovered rPPG signal, interframe ROI selection along a temporal facial sequence has also been explored. One simple solution is to obtain consistent facial skin regions by applying the same spatial ROI selection approach on every frame [2], [6]. However, frame-level ROI localization operators are unstable and easily influenced by head movements and occlusions, which leads to noisy rPPG signals with high-frequency (HF) artifacts. A better solution is to use tracking instead of single-frame ROI detection. In [3], a predefined ROI was tracked through face sequences based on multiple key feature points within the facial ROI region, and the results indicated that tracking is efficient in achieving a continuous and temporally stable ROI for rPPG measurement.

The most fundamental way to alleviate the effects of hardware noise and achieve raw rPPG signals is to average all of the pixels' intensity values within the selected ROIs frame by frame.

rPPG signal recovery

The most fundamental way to alleviate the effects of hardware noise and achieve raw rPPG signals is to average all of the pixels' intensity values within the selected ROIs frame by frame [2]–[6]. Concerning the three color channels, based on the fact that the light wavelength corresponding to the green channel has an absorption peak by (oxy-) hemoglobin, rPPG signals recovered from the green color channel [2], [3] are usually better than those from the red or blue channels.

Raw single/multichannel rPPG signals usually contain mixed sources including the target rPPG signal along with noise fluctuations, such as shading caused by motion and lighting variations. Disentangling the pure and intrinsic pulse curve from irrelevant interference is a basic problem in obtaining a reliable measurement. One solution is to use blind source separation (BSS) methods to remove noise and recover the underlying target signal (i.e., the rPPG signal). Specifically, independent component analysis or principal component analysis can be adopted to decompose the raw rPPG signals into several sources or components, from which the one with the strongest periodicity or energy is selected as the target rPPG signal. Another solution is color-space transfer. Compared with the original red, green, blue (RGB) space, the chrominance [4] subspace is less sensitive to motion and luminance. Thus, it is feasible to refine raw rPPG signals via an exact projection direction on the subspace plane by real-time tuning. In addition, to explore the intrinsic rPPG relationships among multiple ROIs, a spatiotemporal map (STmap) [5] has been used [see Figure 4(b) for a typical example] for each frame. In this method, low-rank-based self-adaptive matrix completion (SMAC) [5] was applied to select high-quality ROIs for rPPG estimation, while those impacted by motion or shading were dropped.

Signal postprocessing

In the final step, signal postprocessing further refines the rPPG quality for subsequent applications. As the frequency of a human heartbeat usually ranges from 0.7 to 4 Hz, a bandpass filter with a corresponding bandwidth [2], [3] has often been used to refine the rPPG signal in the frequency domain. Detrending [6] is another popular postprocessing method; it removes the slow fluctuation of the rPPG signal caused by environmental light variations or auto adjustment of the white balance, for example. Additional postprocessing approaches have also been proposed [3], such as illumination rectification via adaptive filtering and nonrigid motion elimination by excluding low-quality segments to deal with exceptionally challenging data.

To sum up, conventional studies have analyzed factors that could impact the rPPG measurement (e.g., illumination variations and head motions) and proposed some preliminary solutions (e.g., ROI selection, BSS, and illumination rectification) that do not involve supervised learning and usually come with small computational costs. The approaches have the following limitations: 1) they require empirical knowledge to choose the proper parameters for designing the signal processing filters; 2) there is a lack of advanced video processing tools and supervised learning models to counter data variations, especially in challenging environments with a lot of interference.

Emerging supervised rPPG approaches

One main difference from conventional unsupervised approaches (red in Figure 3) is that supervised approaches (blue in Figure 3) can leverage contact-measured ground truths with an efficient supervised learning paradigm.

Video preprocessing for a supervised paradigm

In a supervised rPPG representation paradigm, detected face sequences and extracted STmaps are the two most typical inputs for supervised modeling. Sequences of cropped faces could be directly fed into an end-to-end learning framework without further handcrafted preprocessing. Such a learnable mapping from video-level input to 1D signal output is flexible, but it could also easily overfit on small-scale data. STmaps extracted from multiple predefined ROIs have recently been adopted as a refined form of input for supervised rPPG frameworks. These frameworks focus on learning an underlying mapping from the input feature maps to the target signals. Compared to cropped faces, learning with STmap inputs is more efficient and converges faster because the process of generating STmaps already collects raw rPPG information and excludes major irrelevant elements (e.g., face-shape attributes). Another noticeable video preprocessing approach is the rPPG-dedicated video enhancement technique [7], which deals with highly compressed face videos by enhancing intrinsic rPPG clues and reducing undesired compression artifacts, thus benefiting the subsequent rPPG signal recovery process.

