Remote Respiration Measurement with RGB Cameras: A Review and Benchmark

GIUSEPPE BOCCIGNONE, University of Milan, Milano, Italy
VITTORIO CUCULO, University of Modena and Reggio Emilia, Modena, Italy
ALESSANDRO D'AMELIO, University of Milan, Milano, Italy
GIULIANO GROSSI, University of Milan, Milano, Italy
RAFFAELLA LANZAROTTI, University of Milan, Milano, Italy
SABRINA PATANIA, University of Milan, Milano, Italy

Remote measurement of respiratory behaviour through RGB cameras has gained significant attention in the last couple of decades. Unlike traditional contact-based methods that may cause discomfort and require specialised equipment, contactless physiological measurement techniques offer a non-invasive way to monitor vital signs. In this survey article, we comprehensively review the literature and techniques related to estimating respiratory information from RGB cameras. We categorise the approaches into three main groups: methods utilising respiration-induced body movements, methods extracting respiratory information from blood volume pulse signals obtained via remote photoplethysmography, and deep learning-based techniques for direct respiratory signal extraction. To evaluate these approaches, we perform a comparative assessment using publicly available datasets. As a result, we uncover emerging trends while identifying strengths and weaknesses in the field. Our contributions include a detailed review of the literature, a benchmark of representative methods on multiple datasets, and the introduction of a new Python package called resPyre that implements the benchmarked approaches, making them accessible to the research community. This survey aims at promoting reproducibility, facilitate further research, and guide the development of more accurate and practical methods for remote respiration measurement using RGB cameras.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Computing methodologies → Computer vision;**

Additional Key Words and Phrases: Physiology, contactless monitoring, signal processing, machine learning

ACM Reference Format:
Giuseppe Boccignone, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, Raffaella Lanzarotti, and Sabrina Patania. 2025. Remote Respiration Measurement with RGB Cameras: A Review and Benchmark. *ACM Comput. Surv.* 58, 5, Article 114 (November 2025), 36 pages. https://doi.org/10.1145/3771763

Authors' Contact Information: Giuseppe Boccignone, University of Milan, Milano, Italy; e-mail: giuseppe.boccignone@unimi.it; Vittorio Cuculo, University of Modena and Reggio Emilia, Modena, Emilia-Romagna, Italy; e-mail: vittorio.cuculo@unimore.it; Alessandro D'Amelio (corresponding author), University of Milan, Milano, Italy; e-mail: alessandro.damelio@unimi.it; Giuliano Grossi, University of Milan, Milano, Italy; e-mail: giuliano.grossi@unimi.it; Raffaella Lanzarotti, University of Milan, Milano, Italy; e-mail: raffaella.lanzarotti@unimi.it; Sabrina Patania, University of Milan, Milano, Italy; e-mail: sabrina.patania@unimi.it.

This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2025 Copyright held by the owner/author(s).
ACM 0360-0300/2025/11-ART114
https://doi.org/10.1145/3771763

1 Introduction

The brain constantly regulates body functions by anticipating its needs and attempting to meet them before they arise (allostasis, [140]). Such relentless endeavour to predict changing conditions, evaluate priorities and prepare the organism to satisfy them involves the production of physiological responses of various nature (e.g., variations in heartbeat and respiratory rhythm).

Indeed, the duet of interoception - the set of processes by which the nervous system takes in, integrates, and makes meaning of sensory signals originating within the body - and allostatic regulation is the bulk of the organism's ability to respond to changes in the environment as to preserve its existence [140, 142]. Thus, it is no surprise that the monitoring of physiological signs is of primary importance to a variety of scopes: clinical purposes, aliveness detection, affective computing, and health assessment.

In general, the measurement of vital signs requires invasive equipment to be placed appropriately on the subject's body. In most cases this fact hinders the adoption of such kind of information due to various reasons: from physical discomfort, to the need of employing dedicated equipment and expertise, up to the actual impossibility of placing the required sensors [33]. Under such circumstances, in the last couple of decades, the scientific community has witnessed the rise of a burgeoning wave of novel approaches to measure vital signs in a contactless way. This has led to the development of techniques enabling the non-contact measurement of many different physiological signals, like **Blood Volume Pulse** (**BVP**) signals, Blood Pressure, **Electro-Dermal Activity** (**EDA**), blood Oxygenation levels (SpO_2), **Respiration Waveforms** (**RW**) and **Respiration Rates** (**RR**). This set of methods has been typically referred to as *Contactless physiological measurement* [108]. The flourish of proposals pledges to supersede the contact-based approaches with a negligible loss of reliability and accuracy, but with sensible gains in both unobtrusiveness and wideness of applicability.

Non-contact physiological measurement methods have hitherto been mainly concerned with the proposal of solutions for the measurement of cardiac and respiratory signals. In particular, contactless cardiac measurement has received considerable attention. This trend is witnessed by the increasing number of studies [37, 60, 108, 120, 137, 160], accompanied by the release of publicly available datasets and code [15, 17, 109, 110].

Yet, remote respiration monitoring is gaining more and more consideration, especially after the spread of COVID-19 pandemic [104]. Indeed, respiratory behaviour is considered a sensitive marker of many physiopathological stressors while being one of the earliest vital signs signalling changes in a patient's clinical status [62]. Two recent review articles [6, 103] show that contactless respiration measurement may be performed through many different instruments: acoustic sensors, thermal cameras, depth cameras, radar sensors, laser vibrometry, stereoscopic cameras, time-of-flight sensors, structured light sensors, Wi-Fi and RGB cameras.

RGB cameras are particularly appealing as they represent one of the most ubiquitous and affordable sensors enabling non-contact breathing monitoring. Yet, to the best of our knowledge, a literature review focusing on the explicit adoption of RGB cameras for the non-contact measurement of respiratory signals is still missing, despite the increasing interest in the topic from both the scientific community and the stakeholders. Indeed, although [6] and [103] effectively survey the broad spectrum of approaches adopting any of the aforementioned sensors, the specific literature related to RGB cameras is barely touched.

The chief concern of this work is to carefully review the RGB-camera approach and the possible applications which may result from such technology. Further, we perform a comparative assessment of distinct groups of techniques, by evaluating them on publicly available datasets in order to highlight their strengths and weaknesses. Our motivation is to promote the reproducibility and expansion of this evaluation. In brief, this survey aims at the following contributions:

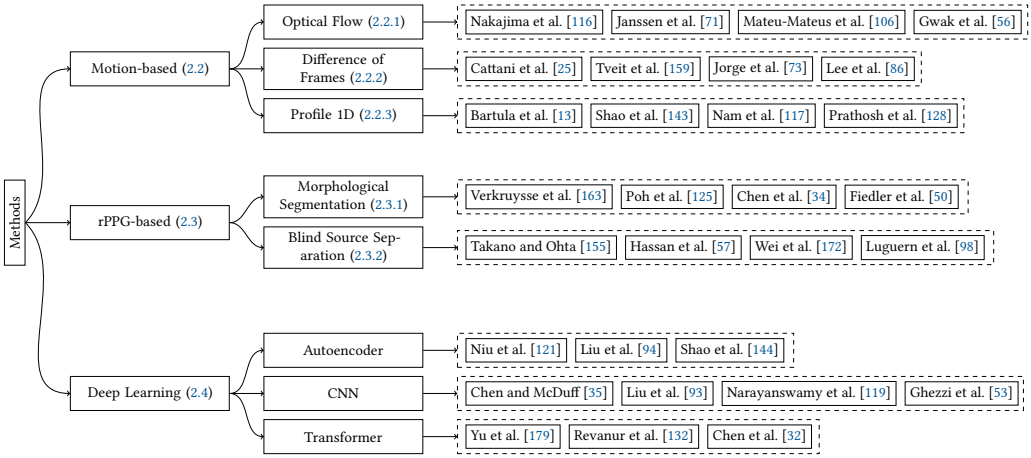


Fig. 1. Overview of the overall organisation of the reviewed literature based on the proposed taxonomy: Motion-based, rPPG-based, and Deep Learning-based approaches. Each group is further divided into appropriate categories based on the methods proposed in the literature. The rightmost part of the figure lists a subset of representative works for each category.

(1) A review of the literature and techniques related to estimation of respiratory information from RGB cameras.
(2) A benchmark of the most representative approaches on 3 publicly available datasets
(3) A novel Python suite, called resPyre [1] implementing the benchmarked approaches, made freely available to the scientific community.

The manuscript unfolds as follows: Section 2 presents the different approaches that can be employed to estimate respiratory information from RGB videos. These can be collected into three main groups, as shown in Figure 1: *(i)* methods that exploit the respiration-induced body movements of subjects for the estimation of RW or RR. *(ii)* Methods that extract respiratory information from BVP signals estimated via **remote photoplethysmography** (**rPPG**) algorithms. *(iii)* Deep Learning-based methods that are either trained end-to-end to extract respiratory signals or employed at some intermediate step. Section 3 presents an in-depth experimental analysis and benchmarking of some of the most representative techniques that have been presented in the literature. Eventually, in Section 4, we discuss the results and draw some concluding remarks.

2 Overview of the Approaches for RR Measurement

2.1 A Data-driven Exploration of the Current Landscape

Historically, respiratory information estimation methods using RGB cameras can be broadly categorised into two distinct branches: those exploiting chest/shoulder movements (*Motion-based* approaches); those related to the extraction of modulations of the cardiac activity eventually estimated via *rPPG-based* approaches.

Earliest attempts of respiratory activity measurement from RGB cameras explicitly relied on the adoption of either one of the two approaches or their fusion. In the last decade, though, the rise of Deep Learning models in computer vision has started to gather momentum in the field of camera-based physiological estimation, respiratory measurement making no exception. In spite of the fact that the physiological and mechanical principles at the foundation of these modern

[1] https://github.com/phuselab/resPyre

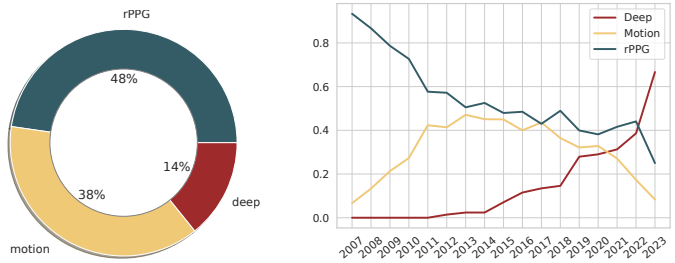


Fig. 2. (a) Percentage of articles belonging to each of the considered categories. (b) Evolution of the percentages over time (3-years moving average).

approaches are the same, their black box nature prevents a sharp categorisation of these methods in either one of the two classes mentioned above. Thus, in this work methods exploiting deep-learning techniques will be aggregated in a third category (*Deep Learning-based* approaches).

Results are summarised in Figure 2(a). The rPPG and motion-based approaches dominate the literature if compared to the Deep Learning-based methods. However, upon scrutinising the temporal evolution of the proportions associated with the three categories, an intriguing yet unsurprising trend emerges. In recent years, there has been a noticeable shift in the field of RGB camera-based respiration measurement. While the initial approaches primarily encompassed rPPG or motion-based methodologies, the past decade has witnessed a significant rise in the adoption of Deep Learning-based techniques, demonstrating a rapid and exponential evolution (see Figure 2(b)).

2.2 Motion-based Approaches

The mechanical activity of the lungs during respiration is reflected on different parts of the body of the subject, thus producing small and periodic spatio-temporal fluctuations in the pixel intensity values recorded by a camera sensor. This evidence has given rise to several motion-based remote respiration measurements that infer the respiration rate on the basis of the movement of some ROI, such as the head, the chest, and the thorax, up to automatically detected ROIs.

Besides the distinction based on the ROI, the motion-based RR approaches can be categorised according to the adopted method of motion estimation. Notably, in [170] a mathematical model describing the algorithmic principles of camera-based respiratory motion extraction has been proposed, together with a fully controllable setup employing a physical phantom apt for investigating different algorithms. In a nutshell, according to [170] the model of respiratory motion extraction from a camera can be defined as follows:

Denote by $F_t(x, y)$ the light intensity of the video frame occurring at time t at the spatial coordinates (x, y). For short intervals, assuming a spatially uniform ambient light , i.e., $I_t(x, y) = I_0$, the intensity of the light contaminated by the component of the respiratory motion hitting the camera sensor can be approximated as follows:

$$F_t(x, y) \approx I_0 P_0 \left(1 + P_m \left(x - v_x t, y - v_y t\right)\right) + N_t(x, y), \tag{1}$$

where the constant P_0 is the DC component of the signal and P_m is a steady and modulating zero-mean shifting pattern, with v_x and v_y velocities along x and y directions, for "small" elements of the chest area captured by the camera. $N_t(x, y)$ is an uncorrelated zero-mean noise accounting

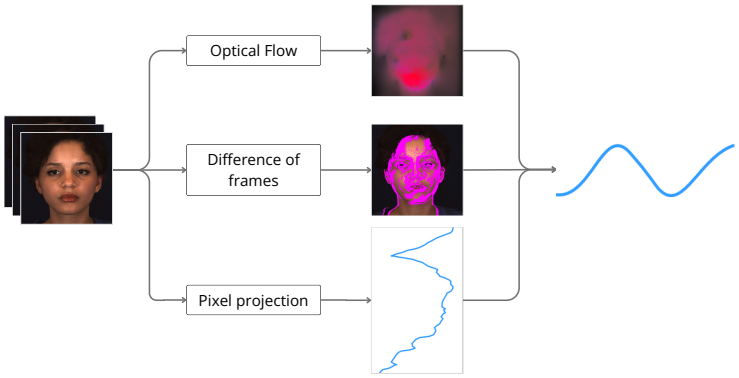


Fig. 3. Motion-based approaches for RGB camera-based respiration estimation at a glance. Motion is estimated from couple of cropped frames containing the desired ROI, using 3 representative approaches: Optical Flow, DoF, Vertical pixels' colour projection cross-correlation. After post-processing, the respiratory waveform is eventually recovered (rightmost part of the figure).

for unequal pixel offsets and signal quantisation. Notably, $P_m\left(x - v_x t, y - v_y t\right)$ speaks for the zero-mean modulating part arising from respiratory induced movements.

Motion-based approach for respiratory information estimation practically implements different strategies to estimate the movement pattern represented by v_x and v_y in P_m.

In the literature, the approaches that deliver such estimates can be split into three main categories: Optical-Flow/Motion tracking based methods, **Difference of Frames** (**DoF**) based methods, RGB/Intensity based methods (see Figure 3). For completeness' sake, a fourth category should be mentioned, which includes approaches that track a specific pattern (e.g., a chessboard) placed on the subject's thorax. Clearly, such approaches suffer from scarce applicability in general. Hence, albeit reviewed here for the sake of completeness, they will not be further considered.

2.2.1 Optical-Flow/Motion Tracking Based Methods. **Optical flow** (**OF**) or motion tracking techniques have been widely exploited as a general strategy to analyse the movement patterns induced by respiration. One of the earliest methods for the estimation of RW has been proposed in [116]. Authors devised an approach for the estimation of both posture changes and the subject's respiration in bed by evaluating chest or blanket movements via OF analysis. RW is obtained by simply averaging OF velocity vectors over a fixed ROI. In the same vein, in [100] RW is obtained as a weighted average of the motion velocities computed via the Lucas-Kanade method. The weights are determined based on the **signal-to-noise ratio** (**SNR**) estimated from the temporal properties of the velocity vector. As a result, pixel blocks exhibiting higher SNR values are selected as those carrying more respiratory information.

The authors of [31] present a solution for estimating the RR by considering two challenges related to respiration. Firstly, respiration generates subtle motion that is captured by a subset of pixels, predominantly along edges. Secondly, different points on the thorax move in distinct directions, resulting in the cancellation of contributions if simply summed. To address these issues, their solution combines OF with image gradient information to reduce the influence of pixels in uniform regions. Additionally, they train a **principal flow field** (**PFF**) on an initial 15-second segment of the video to learn the motion direction. Subsequently, the OF obtained for the rest of the video is projected onto the PFF, effectively solving the problem of summing contributions from vectors with different directions. Finally, RR is determined by identifying the dominant frequency within

small time intervals for each collected signal. Their method does not require the explicit definition of a ROI as long as the video framing captures the chest of the person being monitored.

