Facial Video-based Physiological Signal Measurement: Recent Advances and Affective Applications

Zitong Yu*[1], Xiaobai Li*[1], and Guoying Zhao^[2,1]

(* equal contributions; ^ corresponding author)

[1] Center for Machine Vision and Signal Analysis, University of Oulu, Finland

[2] School of Information and Technology, Northwest University, PRC

Abstract: Monitoring physiological changes (e.g., heart rate, respiration, heart rate variability) are important for measuring human's emotions. Physiological responses are more reliable and harder to alter compared to explicit behaviors (e.g., facial expressions, speech), but require special contact sensors to achieve. Research in the latest decade has shown that photoplethysmograph (PPG) signals can be remotely measured (i.e., rPPG) from facial videos under ambient light, from which physiological changes can be extracted. This promising finding has attracted big interests 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 of unsupervised approaches and recently proposed supervised models, benchmark datasets, and performance evaluation. Furthermore, 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.

1. Introduction

In the recent two decades, affective computing, the study of automatically processing, interpreting and simulating human affects, has attracted more and more

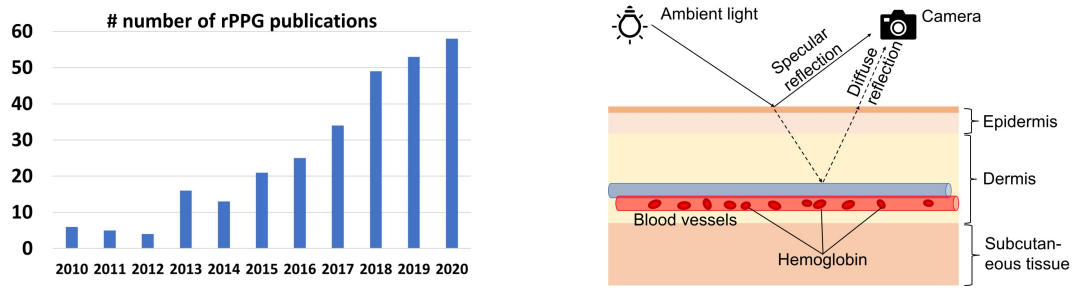
attentions and has been widely applied in everyday applications (e.g., remote education, autonomous driving, and psychotherapy).

There are various emotion theories 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. In the debate, two prevailing emotion models can be summarized, which describe emotions either as 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 models is a key factor for affective data building (i.e., the data labels) which 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, e.g., facial expressions, speech, and physiological responses. While explicit behaviors including facial expressions and speech can be faked, physiological responses (e.g., heart rate (HR), respiration, HR variability (HRV)) modulated by the autonomic nervous system are hard to be voluntarily altered, and are more reliable for affective computing under certain circumstances.

Traditionally, special contact sensors are needed to measure physiological signals, e.g., electrocardiography (ECG) is used for measuring electrical cardiac

activity, a photoplethysmograph (PPG) 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: (i) inconvenience and discomfort especially for long-term monitoring and for human-human/human-computer interactions; and (ii) the constraint 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 remote PPG (rPPG) measurement has attracted great attentions in recent years. Figure 1(a) shows the trend of rPPG related publications in the last ten years.



(a) (b)

Fig. 1 (a) The number of rPPG publications over the past decade. Obtained through Google scholar search with key-words: allintitle: “remote photoplethysmography”, “remote heart rate”, “remote physiological”, “rPPG” and “iPPG”. (b) Reflection model of rPPG.

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 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 to 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.

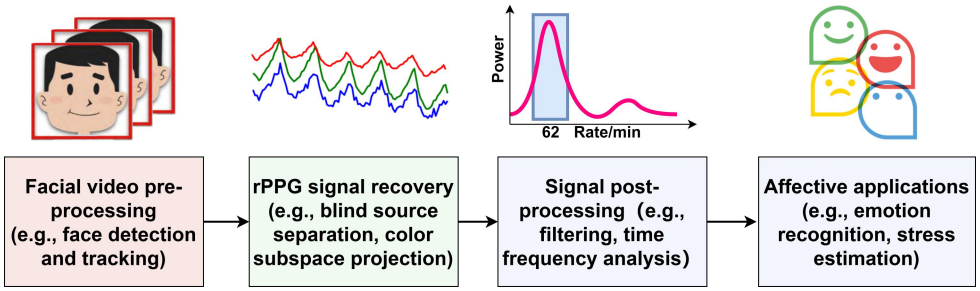


Fig. 2 A general framework for facial rPPG measurement and its affective applications.