Supervised rPPG signal recovery

rPPG signal recovery is the core part of the supervised rPPG approach. We will discuss it from three perspectives: the network architecture (concerning the structure of the model, task-aware inputs, and ground truth), the loss function (for designing suitable constraints to supervise the model for robust feature representation), and the learning strategy (for accuracy, efficiency, and generalization tradeoffs).

From the perspective of network architecture, both 2D (spatial) [8], [9] and 3D (spatiotemporal) [7], [20] models have been explored. A 2D convolutional neural network (CNN) learns spatial rPPG features within each face and aggregates across the frames, while a 3D CNN leverages both spatial and temporal contexts from the input volume. DeepPhys [8] is the first end-to-end rPPG approach with 2D two-stream convolutional attention networks (CANs), in which a motion stream explores facial color changes from the normalized difference of adjacent frames, and an appearance stream generates facial attention maps for rPPG feature refinement. A 3D CNN could explore more efficient spatiotemporal contexts for rPPG representation. In [7], a 3D CNN-based rPPG network (rPPGNet) was designed, which contains a skin-based attention module for adaptively selecting skin areas with stronger rPPG signals. To alleviate the huge number of parameters needed in 3D CNN-based spatiotemporal modeling, an efficient 2D CNN with a temporal shift (TS)-CAN module [9] has been proposed for real-time physiological measurement on mobile platforms, which is more practical for real-world deployment.

Overall, it is practical to consider STmap-like inputs and learnable postprocessing in complex scenarios, and improving the robustness of end-to-end supervised methods is urgent.

In terms of the loss function, both time-domain and frequency-domain constraints have been explored for supervising rPPG models. Supervision in the time domain aims to minimize the intensity of the temporal difference between the predicted rPPG signal and the ground-truth signal. There are two typical time-domain losses: the mean-square error (MSE) loss [8], [9] and the negative Pearson correlation (NegPearson) loss [7], [20]. The MSE compares the mean magnitude difference between the estimated and ground-truth signals, while the NegPearson focuses on their trend similarity. As rPPG signals are recorded in a different way from the ground-truth physiological signals, the curve magnitudes (pixel values) are dependent on the device settings, the environment, and the subject's physical condition. To this end, the NegPearson loss might be a more reasonable option, and it also converges faster, as demonstrated in [7]. A frequency-domain loss assumes that, within a short time span (e.g., <10 s), the power spectrum density (PSD) curve of the rPPG signal should be sharp (with a high amplitude) near the target frequency band (corresponding to the ground-truth HR value) while being comparatively plain (with a low amplitude) in other frequency bands. To improve the periodicity of the rPPG signal, the cross-entropy loss [10] has been used to constrain the PSD distribution for frequency supervision.

Regarding the learning strategy, multiple learning strategies have been explored for rPPG measurement, including multitask learning [7], [9], disentangled learning [10], and metalearning [11]. As the rPPG measurement task is highly related to other tasks, such as facial skin segmentation and respiratory measurement, learning their common features might benefit all and reduce irrelevant interferences. In [7], the rPPGNet employed multitask learning and jointly learned two tasks of regressing rPPG signals and segmenting facial skin regions so that the learned color changes focus more on skin regions. In [9], the joint measurement of multiple physiological signals (rPPG and respiration) was also proven to be efficient in a multitask supervised learning framework. Another strategy is to use disentangled learning to eliminate nonphysiological noise (such as light variation and sensor noise). In [10], a cross-verified disentangling strategy was developed and tested to distill rPPG features from nonphysiological features. Furthermore, the domain-shift issue needs to be considered as practical rPPG measurement can be affected by changes of environment, skin tone, and so on. To counter this issue, Lee et al. [11] introduced a metalearning approach that can adapt and generalize one rPPG model to specific domains.