It is worth noting that several contributions in this field utilise **Eulerian Video Magnification** (**EVM**) [174] to enhance the motion changes in the video. This is the case for [12, 52, 80] where the ROIs to be tracked are first EVM-magnified. The earliest of these three articles [80] requires to draw the ROI manually whilst the other two solutions detect the ROIs automatically.

The distinctive feature proposed in [70] is the adoption of the slow-motion mode built-in smartphone cameras for capturing body movement. This allows them to have 240 fps useful to capture even very small movements. Apart from this, the method is straightforward, tracking several ROIs and applying a standard frequency analysis, requiring the persons to be still and lying on their side.

The classic **Horn and Shunk** (**HS**) OF algorithm followed by the extraction of the y-axis velocity component has been adopted in [90, 115, 135]. A single velocity vector is then obtained either by averaging [115, 135] or computing the median [90] of the vertical velocities within a ROI located over the chest.

In [106] RR is estimated by tracking the motion of the intercostal and abdominal muscles by the means of **Dense Inverse Search** (**DIS**) OF method [81]; a specific setting is adopted in which subjects are filmed from a lateral perspective while seated. The respiratory signal is recovered from the phase of the OF, while extracting at the same time a quality index from its modulus. DIS-OF is adopted also in [44] for estimating motion on patient's own and forced breathing during mechanical ventilation in intensive care unit. Displacement OF vectors computed on three manually localised ROIs in zones of the lungs and abdomen are averaged and then filtered, thus obtaining the respiration signal. In [71], RR is conducted by employing vertical-OF feature and subsequently processed using motion factorisation techniques (SVD or PCA). These techniques are based on the observation that the motion related to chest/abdomen movement caused by respiration can be treated as independent motion within a video. ROIs are automatically detected by leveraging inherent characteristics of respiration. Respiratory signal is recovered via integration of the flow vectors. The same method was later employed by [95] for RR and cessation of breathing detection.

In [56] motion along the y-axis is derived by applying the **Kanade–Lucas–Tomasi** (**LKT**) algorithm on face and chest landmarks. Moreover, to limit the method's vulnerability to body movement, a motion artefact removal step is applied adopting the kurtosis to remove sudden motion artefacts. RR is calculated via spectral analysis of the cleaned signal. Table 1 summarises the solutions relying on respiratory motion estimation at a glance. Approaches are grouped based on the specific motion estimation method adopted, and the considered region of interest.

2.2.2 Difference of Frames (DoF) Based Methods. To detect the motion induced by respiration, one possible approach involves the simple computation of the DoF, as the fluctuation of this feature effectively describes the respiration waveform.

In [156] an early solution was proposed computing the subtraction of consecutive frames, so to detect chest movement. Specifically, once computed the DoF at time t, the latter is binarised and the pixel equal to 1 is summed thus capturing the global movement at that time. The temporal series obtained by applying this process to the whole video is then smoothed via an average filter, and the respiratory rate is determined by calculating the gradient of the respiratory wave so far derived. Another very simple solution is proposed in [122] where RR is estimated from respiratory movements while sleeping. To this end, a suitably developed bed cover exhibiting specific patterns is employed. Motion is still estimated via the difference of frames.

Lee et al. [86] propose a more robust solution that can adaptively detect motion in subjects without any positional constraint. This is achieved by utilising the **Persistent Luminous Impression**

Table 1. Motion-based Methods

Article	Method	ROI	Additional notes
[100, 116]	OF	Whole Frame (person in bed)	Output: RW
[31]	OF	Whole Frame (Thorax)	Processing includes: PFF
[80]	OF	Manual Selection of the Chest	Processing includes: ICA, PCA, EVM
[12]	Vertical-OF	Chest	Processing includes: EVM
[52]	Tracking	Detected Chest	Processing includes: PCA, EVM
[70]	Tracking	Abdomen/Shoulders (manual)	Video acquired in Slow-Motion mode
[115, 135]	Vertical-OF	Chest	-
[90]	Vertical-OF	Upper body	Processing includes: Median, Zero Crossing
[106]	Phase-OF	Side of the thorax	Processing includes: Quality Score
[44]	OF	Manual Selection of Abdomen	-
[71]	Vertical-OF	Chest	Processing includes: PCA
[56]	Tracking	Face and Chest landmarks	Processing includes: Kurtosis
[173]	Tracking	Markers on the Abdomen	-
[107]	Tracking	Markers on the Abdomen	Output: RW
[156]	DoF	Whole Frame (Abdomen)	-
[122]	DoF	Whole Frame (person in bed)	Acquisition: employing a patterned bed cover
[86]	DoF	Whole Frame (person in bed)	Processing includes: PLIM
[73]	DoF	Skin segmented (infant in bed)	Processing includes: skin detection
[159]	DoF	Manual selection of Abdomen	Processing includes: Riesz Transform
[25]	DoF	Whole frame (infant in bed)	Processing includes: EVM, ML; Output: RW
[26]	DoF	Whole frame (infant in bed)	Processing includes: EVM, ML; Output: RW
[7]	DoF-EVM	Chest	Processing includes: ML
[8]	DoF	Chest	Processing includes: ML
[9]	DoF	Chest	Processing includes: phase-based-EVM, ML
[2]	DoF	Manual selection of Chest	Processing includes: WPD-EVM
[136]	DoF	Whole frame (infant in bed)	Processing includes: PFF, EVM
[13]	Projection	Whole frame (person in bed)	-
[21]	Projection	Whole frame (upper body)	Processing includes: ML
[143]	Image Gradient	Manual selection of Shoulders	Processing includes: Tracking
[134]	Intensity	Manual selection of Chest	-
[145]	Intensity	Manual selection of Chest	Processing includes: EVM
[102, 105]	Intensity	Manual selection of Chest	-
[146]	Intensity	Chest	Processing includes: EVM
[117]	Spectral Analysis	Abdomen (manual)	Acquisition: front and rear smartphone cameras
[128]	Blind Decomp.	Chest	Output: RW
[135]	Intensity	Chest	Processing includes: PCA

OF: optical flow; RW: Respiration Waveform; PFF: Principal Flow Field; ICA: Independent Component Analysis; PCA: Principal Component Analysis; EVM: Eulerian Video Magnification; DoF: difference of frames; PLIM: Persistent Luminous Impression Model; ML: Maximum-likelihood; WPD: Wavelet Pyramid Decomposition.

Model (PLIM) previously introduced in [165]. PLIM implements background subtraction, so that the slow adaptation of the reference frame compensates for motion unrelated to respiration. In [73] a method is proposed for automatic respiration monitoring of neonatal infants. The method consists of an automatic ROI selection scheme based on a skin/non-skin classification procedure and a motion analysis step. The latter accomplishes two tasks: detection of spurious movements (e.g., nursing interventions), and detection of subtle movements of the thorax and the abdomen.

In [159] respiration is detected on the basis of the variations in the local phase, computed using an efficient approximation of the Riesz transform. Movement at time t is characterised by summing up the local phases weighted by the local amplitude, and differentiated with respect to the analogous feature obtained on the Reference frame (Difference of Phases, DoP). Eventually, RR is obtained by windowing the obtained signal, and by selecting the largest frequency component of the Fourier transformed signal.

Also in this category, several works exploit the EVM capability to magnify intensity video frames, facilitating the subsequent task of the DoF. For example, in [25] DoF on EVM magnified intensity video frames is utilised. Aiming at detecting potential apnea episodes in the neonatal population, the **Maximum Likelihood** (**ML**) approach is then adopted to determine the presence/absence of periodic motion. The same authors propose an extension of their work to support multi-cameras (or depth-sensors) [26]. This way the system becomes more robust by exploiting the correspondence of the periodic components estimated by each signal. Again, in [7] the same authors propose a variant to obtain the effect of DoF directly with the EVM magnification algorithm (DoF-EVM). First, they integrate the motion signal extraction with the EVM spatio-temporal processing, designing a task driven band-pass filter so to include normal periodicity movements (cut-off frequencies adopted: 0.25 and 1.05 Hz). Second, they do not reconstruct the magnified video while separately considering the filtered and amplified levels. Each level is binarised and then the average value computed for each frame and each level. The final RR is estimated by jointly analysing the obtained temporal signal per level, using the Maximum Likelihood criterion.

In [8] and [9] Alinovi et al. make their approach proposed in [7] more robust by tackling the problem of sensitivity to large body movements. The proposed solution operates at two levels: global exclusion of frames that are significantly affected by large changes, and local identification of ROIs that are affected by local movements. Specifically, in [8] no video magnification is adopted, whilst in [9] a phase-based motion magnification is applied.

In [2] the chest motion of newborns is detected by computing the DoF of magnified video frames, which is obtained via **Wavelet Pyramid Decomposition** (**WPD**) and an elliptic band pass filter. This demonstrates smaller errors in RR estimation than using traditional Laplacian decomposition.

The last article we consider that adopts EVM is [136]. Here the goal is the measurement of the RR of preterm infants placed in open cribs, and monitored via an RGB camera. Once applied the EVM, the motion gradient is obtained, followed by noise reduction steps via the PFF method.

2.2.3 RGB/Intensity Based Methods. The process of respiration leads to observable alterations in appearance which, if accurately characterised, can enable the modelling of respiration dynamics.

An efficient way to characterise the visual variations induced by the respiration movement is the computation of the vertical colour projection, as done in [13] and in [21]. In [13] the integration is extended to the whole frame, while in [21] multi sub-blocks are considered and the ML criterion is adopted to select the block containing valid respiratory motion.

Alternatively, in [143] breathing patterns are estimated calculating the image vertical derivative within a small ROI located on the subjects' shoulder. The shoulder's edge is then used to divide the ROI into two sections (above and below the border line) so that the vertical movement associated with breathing is estimated by comparing the displacement of the two sections. In addition, motion tracking is applied to compensate the large body movements. Even simpler, the bare fluctuation of the mean colour/intensity over a ROI has been pointed out in [134, 145] as a possible strategy to capture the respiration waveform. In [102, 105] intensity changes in correspondence to the upper chest are monitored for extracting the respiratory rate. The novelty here lies in the fact that, among all the lines within the considered ROI, only the five with the highest standard deviations are addressed for the RR estimate. In [146] EVM is employed to amplify the motion changes in the video before extracting and integrating the intensity values within the ROI. In [117] the spectral analysis of the intensity values in the chest and abdominal ROIs is carried out for the RR estimation. The most reliable ROI is identified by selecting the signal with the greatest absolute value of the mean auto-correlation. The method also captures the fingertip with the smartphone's rear camera for the **Heart Rate** (**HR**) estimation.

More robust solutions have been proposed, addressing the distortion caused by the presence of noise in the colour/intensity signal. In [128] each pixel is considered as a noisy realisation of a **linear time-invariant** (**LTI**) system composed of multiple channels connected in parallel with unknown dynamics, but driven by the same process, namely the respiratory signal. RW is hence recovered by solving a blind deconvolution problem. In [135] the intensity signals of pixels belonging to the subject's chest are extracted and combined via two different approaches, namely PCA and 5% method. The PCA selects the signals that constitute 95% of the variance explained, whereas the 5% method selects 5% of the signals with the highest standard deviation. RR is determined in the time domain on the basis of the peak-to-peak distances.

2.2.4 Pattern Matching Based Methods. Pattern matching methods are those relying on the detection and tracking of the movement of predetermined patterns that are discernible in the recorded video. The advantages are clearly related to the surmised higher reliability of algorithms that are specifically tailored for the detection of a given pattern; yet such methods are much less serviceable in more general and unconstrained settings.

Wiesner et al. [173] proposed one of the earliest approaches in this direction, estimating RR by measuring abdominal motion through the tracking of plastic cubes with coloured faces placed on the subject's abdomen.

More recently, in [107] RW is derived by the detection and tracking of a custom pattern on the thorax of the subjects. Such pattern is detected using standard image processing methods and then tracked via the LKT method.

2.2.5 Summary. This paragraph has introduced and reviewed the literature on motion-based approaches for respiratory information estimation. The reviewed techniques aim at exploiting the subtle motion induced by respiration in various body regions. Different sub-categories have been identified, primarily differing in their chosen motion estimation method.

Advantages: Motion-based techniques can be adapted effortlessly to various **regions of interest** (**ROIs**), such as the head, chest, and thorax, or automatically detected ROIs. Notably, no strict video specifications must be ensured: video compression, lower resolution, or low frame rates are acceptable as long as reasonable quality is maintained. Specifically, many of the reviewed articles employed lower-resolution video frames to speed up computations. Similarly, frame rate requirements are extremely lax; as respiratory cycles occur at very low frequencies (0.1 Hz to 1 Hz), the Nyquist theorem dictates that sampling frequencies should be above 2 Hz–far below typical video frame rates.

Limitations: Some motion-based methods, especially those that rely on deep learning for computing optical flow, can be computationally intensive, requiring substantial processing resources. Moreover, all challenges related to motion estimation (e.g., the aperture problem, occlusion, and rigid motion assumptions) must be taken into account. Most importantly, when selecting a ROI, uniform or textureless regions should be avoided, as they offer minimal useful information for motion estimation.

2.3 Remote-PPG-based Approaches

Photoplethysmography (**PPG**) is an opto-electronic technology usually employed for the measurement of cardiac activity and Heart Rate [59]. It basically employs a sensor able to capture the reflected light skin variations due to the blood volume changes. rPPG is the name usually attributed to the contactless technique able to measure reflected light skin variations by using an RGB-video camera as a virtual sensor. Notably, beyond contactless measurement of heart rates, rPPG signals can be employed for a variety of downstream tasks such as the assessment of health related conditions

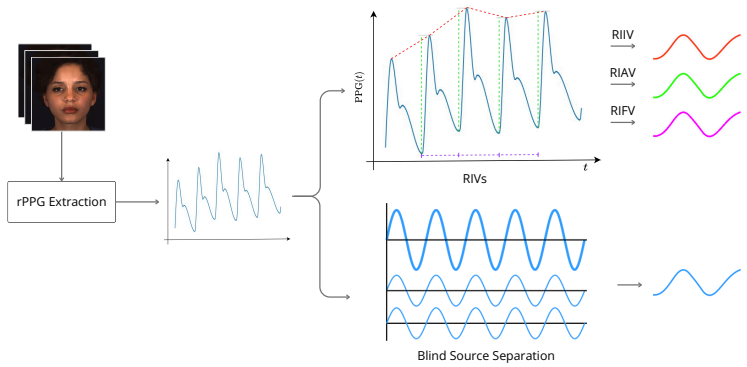


Fig. 4. Remote-PPG-based methods. The upper part of the figure depicts the derivation of the respiratory induced variations capturing the amplitude (RIAV), the frequency (RIFV) and the intensity (RIIV) variations from the segmentation of the PPG waveform. The lower part of the figure depicts the respiratory waveform as recovered by BSS.

(e.g., stress [185] or depression [23]), to reveal DeepFakes in videos (e.g., [14, 38, 41]) or for the estimation of other physiological signals [18, 65].

Indeed, due to the tight bond between cardiac and respiratory activities, signals coming from PPG waveforms (and to some extent rPPG waveforms, too) may be employed to extract respiratory related information, the so called **Respiratory Induced Variations** (**RIVs**). According to the related literature, there exists 3 different kinds of RIVs: [18, 58, 113]:

(1) **Respiratory-Induced Intensity Variation** (**RIIV**): It is an amplitude modulation of the BVP signal originating from changes in venous return due to changes in intra-thoracic pressure throughout the respiratory cycle. This causes a baseline modulation of the PPG signal. During inspiration, intra-thoracic pressure decreases, thus producing a small decrease in central venous pressure increasing venous return. The opposite happens during expiration.
(2) **Respiratory-Induced Amplitude Variation** (**RIAV**): Inspiration causes left ventricular stroke volume decrease due to changes in intra-thoracic pressure. This produces a decrease in cardiac output and hence in peripheral pulse strength. The opposite occurs during expiration.
(3) **Respiratory-Induced Frequency Variation** (**RIFV**): It is a variation of HR in the course of the respiratory cycle; HR increases during inspiration and decreases during expiration. This produces a periodic change in pulse rate caused by an autonomic nervous system response. Such phenomenon is known as **respiratory sinus arrhythmia** (**RSA**).

The literature concerning the estimation of respiratory information from rPPG can be broadly divided into two categories: methods that extract RIVs after a morphological segmentation of the estimated rPPG signal, and methods aimed at separating the cardiac vs. respiratory signals (the last implicitly connected to RIVs) via **Blind Source Separation** (**BSS**) techniques (see Figure 4).