One major challenge is that the recorded skin color changes indicating rPPG signals are very subtle, which can be easily affected by noises such as environmental light variations and the subjects' head movements. In previous rPPG studies, efforts 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, which have shown the great potential of the rPPG technology. The article is organized as follows: Section 2 reviews impactful rPPG approaches, datasets, and performance evaluation. Section 3 introduces rPPG-based affective applications and comparison with other affective modalities (e.g., speech and face). At last, we discuss existing challenges and future directions in Section 4.

2. Approaches to rPPG measurement

Despite each approach's specialties, a general framework of prevailing rPPG measurement approaches can be summarized in three steps, including 'video pre-processing', 'rPPG signal recovery', and 'signal post-processing'. As shown in

Figure 3, each step may involve multiple processes and the most frequently employed ones are shown in the figure. Prevailing unsupervised (red) and supervised (blue) rPPG approaches will be introduced subsequently.

2.1 Early-stage unsupervised rPPG approaches

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

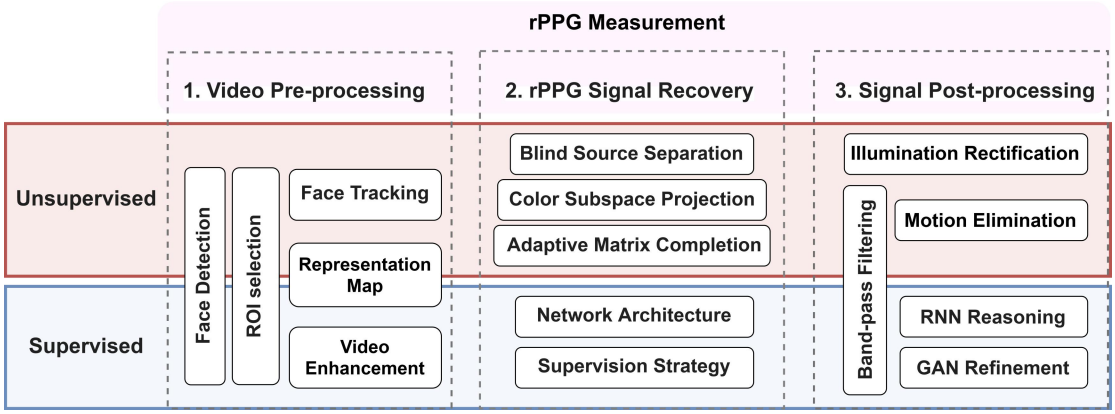
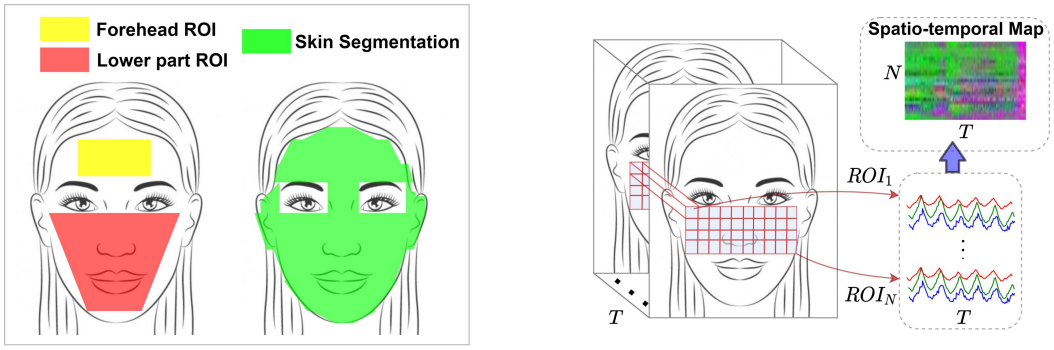


Fig. 3 A general rPPG measurement framework: three steps (column) and two groups (row) of approaches.