Supervised signal postprocessing

Supervised signal postprocessing aims to adaptively exploit the temporal contexts to refine the estimated rPPG signals or features. One approach is long-range temporal modeling between adjacent face video clips as their physiological parameters should be highly related. In [12], one temporal reasoning technique (a gated recurrent unit) was applied to adaptively refine rPPG features according to clip-level temporal contexts. Another study [13] also used a generative model with adversarial learning to postprocess estimated rPPG signals to reduce noise and improve output quality.

One thing to mention is that not all supervised methods contain explicit preprocessing and postprocessing steps. Some studies [7]–[9], [11], [20] preferred an integrated end-to-end approach, which takes face frames as the inputs and outputs rPPG signals directly. End-to-end rPPG approaches are less dependent on task-related prior knowledge and handcrafted engineering (e.g., STmap generation) but rely on diverse and large-scale data to alleviate the problem of overfitting.

Benchmark data sets and evaluations

Before 2012 there were no public data sets for rPPG measurement; therefore, most studies used self-collected, small-scale data sets that were not shared. It is a waste of time to repetitively collect data, and unshared data make it impossible for fair comparisons among different algorithms. Later, several public data sets were released to fulfill the needs of rPPG measurement studies. A summary of benchmark data sets for rPPG measurement is shown in Table 1.

Table 1. Public data sets for remote physiological measurement.

Data Set	Year	Subjects	Videos	Physiological Signal	Affective Application
MAHNOB [14]	2012	27	527	ECG, EEG	Emotion recognition
BioVid [15]	2013	90	8,700	ECG, EEG, EMG, SC	Pain estimation
MMSE-HR [5]	2016	40	102	HR	HR estimation
OBF [16]	2018	100	200	BVP, ECG, BR	HR estimation
VIPL-HR [12]	2019	107	2,378	HR, BVP, SpO2	HR estimation
UBFC-rPPG [17]	2019	42	42	BVP	HR estimation
uulmMAC [18]	2020	57	95	ECG, BR, SC, EMG	Emotion recognition
UBFC-Phys [19]	2021	56	168	BVP, SC, EDA	Stress recognition

MMSE: minimum MSE; OBF: Oulu Bio-Face; EEG: electroencephalography; EMG: electromyography; SC: skin conductance; BR: breathing rate; SpO2: oxygen saturation; EDA: electrodermal activity.

The data sets contain facial videos and corresponding physiological signals as the ground truth for performance evaluation. Concerning the scale of the data sets, the BioVid and VIPL-HR data sets contain a much larger number of samples (thousands) than the other data sets (about 500 or less). Concerning the diversity of the data, most of the data sets were recorded indoor with one fixed scenario setup, while the VIPL-HR data set involves various scenarios with different illuminations and camera setups. From these aspects, these data sets are still not sufficient for training large and deep networks. In terms of video quality, 1) most data sets contain videos that are compressed via modern standards (e.g., MPEG-4 and H.264) except UBFC-rPPG and UBFC-Phys, which contain lossless videos without compression; 2) videos of most data sets are of high definition resolution (1,920 × 1,080), except BioVid (1,280 × 1,024) and UBFC-rPPG (640 × 480). It is worth mentioning that, besides color videos, Oulu Bio-Face (OBF), VIPL-HR, and uulmMAC also provide near-infrared videos. As for the ground-truth signals, minimum MSE (MMSE)-HR and UBFC-rPPG only provide one single ground-truth signal about the HR, while the other data sets provide multiple physiological signals, including ECG, electroencephalography (EEG), electromyography (EMG), skin conductance (SC), oxygen saturation (SpO2), and electrodermal activity. Some of the data sets were designed for affective applications and provide special affective labels, e.g., MAHNOB [14] and uulmMAC [18] for emotion recognition, BioVid [15] for pain-level estimation, and UBFC-Phys [19] for stress recognition.

Performance evaluation

Most existing methods compare performance on the average HR of each input video in beats per minute (bpm). Several common evaluation metrics have been used, such as the standard deviation of the error (SD), the mean absolute error (MAE), the root MSE (RMSE), the mean error rate percentage, and Pearson's correlation coefficient (*r*).