2.3.1 Methods Relying on rPPG Morphological Segmentation. An early method proving the effectiveness of RIVs for RR estimation is [163], where Verkruysse *et al.* extracts the baseline modulation (RIIV) from the rPPG signal estimated by selecting the G or B channels from the RGB traces; the latter are obtained via a simple forehead ROI pixels intensity average. Similarly, in [162] authors extract the baseline wandering (RIIV) from rPPG signals estimated via the CHROM [42] and PBV [43] methods. The weights defining the CHROM and PBV projections from RGB traces to rPPG estimate are computed for the filtered traces of normalised pixel differences that include the range of pulsatile frequencies, and then applied to the differently filtered pixel differences containing

only respiratory frequencies. Conversely, in [125] authors retrieve the RIFV-related respiration rate. To this end, the rPPG signal is estimated by the ICA method [126]. **Inter-Beat-Intervals** (**IBI**) are computed by detecting the peaks in the estimated BVP signal, thus obtaining the **Heart Rate Variability** (**HRV**) signal. HRV is analysed via Power Spectral Density estimation using the Lomb periodogram. The **high frequency** (**HF**) powers only ($[0.15 - 0.4]Hz$) are considered as reflecting parasympathetic influence on the heart, which is connected to RSA. The RR is hence estimated by selecting the HF peak in the HRV PSD derived via rPPG. Similarly, in [34] authors compute the HRV from the IBIs of the BVP signal, which is estimated via the POS [171] rPPG method. HRV samples with values exceeding three times the standard deviation of the HRV signal are considered outliers and are therefore removed. RR is calculated by selecting the frequency with the maximal energy peak within the range of 5 to 30 breaths per minute relying on the HRV's PSD estimate yielded by the Lomb-Scargle periodogram. In the same vein, in [114] the instantaneous respiratory rate was estimated from the rPPG IBIs, by exploiting a bank of short FIR notch filters. In [19] both pulse and breathing rates are recovered by analysing the periodicity of the face skin colour. After having detected the face and the skin pixels, the methods consider the u^* plane in the $CIEL^*u^*v^*$ colour space; the u^* plane suitably covers the $[540 - 577]nm$ wavelength band that best accounts for the oxy-deoxyemoglobin fluctuation. The 1D signal obtained by averaging the $u*$ value corresponding to all the skin pixels in a frame is then converted via the **Continuous Wavelet Transform** (**CWT**) within the $[0.15 - 0.4]Hz$ band of frequency. The RIFV is obtained by detecting the respiratory cycles and measuring the intercycle intervals.

Actually, RIVs information can be further processed to derive amplitude and frequency modulations as reviewed and benchmarked by Charlton *et al.* [30] as to the extraction of respiratory information from contact PPG signals, and by [99] for rPPG signals. Specifically, in [99] the rPPG signal is obtained by adopting four rPPG methods, namely CHROM [42], PBV [43], PVM [101], and Energy Variance Maximisation [97]. Then, the **Incremental Merge Segmentation** (**IMS**) algorithm is applied, and the resulting signals are processed using 12 algorithms to extract a comprehensive set of features characterising the amplitude and frequency of the rPPG. RR is eventually estimated by using different methods, either frequency-domain or time-domain based. The RRs estimated from different features are finally merged by exploiting different fusion techniques. Results show that RIIV and RIFV are the features that deliver the best RR estimates; fusion of multiple RIVs yields more robust estimates. Energy Variance Maximisation and CHROM turned out to be the best methods to estimate rPPG in the service of the extraction of respiratory rates.

Several other works investigated the opportunity of extracting different features and merging them to obtain a more reliable estimate. Notably, in [50], the authors proposed a method (FuseMod) for the extraction of respiratory information from rPPG computed either referring to the whole face or the forehead only, both automatically detected. The method relies on 7 different signals, obtained from the RIVs, to characterise the amplitude and frequency modulations of the rPPG signals. The RW signals are then opportunely post-processed via normalisation, artefact removal and signal differentiation procedures, and fused using different heuristics (simple mean, α-trimmed mean, median). The effectiveness of different rPPG methods (GREEN [163], CHROM [42], Hue [138], normG [150]) is investigated, too. Results show that RR can be more effectively estimated by employing CHROM as rPPG method and by using the median of the 7 RIVs as the fusion strategy. Importantly, experiments are conducted on a publicly available dataset (BP4D+ [182]).

A further feature investigated is the RR estimation from the image of the fingertip placed on the lens of a mobile phone, notably affected by artefacts. In [75] two approaches are proposed. The first approach involves utilising a smart fusion algorithm (previously proposed for PPG in [76]) to detect artefacts and merge the RIVs components obtained after applying the IMS technique to the rPPG signal. The second concerns the adoption of the EMD that decomposes the rPPG signal

in **Intrinsic Mode Functions** (**IMF**); the RR is identified through the frequency peak with the highest power among all IMF.

Alternatively, in [84] RR is obtained by fusing the estimates derived from the RIVs, three *pulse-to-pulse* respiration signals (the pulse width variability, PWV, the pulse amplitude variability, PAV, and the pulse rate variability, PRV), and the respiration-induced amplitude and frequency modulations (AM and FM) extracted from a time-frequency spectrum; the latter is obtained by exploiting the **variable frequency complex demodulation method** (**VFCDM**) [167]. Again, the extraction of amplitude and frequency modulations from the rPPG signal has been pursued by [118] and [141]. Nam *et al.* [118] presents a method for RR estimation using the rPPG signal. An **autoregressive** (**AR**) model, VFCDM, and the CWT were used to identify frequency and amplitude modulations containing respiratory-related information. According to the obtained results, the VFCDM method provided the most accurate results.

2.3.2 Methods Relying on Blind Source Separation of rPPG. Within this subgroup, we encounter techniques that do not focus directly on characterising the RIVs, but rather recover it from the rPPG signal via Blind Source Separation.

In [57], multiple patches are detected on the subject's face. The GREEN rPPG method is applied and RR is estimated after performing a denoising procedure (Wavelet-based Multivariate Denoising). According to the authors, this approach enables the detection of the effects of respiratory motion, which directly modulates the baseline of the PPG signal. The frequency determined in the range between 9 to 40 respirations per minute with the highest number of votes among patches yields the predicted RR. In [127, 138] a very simple approach is pursued. After performing face and landmark detection, a ROI on the subject's forehead is selected. RGB pixels belonging to the ROI are converted to the HSV colour space. The average Hue intensity trace in the ROI is considered as the rPPG estimate. Respiration rate is recovered by selecting the peak in the spectrum of the estimated rPPG in the $[0.18 - 0.5]Hz$ range.

Conversely, in [155] RR is recovered from the rPPG signal obtained by measuring brightness changes in a ROI. More precisely, authors perform an AR spectral estimation on the derivative of the average brightness values measured on the cheeks of participants.

In [39], once the ROI corresponding to the newborn's chest is manually localised, the average RGB color for each frame is computed, and the values are stacked into 6-second-long windows. A pre-filtering step is applied to reduce the signal offset: the linear function that best fits the signal is subtracted from the original signal, thereby reducing low-frequency components. A narrow bandpass filter $[0.50 - 1]Hz$ is applied, and the RR is found via peak detection in the spectrum of the obtained signals. In [10, 11], video of a naked newborn is acquired under light sources that ensure uniform illumination. The ROI of interest, located on the thorax, is manually selected and magnified using the EVM. The average RGB color of the ROI is computed for each frame, summarised into a single signal, and used for peak extraction from the PSD.

In [97, 98] the goal is to define a novel rPPG method by exploiting the combination of RGB traces that maximise the respiratory information over the cardiac one. In [97] an algorithm called Energy Variance Maximisation is proposed in order to effectively combine the RGB traces to obtain a signal that maximises the SNR of the output trace, while in [98] the continuous wavelet transform is employed in order to deal with non-stationarities of the respiratory signal (Wavelet Variance Maximisation algorithm).

In [172] the respiratory motion artefact is recovered via BSS. Specifically, two regions of interest are selected on subjects' faces; the 6 channels obtained by stacking the RGB average color intensity variation traces belonging to the ROIs are fed as input to the **Second-Order Blind Identification** (**SOBI**) BSS algorithm. The separated output channels are surmised to contain both the BVP signal

and the respiratory motion artefact. A kurtosis-based identification strategy is then employed for the automatic selection of the Respiration Waveform (and BVP).

In [157], RR is estimated by exploiting the low-frequency amplitude variations of the rPPG signal. To this end, the breathing-induced amplitude variations are extracted using a band-pass filter with an upper cut-off frequency of $0.7Hz$. Subsequently, an AR model is applied to identify the RR.

In [112] authors propose a method for the estimation of RR by tracking the nostrils' movements in facial videos. To this end, the rPPG POS method is applied on two ROIs. One around the nostrils, the other on the forehead. The two spectra derived from the corresponding BVP estimates are subtracted. The authors claim that this procedure isolates the nostril movement related to respiration from BVP-related information and noise.

It should be noted that certain contributions have investigated the utilisation of both rPPG-based and motion-based methods or, at least, they have proposed a direct comparison.

In [3] and similarly in [5], the authors propose and compare the rPPG-based solution (based on the GREEN method [163]) with a motion-based one (by computing the vertical motion on a feature point in the forehead). In both cases a face detector and EVM are applied following [4]. Denoising is obtained by using the **complete ensemble empirical mode decomposition with adaptive noise** (**CEEDAM**) method and the **canonical correlation analysis** (**CCA**). Interestingly enough, the solution is suitable for monitoring multiple subjects simultaneously (up to 6). Experiments show that the rPPG-method obtains better performances than the motion-based one.

In [139] another comparison is put in place, by processing videos of head and upper part of the chest in three ways: deriving the rPPG from the face (based on the GREEN method [163] coupled with EMD and IMF analysis), the head and the chest motion (detecting the vertical-OF). This study produced the best results referring to the chest motion, followed by the rPPG method.

To fully leverage the potential of multiple features, methods that involve their fusion have been proposed. For example, in [69] authors estimate respiration waveforms by exploiting both RIVs and head movements. Specifically, PWV, RIFV and RIAV are derived from the estimated rPPG signal [83] over M different ROIs selecting the candidate rPPG trace that maximises the SNR among the $3 \times M$ traces. A fourth RW is obtained by considering the signal recovered from the face movement on the y-axis as returned by the KLT tracking algorithm. The fusion of RIVs and motion-related respiratory information is obtained in the shape of a weighted combination of their spectra.

Alternatively, in [164] authors propose a method for the camera-based estimation of physiological signals (HR and RR) in the Neonatal Intensive Care Unit. A multi-task deep **convolutional neural network** (**CNN**), previously presented in [28] is devised in order to automatically perform patient skin segmentation and detect its presence/absence. Given the segmented skin region, 3 rPPG signals were estimated by averaging the RGB colour intensity channels, and 9 motion-related respiratory signals were extracted by tracing the waveforms of geometrical information describing the skin area (area, perimeter, centroid, fitting ellipse). The Multiple Kalman filter approach is eventually adopted to fuse the RR estimates obtained for each of the 12 above-mentioned respiratory signals. A similar, yet simplified version of the above procedure has been later presented in [74] and [29].

Table 2 provides an overview of solutions that extract respiratory-induced variations from rPPG signals. Approaches relying solely on estimated rPPG signals are marked with an "R" in the "Approach" column, while those utilising hybrid techniques (i.e., combining rPPG estimates with motion information) are marked with an "H." The "Method" column specifies the rPPG technique used for blood volume pulse estimation, while the "ROI" column indicates the considered Region of Interest.

2.3.3 Summary. This paragraph has introduced and reviewed the literature on rPPG-based respiratory estimation methods. These approaches rely on extracting **Blood Volume Pulse** (**BVP**)

Table 2. rPPG-based (R), Hybrid (H) Methods

Article	Approach	Method	ROI	Processing includes
[163]	R	G, B channel average + RIIV	Forehead	
[162]	R	CHROM, PBV + RIIV	Face	
[125]	R	ICA + RIFV	Face	
[34]	R	POS + RIFV	Face	
[114]	R	Green + RIFV	Face	
[19]	R	u^* channel average, CWT + RIFV	Face	
[99]	R	CHROM, PBV, PVM, EVM + RIVs+	Face	IMS, fusion
[50]	R	Green, CHROM, Hue, normG + RIVs+	Face, Forehead	Fuse-Mod
[75]	R	intensity average + RIVs+	Fingertip	IMS, smart fusion, EMD
[84, 118, 141]	R	Green + RIVs+	Fingertip	
[57]	R	Green + most voted frequency peak detection	Face Multi-Patches	Wavelet denoising
[127, 138]	R	Hue + frequency peak detection (FPD)	forehead	
[155]	R	AR spectral estimation/intensity average	Cheeks	
[39]	R	color average + FPD	Newborn Chest	
[10, 11]	R	color average + FPD	Newborn Chest	EVM
[97]	R	Energy Variance Maximization + FPD	Face	
[98]	R	Wavelet Variance Maximization + FPD	Face	
[172]	R	color average, BSS + FPD	2 ROIs on the face	Kurtosis
[157]	R	Green + auto-regressive	Face	
[112]	R	POS + spectra subtraction	Nostrils, Forehead	
[3]	H	Green + FPD / vertical-OF	Face / Forehead	FD, EVM, CEEDAM, CCA
[139]	H	Green + FPD / Motion	Face / Head, Chest	EMD
[69]	H	color average + RIVs / vertical-OF	Multi-ROIs on Face / Head	
[164]	H	3 color intensity average / 9 motion estimations	Face	CNN for Skin Segmentation, multiple Kalman filters

The column "Method" includes both the technique adopted to compute the rPPG and the subsequent method adopted to derive the respiration rate. In case of hybrid approaches, also the motion-based method is specified after the "/". Analogously, the column "ROI" specifies the area considered by the rPPG-based method and, in case of multi or hybrid method, also the ROI considered by the motion-based branch. Acronyms used in the table: CHROM, PBV, POS: rPPG methods; ICA: Independent Component Analysis; RIV: Respiratory Induced Variations; RIFV: RIV capturing the frequency (RIFV); RIIV: RIV capturing the intensity; RIV+: extended set of features derived from the rPPG signals; CWT: continuous wavelet transform; FD: Face Detection; IMS: Incremental Merge Segmentation; EMD: Empirical Mode Decomposition; AR: auto-regressive; EVM Eulerian Video Magnification; BSS: Blind Source Separation; CEEDAM: Complete Ensemble Empirical Mode Decomposition with Adaptive Noise; CCA: Canonical Correlation Analysis; CNN: Convolutional Neural Network; EVM: Energy Variance Maximization.

signals using rPPG, from which respiratory information is derived using either morphological segmentation or blind source separation.

Advantages: The considerations for motion-based methods (see Section 2.2.5) regarding video resolution and frame rates also apply to rPPG-based approaches. Low resolution is not an issue when estimating rPPG signals, but it can help reduce quantisation noise if intended as a means of performing pixel averaging [36, 171]. Similarly, typical video frame rates (e.g., 20 fps) are perfectly reasonable sampling frequencies for representing respiratory information in rPPG signals. Unlike motion-based approaches, however, a different source of information–respiratory-induced BVP variations–is used. Since rPPG does not explicitly rely on motion information, the issue of uniform or textureless regions does not apply here.

Limitations: The most important issue in rPPG-based solutions is that the accuracy of the estimated respiratory signal heavily depends on the quality of the underlying rPPG signal. Notably, it can be influenced by factors such as video compression, motion artifacts, and lighting conditions. Heavily compressed videos can distort the rPPG signal [111], thus limiting the applicability of rPPG-based solutions in general scenarios.