Video pre-processing and ROI selection: The density distribution of blood vessels varies in different facial regions, and it is important to select effective ROIs with rich PPG clues. Face detection is usually firstly 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 intra-frame facial ROIs in the spatial domain. One study [2] found the forehead region (see Fig. 4 (a) yellow) works better than other facial regions, as the recovered rPPG signals have a higher signal-to-noise ratio (SNR). However, the forehead might be occluded by hair or a hat. Other studies [3] preferred to use the lower facial regions (see Fig. 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 which are less affected by random noise. Skin segmentation is employed to segment all skin pixels within the face region for rPPG measurement (see Fig. 4 (a) green) according to the color contrast between skin and non-skin parts. Besides defining one single ROI, one approach [5] employed multiple small ROIs divided from a large ROI (see Figure 4(b)) for rPPG measurement. Such local ROI banks could provide more synergic rPPG clues from various facial regions and mitigate the impact of occluded facial regions.



(a) (b)

Fig. 4 (a) Different facial ROIs for rPPG recovery. (b) Spatio-temporal representation of multiple facial ROIs, i.e., the STmap.

On the other hand, to eliminate the noisy fluctuations of the recovered rPPG signal, inter-frame ROI selection along a temporal facial sequence have 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 lead to noisy rPPG signals with high frequency artifacts. A better solution is to use tracking instead of single frame ROI detection. In [3] a pre-defined ROI was tracked through face sequences based on multiple key feature points within

the facial ROI region, and the results showed that tracking is efficient to achieve continuous and a temporally stable ROI for rPPG measurement.

rPPG signal recovery: The most fundamental way to alleviate the effects of hardware noise and achieve raw rPPG signals is to average all 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/multi-channel rPPG signals usually contain mixed sources including the target rPPG signal along with noise fluctuations such as shading caused by motion and lighting variations. To disentangle the pure and intrinsic pulse curve from irrelevant interference is an essential problem for 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). To be specific, independent component analysis (ICA) or principal component analysis (PCA) 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 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 spatio-temporal map (STmap) [5] has been used (see Fig. 4(b) for a typical example) for each frame, on which low-rank based self-adaptive matrix completion [5] was applied to select high-quality ROIs for rPPG estimation, while those impacted by motion or shading were dropped.

Signal post-processing: In the final step, signal post-processing further refines the rPPG quality for subsequent applications. As the frequency of a human heartbeat usually ranges from 0.7 to 4 Hz, a band-pass filter with the corresponding bandwidth [2,3] has often been used to refine the rPPG signal in the frequency domain. Detrending [6] is another popular post-processing method which removes the slow fluctuance of the rPPG signal caused by environmental light variations or auto adjustment of the white balance, for example. Additional post-processing approaches have also been proposed [3] such as illumination rectification via adaptive filtering, and non-rigid 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 have proposed some preliminary solutions (e.g., ROI selection, BSS and illumination rectification) that do not involve supervised learning and which usually come with small computational costs. The approaches have the following limitations: 1) they require empirical knowledge to choose proper parameters for designing the signal processing filters; 2) lack of advanced video processing tools and supervised learning models to counter data variations especially in challenging environments with a lot of

interference.

2.2 Emerging supervised rPPG approaches

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

Video pre-processing for 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 hand-crafted pre-processing. Such learnable mapping from video-level input to one dimensional signal output is flexible but could also easily overfit on small-scale data. STmaps extracted from multiple pre-defined ROIs have recently been adopted as a refined form of input for supervised rPPG frameworks. The 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 pre-processing 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, and we will discuss it from three perspectives, including 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 trade-offs).

From the perspective of **network architecture**, both 2D (spatial) [8,9] and 3D (spatio-temporal) [7,20] models have been explored. A 2D convolutional neural network (CNN) learns spatial rPPG features within each face and aggregate 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, 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. 3D CNN could explore more efficient spatio-temporal contexts for rPPG representation. In [7], a 3D CNN based 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 spatio-temporal modeling, an efficient 2D CNN with a temporal shift module [9] has been proposed for real-time physiological measurement on mobile platforms, which is more practical for real-world deployment.