Table 2 summarizes the performance comparison of popular rPPG measurement methods. The state-of-the-art methods (PulseGAN [13] and rPPGNet [7]) can achieve, respectively, a satisfactory performance (MAE = 1.19 bpm and RMSE = 1.8 bpm) on the high-quality data sets OBF and UBFC-rPPG, while the performance is yet to be improved on other more challenging data sets (e.g., VIPL-HR and MAHNOB-HCI). Compared with unsupervised methods (e.g., Verkruysse et al. [2] and CHROM [4]), supervised learning-based methods can predict more accurate HRs on OBF and UBFC-rPPG data sets because of the efficient feature representation learning. In terms of the inputs, STmap-based methods (RhythmNet [12] and CVD [10]) outperform the end-to-end method DeepPhys [8] with face inputs by a large margin (>5 bpm RMSE) on the VIPL-HR data set as the latter is sensitive to head movements. It is worth noting that, with learnable and adaptive postprocessing for coarse rPPG signal refinement, PulseGAN [13] outperforms the other three methods with fixed and straightforward postprocessing on the UBFC-rPPG data set.

Overall, it is practical to consider STmap-like inputs and learnable postprocessing in complex scenarios, and improving the robustness of end-to-end supervised methods is urgent. Table 2 summarizes the performance of recent approaches,

Table 2. Performance evaluation of rPPG methods for average HR estimation.

Method	Data Set	SD$_{(bpm)}$↓	MAE$_{(bpm)}$↓	RMSE$_{(bpm)}$↓	*r*↑
Verkruysse et al. [2]	UBFC-rPPG	–	7.5	14.41	0.62
Meta-rPPG [11]	UBFC-rPPG	7.12	5.97	7.42	0.53
TS-CAN [9]	UBFC-rPPG	–	4.68	–	0.74
PulseGAN [13]	UBFC-rPPG	–	1.19	2.1	0.98
CHROM [4]	OBF	2.73	–	2.73	0.98
PhysNet [20]	OBF	–	–	1.81	0.992
rPPGNet [7]	OBF	1.76	–	1.8	0.992
Li et al. [3]	MAHNOB-HCI	6.88	–	7.62	0.81
SMAC [5]	MAHNOB-HCI	5.81	–	6.23	0.83
DeepPhys [8]	VIPL-HR	13.6	11	13.8	0.11
RhythmNet [12]	VIPL-HR	8.11	5.3	8.14	0.76
CVD [10]	VIPL-HR	7.92	5.02	7.97	0.79

but we mention that some studies used different validation protocols or data partitions in the training and testing phases. To provide a fair comparison platform, the RePSS Challenge has been organized as an annual competition series since 2020 (RePSS 2020: https://competitions.codalab.org/competitions/22287#; RePSS 2021: https://competitions.codalab.org/competitions/30855). This challenge specifically focuses on the fair evaluation of rPPG measurement approaches.

Applications in affective computing

In this section, we first review studies that use rPPG signals for affective computing applications. Then we compare rPPG with other modalities and discuss their strengths and limitations.

Remote physiological signal measurement for affective computing

Emotion understanding is a focused area of research in physiological signal analysis. Multiple kinds of physiological signals are related to emotional status, including those from ECG, EMG, SC, PPG, and EEG, among others. ECG and PPG both measure cardiac activities. For measuring affective status, the average HR alone is not sufficient. The ECG and PPG signals are usually further processed to compute interbeat intervals and conduct an HRV analysis to obtain more sophisticated features. Common HRV features include low-frequency (LF) and HF features, the normalized ratio, and other characteristics in both the time and frequency domains.

Using rPPG for emotion understanding is an emerging new topic, and not many articles have been published so far on this subject. These studies build upon traditional ECG- and PPG-based affective computing studies with an essential extra challenge: to reconstruct rPPG signals from facial videos. Considering the coarse-designed statistical attributes, finding proper features from imperfectly measured rPPGs and training a model to measure the target affective status are key. Here we review representative works that use rPPG signals for affective computing under different scenarios.

Emotion recognition in HCI scenarios

HCI is one of the most common scenarios for affective computing studies; e.g., it can be used to measure emotions while a subject is watching a movie or playing a video game. Yu et al. [20] designed a spatiotemporal network to recover rPPG signals from movie watchers' faces, and then 10D HRV features were extracted for emotion recognition. The study explored emotion recognition in nine categories and also in the valence and arousal dimensions. Besides direct emotion measurement, Gupta et al. [21] also used rPPG to detect the onset of emotional behaviors. The extracted features from recovered rPPG signals were used for facial microexpression spotting under HCI scenarios.