2.4 Deep-learning based Approaches

Recently, deep learning has gained widespread adoption across various fields such as computer vision [27], natural language processing [123], and autonomous vehicles [82]. In the realm of

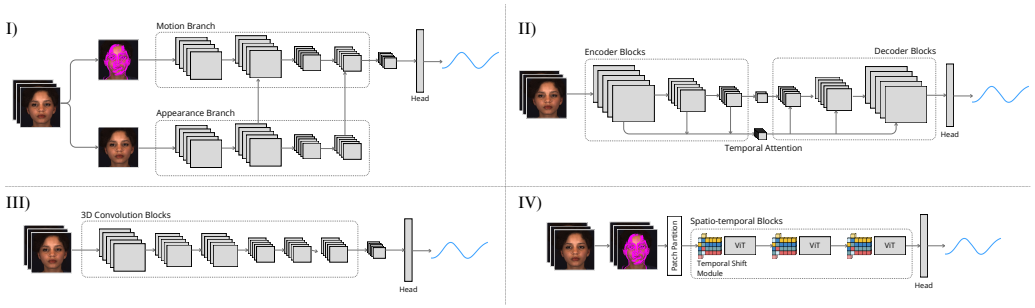


Fig. 5. Deep-learning methods for RGB camera-based respiration estimation at a glance. Given a sequence of multiple frames, a variety of neural architectures have been employed to learn to predict respiratory waveforms. Some notable examples are: (I) Dual branch Convolutional Neural Networks [93, 119] - (II) Autoencoders [144] - (III) 3D Convolutional Neural Networks [53, 180] and (IV) Transformers with Temporal Shift Modules [178].

contactless estimation of physiological data, deep learning is also garnering significant interest and achieving remarkable success [49]. Some articles adopt machine learning techniques in the preprocessing phase to locate regions of interest [51, 64, 77, 79, 87, 89]. More interestingly, in the specific domain of RR via RGB cameras, very recent publications provide further evidence for the effectiveness of deep learning-based approaches in this field. They are characterised by the architecture adopted, together with the feature extracted and the considered ROI. A variety of different neural architectures have been employed in order to learn to predict respiratory information from videos; they are summarised at a glance in Figure 5.

Specifically, in [121] autoencoders are used to disentangle physiological from non-physiological features. In particular, a multi-scale spatiotemporal map is carried out to characterise the RGB and YUV channels in correspondence to facial ROIs. Then an autoencoder architecture with two encoders is designed, aiming at characterising the physiological/non-physiological signals, using a cross-verified scheme. The method predicts both HR and RR, and concerning RR it has been tested on the OBF dataset.

Undoubtedly, CNN architectures are the most extensively utilised in this domain. In [22] the Eulerian motion magnification method through Hermite transform is applied to the whole video capturing the person's chest. The magnified video is then used to train a CNN to classify inhalation and exhalation frames. Then, the average of the breathing cycle length (in number of frames) is computed, and the RR is derived accordingly.

In [168] a two-stage pipeline based on the CNN architecture is used for the RR and HR estimations. First, a preprocessing is adopted, consisting in the automatic cheek ROI detection and the **phase based video motion processing** (**PVMP**), aiming at attenuating motion changes while magnifying the colour ones in the frequency bands corresponding to respiration and pulse rates. Second, this filtered ROI is processed by a CNN constituted by three parts: a shared one, and two branches specialised for the RR and HR estimation respectively. The network is a pipeline including 3 types of convolutions: the standard one, the depth-wise convolution, and the 1D convolutional layers.

The method presented in [68] is a CNN model based on residual blocks and the 1D convolution, aiming at identifying the pixels having an appearance variance correlated to the respiration information. Mainly due to the pixel-based nature of this step, the produced binary map is noisy. The evidence used to denoise this map is that the "good" pixels are either in-phase or out-of-phase, exhibiting symmetrical behaviour, whereas noisy pixels do not display this characteristic.

In [154], a spatio-temporal CNN is employed to estimate the rPPG signal, taking as input the automatically obtained ROI encompassing the forehead, cheeks, and nose, which is achieved through the Haar cascade detector. HR and RR are subsequently derived from the rPPG signal through post-processing.

A research trend has emerged, introducing dual-branch architectures that aim at modelling both motion and appearance simultaneously. A seminal work in this line is represented by [35], where a dual branch CNN, namely DeepPhys, is adopted to estimate both HR and RR in an end-to-end fashion. The motion model branch is a VGG-style CNN that takes as input the difference between two consecutive frames' face crops to capture the motion and produces a **feature map** (**FM**). The appearance model takes as input the original frame's face crop, aiming at learning a **soft-attention mask** (**SAM**) that assigns high values to skin areas relevant for physiological estimation. The final masked feature map is obtained by multiplying the FM with the SAM, and by processing it through fully connected layers to output the continuous signal to be used for the HR and RR estimation. During post-processing, a band-pass filter selects the frequency range of interest, the power spectrum is derived, and the highest peak is detected. An improvement of this approach is proposed in [93], where a **multi-task temporal shift convolutional attention network** (**MTTS-CAN**) is presented to predict cardiovascular and respiratory measurements simultaneously. The main novelty here is the introduction of the **Temporal Shift Module** (**TSM**) in the motion branch, so to overcome limitations of the previous work in capturing long temporal dependencies. TSM enables the creation of temporal dependencies beyond consecutive frames by temporally shifting part of the channels. This approach simulates the behaviour of a 3D CNN while maintaining the computational efficiency of a 2D CNN. Additionally, to mitigate the undesired introduction of artifacts from the TSM module, a soft-attention mask is applied in the motion branch. The method's efficiency is further enhanced, as most of the computations serve both the BVP and respiration estimates, with the loss function designed to account for both tasks.

MTTS-CAN has achieved great success, to the extent that several variants have been proposed to further improve its performance: in [130] the architecture proposed in [93] is enriched by the channel-wise attention, the rationale being that while the spatial attention highlights the most relevant pixels to be considered for the physiological estimation, the channel-wise attention weights differently the features along the channel dimension, further cleaning the feature map. In [85], a lighter model inspired by MTTS-CAN is designed to predict PPG and respiratory signals. It is a multi-input, multi-task Siamese network model, without attention or time-shift modules. The two siamese branches are VGG-like and take as input the forehead and the lower part of the face, respectively. These ROIs are automatically detected in the first frame of a 600-frame time window and remain fixed throughout the entire video. The features produced by the two branches are then added and used to estimate the PPG and respiratory signals. For what concerns the RR estimation, the method applies to the produced respiratory signal a band-pass filter with cutoff frequencies $[0.1 - 0.4]Hz$, then it searches for the peaks and computes the average peak-to-peak distance. The model is lightweight, thus suitable for mobile devices. The method proposed in [131] also adopts MTTS-CAN as the backbone, while introducing a novel loss function, the **Pulse-Respiration Quotient** (**PRQ**), which integrates the frequency correlation between HR and RR to enhance the robustness of both estimates. PRQ information is further utilised in post-processing filtering as a correlation measure between HR and RR to improve prediction accuracy.

More recently, in [119] another advancement of MTTS-CAN, namely BigSmall, has been proposed. It is a unified model to derive simultaneously facial action, cardiac, and pulmonary measurements. It is characterised by a high-resolution branch (Big) aiming at deriving spatial texture features, and a low-resolution branch (Small) conceived to model the temporal dynamics. Another

novelty consists in the adoption of the wrapping temporal shift module in order to deal efficiently with short frame sequences.

Differently, in [32] Chen et al. propose **dual branch architecture** (**ACTNet**) specifically designed for respiratory rate estimation from facial video data. It integrates a CNN branch to capture subtle facial color changes with local attention mechanisms and a Transformer branch to model long-term temporal dependencies in video sequences. A feature coupling unit fuses local and global features to enhance performance. The authors report improved performance on the COHFACE and DEAP datasets compared to other neural approaches.

In [132] transformers are used for generating instantaneous evaluation of physiological signals. The proposed method consists of a spatial backbone to characterise each single frame (DeepPhys-based architecture is adopted to this aim), and addresses temporal aggregation via a Transformer encoder focused over N frames. The authors adopt the maximum cross-correlation loss applied in the frequency domain to overcome the limitations of the MSE loss. Experiments have been conducted on the V4V dataset obtaining a MAE of 5.4 BPM.

Many recent solutions have focused on the robust estimation of rPPG signals and the extraction of **respiratory frequencies** (**RF**) from processed rPPG signals [66, 91, 92, 184]. However, the reported results are validated not against a respiratory ground truth, but rather by computing RFs from contact PPG signals. Notable examples include the architectures proposed in [178, 179], namely PhysFormer and PhysFormer++. These models utilise transformers to generate rPPG signals. The respiration rate is subsequently derived from IBIs. PhysFormer is characterised by cascaded temporal difference transformer blocks, while PhysFormer++ takes this analysis further by introducing a two-pathway SlowFast temporal difference transformer that implements a periodic- and cross-attention transformer. This approach allows for capturing both the fine-grained and long-range spatio-temporal relationships for rPPG measurement.

Similarly, in [96], the authors propose a **Dual Generative Adversarial Network** (**Dual-GAN**) to enhance rPPG-based physiological measurements. RF estimates are obtained through signal processing of the recovered rPPG signals.

More recently, some approaches have employed self-supervised learning techniques to learn to estimate physiological signals from facial videos without the need of a synchronised ground truth. In line with the aforementioned methods, most of these approaches do not explicitly predict a respiratory waveform, but extract breathing rates from rPPG signals [94, 180]. Alternatively, the approach presented in [1] (CalibrationPhys) leverages multiple synchronised cameras to train deep learning models, using contrastive learning to distinguish between correct and incorrect respiratory signals. By incorporating data augmentation and pre-trained models, it enhances robustness and adaptability across different camera types to directly predict BVP and respiratory waveforms. Results on a self-collected multi camera dataset demonstrate the efficacy of the contrastive pre-training stage with a sensible performance boost if compared with other DL-based approaches and slight improvements over a method relying solely on optical flow.

In a different vein, in [53] the authors introduce CliffPhys, a family of models adopting hypercomplex neural architectures for camera-based respiratory measurement. Their approach extracts respiratory motion from RGB videos using optical flow and monocular depth estimation to obtain motion and depth fields. By employing Clifford Neural Layers [20], the approach aims at modelling geometric relationships within these fields to estimate respiratory waveforms. CliffPhys is trained on the SCAMPS and COHFACE datasets. The authors report results on a COHFACE test split (MAE = 0.83) and on the BP4D+ dataset (MAE = 3.50), showing a significant performance boost compared to other deep learning-based solutions such as MTTS-CAN and BigSmall.

Eventually, the work presented in [144] aims at disentangling the contribution of various physiological signals for robust estimation of respiration. The proposed model is a two-stage network for

Table 3. Deep-learning Based Methods

Article	Architecture	ROI	Additional notes
[121]	Autoencoder	FH, C	RR as a postprocessing of DL; DL characterised by: Spatio-Temporal Map over RGB and YUV color space
[161]	LSTM	CH	RR derived directly with DL; Processing includes: Optical Flow
[168]	CNN	C	RR derived directly with DL; DL characterised by: 2-stage processing
[68]	CNN	F	RR derived directly with DL; DL characterised by: Residual block and pixel selection
[22]	CNN	CH	RR derived directly with DL; Processing includes: EMV
[154]	CNN	FH, C, N	RR as a postprocessing of DL; Processing includes: Haar cascade face detector
[35]	CNN - Dual branch	F	DeepPhys RR as a postprocessing of DL
[93]	CNN - Dual branch	F	MTTS-CAN; DL characterised by: Multi-task Temporal Shift Convolution Attention network; RR derived directly with DL
[130]	CNN - Dual branch	F	DL characterised by: channel-wise attention added to MTTS-CAN; RR derived directly with DL
[85]	CNN - Dual branch	FH, F	DL characterised by: lighter version of MTTS-CAN eliminating Attention and MTS module; RR derived directly with DL
[131]	CNN - Dual branch	F	DL characterised by: Pulse-Respiration Quotient loss used in the MTTS-CAN architecture; RR derived directly with DL
[119]	CNN - Dual branch	F	BigSmall; characterised by: two resolution branches and wrapping temporal shift module; Predicts Respiratory Waveform
[132]	Transformers	F	RR as a postprocessing of rPPG; DL characterised by: Maximum cross-correlation Loss
[179]	Transformers	F	Physformer; RR as a postprocessing of rPPG; DL characterised by: temporal difference transformer
[178]	Transformers	F	Physformer++; RR as a postprocessing of rPPG; DL characterised by: SlowFast temporal difference transformer
[96]	GAN	F	Dual-GAN; RR as a postprocessing of rPPG
[180]	3D/1D Conv + BiLSTM	F	rPPG estimation via a frequency-inspired self-supervised framework; RR as a postprocessing of rPPG
[94]	Masked Autoencoder	F	rPPG estimation as a self-supervised reconstruction task; RR as a post-processing of rPPG
[53]	Hypercomplex 3D CNN	F	CliffPhys; strong 3D motion and geometric inductive biases for respiratory waveform prediction
[32]	CNN + Transformer	F	ACTNet; Exploits local color changes between facial frames (CNN) while the Transformer captures long-term temporal relationships
[1]	2D CNN	F	CalibrationPhys; Employs data augmentation and contrastive learning to build pre-trained models for Heart and Respiratory Rate Measurements
[144]	2D CNN Enc./Dec. + 3D CNN	F	It implements a Time-Varying Interference Disentanglement module followed by Biosignal Regression with Long-Term Temporal Attention

Acronyms used in the table: RR - respiration rate; DL - Deep Learning; F - Face; FH - Forehead; C - Cheeks; CH - Chest; N - Nose.

remote respiratory rate estimation using facial video under natural light. It addresses challenges like specular reflections from head movements and entangled physiological signals by first using an encoder-decoder structure to extract respiratory motion from video frame differences. This re-characterised signal, disentangled from interferences, is then combined with facial appearance in the second stage. A long-term temporal attention module enhances the model's ability to track the extended breathing cycle. The proposed model is trained, validated, and tested on the COHFACE and MAHNOB-HCI datasets. The authors report an MAE of 5.025 on the COHFACE dataset and 8.191 on the MAHNOB-HCI dataset for the best model.

The reviewed deep-learning based solutions are summarised at a glance in Table 3. Methods are grouped based on the employed architecture and the adopted Region of Interest.

2.4.1 Summary. This paragraph has introduced and reviewed the literature on deep learning-based remote respiratory estimation methods. These approaches rely on different neural

Table 4. Datasets

Dataset	Subjects	Videos	Resolution	FPS	Signals	Additional notes
MAHNOB [147]	27	527	780 × 580	61	ECG; EEG, EDA, RW, temperature, AU, eye gaze, body movements	Videos are quite heavily compressed, and thought for implicit media tagging
OBF [88]	106	212	1920 × 1080	60	ECG, BVP, RW	It includes near infrared videos
AFRL [47]	25	300	658 × 492	30	PPG, RW	Videos capture head motion
MMSE/BP4D+/V4V [133, 182]	140	1400	1040 × 1392	25	BVP, RW, EDA, AU	Thought for Multimodal affect analysis. Includes thermal videos.
COHFACE [61]	40	160	640 × 480	20	PPG, RW	Heavily compressed videos
DEAP [78]	22	160	720 × 578	50	EEG, EDA, BVP, RW, temperature	Thought for Multimodal affect analysis
MSPM [149]	103	-	1920 × 1080	90	PPG (multiple sites), BP, RW	The respiratory GT is represented by a sinusoidal pattern displayed to the subjects on a screen
SCAMPS [110]	2800	2800	320 × 240	30	PPG, PR, RW, AU	Videos are synthetically produced

Acronyms used in the table: ECG - Electrocardiogram waveform, EEG - Electroencephalogram waveforms, RW - Respiratory waveform, PPG - Photoplethysmogram waveform, BVP - Blood Volume Pulse Signal, EDA - Electrodermal activity, AU - Action Units, BP - Blood Pressure.

architectures (convolutional, recurrent, transformer) to predict respiratory waveforms using end-to-end training.

Advantages: General considerations on video resolution and frame rates made in Section 2.2.5 also apply to DL-based methods. Interestingly, spatial subsampling is deliberately employed by some methods [35, 93, 119] to reduce computational cost and quantisation noise. Moreover, most of the proposed approaches can be trained end-to-end to directly extract respiratory signals from videos, allowing for the avoidance of specific preprocessing steps. Furthermore, the power of cutting-edge DL methods can be exploited to improve the estimation of respiratory signals (self/semi-supervision, large-scale pre-training, multi-task learning).

Limitations: DL methods can be computationally demanding during training, with some deep models requiring significantly more processing time than simpler methods in the evaluation phase. Moreover, they typically require large and diverse datasets for training and validation; the lack of such datasets can lead to issues with accuracy and generalisation. Additionally, some models are trained in a multi-task manner to leverage both the respiratory-induced variations in rPPG signals as well as motion information. Consequently, as discussed in Section 2.3.3, these models may still need to rely on uncompressed or lightly compressed videos.