In terms of **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, i.e., 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 to the ground truth physiological signals, the curve magnitudes (pixel values) are dependent on device settings, the environment and the subject's physical condition. To this end, the NegPearson loss might be a more reasonable option, which also converges faster as demonstrated in [7]. Frequency domain loss assumes that within a short time span (e.g., <10s), 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 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.

As for the **learning strategy**, multiple learning strategies have been explored for rPPG measurement, including multi-task learning [7,9], disentangled learning [10], and meta-learning [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 multi-task 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 (i.e., rPPG and respiration) was also proven to be efficient in a multi-task supervised learning framework. Another strategy is to use disentangled learning to eliminate non-physiological noise (e.g., light variation and sensor noise). In [10], a cross-verified disentangling strategy was developed and tested to distil rPPG features from non-physiological 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 [11] introduced a meta-learning approach, which can adapt and generalize one rPPG model to specific domains.

Supervised signal post-processing: Supervised signal post-processing 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 (GRU)), 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 post-process estimated rPPG signals to reduce noise and improve output quality.

One thing to mention is that, not all supervised methods contain explicit pre-processing and post-processing 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 prior

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.

2.3 **Benchmark datasets and evaluations**

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

Table 1 Public datasets for remote physiological measurement.

Dataset	Year	Subjects	Videos	Physiological signal*	Affective Application
MAHNOB [14]	2012	27	527	ECG, EEG	Emotion recognition
BioVid [15]	2013	90	8700	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	2378	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

* ECG: Electrocardiogram; EEG: Electroencephalogram; EMG: Electromyography; SC: Skin conductance; BVP: Blood volume pulse; BR: Breathing rate; SpO2: Oxygen saturation.

The datasets contain facial videos and corresponding physiological signals as the ground truth for performance evaluation. Concerning the scale of the datasets, the BioVid and VIPL-HR datasets contain a much larger number of samples (thousands) than the other datasets (five hundred or less). Concerning the diversity of the data, most of the datasets were recorded indoor with one fixed scenario setup, while the

VIPL-HR dataset concerns various scenarios with different illuminations and camera setups. From these aspects, these datasets are still not sufficient for training large and deep networks. In terms of video quality, 1) most datasets contain videos that are compressed via modern standards (e.g., MPEG4 and H.264) except UBFC-rPPG and UBFC-Phys, which contain lossless videos without compression; 2) videos of most datasets are with HD resolution (1920x1080), except BioVid (1280x1024) and UBFC-rPPG (640x480). It is worth mentioning that besides color videos, OBF, VIPL-HR, and uulmMAC also provide near infra-red (NIR) videos. As for the ground truth signals, MMSE-HR and UBFC-rPPG only provide one single ground truth signal about the HR, while the other datasets provide multiple physiological signals, including ECG, EEG, EMG, SC, SpO2, and EDA. Some of the datasets 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.

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

Method	Dataset	$SD_{(bpm)}$ ↓	$MAE_{(bpm)}$ ↓	$RMSE_{(bpm)}$ ↓	r ↑
Verkruysse et al. [2]	UBFC-rPPG	-	7.50	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	5.30	8.14	0.76

CVD [10]	VIPL-HR	7.92	5.02	7.97	0.79

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 mean square error (RMSE), the mean error rate percentage (MER), and the 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 satisfactory performance (MAE=1.19 bpm and RMSE=1.8 bpm) on high-quality datasets OBF and UBFC-rPPG, while the performance is yet to be improved on other more challenging datasets (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 HR on OBF and UBFC-rPPG datasets due to 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 VIPL-HR dataset as the latter is sensitive to head movements. It is worth noting that with learnable and adaptive post-processing for coarse rPPG signals refinement, PulseGAN [13] outperforms the other three methods with fixed and straightforward post-processing on UBFC-rPPG dataset. Overall, it is practical to consider STmap-like inputs and learnable post-processing in complex scenarios while improving the robustness of end-to-end supervised methods is urgent. Table 2 concludes the performance of recent approaches, but it is worth mentioning 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 is organized as an annual competition series since 2020 (RePSS 2020: https://competitions.codalab.org/competitions/22287#; RePSS 2021: https://competitions.codalab.org/competitions/30855) which specifically focuses on fair evaluation of the rPPG measurement approaches.