Cognitive stress estimation

McDuff et al. [22] showed that remotely measured physiological changes with a camera could be used for cognitive stress estimation. It is a fact that, when people are under cognitive stress, their autonomic nervous system activity changes, and this can be reflected in some HRV features. McDuff's study demonstrated that the remote rPPG measurement was not 100% accurate, and their model could achieve an accuracy of 85% for cognitive stress estimation. The LF component and breathing rate (BR) are the most indicative features. A recent study [23] showed that peripheral hemodynamics and vasomotion power extracted from rPPG amplitudes are also important indicators for cognitive stress estimation. A multimodal data set was established in [19] for stress estimation with video-based rPPG modality.

One major challenge is that the recorded skin-color changes indicating rPPG signals are very subtle and can be easily affected by noises such as environmental light variations and the subjects' head movements.

Driver status monitoring

Driver status monitoring is one of the focused topics in autonomous driving. Future intelligent driving systems should be able to detect a dangerous status of a driver, e.g., fatigue or sleepiness, to improve safe vehicle operation. Tsai et al. [24] proposed a remote physiological measurement system to instantly monitor driver fatigue without contact devices. Statistical HR and HRV features were extracted from measured rPPG signals and then fed to a regressor to predict the driver's fatigue level.

Pain estimation

Pain is a topic of major research as it not only causes physiological discomfort but also impacts people's mental status (e.g., it causes stress or depression). Kessler et al. [25] used remotely measured rPPG signals as a new approach to pain-level estimation as the heartbeat and breathing patterns are altered when people are in pain. RGB facial videos were evaluated for rPPG signal recovery and HRV analysis, and then the pain level was estimated with a support vector machine or random forest classifier. One main finding is that the LF component of HRV features is important for estimating the pain level.

Engagement measurement in educational activities

Education is a major application field for affective computing technologies. By analyzing teachers' and learners' status in educational activities, e.g., whether students are engaged or not, we can evaluate the effectiveness of educational approaches. Monkaresi et al. [26] estimated students' engagement levels by measuring HRs from facial videos while they were conducting a structured writing task. Seven statistical features were extracted from instantaneous HR signals and cascaded to train a supervised learning model for engagement-level estimation. Unlike typical emotions, engagement cannot be measured as a prototype facial expression; thus, remotely measured physiological signals could be a novel, yet convenient, approach for engagement measurement.

Comparison with other affective computing modalities

As humans can perceive emotions from multiple sources, affective computing can be achieved from different modalities, e.g., text, audio speech, visual clues, and physiological signals concerning the input types. Here we mainly compare the audio and visual modalities because of their widespread use in various applications.

Audio modalities

Audio modalities concern an affective analysis from various acoustic inputs, among which one major research area is speech emotion recognition (SER) [27]. SER focuses on recognizing emotions conveyed by speech signals. Speech signals are segmented as a "unit of analysis" for feature extraction and model learning. Features for SER can be summarized into two groups. 1) Textual features are related to the speech content, e.g., the occurrence of some keywords (or word groups). Automatic speech recognition (ASR) is needed to extract the textual features. Culture and language must be involved when using textual features for SER. 2) Audio features indicate lower-level acoustic features such as the energy or spectral information, speed, and rhythm, which do not require ASR and are more robust across different languages and cultures.

Visual modalities

Visual modalities include both images and videos as the inputs for affect analysis. Major research areas include facial expression recognition (FER), emotion body gesture recognition (EBGR), and affective image content analysis (AICA). One key assumption of FER [28] is that each emotion category corresponds to one or more prototypical facial expressions, e.g., wide-opened eyes with a lowered jaw when one is surprised. Besides recognizing general emotion categories, some studies also focused on differentiating genuine and fake expressions as facial behaviors can be voluntarily altered for various purposes. EBGR [29] focuses on analyzing body behaviors (other than those of the face) for affect measurement.

A human body can be modeled as either a composition of multiple local parts or a kinematic chain model of skeleton joints for action (or posture) recognition, e.g., sitting, walking, or jumping. Then relevant representations can be extracted from the actions (postures) for emotion measurement. Compared with facial expressions, body gestures are more complex and diverse in terms of emotion representation. While FER and EBGR focus on human behaviors, AICA [30] concerns all images of any content, i.e., the emotions that an image can induce when it is shown to a person. Various features are explored for the task, including low-level ones, such as colors and edges; midlevel features, such as materials and eigenfaces; and high-level semantic features, such as facial expressions.