3 Experimental Evaluation

By and large, the articles surveyed here primarily focus on describing the proposed approach, while the experimental analysis and comparison with both baselines and competing methods are often relegated to the background, with a few exceptions. Moreover, the shortage of viable datasets available for such a purpose appears to be a long-standing issue of this research field that exacerbates such a tendency. As a matter of fact, many works rely on a handful of non-public videos to test their proposals, thus making it difficult to achieve a fair evaluation. This section aims at filling this gap by setting out an extensive experimental analysis of the literature on multiple publicly available datasets. Table 4 lists and describes the datasets proposed in the literature that can serve as a benchmark for testing and benchmarking camera-based respiratory measurement algorithms.

Clearly, an exhaustive evaluation of all the methods surveyed in Section 2 is not feasible. However, it is worth noticing that many approaches in the literature often present little novelty when compared with each other. As a matter of fact, they basically differ for the particular type of pre/post-processing applied, or for a heuristic specifically designed for the condition at hand. Therefore, here we deliberately adopt a simplified experimental pipeline that aims at highlighting

the eventual empirical differences between the variety of approaches. In a nutshell, and at the highest level of abstraction, it can be summarised via the following chain of operations:

$$\text{video} \rightarrow \text{ROI Selection} \rightarrow \text{Method} \rightarrow \text{Band-pass Filtering} \rightarrow \text{RW} \rightarrow \text{RR}$$

A representative (albeit plainer) version of the diverse procedures surveyed here has been reimplemented and tested on three publicly available datasets, namely BP4D+, COHFACE, and MAHNOB. This resulted in the implementation of a software package called resPyre , designed to standardise the evaluation of datasets and models in this domain. To ensure reproducibility and facilitate future research, resPyre is released as an open-source Python package, available at https://github.com/phuselab/resPyre. It provides a unified benchmarking framework with modular implementations of various approaches, allowing users to easily compare methods under a common experimental setup. The package includes documentation and example scripts to guide users through installation, dataset integration, and method evaluation. Additionally, resPyre is designed to be extensible, enabling researchers to incorporate new models and datasets seamlessly.

The choice of the approaches to test closely follows from the taxonomy proposed in Section 2.

In summary, the original RGB video frames are cropped so as to obtain a ROI. Two different ROIs can be defined depending on the estimation method at hand:

(1) *Face ROI:* The Mediapipe [55] face detector is adopted to obtain a sequence of frames centre-cropped on the subject's face. Subsequently, these ROIs are fed as input to either an rPPG method or a deep learning-based approach to produce a respiration waveform.
(2) *Chest ROI:* The Mediapipe [55] pose landmarker is adopted to obtain a sequence of frames centre-cropped on the subject's chest. The ROIs are then used by various motion-based respiratory information estimation methods to produce a respiration waveform.

The ROIs obtained are then fed as input into subsequent processing stages.

3.1 Evaluation Procedure and Metrics

Each method produces an estimate of the respiratory waveform. Subsequently, these are bandpass filtered using a Butterworth bandpass filter with cut-off frequencies set at $[0.1-0.5]Hz$. Respiratory rates can henceforth be derived via spectral analysis; specifically, the Welch periodogram is used to estimate the spectrum of the estimated respiratory waveform. The peak in the estimated spectrum yields the estimated RR $\hat{h}$. In order to measure the accuracy of the estimate $\hat{h}$, this is compared to the reference RR h recovered from contact respiration sensors. The same procedure (band-pass filtering and maximisation of the Welch periodogram) is employed for the reference RW.

We adopt common metrics to evaluate the performance of one or more methods in estimating the correct RR, namely **Mean Absolute Error** (**MAE**), **Mean Absolute Percentage Error** (**MAPE**), **Pearson Correlation Coefficient** (**PCC**), and **Concordance Correlation Coefficient (CCC**).

3.2 Datasets

Among the datasets briefly summarised in Table 4, we selected three: BP4D+ [182], COHFACE [60], and MAHNOB-HCI [147]. This selection was based on their public availability, widespread adoption, ease of access, and coverage of diverse acquisition conditions, including variations in video quality, lighting environments, camera perspectives, subject demographics (e.g., different ages, genders, and ethnic backgrounds), elicitation stimuli, and movement patterns (e.g., stationary subjects, head movements, or talking). While the remaining datasets in Table 4 contain valuable information and could be considered for more in-depth analyses, some were excluded from the current preliminary benchmark due to various factors such as: limited accessibility or their primary use as training sets for deep learning-based methods (e.g., OBF and AFRL), overlapping acquisition conditions (e.g.,

DEAP), absence of a *contact* respiratory ground truth (e.g., MSPM) or their synthetic nature (e.g., SCAMPS). The selected datasets are briefly described below.

3.2.1 BP4D+ [182]. BP4D+ includes 3D, 2D, thermal, and physiological data recordings, such as heart rate, blood pressure, skin conductance (EDA), and respiration rate, along with metadata like facial features and FACS codes. The dataset comprises recordings from 140 individuals (58 men and 82 women) between the ages of 18 and 66, representing diverse ethnic backgrounds, including East Asian, Middle Eastern, Hispanic/Latino, and Native American participants. Each subject completed 10 tasks designed to elicit various emotional responses. High quality videos were captured at 25 frames per second, while physiological signals were recorded at a sampling rate of 1000 Hz. Recently, the V4V dataset [133] has been constructed by selecting subjects from BP4D+.

3.2.2 COHFACE [60]. This dataset comprises 160 RGB video sequences, each lasting one minute, captured alongside synchronised heart rate and breathing rate data from 40 participants (12 women and 28 men). During recording, subjects remained seated and motionless in front of a webcam to ensure full facial visibility. The dataset includes recordings under two lighting conditions: controlled studio lighting with a spotlight and natural ambient light. The videos are encoded using the MPEG-4 Visual format (MPEG-4 Part 2) with a bitrate of approximately 250 kbps. They have a resolution of 640 × 480 pixels and a frame rate of 20 frames per second.

3.2.3 MAHNOB-HCI [147]. Originally designed for emotion analysis, this database has also been utilized for evaluating remote physiological sensing algorithms. A total of 30 participants (17 women and 13 men, aged 19 to 40) were monitored while watching selected movie clips and images. Data collection involved six video cameras capturing different viewpoints, a head-worn microphone, an eye gaze tracker, and various physiological sensors including respiration amplitude. The MAHNOB-HCI dataset videos are encoded using H.264/MPEG-4 AVC compression, with a bitrate of 4200 kbps, a frame rate of 61 fps, and a resolution of 780 × 580 pixels.

3.3 Motion-based Methods

3.3.1 Difference of Frames (DoF). The DoF approach is probably the simplest method to detect motion in videos and is adopted here as the baseline model. As the name suggests, it basically computes the difference between consecutive frames (chest ROIs) [73, 159].

Denote $\mathcal{F}_t \in \mathbb{R}^{w \times h}$ (with w and h the ROI's width and height) the gray-scale cropped video frame, then the DoF can be seen as a simple pixel-wise FIR filter expressed by $DoF_t = \mathcal{F}_t - \mathcal{F}_{t-\Delta t}$, where Δt is the time step between consecutive frames. The cumulative motion signal derived by summing the intensity of the pixels in the DoF sequence of ROIs, constitutes the estimated raw respiratory waveform RW: $r_t = \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} DoF_t(x, y), \ \forall t$. Results of the DoF approach on the BP4D+, COHFACE and MAHNOB-HCI datasets are reported in Table 5.

3.3.2 RGB/Intensity-based Methods (Pixels Projection). A variety of approaches exploiting pixels intensity/colour variations caused by the respiratory induced movements have been proposed. Here we implement a simple yet effective approach that estimates motion by correlating one dimensional image profiles (1D-PROFILE) obtained by aggregating the pixels of the current cropped frame onto its horizontal axis [13, 21, 170]. Such approach assumes that the relevant information (respiration) appears by and large as rigid vertical motion. Subsequently, the obtained one-dimensional profile is cross-correlated with that of a previous frame, and the position of the maximum in the cross-correlation function yields an estimate of the vertical displacement ($v_y t$ in Equation (1)).

Table 5. Comparative Performance of Respiratory Rate Estimation Methods Across Multiple Datasets

Method Type	Method	BP4D+				COHFACE				MAHNOB-HCI			
		MAE↓	MAPE↓	PCC↑	CCC↑	MAE↓	MAPE↓	PCC↑	CCC↑	MAE↓	MAPE↓	PCC↑	CCC↑
Difference of Frames	DoF	2.52	14.59	0.44	0.41	2.48	18.42	0.30	0.25	3.18	17.95	0.08	0.07
Profile 1D	Linear	3.07	16.27	0.39	0.38	1.29	9.45	0.58	0.57	3.24	20.14	-0.16	-0.16
	Quadratic	3.80	19.64	0.13	0.12	0.85	6.36	0.74	0.71	2.23	13.99	0.33	0.33
	Cubic	3.92	20.18	0.06	0.06	0.84	6.41	0.73	0.71	2.51	16.16	0.31	0.29
Optical Flow	FNBK	1.87	11.78	0.54	0.49	1.02	7.89	0.72	0.69	2.71	16.74	0.26	0.25
	CRAFT	1.59	9.39	0.62	0.60	0.82	6.38	0.74	0.71	**1.81**	**11.14**	0.44	0.44
	lcv-RAFT	1.43	**8.37**	0.69	0.68	0.68	5.18	0.82	0.80	1.97	12.20	0.47	0.47
	irr-PWC	1.53	9.16	0.66	0.64	0.69	5.39	0.77	0.75	1.84	11.51	**0.50**	**0.50**
	GMA	**1.39**	8.38	0.69	0.67	0.66	5.08	0.83	0.81	2.10	13.08	0.35	0.34
	RAFT-s	1.42	8.52	0.70	0.69	0.72	5.66	0.76	0.74	1.95	12.25	0.42	0.42
	RAFT	1.40	8.48	**0.71**	**0.70**	**0.64**	**4.84**	**0.84**	**0.82**	2.06	12.68	0.43	0.43
rPPG Morph. Feat.	IMS	4.40	21.20	0.07	0.04	3.76	27.44	-0.03	-0.02	2.53	15.24	0.21	0.20
	Peaks	4.69	22.52	0.16	0.06	3.13	22.35	-0.04	-0.03	2.78	16.37	-0.17	-0.13
BSS of rPPG	EMD	6.26	29.76	0.03	0.01	2.54	16.44	0.22	0.22	3.63	21.12	0.04	0.04
	SSA	6.36	30.24	-0.01	0.00	2.53	16.16	0.16	0.16	2.98	17.05	0.25	0.22
Deep Learning	MTTS-CAN	3.62	19.16	0.23	0.21	2.31	17.50	0.43	0.35	3.32	19.94	-0.07	-0.07
	BigSmall	3.17	16.71	0.30	0.29	3.92	28.70	0.21	0.15	3.36	19.54	-0.12	-0.12

Best scores are in bold, best for each method type are underlined.

In our implementation, the 1D-profile $p_t \in \mathbb{R}^h$ is obtained as a combination of mean and standard deviation of the gray-scaled pixels intensity of the cropped frame, i.e.,

$$p_t = 0.5 \times \mu(\mathcal{F}_t) + 0.5 \times \sigma(\mathcal{F}_t), \tag{2}$$

where $\mu(\mathcal{F}_t) \in \mathbb{R}^h$ and $\sigma(\mathcal{F}_t) \in \mathbb{R}^h$ represent the average and standard deviation of pixels' intensity on the horizontal direction, respectively. The raw respiratory waveform is then obtained as the position of the maximum of the cross-correlation function of the current 1D-profile p_t with the previous $p_{t-\Delta t}$. In practice, as respiratory motion may happen at the sub-pixel level, the cross-correlation function is typically interpolated before determining the maximum [13, 21, 170]. More formally, denote $A_{p_{t-\Delta t}, p_t}(\tau)$ the interpolated cross-correlation between the 1D-profiles at times t and $t - \Delta t$ as a function of the displacement τ; the RW at time t is obtained as

$$r_t = \frac{1}{\Delta t} \arg\max_{\tau} \; A_{p_{t-\Delta t}, p_t}(\tau). \tag{3}$$

Results are reported in Table 5 varying the types of interpolation (Linear, Quadratic, Cubic) of the profiles cross-correlation function.

3.3.3 Optical Flow. Another popular approach to estimate the respiratory related velocities vectors is given by the computation of the OF [71, 90, 170]. It allows to extract the motion of objects between consecutive frames of a video, caused by the relative movement of a subject w.r.t. the camera. Call $\mathcal{F}_t$ the RGB video frame at time t, at the most general level the OF equation writes $\frac{\partial \mathcal{F}_t}{\partial x} v_x + \frac{\partial \mathcal{F}_t}{\partial y} v_y + \frac{\partial \mathcal{F}_t}{\partial t} = 0$, where $v_x = \frac{\partial x}{\partial t}$ and $v_y = \frac{\partial y}{\partial t}$. The problem of OF is solving v_x and v_y to determine movement over time. This is, in general, unfeasible but a variety of approaches have been developed as approximate methods. OF can be computed on a subset of "interesting" pixels (sparse OF) or on the entire frame (dense OF). While the former is less computationally expensive, the latter typically exhibits higher accuracy, hence in our experiments dense OF has been used.

OF has traditionally been approached as an optimisation challenge usually leveraging variational techniques, which can be solved as an energy minimisation process. However, recent methods employing deep learning have yielded remarkable outcomes, typically surpassing "knowledge-driven" approaches on benchmarks [181]. Typically, these methods take a pair of video frames as input and produce OF as output. Formally: $(v_x, v_y) = \mathrm{NN}(\mathcal{F}_{t-\Delta t}, \mathcal{F}_t)$, where NN represents a Neural Network. Similarly to the *pixels projection* case (and following the vast majority of approaches presented in the literature), we assume that the respiratory information appears mainly as vertical

motion. The RW can be hence recovered by computing the median of all the vertical displacements vectors estimated by a given OF method at every frame.

We compare results for respiratory rate estimation using 6 Deep OF models, namely CRAFT [151], LCV-RAFT [175], IRR-PWC [67], GMA [72], RAFT-Small and RAFT [158]. In addition the Farneback [48] knowledge-based OF method is employed as a baseline. Results for all the adopted metrics are reported in Table 5.

3.4 rPPG-based Methods

To extract BVP signals from video, we implement the rPPG estimation pipeline as implemented in the pyVHR package [15, 17]. pyVHR provides a variety of different rPPG methods [16, 24, 42, 124, 171]; in our experiments the CHROM [42] rPPG approach is employed. Differently from the original settings, the BVP signal band-pass filtering cut-off frequencies have been set to the range $[0.1 - 4.0]Hz$ in order to keep the lower frequencies eventually conveying respiratory information.

The obtained estimate is subsequently processed to extract RIVs. In the literature, this has been accomplished either through morphological segmentation of the rPPG signal or via Single-Channel Blind Source Separation techniques.

3.4.1 rPPG Morphological Segmentation. As summarised in Section 2.3, respiratory information can be mined from (r)PPG signals by tracking its local minima/maxima. In the biomedical signal processing literature, lots of efforts have been devoted to the development of algorithms to extract such quantities that are specifically tailored to the characteristics of the BVP signals (e.g., [45, 75]); however some approaches employ "general purpose" peak finding methods [50].

In our experiments, two approaches have been implemented to extract local minima/maxima from the estimated rPPG; the first one relies on a morphological segmentation of the rPPG signal, called IMS algorithm [75]. For comparison, a simple local minima/maxima finding algorithm as proposed in [50] has been implemented too.

In both cases, three respiratory induced variations (namely RIIV, RIAV, RIFV) have been recovered (see Figure 4). Each RIV represents a RW estimate; the corresponding respiration rates are determined via maximisation of their Power Spectral Density. The final RR value is then obtained via averaging. Results are reported in Table 5.

3.4.2 Blind Source Separation. An alternative approach to recover respiratory information embedded into rPPG signals is to consider the latter as the result of a mixed number of sources (one of which is respiratory modulation). **Single Channel Blind Source Separation** (**SC-BSS**) techniques may hence be employed to extract such information.