3. Applications in affective computing

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

3.1 Remote physiological signal measurement for affective computing

Emotion understanding is one focus area of physiological signal analysis. Multiple kinds of physiological signals are related to emotional status, including ECG, EMG, SC, PPG, EEG, and so on. ECG and PPG both measure the cardiac activities. For measuring affective status, the average HR alone is not sufficient. The ECG and PPG signals are usually further processed to compute inter-beat-intervals (IBI) and conduct an HRV analysis to achieve more sophisticated features. Common HRV features include low frequency (LF), high frequency (HF), their normalized ratio, and other features in both time and frequency domains.

Using rPPG for emotion understanding is a new rising topic and not many papers have been published so far. These studies build upon traditional ECG- and PPG-based affective computing studies, with an essential extra challenge, i.e., to

reconstruct rPPG signals from facial videos. In consideration of the coarse-designed statistical features, finding proper features from imperfectly measured rPPGs and training a model to measure the target affective status are key. Here we review representative works using rPPG signals for affective computing under different scenarios.

Emotion recognition in human-computer-interaction (HCI) scenarios: HCI is one of the most common scenarios for affective computing studies, e.g., to measure emotions while the subject is watching a movie or playing a video game. Yu et al [20] designed a spatio-temporal network to recover rPPG signals from movie watchers' faces, and then ten dimensional HRV features were extracted for emotion recognition. The study explored emotion recognition in nine categories and also in the valence and arousal dimensions. Beside direct emotion measurement, Gupta et al. [21] also used rPPG to detect the on-set of emotional behaviors. The extracted features from recovered rPPG signal are used for facial micro-expression (ME) 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. One fact is that when people are under cognitive stress, their autonomic nervous system activity changes which can be reflected in some HRV features. McDuff's study showed that even the remote rPPG measurement was not 100% accurate and their model could achieve accuracy of 85% for cognitive stress estimation. The LF component and breathing rate are the most indicative features. A recent study [23] showed peripheral hemodynamics and vasomotion power extracted from rPPG

amplitudes are also important indicators for cognitive stress estimation. One multimodal dataset was established in [19] for stress estimation with video based rPPG modality.

Driver's status monitoring: Driver's status monitoring is one of the focused topics in autonomous driving. Future intelligent driving system should be able to detect the dangerous status of a driver, e.g., fatigue or feeling sleepy, to improve safe driving. 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 drivers' fatigue level.

Pain estimation: Pain is a major research focus as it not only causes physiological discomfort but also impacts people's mental status (e.g., 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, then the pain level was estimated with an SVM or a 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 one major application field for affective computing technologies. By analyzing teachers' and learners' status in educational activities, e.g., 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.

3.2 **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 with the audio and visual modalities due to their wide use in various applications.

Audio modalities: Audio modalities concern an affective analysis from various acoustic inputs, among which one major research area is the 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 key words (or word groups). Automatic speech recognition (ASR) is needed to extract textual features. Culture and language need to be concerned 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 expression, 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] focus on analyzing body behaviors (other than the face) for affect measurement. A human body can be modelled either as 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., what emotions an image can induce when it is shown to a person. Various features were explored for the task, including low-level features such as colors and edges, mid-level features such as materials and eigenfaces, and high-level semantic features such as facial expressions.

Modality comparison: The rPPG is a unique one among all modalities for affective computing. On one side, it is one of the visual modalities (along with facial expression, body gesture, etc.) which may share some common advantages and challenges, e.g.,

related to video quality, lighting, occlusion, etc. On the other side, the rPPG also intersects with physiological modalities and possesses some of their characteristics. Compared with other modalities, the rPPG signals have two main advantages for emotion measurement. The first advantage is inherent to the general physiological domain, that physiological signals might be the most reliable source among all modalities, as it is difficult to intentionally control or alter one's physiological responses. On the other hand, people can control their facial expressions, body gestures and speeches to hide or convey fake emotions if needed. From this perspective, the physiological modality (including rPPG) is 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 only requires 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, comparing to other visual modalities, e.g., facial expressions or body gestures, rPPG signals are weaker signals 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 it is not likely for e.g., SC

and EEG signals so far.