Modality comparison

rPPG is a unique technique among all modalities for affective computing. On one hand, it is one of the visual modalities (along with facial expression, body gesture, and so on) and so may share some of their common advantages and challenges, e.g., related to video quality, lighting, occlusion, and so on. On the other hand, rPPG also intersects with physiological modalities and possesses some of their characteristics.

Compared with other modalities, rPPG signals have two main advantages for emotion measurement. The first is inherent to the general physiological domain: physiological signals might be the most reliable source among all modalities as it is difficult to intentionally control or alter one's physiological responses. People can control their facial expressions, body gestures, and speech to hide emotions or convey fake ones if needed. From this perspective, physiological modalities (including rPPG) are essential for measuring suppressed emotions when limited movement or speech is presented. However, traditional physiological monitoring requires contact sensors, which could be a major drawback in practical applications before rPPG techniques appear. The second advantage of rPPG is that it requires only one color camera, and the captured facial videos can be processed for both rPPG measure and facial expression analysis for emotion recognition.

There are also disadvantages to using rPPG for emotion recognition. First, compared to other visual modalities, e.g., facial expressions or body gestures, rPPG signals are weaker and are more easily affected by lighting changes and motion. Current methods still need to be improved to increase their robustness. Second, rPPG only measures heartbeat, which is limited for emotion recognition. It would be better if more physiological signals could be remotely measured and combined for the task. Some work [23] has explored novel rPPG-related physiological indexes from facial videos, but so far it is not likely to involve, e.g., SC and EEG signals.

Open challenges and future directions

In this article, we introduced facial-video-based remote physiological measurement approaches, data sets, and applications in affective computing. Despite the great progress in recent years, there are the following noteworthy challenges:

1) The robustness and generalization ability of the current methods are limited for practical applications. Video quality, human attributes and behaviors, and environmental changes all influence the accuracy. The approaches also do not generalize well to novel data sets because of the large data differences in the domain shift.
2) There are insufficient data with limited scale and diversity for deep learning models. The acquisition of ground-truth physiological signals requires medical equipment and professional operation, which limits the data set's scale.
3) Remote measurement of physiological signals other than those of rPPG need to be explored for affective computing.

More effort will be needed to fill in the gaps. Potential future directions include the following:

1) Robust, efficient, and interpretable approaches for rPPG feature representation should be designed. Firstly, more

informative representations from both the time and frequency domains could be designed. Secondly, lightweight CNNs could be explored for real-time rPPG applications.

2) It would be useful to learn from limited or unlabeled data. Data augmentation or synthesis methods would be helpful to achieve more data samples. Self-supervised pretraining or semisupervised methods could be explored to use unlabeled data from the Internet.

3) It would also be helpful to explore multimodality approaches that fuse rPPG signals with other modalities for more reliable affective measurement.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant 61772419) and the Academy of Finland (grants 316765 and 323287). Zitong Yu and Xiaobai Li contributed equally. Guoying Zhao is the corresponding author.

Authors

Zitong Yu (zitong.yu@oulu.fi) received his M.S. degree in multimedia from the University of Nantes. He is currently a Ph.D. candidate in the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, 90014, Finland. He is a Student Member of IEEE.

Xiaobai Li (xiaobai.li@oulu.fi) received her Ph.D. in computer science and engineering from the Center for Machine Vision and Signal Analysis, University of Oulu. She is currently an assistant professor in the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, 90014, Finland. She is a Member of IEEE.

Guoying Zhao (guoying.zhao@oulu.fi) received her Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences. She is with the School of Information and Technology, Northwest University, China, and is currently a professor with the Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, 90014, Finland. She is a fellow of the International Association for Pattern Recognition and a Senior Member of IEEE.

References

[1] J. M. F. Dols and J. A. Russell, *The Science of Facial Expression*. Oxford, U.K.: Oxford Univ. Press.

[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. doi: 10.1364/OE.16.021434.

[3] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2014, pp. 4264–4271.

[4] G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2878–2886, 2013. doi: 10.1109/TBME.2013.2266196.

[5] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2016, pp. 2396–2404.

[6] M. Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics Expr.*, vol. 18, no. 10, pp. 10,762–10,774, 2010. doi: 10.1364/OE.18.010762.