In the literature, articles pursuing this route have mainly adopted **Empirical Mode Decomposition** (**EMD**) as the SC-BSS method to extract RIVs. In a nutshell, EMD is an analytical technique employed for effectively characterising time series that are both non-linear and non-stationary in nature. This method entails the projection of the time series onto a basis within a space that consists of IMFs [63]. EMD segregates a limited set of the IMFs and directly derives the frequency and amplitude dynamics from these functions [129]. In this work, and following [18], we compare EMD with **Singular Spectrum Analysis** (**SSA**) as an alternative SC-BSS technique. SSA [46] represents a decorrelation method that projects a singular combination of sources with zero mean (time series) onto an orthonormal basis within a space. This technique entails breaking down the time series (in our context, rPPG) into an assortment of **empirical orthogonal functions** (**EOFs**) by constructing a trajectory matrix using the original data and subsequently employing **Singular Value Decomposition** (**SVD**) on this matrix. The outcome entails a collection of **principal components** (**PCs**) that effectively capture the most prevalent patterns present within the data. Subsequently, these

principal components are harnessed to reconstruct the initial time series, effectively disentangling the signal from noise.

Results of correct RR recovering using either EMD or SSA are reported in Table 5.

3.5 Deep Learning-based Methods

Among the various neural architectures proposed for predicting respiratory information, we have selected those with publicly available implementation code and pre-trained model weights.

For our benchmark, two approaches have been selected: MTTS-CAN and BigSmall. We excluded from the experimental analysis methods that do not explicitly learn to predict respiratory waveforms, but instead derive them subsequently from predicted cardiac signals. Although, from a taxonomic standpoint, such methods are technically Deep-Learning-based, they are primarily trained to leverage only cardiac information, from which respiration is later recovered (in a sense similar to an rPPG-based approach).

As discussed in Section 2.4, both MTTS-CAN and BigSmall are two recent expressions of probably the most successful and studied neural architecture in this field (dual-branch Convolutional Attention Neural Network) and training protocol (Multi-task learning).

In our implementation, we have adapted the original code and used the pre-trained model weights released by the authors for both MTTS-CAN [2] and BigSmall. [3]

RR prediction results are reported in Table 5. It is worth remarking that the BigSmall model was originally trained and tested on the BP4D+ dataset. Notably, our results are comparable to those published by the authors, and for some metrics, even better. This improvement is likely due to the prediction on some videos that were originally used for training, which means the results for this model are probably slightly overestimated.

3.6 Discussion

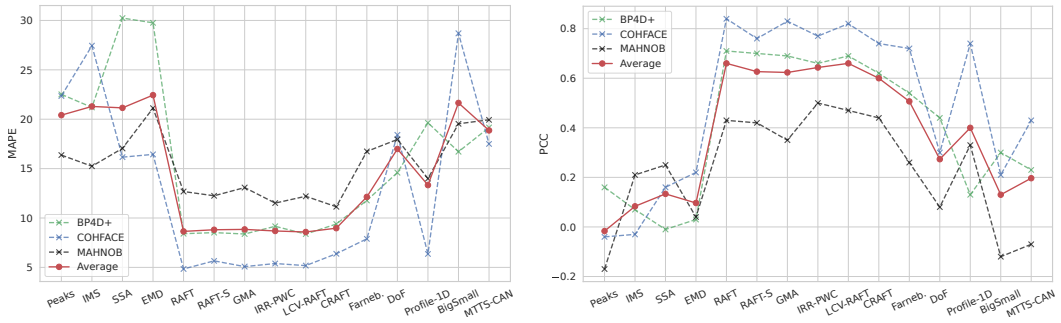
The above experimental results show how a marked difference exists in the accuracy of the predictions when it comes to the remote respiration monitoring via RGB cameras. First thing to notice is how such differences are specific to the dataset and the experimental conditions at hand. Figure 6 clearly indicates that, in our benchmark, the MAHNOB-HCI dataset resulted as the most challenging for almost all the tested approaches. Conversely, COHFACE showed as the simplest corpus w.r.t. the prediction of respiratory information, with BP4D+ sitting somewhat in the middle.

Most importantly, some approaches exhibited stark differences in performance w.r.t. the others. Optical Flow-based approaches clearly outperform all the others by a large margin. Notably, this trend is consistent within all the datasets.

Interestingly enough, the OF estimation method plays a crucial role; when comparing the "baseline" Farneback OF estimation method with state-of-the-art Deep OF approaches, a clear divergence in terms of both MAPE and PCC can be noticed in favour of the methods relying on Deep OF estimation (see Figure 6). Similar considerations hold for the MAE and CCC metrics. On average, the best performing optical flow-based method (LCV-RAFT) achieves a PCC value of 0.66 (strong correlation) and a MAPE of 8.58. To the best of our knowledge, these results can be considered the **state of the art** (**SOTA**) for this task.

The superior performance of OF-based methods can be attributed to their ability to explicitly capture motion patterns and directly model the chest and abdominal movements induced by breathing. This makes them more robust and generalisable across different conditions and datasets. Additionally, our results indicate that deep learning-based OF methods (e.g., LCV-RAFT) outperform

[2] https://github.com/xliucs/MTTS-CAN
[3] https://github.com/girishvn/BigSmall



(a) (b)

Fig. 6. MAPE (a) and PCC (b) metrics results for each method and dataset.

traditional algorithms (e.g., Farneback), likely due to their superior capability in accurately capturing fine-grained motion patterns and handling complex real-world scenarios. However, this increased accuracy comes at the cost of higher computational demands, which raises considerations regarding real-time deployment.

rPPG-based approaches showed as the least accountable for predicting respiration rates from RGB cameras in our experiments. This is not surprising as respiratory induced variation extraction heavily relies on the quality of the underlying (r)PPG signal. Notoriously, this depends on the amount of compression of the original video, with heavily compressed videos (e.g., COHFACE or MAHNOB-HCI) significantly disrupting the rPPG signal. Clearly, such limitation hinders the adoption of rPPG-based solutions in the most general case. In our benchmark, all rPPG-based approaches reached weak correlation (PCC < 0.2) and high errors (MAPE > 20) with SC-BSS (specifically SSA) slightly outperforming approaches based on morphological segmentation.

Quite surprisingly, cutting edge approaches based on Convolutional Attention Networks, albeit beating rPPG-based methods, ranked unfavourably if compared against motion-based approaches, specifically with those grounded on (deep) OF estimation. Unexpectedly, simple motion-based baseline methods such as DoF or Profile-1D yielded more robust estimations of RRs. The underperformance of attention-based CNN methods relative to motion-based approaches suggests that these models may struggle to effectively disentangle respiratory-related motion from other facial and body movements. We hypothesise that their reliance on learned feature representations, rather than an explicit modelling of either motion or respiratory induced cardiac variations, could make them more susceptible to domain shifts across datasets.

As to the computational demands of the benchmarked approaches, Figure 7 depicts the relationship between either the MAPE (Figure 7(a)) or PCC (Figure 7(b)) and the time (in seconds) required by each method to process one frame of a video, on average. Clearly, there exists a bold gap in computational demands when comparing the baseline DoF approach with the most computationally expensive IRR-PWC Deep OF-based method, with almost three orders of magnitude of difference. However, by glimpsing at Figure 7, it is clear that an increase in resource requirements does not always correlate with an improvement in performance. If, on the one hand, the most accurate approaches are the heaviest in terms of computation (i.e., Deep OF-based methods), on the other hand, the lightest procedure (DoF) is far from being the least accurate. In this respect, it is worth mentioning the excellent speed/accuracy tradeoff exhibited by the Profile-1D procedure, significantly outperforming all the approaches except the OF-based, but with tremendously lower computational requirements.

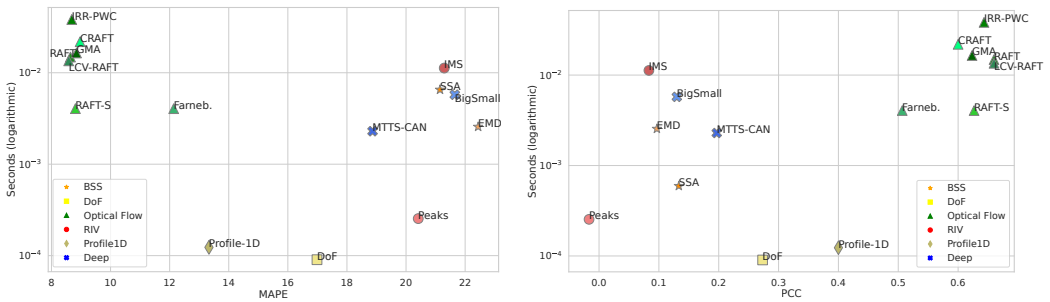


Fig. 7. MAPE (a) and PCC (b) values *vs.* average wall time requirements (seconds) of each method to process a video frame. y axis is on a logarithmic scale.

3.6.1 Open Challenges and Future Directions. Despite the growing interest in RGB camera-based respiratory estimation, several open challenges may hinder its widespread adoption and reliability. One of the primary limitations is the lack of large and diverse publicly available datasets that include physiological ground truth for respiration. Existing datasets are often constrained in size, demographic diversity, and recording conditions, limiting the generalisability and proper evaluation of proposed methods. In fact, current approaches cannot be adequately tested for their robustness to challenging conditions such as camera motion, illumination variations, significant subject motion (e.g., during physical exercise), and, more generally, any out-of-laboratory scenario.

Another critical challenge is video compression: to fully exploit the available information (i.e., both motion and respiratory-induced cardiac variations, RIVs), it is essential that data remain uncompressed, as compression artifacts degrade the quality of rPPG signals [111], from which RIVs are extracted. Moreover, the absence of well-established and reproducible benchmarks further complicates the objective evaluation and comparison of different approaches.

On the modelling side, it is evident that end-to-end deep learning-based solutions have yet to emerge as dominant and reliable approaches for this task, likely–at least in part–due to the aforementioned data limitations. While some models attempt to integrate both rPPG and motion-based cues for respiratory estimation (e.g., [93, 119]), the use of compressed videos during training may render such efforts ineffective. Conversely, training on uncompressed videos only partially addresses the problem, as the use of widely available compressed videos during inference would still force these models into an out-of-distribution setup, where rPPG signals are severely degraded, leaving motion as the primary, yet often unreliable, source of information.

Accurately estimating motion remains a challenging problem, as demonstrated by the extensive research on OF estimation using deep learning. Notably, current deep learning solutions for respiratory estimation tend to be relatively shallow compared to architectures designed for OF estimation [72, 158], largely due to data scarcity. This limitation in representational power may lead to suboptimal results.

Future advancements in this field should focus on integrating synthetic data [110] along with specifically designed unsupervised [148], self-supervised [94, 166, 176, 180] or contrastive learning [1, 54, 152, 153] procedures to mitigate labelled data limitations. While this has recently become a vibrant research direction for rPPG estimation, approaches that explicitly apply the same concepts to other physiological signals, including respiration, remain relatively uncommon. Additionally, explicit inductive biases, particularly in the modelling of motion, could improve robustness and

generalisation. For instance, current dual-branch architectures (e.g., [93, 119]) initialise the "motion branch" by computing frame differences; however, our experimental analysis suggests that, while efficient, this approach may be suboptimal. Instead, leveraging the improved motion estimation capabilities of (pre-trained) OF models could enhance robustness to distribution shifts for motion modelling. Similarly, with regard to the exploitation of respiratory-induced cardiac variations, future advancements should build upon recent efforts in rPPG estimation from compressed videos [40, 169, 177, 183], which continues to be a compelling yet challenging research problem.

Nonetheless, a concerted effort is also needed to expand the availability of diverse and open datasets featuring uncompressed or minimally compressed videos with corresponding respiratory ground truth, mirroring recent progress in the rPPG domain [108]. Lastly, the establishment of standardised benchmarking protocols is essential to ensure fair comparisons and facilitate progress in this research area. This work aims at taking a first step in this direction.

4 Conclusion

Estimating respiratory information from RGB cameras is an evolving and continuously expanding area of research. In many aspects, it also qualifies as a well-established field that has amassed a substantial body of outcomes, including the introduction of algorithmic principles and the accumulation of knowledge over time. However, beneath the progress hitherto achieved, there persist several critical issues that quest for attention.

Specifically, in the examination of the existing literature, it becomes apparent a noticeable trend when it comes to the experimental evaluation of proposed algorithms.In many cases, there seems to be a lack of careful consideration when comparing newly proposed techniques with already established ones.

Moreover, although a few recent proposals have taken advantage of publicly available datasets for performance evaluation, the use of limited or non-public datasets hinders significant advancements and tends to favor less effective methods.

The in-depth experimental, taxonomy-based analysis conducted here has shown how the overall efficacy of remote respiration measurement via RGB cameras delivered by the methods proposed in the literature, is on average reasonably high. However, there are significant differences in outcomes when multiple methods are benchmarked. In our experiments, at least, approaches exploiting chest/shoulder/torso movements qualified as the most reliable. Specifically, the estimation of velocity vectors via Optical Flow yielded the most robust outcomes; in this respect, the accuracy of such estimation appears to be of paramount importance, with SOTA Deep OF estimators delivering the best results.

Conversely, those methods relying on the processing of Blood Volume Pulse signals to recover respiratory modulations appear to deliver sub-optimal results with increased computational demands. It is worth noting, however, that while a variety of datasets can serve the purpose of benchmarking RR estimation algorithms (see Table 4), only a handful of them are easily accessible and provide uncompressed (or lightly compressed) videos, which are essential for the proper extraction of cardiac (and consequently respiratory) information.

The analysis of the literature unsurprisingly reveals a general trend in the community toward shifting to end-to-end deep learning-based solutions. Notwithstanding that it is reasonable to assume that in a more or less near future such approaches may take over in terms of prediction accuracy (similarly to many other fields), current empirical evidence suggests a cautious attitude. Indeed, for what concerns RR estimation, end-to-end solutions appear to suffer from scarce accuracy and poor generalisation abilities. This is an issue that in this specific field (and probably, more in general, in the broader case of remote physiological measurement) appears to be exacerbated due to the dearth of large and diverse datasets for training and validation.

In this article, we have reviewed and analysed the literature and techniques for estimating respiratory information in a contactless manner using RGB cameras. Additionally, based on the analysis, a benchmark for the most representative approaches has been conducted on three publicly available datasets. The code to reproduce the experiments is made publicly available to the scientific community in the form of a Python package named resPyre . We hope this work serves as a valuable guide for practitioners facing the challenges of RGB-based respiratory information estimation, its understanding, and benchmarking.

References

[1] Yusuke Akamatsu, Terumi Umematsu, and Hitoshi Imaoka. 2023. Calibrationphys: Self-supervised video-based heart and respiratory rate measurements by calibrating between multiple cameras. *IEEE Journal of Biomedical and Health Informatics* 28, 3 (2023), 1460–1471.

[2] Ali Al-Naji and Javaan Chahl. 2016. Remote respiratory monitoring system based on developing motion magnification technique. *Biomedical Signal Processing and Control* 29 (2016), 1–10. DOI: https://doi.org/10.1016/j.bspc.2016.05.002

[3] Ali Al-Naji and Javaan Chahl. 2017. Simultaneous tracking of cardiorespiratory signals for multiple persons using a machine vision system with noise artifact removal. *IEEE Journal of Translational Engineering in Health and Medicine* 5 (2017), 1–10. DOI: https://doi.org/10.1109/JTEHM.2017.2757485

[4] Ali Al-Naji, Sang-Heon Lee, and Javaan Chahl. 2017. Quality index evaluation of videos based on fuzzy interface system. *IET Image Processing* 11, 5 (2017), 292–300.

[5] Ali Al-Naji, Asanka G. Perera, and Javaan Chahl. 2017. Remote monitoring of cardiorespiratory signals from a hovering unmanned aerial vehicle. *Biomedical Engineering Online* 16, 1 (2017), 1–20.

[6] Mohamed Ali, Ali Elsayed, Arnaldo Mendez, Yvon Savaria, and Mohamad Sawan. 2021. Contact and remote breathing rate monitoring techniques: A review. *IEEE Sensors Journal* 21, 13 (2021), 14569–14586.

[7] Davide Alinovi, Luca Cattani, Gianluigi Ferrari, Francesco Pisani, and Riccardo Raheli. 2015. Spatio-temporal video processing for respiratory rate estimation. In *Proceedings of the 2015 IEEE International Symposium on Medical Measurements and Applications*.