4. Open Challenges and Future Directions

In this article, we introduced facial video-based remote physiological measurement approaches, datasets and applications in affective computing. Despite great progress in recent years, there are noteworthy challenges. 1) The robustness and generalization ability of current methods are limited for practical applications. Video quality, human attributes and behaviors, and environmental changes all influence the accuracy. Approaches do not generalize well to novel datasets due to large data differences in the domain shift. 2) There is 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 dataset's scale. 3) Remote measurement of other physiological signals than the rPPG need to be explored for affective computing.

More effort is needed in future to fill in the gaps. Potential future directions include: 1) to design robust, efficient, and interpretable approaches for rPPG feature representation. Firstly, more informative representations from both time domain and frequency domain could be designed. Secondly, lightweight CNNs could be explored for real-time rPPG applications. 2) It would be useful to learn from limited or unlabelled data. Data augmentation or synthesis methods would be helpful to achieve more data samples. Self-supervised pre-training or semi-supervised methods could be explored to use unlabelled data from the internet. 3) It would be useful to explore multimodality approaches that fuse rPPG signals with other modalities for more

reliable affective measurement.

5. Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant 61772419) and the Academy of Finland (Grant 316765 and 323287).

6. Authors

Zitong Yu (zitong.yu@oulu.fi) and ***Xiaobai Li*** (xiaobai.li@oulu.fi) contribute equally.

Guoying Zhao (guoying.zhao@oulu.fi) is the corresponding author.

Zitong Yu is currently a PhD candidate in the Center for Machine Vision and Signal Analysis, University of Oulu.

Xiaobai Li is currently an assistant professor in the Center for Machine Vision and Signal Analysis, University of Oulu. She is a Member of IEEE.

Guoying Zhao is with School of Information and Technology, Northwest University, PRC, and currently a Professor with the Center for Machine Vision and Signal Analysis, University of Oulu. She is a Senior Member of IEEE and IAPR Fellow.

7. References

[1] Dols, J. M. F., & Russell, J. A. (Eds.). (2017). The science of facial expression. Oxford University Press.

[2] Verkruysse, W., Svaasand, L. O., & Nelson, J. S. (2008). Remote plethysmographic imaging using ambient light. Optics express, vol. 16, no. 26, pp. 21434-21445.

[3] Li, X., Chen, J., Zhao, G., & Pietikainen, M. (2014). Remote heart rate measurement from face videos under realistic situations. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4264-4271).

[4] De Haan, G., & Jeanne, V. (2013). Robust pulse rate from chrominance-based rPPG. IEEE Transactions on Biomedical Engineering, vol. 60, no. 10, pp. 2878-2886.

[5] Tulyakov, S., Alameda-Pineda, X., Ricci, E., Yin, L., Cohn, J. F., & Sebe, N. (2016).

Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions. IEEE conference on computer vision and pattern recognition (pp. 2396-2404).

[6] Poh, M. Z., McDuff, D. J., & Picard, R. W. (2010). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Optics express, vol. 18, no. 10, pp. 10762-10774.

[7] Yu, Z., Peng, W., Li, X., Hong, X., & Zhao, G. (2019). Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement. IEEE International Conference on Computer Vision (pp. 151-160).

[8] Chen, W., & McDuff, D. (2018). Deepphys: Video-based physiological measurement using convolutional attention networks. European Conference on Computer Vision (pp. 349-365).

[9] Liu, X., Fromm, J., Patel, S., & McDuff, D. (2020). Multi-task temporal shift attention networks for on-device contactless vitals measurement. Advances in Neural Information Processing Systems, vol. 33, pp. 19400-19411.

[10] Niu, X., Yu, Z., Han, H., Li, X., Shan, S., & Zhao, G. (2020). Video-based remote physiological measurement via cross-verified feature disentangling. European Conference on Computer Vision (pp. 295-310).