[7] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 151–160.

[8] W. Chen and D. McDuff, "DeepPhys: Video-based physiological measurement using convolutional attention networks," *European Conf. Comput. Vision,* pp. 349–365, 2018.

[9] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-task temporal shift attention networks for on-device contactless vitals measurement," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 19,400–19,411, Dec. 2020.

[10] X. Niu, Z. Yu, H. Han, X. Li, S. Shan, and G. Zhao, "Video-based remote physiological measurement via cross-verified feature disentangling," in *Proc. European Conf. Comput. Vision*, 2020, pp. 295–310.

[11] E. Lee, E. Chen, and C. Y. Lee, "Meta-RPPG: Remote heart rate estimation using a transductive meta-learner," in *Proc. European Conf. Comput. Vision*, 2020, pp. 392–409.

[12] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Trans. Image Process.*, vol. 29, pp. 2409–2423, Oct. 2019. doi: 10.1109/TIP.2019.2947204.

[13] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, "PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1373–1384, 2021. doi: 10.1109/JBHI.2021.3051176.

[14] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 42–55, 2011. doi: 10.1109/T-AFFC.2011.25.

[15] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C.Traue, P. Werner, and G. M. da Silva, "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in *Proc. IEEE Int. Conf. Cybernetics*, 2013, pp. 128–131.

[16] X. Li, I. Alikhani, J. Shi, T. Seppanen, J. Junttila, K. Majamaa-Voltti, and G. Zhao, "The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recogn.*, 2018, pp. 242–249.

[17] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recog. Lett.*, vol. 124, pp. 82–90, 2019. doi: 10.1016/j.patrec.2017.10.017.

[18] D. Hazer-Rau, S. Meudt, A. Daucher, J. Spohrs, H. Hoffmann, F. Schwenker, and H. C. Traue, "The uulmMAC database—A multimodal affective corpus for affective computing in human-computer interaction," *Sensors*, vol. 20, no. 8, p. 2308, 2020. doi: 10.3390/s20082308.

[19] R. Meziatisabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang, "UBFC-Phys: A multimodal database for psychophysiological studies of social stress," *IEEE Trans. Affective Comput.*, 2021. doi: 10.1109/TAFFC.2021.3056960.

[20] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," *Proc. British Machine Vision Conf.*, 2019, pp. 277–286.

[21] P. Gupta, B. Bhowmick, and A. Pal, "Exploring the feasibility of face video based instantaneous heart-rate for micro-expression spotting," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn. Workshops*, 2018, pp. 1316–1323.

[22] D. McDuff, S. Gontarek, and R. Picard, "Remote measurement of cognitive stress via heart rate variability," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2014, pp. 2957–2960.

[23] D. McDuff, I. Nishidate, K. Nakano, H. Haneishi, Y. Aoki, C. Tanabe, K. Niizeki, and Y. Aizu, "Non-contact imaging of peripheral hemodynamics during cognitive and psychological stressors," *Sci. Rep.*, vol. 10, no. 1, pp. 1–13, 2020. doi: 10.1038/s41598-020-67647-6.

[24] Y. C. Tsai, P. W. Lai, P. W. Huang, T. M. Lin, and B. F. Wu, "Vision-based instant measurement system for driver fatigue monitoring," *IEEE Access*, vol. 8, pp. 67,342–67,353, Apr. 2020. doi: 10.1109/ACCESS.2020.2986234.

[25] V. Kessler, P. Thiam, M. Amirian, and F. Schwenker, "Pain recognition with camera photoplethysmography," in *Proc. Int. Conf. Image Process. Theory, Tools Appl.*, 2017, pp. 1–5.

[26] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Trans. Affective Computing*, vol. 8, no. 1, pp. 15–28, 2016. doi: 10.1109/TAFFC.2016.2515084.

[27] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 107–129, 2017. doi: 10.1109/MSP.2017.2699358.

[28] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Comput.*, early access, Mar. 17, 2020. doi: 10.1109/TAFFC.2020.2981446.

[29] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affective Comput.*, vol. 12, no. 2, pp. 505–523, 2018. doi: 10.1109/TAFFC.2018.2874986.

[30] S. Zhao, G. Ding, Q. Huang, T. S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: A comprehensive survey," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 5534–5541.

SP