[8] Davide Alinovi, Gianluigi Ferrari, Francesco Pisani, and Riccardo Raheli. 2016. Respiratory rate monitoring by maximum likelihood video processing. In *2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. 172–177. DOI: https://doi.org/10.1109/ISSPIT.2016.7886029

[9] Davide Alinovi, Gianluigi Ferrari, Francesco Pisani, and Riccardo Raheli. 2018. Respiratory rate monitoring by video processing using local motion magnification. In *Proceedings of the 2018 26th European Signal Processing Conference*.

[10] Luca Antognoli, Paolo Marchionni, Stefano Nobile, Virginio Paolo Carnielli, and Lorenzo Scalise. 2018. Assessment of cardio-respiratory rates by non-invasive measurement methods in hospitalized preterm neonates. In *Proceedings of the 2018 IEEE International Symposium on Medical Measurements and Applications*.

[11] Luca Antognoli, Paolo Marchionni, Susanna Spinsante, Stefano Nobile, Virgilio Paolo Carnielli, and Lorenzo Scalise. 2019. Enanced video heart rate and respiratory rate evaluation: Standard multiparameter monitor vs clinical confrontation in newborn patients. In *Proceedings of the 2019 IEEE International Symposium on Medical Measurements and Applications*.

[12] Sean Bae, Silviu Borac, Yunus Emre, Jonathan Wang, Jiang Wu, Mehr Kashyap, Si-Hyuck Kang, Liwen Chen, Melissa Moran, Julie Cannon, et al. 2022. Prospective validation of smartphone-based heart rate and respiratory rate measurement algorithms. *Communications Medicine* 2, 1 (2022), 1–10.

[13] Marek Bartula, Timo Tigges, and Jens Muehlsteff. 2013. Camera-based system for contactless monitoring of respiration. In *Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

[14] Giuseppe Boccignone, Sathya Bursic, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, Raffaella Lanzarotti, and Sabrina Patania. 2022. Deepfakes have no heart: A simple rppg-based method to reveal fake videos. In *Proceedings of the International Conference on Image Analysis and Processing*. Springer, 186–195.

[15] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, and Raffaella Lanzarotti. 2020. An open framework for remote-PPG methods and their assessment. *IEEE Access* 8 (2020), 216083–216103. DOI: https://doi.org/10.1109/ACCESS.2020.3040936

[16] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, and Raffaella Lanzarotti. 2025. Enhancing rPPG pulse-signal recovery by facial sampling and PSD Clustering. *Biomedical Signal Processing and Control* 101 (2025), 107158. DOI: https://doi.org/10.1016/j.bspc.2024.107158

[17] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, Raffaella Lanzarotti, and Edoardo Mortara. 2022. pyVHR: A Python framework for remote photoplethysmography. *PeerJ Comput. Sci.* 8 (2022), e929. DOI: https://doi.org/10.7717/PEERJ-CS.929

[18] Giuseppe Boccignone, Alessandro D'Amelio, Omar Ghezzi, Giuliano Grossi, and Raffaella Lanzarotti. 2023. An evaluation of non-contact photoplethysmography-based methods for remote respiratory rate estimation. *Sensors* 23, 7 (2023), 3387.

[19] Frédéric Bousefsaf, Choubeila Maaoui, and Alain Pruski. 2013. Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate. *BSPC* 8, 6 (2013), 568–574.

[20] Johannes Brandstetter, Rianne van den Berg, Max Welling, and Jayesh K. Gupta. 2023. Clifford neural layers for PDE modeling. In *Proceedings of the 11th International Conference on Learning Representations*.

[21] Fabian Braun, Alia Lemkaddem, Virginie Moser, Stephan Dasen, Olivier Grossenbacher, and Mattia Bertschi. 2018. Contactless respiration monitoring in real-time via a video camera. In *EMBEC & NBC 2017*. Springer Singapore, Singapore, 567–570.

[22] Jorge Brieva, Hiram Ponce, and Ernesto Moya-Albor. 2020. A contactless respiratory rate estimation method using a hermite magnification technique and convolutional neural networks. *Applied Sciences* 10, 2 (2020), 607.

[23] Constantino Álvarez Casado, Manuel Lage Cañellas, and Miguel Bordallo López. 2023. Depression recognition using remote photoplethysmography from facial videos. *IEEE Transactions on Affective Computing* 14, 4 (2023), 3305–3316.

[24] Constantino Alvarez Casado and Miguel Bordallo López. 2023. Face2PPG: An unsupervised pipeline for blood volume pulse extraction from faces. *IEEE Journal of Biomedical and Health Informatics* 27, 11 (2023), 5530–5541.

[25] L. Cattani, D. Alinovi, Giorgio Ferrari, R. Raheli, E. Pavlidis, C. Spagnoli, and F. Pisani. 2014. A wire-free, non-invasive, low-cost video processing-based approach to neonatal apnoea detection. In *Proceedings of the 2014 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications*. 67–73.

[26] Luca Cattani, Davide Alinovi, Gianluigi Ferrari, Riccardo Raheli, Elena Pavlidis, Carlotta Spagnoli, and Francesco Pisani. 2017. Monitoring infants by automatic video processing: A unified approach to motion analysis. *Computers in Biology and Medicine* 80 (2017), 158–165. DOI: https://doi.org/10.1016/j.compbiomed.2016.11.010

[27] Junyi Chai, Hao Zeng, Anming Li, and Eric W.T. Ngai. 2021. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications* 6 (2021), 100134. DOI: https://doi.org/10.1016/j.mlwa.2021.100134

[28] Sitthichok Chaichulee, Mauricio Villarroel, Joao Jorge, Carlos Arteta, Gabrielle Green, Kenny McCormick, Andrew Zisserman, and Lionel Tarassenko. 2017. Multi-task convolutional neural network for patient detection and skin segmentation in continuous non-contact vital sign monitoring. In *Proceedings of the 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition*. 266–272.

[29] Sitthichok Chaichulee, Mauricio Villarroel, João Jorge, Carlos Arteta, Kenny McCormick, Andrew Zisserman, and Lionel Tarassenko. 2019. Cardio-respiratory signal extraction from video camera data for continuous non-contact vital sign monitoring using deep learning. *Phys. Measurement* 40, 11 (2019), 115001.

[30] Peter H. Charlton, Timothy Bonnici, Lionel Tarassenko, David A. Clifton, Richard Beale, and Peter J. Watkinson. 2016. An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. *Phys. Measurement* 37, 4 (2016), 610.

[31] Avishek Chatterjee, AP Prathosh, and Pragathi Praveena. 2016. Real-time respiration rate measurement from thoracoabdominal movement with a consumer grade camera. In *Proceedings of the 2013 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

[32] Huahua Chen, Xiang Zhang, Zongheng Guo, Na Ying, Meng Yang, and Chunsheng Guo. 2024. ACTNet: Attention based CNN and transformer network for respiratory rate estimation. *Biomedical Signal Processing and Control* 96 (2024), 106497. DOI: https://doi.org/10.1016/j.bspc.2024.106497

[33] Jerry Chen, Maysam Abbod, and Jiann-Shing Shieh. 2021. Pain and stress detection using wearable sensors and devices–A review. *Sensors* 21, 4 (2021), 1030.

[34] Mingliang Chen, Qiang Zhu, Harrison Zhang, Min Wu, and Quanzeng Wang. 2019. Respiratory rate estimation from face videos. In *Proceedings of the 2019 IEEE EMBS International Conference on Biomedical and Health Informatics*. 1–4.

[35] Weixuan Chen and Daniel McDuff. 2018. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision*.

[36] Weixuan 'Vincent' Chen and Daniel J. McDuff. 2018. DeepPhys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the European Conference on Computer Vision*.

[37] Chun-Hong Cheng, Kwan-Long Wong, Jing-Wei Chin, Tsz-Tai Chan, and Richard H. Y. So. 2021. Deep learning methods for remote heart rate measurement: A review and future research agenda. *Sensors* 21, 18 (2021), 6296.

[38] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2020. FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. DOI: https://doi.org/10.1109/TPAMI.2020.3009287

[39] Juan-Carlos Cobos-Torres, Mohamed Abderrahim, and José Martínez-Orgado. 2018. Non-contact, simple neonatal monitoring by photoplethysmography. *Sensors* 18, 12 (2018), 4362.

[40] Joaquim Comas, Adria Ruiz, and Federico Sukno. 2024. Deep pulse-signal magnification for remote heart rate estimation in compressed videos. arXiv:2405.02652. Retrieved from https://arxiv.org/abs/2405.02652

[41] Alessandro D'Amelio, Raffaella Lanzarotti, Sabrina Patania, Giuliano Grossi, Vittorio Cuculo, Andrea Valota, and Giuseppe Boccignone. 2023. On using rPPG signals for DeepFake detection: A cautionary note. In *Proceedings of the International Conference on Image Analysis and Processing*.

[42] Gerard De Haan and Vincent Jeanne. 2013. Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering* 60, 10 (2013), 2878–2886.

[43] G. de Haan and A. van Leest. 2014. Improved motion robustness of remote-PPG by using the blood volume pulse signature. *Physics and Measurement* 35, 9 (2014), 1913–1926.

[44] Alexandra Dunaeva, Daria Konovalova, and Victor Kostousov. 2020. Video analysis methods for remote measurement of respiration characteristics and heart rate variability. In *Proceedings of the 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology*. IEEE, 0171–0174.

[45] Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, and Dale Schuurmans. 2013. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PloS one* 8, 10 (2013), e76585.

[46] James B. Elsner and Anastasios A. Tsonis. 1996. *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Springer Science and Business Media.

[47] Justin R. Estepp, Ethan B. Blackford, and Christopher M. Meier. 2014. Recovering pulse rate during motion artifact with a multi-imager array for non-contact imaging photoplethysmography. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 1462–1469. DOI: https://doi.org/10.1109/SMC.2014.6974121

[48] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Proceedings of the Scandinavian Conference on Image Analysis*.

[49] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U. Rajendra Acharya. 2018. Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine* 161 (2018), 1–13. DOI: https://doi.org/10.1016/j.cmpb.2018.04.005

[50] Marc-André Fiedler, Micha Rapczyński, and Ayoub Al-Hamadi. 2020. Fusion-based approach for respiratory rate recognition from facial video images. *IEEE Access* 8 (2020), 130036–130047. DOI: https://doi.org/10.1109/ACCESS.2020.3008687

[51] Marc-André Fiedler, Philipp Werner, Michał Rapczyński, and Ayoub Al-Hamadi. 2023. Deep face segmentation for improved heart and respiratory rate estimation from videos. *Journal of Ambient Intelligence and Humanized Computing* 14, 7 (2023), 9383–9402.

[52] Gaddisa Olani Ganfure. 2019. Using video stream for continuous monitoring of breathing rate for general setting. *Signal, Image and Video Processing* 13, 7 (2019), 1395–1403.

[53] Omar Ghezzi, Giuseppe Boccignone, Giuliano Grossi, Raffaella Lanzarotti, and Alessandro D'Amelio. 2025. CliffPhys: Camera-based respiratory measurement using clifford neural networks. In *Proceedings of the European Conference on Computer Vision*.

[54] John Gideon and Simon Stent. 2021. The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[55] Google. 2021. *MediaPipe Face Mesh*. Retrieved from https://google.github.io/mediapipe/solutions/face_mesh. Access Date: 22/10/2025.

[56] Migyeong Gwak, Korosh Vatanparvar, Jilong Kuang, and Alex Gao. 2022. Motion-based respiratory rate estimation with motion artifact removal using video of face and upper body. In *Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*.

[57] M. A. Hassan, A. S. Malik, D. Fofi, N. Saad, and F. Meriaudeau. 2017. Novel health monitoring method using an RGB camera. *Biomedical Optics Express* 8, 11 (2017), 4838–4854.

[58] Alberto Hernando, María Dolores Peláez-Coca, M. T. Lozano, J. Lázaro, and E. Gil. 2019. Finger and forehead PPG signal comparison for respiratory rate estimation. *Physics and Measurement* 40, 9 (2019), 095007.

[59] Alrick B Hertzman. 1937. Photoelectric plethysmography of the fingers and toes in man. *Society for Experimental Biology and Medicine* 37, 3 (1937), 529–534.

[60] Guillaume Heusch, André Anjos, and Sébastien Marcel. 2017. A reproducible study on remote heart rate measurement. arXiv:1709.00962. Retrieved from https://arxiv.org/abs/1709.00962

[61] Guillaume Heusch, André Anjos, and Sébastien Marcel. 2017. A reproducible study on remote heart rate measurement. arXiv:1709.00962. Retrieved from https://arxiv.org/abs/1709.00962

[62] Weili Hong, Arul Earnest, Papia Sultana, Zhixiong Koh, Nur Shahidah, and Marcus Eng Hock Ong. 2013. How accurate are vital signs in predicting clinical outcomes in critically ill emergency department patients. *European Journal of Emergency Medicine* 20, 1 (2013), 27–32.

[63] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Series A: Mathematical, Physical and Engineering Sciences* 454, 1971 (1998), 903–995.

[64] Ruqiang Huang, Weihua Su, Rui Jian, Shiguo Li, Pengfa Xie, Shiyue Zhang, and Wei Qin. 2019. Non-contact vital signals measurement on a mobile rescue robot. In *Proceedings of the 2019 3rd International Conference on Data Science and Business Analytics*.

[65] Yukai Huang, Dongmin Huang, Jia Huang, Guowei Wang, Liping Pan, Hongzhou Lu, Min He, and Wenjin Wang. 2024. Camera-based blood pressure monitoring based on multisite and multiwavelength pulse transit time features. *IEEE Transactions on Instrumentation and Measurement* 73 (2024), 1–14. DOI: https://doi.org/10.1109/TIM.2024.3457944

[66] Chaoyang Huo, Pengbo Yin, and Bo Fu. 2024. MultiPhys: Heterogeneous fusion of mamba and transformer for video-based multi-task physiological measurement. *Sensors* 25, 1 (2024), 100.

[67] Junhwa Hur and Stefan Roth. 2019. Iterative residual refinement for joint optical flow and occlusion estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[68] Hyeon-Sang Hwang and Eui-Chul Lee. 2021. Non-contact respiration measurement method based on RGB camera using 1D convolutional neural networks. *Sensors* 21, 10 (2021), 3456.

[69] Luca Iozza, Jesús Lázaro, Luca Cerina, Davide Silvestri, Luca Mainardi, Pablo Laguna, and Eduardo Gil. 2019. Monitoring breathing rate by fusing the physiological impact of respiration on video-photoplethysmogram with head movements. *Physics and Measurement* 40, 9 (2019), 094002.

[70] Prasara Jakkaew and Takao Onoye. 2019. An approach to non-contact monitoring of respiratory rate and breathing pattern based on slow motion images. In *Proceedings of the 2019 IEEE International Conference on Consumer Electronics-Asia*. 47–51.

[71] Rik Janssen, Wenjin Wang, Andreia Moço, and Gerard De Haan. 2015. Video-based respiration monitoring with automatic region of interest detection. *Physics and Measurement* 37, 1 (2015), 100.

[72] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. 2021. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[73] Joao Jorge, Mauricio Villarroel, Sitthichok Chaichulee, Alessandro Guazzi, Sara Davis, Gabrielle Green, Kenny McCormick, and Lionel Tarassenko. 2017. Non-contact monitoring of respiration in the neonatal intensive care unit. In *Proceedings of the 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition*. 286–293.

[74] João Jorge, Mauricio Villarroel, Sitthichok Chaichulee, Kenny McCormick, and Lionel Tarassenko. 2018. Data fusion for improved camera-based detection of respiration in neonates. In *Optical Diagnostics and Sensing XVIII: Toward Point-of-Care Diagnostics*. SPIE, 215–224.

[75] Walter Karlen, Ainara Garde, Dorothy Myers, Cornie Scheffer, J. Mark Ansermino, and Guy A. Dumont. 2015. Estimation of respiratory rate from photoplethysmographic imaging videos compared to pulse oximetry. *IEEE Journal of Biomedical and Health Informatics* 19, 4 (2015), 1331–1338.

[76] Walter Karlen, Srinivas Raman, J. Mark Ansermino, and Guy A. Dumont. 2013. Multiparameter respiratory rate estimation from the photoplethysmogram. *IEEE Trans. Biomedical Engineering* 60, 7 (2013), 1946–1953.