[11] Lee, E., Chen, E., & Lee, C. Y. (2020). Meta-rppg: Remote heart rate estimation using a transductive meta-learner. European Conference on Computer Vision (pp. 392-409).

[12] Niu, X., Shan, S., Han, H., & Chen, X. (2019). Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Transactions on Image Processing, vol. 29, pp. 2409-2423.

[13] Song, R., Chen, H., Cheng, J., Li, C., Liu, Y., & Chen, X. (2021). PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography. IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 5, pp. 1373-1384.

[14] Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. IEEE transactions on affective computing, vol. 3, no. 1, pp. 42-55.

[15] Walter, S., Gruss, S., Ehleiter, H., Tan, J., Traue, H. C., Werner, P., ... & da Silva, G. M. (2013). The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. IEEE international conference on cybernetics (pp.

128-131).

[16] Li, X., Alikhani, I., Shi, J., Seppanen, T., Junttila, J., Majamaa-Voltti, K., ... & Zhao, G. (2018). The OBF database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. IEEE International Conference on Automatic Face & Gesture Recognition (pp. 242-249)..

[17] Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., & Dubois, J. (2019). Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters, vol. 124, pp. 82-90.

[18] Hazer-Rau, D., Meudt, S., Daucher, A., Spohrs, J., Hoffmann, H., Schwenker, F., & Traue, H. C. (2020). The uulmMAC database—A multimodal affective corpus for affective computing in human-computer interaction. Sensors, vol. 20, no. 8, pp. 2308.

[19] Meziatisabour, R., Benezeth, Y., De Oliveira, P., Chappe, J., & Yang, F. (2021). UBFC-Phys: A Multimodal Database For Psychophysiological Studies Of Social Stress. IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2021.3056960.

[20] Yu, Z., Li, X., & Zhao, G. (2019). Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. The British Machine Vision Conference (pp. 277-286).

[21] Gupta, P., Bhowmick, B., & Pal, A. (2018). Exploring the feasibility of face video based instantaneous heart-rate for micro-expression spotting. IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 1316-1323).

[22] McDuff, D., Gontarek, S., & Picard, R. (2014). Remote measurement of cognitive stress via heart rate variability. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 2957-2960).

[23] McDuff, D., Nishidate, I., Nakano, K., Haneishi, H., Aoki, Y., Tanabe, C., ... & Aizu, Y. (2020). Non-contact imaging of peripheral hemodynamics during cognitive and psychological stressors. Scientific Reports, vol. 10, no.1, pp. 1-13.

[24] Tsai, Y. C., Lai, P. W., Huang, P. W., Lin, T. M., & Wu, B. F. (2020). Vision-based instant measurement system for driver fatigue monitoring. IEEE Access, vol. 8, pp. 67342-67353.

[25] Kessler, V., Thiam, P., Amirian, M., & Schwenker, F. (2017). Pain recognition with camera photoplethysmography. International Conference on Image Processing Theory, Tools and

Applications (pp. 1-5).

[26] Monkaresi, H., Bosch, N., Calvo, R. A., & D'Mello, S. K. (2016). Automated detection of engagement using video-based estimation of facial expressions and heart rate. IEEE Transactions on Affective Computing, vol. 8, no.1, pp. 15-28.

[27] Zhang, Z., Cummins, N., & Schuller, B. (2017). Advanced data exploitation in speech analysis: An overview. IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 107-129.

[28] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. IEEE transactions on affective computing, doi: 10.1109/TAFFC.2020.2981446.

[29] Noroozi F, Kaminska D, Corneanu C, et al. Noroozi, F., Kaminska, D., Corneanu, C., Sapinski, T., Escalera, S., & Anbarjafari, G. (2018). Survey on emotional body gesture recognition. IEEE Transactions on Affective Computing, vol. 12, no. 2, pp. 505-523.

[30] Zhao, S., Ding, G., Huang, Q., Chua, T. S., Schuller, B. W., & Keutzer, K. (2018). Affective Image Content Analysis: A Comprehensive Survey. International Joint Conference on Artificial Intelligence (pp. 5534-5541).