[77] Fatema-Tuz-Zohra Khanam, Asanka G .Perera, Ali Al-Naji, Kim Gibson, and Javaan Chahl. 2021. Non-contact automatic vital signs monitoring of infants in a neonatal intensive care unit based on neural networks. *Journal of Imaging* 7, 8 (2021), 122.

[78] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. on Affective Computing* 3, 1 (2011), 18–31.

[79] Dimitrios Kolosov, Vasilios Kelefouras, Pandelis Kourtessis, and Iosif Mporas. 2023. Contactless camera-based heart rate and respiratory rate monitoring using AI on hardware. *Sensors* 23, 9 (2023), 4550.

[80] Ninah Koolen, Olivier Decroupet, Anneleen Dereymaeker, Katrien Jansen, Jan Vervisch, Vladimir Matic, Bart Vanrumste, Gunnar Naulaers, Sabine Van Huffel, and Maarten De Vos. 2015. Automated respiration detection from neonatal video data. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods - Volume 2 (ICPRAM 2015)*, SCITEPRESS - Science, Lisbon, Portugal, 164–169. DOI: https://doi.org/10.5220/0005187901640169

[81] Till Kroeger, Radu Timofte, Dengxin Dai, and Luc Van Gool. 2016. Fast optical flow using dense inverse search. In *Proceedings of the European Conference on Computer Vision*.

[82] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. 2020. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems* 22, 2 (2020), 712–733.

[83] Jesús Lázaro, Eduardo Gil, Raquel Bailón, Ana Mincholé, and Pablo Laguna. 2013. Deriving respiration from photoplethysmographic pulse width. *Medical and Biological Engineering and Computing* 51, 1 (2013), 233–242.

[84] Jesús Lázaro, Yunyoung Nam, Eduardo Gil, Pablo Laguna, and Ki H. Chon. 2015. Respiratory rate derived from smartphone-camera-acquired pulse photoplethysmographic signals. *Physics and Measurement* 36, 11 (2015), 2317.

[85] Heejin Lee, Junghwan Lee, Yujin Kwon, Jiyoon Kwon, Sungmin Park, Ryanghee Sohn, and Cheolsoo Park. 2022. Multitask siamese network for remote photoplethysmography and respiration estimation. *Sensors* 22, 14 (2022), 5101.

[86] Yu-Ching Lee, Abdan Syakura, Muhammad Adil Khalil, Ching-Ho Wu, Yi-Fang Ding, and Ching-Wei Wang. 2021. A real-time camera-based adaptive breathing monitoring system. *Medical and Biological Engineering and Computing* 59, 6 (2021), 1285–1298.

[87] Shiqi Li, Haipeng Wang, Shuze Wang, and Shuai Zhang. 2020. Life detection and non-contact respiratory rate measurement in cluttered environments. *Multimedia Tools Appl.* 79, 43–44 (November 2020), 32065–32077. DOI: https://doi.org/10.1007/s11042-020-09510-4

[88] Xiaobai Li, Iman Alikhani, Jingang Shi, Tapio Seppanen, Juhani Junttila, Kirsi Majamaa-Voltti, Mikko Tulppo, and Guoying Zhao. 2018. The obf database: A large face video database for remote physiological signal measurement and atrial fibrillation detection. In *Proceedings of the Face and Gestures*.

[89] Zhengzheng Li, Jiancheng Zou, Peizhou Yan, and Don Hong. 2021. Non-contact real-time monitoring of driver's physiological parameters under ambient light condition. *Intelligent Automation and Soft Computing* 28, 3 (2021), 811–822.

[90] Kuan-Yi Lin, Duan-Yu Chen, and Wen-Jiin Tsai. 2016. Image-based motion-tolerant remote respiratory rate evaluation. *IEEE Sensors Journal* 16, 9 (2016), 3263–3271.

[91] Tianqi Liu, Hanguang Xiao, Yisha Sun, Yulin Li, Shiyi Zhao, Zhenyu Yi, and Aohui Zhao. 2025. Style-rPPG: Exploration and analysis of style transfer in unsupervised remote physiological measurement. *Expert Systems with Applications* 269 (2025), 126310. DOI: https://doi.org/10.1016/j.eswa.2024.126310

[92] Tianqi Liu, Hanguang Xiao, Yisha Sun, Kun Zuo, Zhipeng Li, Zhiying Yang, and Shihong Liu. 2025. PhysKANNet: A KAN-based model for multiscale feature extraction and contextual fusion in remote physiological measurement. *Biomedical Signal Processing and Control* 100 (2025), 107111. DOI: https://doi.org/10.1016/j.bspc.2024.107111

[93] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. 2020. Multi-task temporal shift attention networks for on-device contactless vitals measurement. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*. Curran Associates Inc., Vancouver, BC, Canada.

[94] Xin Liu, Yuting Zhang, Zitong Yu, Hao Lu, Huanjing Yue, and Jingyu Yang. 2024. rPPG-MAE: Self-supervised pretraining with masked autoencoders for remote physiological measurements. *IEEE Transactions on Multimedia* 26 (2024), 7278–7293. DOI: https://doi.org/10.1109/TMM.2024.3363660

[95] Ilde Lorato, Sander Stuijk, Mohammed Meftah, Wim Verkruijsse, and Gerard De Haan. 2019. Camera-based on-line short cessation of breathing detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

[96] Hao Lu, Hu Han, and S. Kevin Zhou. 2021. Dual-gan: Joint bvp and noise modeling for remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[97] Duncan Luguern, Yannick Benezeth, Virginie Moser, L. Andrea Dunbar, Fabian Braun, Alia Lemkaddem, Keisuke Nakamura, Randy Gomez, and Julien Dubois. 2020. Remote photoplethysmography combining color channels with SNR maximization for respiratory rate assessment. In *Proceedings of the 2020 14th International Symposium on Medical Information Communication Technology*. 1–6.

[98] Duncan Luguern, Richard Macwan, Yannick Benezeth, Virginie Moser, L. Andrea Dunbar, Fabian Braun, Alia Lemkaddem, and Julien Dubois. 2021. Wavelet Variance Maximization: A contactless respiration rate estimation method based on remote photoplethysmography. *Biomedical Signal Processing and Control* 63 (2021), 102263. DOI: https://doi.org/10.1016/j.bspc.2020.102263

[99] Duncan Luguern, Simon Perche, Yannick Benezeth, Virginie Moser, L. Andrea Dunbar, Fabian Braun, Alia Lemkaddem, Keisuke Nakamura, Randy Gomez, and Julien Dubois. 2020. An assessment of algorithms to estimate respiratory rate from the remote Photoplethysmogram. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1232–1241. DOI: https://doi.org/10.1109/CVPRW50498.2020.00160

[100] Tomáš Lukáč, Jozef Púčik, and Lukáš Chrenko. 2014. Contactless recognition of respiration phases using web camera. In *2014 24th International Conference Radioelektronika*. 1–4. DOI: https://doi.org/10.1109/Radioelek.2014.6828427

[101] Richard Macwan, Serge Bobbia, Yannick Benezeth, Julien Dubois, and Alamin Mansouri. 2018. Periodic variance maximization using generalized eigenvalue decomposition applied to remote photoplethysmography estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[102] Carlo Massaroni, Daniela Lo Presti, Domenico Formica, Sergio Silvestri, and Emiliano Schena. 2019. Non-contact monitoring of breathing pattern and respiratory rate via RGB signal measurement. *Sensors* 19, 12 (2019), 2758.

[103] Carlo Massaroni, Andrea Nicolo, Massimo Sacchetti, and Emiliano Schena. 2020. Contactless methods for measuring respiratory rate: A review. *IEEE Sensors Journal* 21, 11 (2020), 12821–12839.

[104] Carlo Massaroni, Andrea Nicolò, Emiliano Schena, and Massimo Sacchetti. 2020. Remote respiratory monitoring in the time of COVID-19. *Frontiers in Physiology Volume* 11-2020, (2020). DOI: https://doi.org/10.3389/fphys.2020.00635

[105] Carlo Massaroni, Emiliano Schena, Sergio Silvestri, Fabrizio Taffoni, and Mario Merone. 2018. Measurement system based on RBG camera signal for contactless breathing pattern and respiratory rate monitoring. In *Proceedings of the 2018 IEEE International Symposium on Medical Measurements and Applications*.

[106] Marc Mateu-Mateus, Federico Guede-Fernández, Miguel ángel García-González, Juan José Ramos-Castro, and Mireya Fernández-Chimeno. 2020. Camera-based method for respiratory rhythm extraction from a lateral perspective. *IEEE Access* 8 (2020), 154924–154939. DOI: https://doi.org/10.1109/ACCESS.2020.3018616

[107] M. Mateu-Mateus, F. Guede-Fernández, N. Rodriguez-Ibáñez, M.A. García-González, J. Ramos-Castro, and M. Fernández-Chimeno. 2021. A non-contact camera-based method for respiratory rhythm extraction. *Biomedical Signal Processing and Control* 66 (2021), 102443. DOI: https://doi.org/10.1016/j.bspc.2021.102443

[108] Daniel McDuff. 2023. Camera measurement of physiological vital signs. *Computing Surveys* 55, 9 (2023), 1–40.

[109] Daniel McDuff and Ethan Blackford. 2019. iPhys: An open non-contact imaging-based physiological measurement toolbox. In *Proceedings of the 2019 41th Annual International Conference of the IEEE Engineering in Medicine & Biology Society*.

[110] Daniel McDuff, Miah Wander, Xin Liu, Brian L. Hill, Javier Hernandez, Jonathan Lester, and Tadas Baltrusaitis. 2022. SCAMPS: synthetics for camera measurement of physiological signals. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS'22)*. Curran Associates Inc., New Orleans, LA, USA.

[111] Daniel J. McDuff, Ethan B. Blackford, and Justin R. Estepp. 2017. The impact of video compression on remote cardiac pulse measurement using imaging photoplethysmography. In *Proceedings of the Face and Gestures*.

[112] Arya Deo Mehta and Hemant Sharma. 2020. Tracking nostril movement in facial video for respiratory rate estimation. In *Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies*. 1–6.

[113] David J. Meredith, D. Clifton, Peter Charlton, J. Brooks, C. W. Pugh, and L. Tarassenko. 2012. Photoplethysmographic derivation of respiratory rate: A review of relevant physiology. *Journal of Medical Engineering and Technology* 36, 1 (2012), 1–7.

[114] Leila Mirmohamadsadeghi, Sibylle Fallet, Virginie Moser, Fabian Braun, and Jean-Marc Vesin. 2016. Real-time respiratory rate estimation using imaging photoplethysmography inter-beat intervals. In *Proceedings of the 2016 Computing in Cardiology Conference*. 861–864.

[115] Nunzia Molinaro, Emiliano Schena, Sergio Silvestri, and Carlo Massaroni. 2022. Multi-ROI spectral approach for the continuous remote cardio-respiratory monitoring from mobile device built-in cameras. *Sensors* 22, 7 (2022), 2539.

[116] Kazuki Nakajima, Yoshiaki Matsumoto, and Toshiyo Tamura. 2001. Development of real-time image sequence analysis for evaluating posture change and respiratory rate of a subject in bed. *Physics and Measurement* 22, 3 (2001), N21.

[117] Yunyoung Nam, Youngsun Kong, Bersain Reyes, Natasa Reljin, and Ki H. Chon. 2016. Monitoring of heart and breathing rates using dual cameras on a smartphone. *PloS One* 11, 3 (2016), e0151013.

[118] Yunyoung Nam, Jinseok Lee, and Ki H. Chon. 2014. Respiratory rate estimation from the built-in cameras of smartphones and tablets. *Annals of Biomedical Engineering* 42, 4 (2014), 885–898.

[119] Girish Narayanswamy, Yujia Liu, Yuzhe Yang, Chengqian Ma, Xin Liu, Daniel McDuff, and Shwetak Patel. 2024. BigSmall: Efficient multi-task learning for disparate spatial and temporal physiological measurements. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE Computer Society, Los Alamitos, CA, USA, 7899–7909. DOI: https://doi.org/10.1109/WACV57701.2024.00773

[120] Aoxin Ni, Arian Azarang, and Nasser Kehtarnavaz. 2021. A review of deep learning-based contactless heart rate measurement methods. *Sensors* 21, 11 (2021), 3719.

[121] Xuesong Niu, Zitong Yu, Hu Han, Xiaobai Li, Shiguang Shan, and Guoying Zhao. 2020. Video-based remote physiological measurement via cross-verified feature disentangling. In *Proceedings of the European Conference on Computer Vision*.

[122] Teruaki Nochino, Yuko Ohno, and Shima Okada. 2017. Development of noncontact respiration monitoring method with web-camera during sleep. In *Proceedings of the 2017 IEEE 6th Global Conference on Consumer Electronics* 1–2.

[123] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* 32, 2 (2020), 604–624.

[124] Christian S. Pilz, Sebastian Zaunseder, Jarek Krajewski, and Vladimir Blazek. 2018. Local group invariance for heart rate estimation from face videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[125] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. 2010. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering* 58, 1 (2010), 7–11.

[126] Ming-Zher Poh, Daniel J. McDuff, and Rosalind W. Picard. 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express* 18, 10 (2010), 10762–10774.

[127] Jafar Pourbemany, Almabrok Essa, and Ye Zhu. 2021. Real-time video-based heart and respiration rate monitoring. In *Proceedings of the NAECON 2021-IEEE National Aerospace and Electronics Conference*. 332–336.

[128] A. P. Prathosh, Pragathi Praveena, Lalit K. Mestha, and Sanjay Bharadwaj. 2017. Estimation of respiratory pattern from video using selective ensemble aggregation. *IEEE Transactions on Signal Processing* 65, 11 (2017), 2902–2916.

[129] Andrew J. Quinn, Vitor Lopes-dos Santos, David Dupret, Anna Christina Nobre, and Mark W. Woolrich. 2021. EMD: Empirical mode decomposition and hilbert-huang spectral analyses in python. *Journal of Open Source Software* 6, 59 (2021), 2977. DOI: https://doi.org/10.21105/joss.02977

[130] Yuzhuo Ren, Braeden Syrnyk, and Niranjan Avadhanam. 2021. Dual attention network for heart rate and respiratory rate estimation. In *Proceedings of the 2021 IEEE 23rd International Workshop on Multimedia Signal Processing*. IEEE, 1–6.

[131] Yuzhuo Ren, Braeden Syrnyk, and Niranjan Avadhanam. 2022. Improving video-based heart rate and respiratory rate estimation via pulse-respiration quotient. In *Proceedings of the Workshop on Healthcare AI and COVID-19*. PMLR, 136–145.

[132] Ambareesh Revanur, Ananyananda Dasari, Conrad S. Tucker, and László A. Jeni. 2022. Instantaneous physiological estimation using video transformers. In *Proceedings of the Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*. Springer, 307–319.

[133] Ambareesh Revanur, Zhihua Li, Umur A. Ciftci, Lijun Yin, and László A Jeni. 2021. The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[134] Bersain A. Reyes, Natasa Reljin, Youngsun Kong, Yunyoung Nam, and Ki H. Chon. 2016. Tidal volume and instantaneous respiration rate estimation using a volumetric surrogate signal acquired via a smartphone camera. *IEEE Journal of Biomedical and Health Informatics* 21, 3 (2016), 764–777.

[135] Chiara Romano, Emiliano Schena, Sergio Silvestri, and Carlo Massaroni. 2021. Non-contact respiratory monitoring using an RGB camera for real-world applications. *Sensors* 21, 15 (2021), 5126.

[136] Scott L. Rossol, Jeffrey K. Yang, Caroline Toney-Noland, Janine Bergin, Chandan Basavaraju, Pavan Kumar, and Henry C. Lee. 2020. Non-contact video-based neonatal respiratory monitoring. *Children* 7, 10 (2020), 171.

[137] Philipp V. Rouast, Marc T. P. Adam, Raymond Chiong, David Cornforth, and Ewa Lux. 2018. Remote heart rate measurement using low-cost RGB face video: A technical literature review. *Frontiers of Computer Science* 12, 5 (2018), 858–872.

[138] Shourjya Sanyal and Koushik Kumar Nundy. 2018. Algorithms for monitoring heart rate and respiratory rate from the video of a user's face. *IEEE Journal of Translational Engineering in Health and Medicine* 6 (2018), 1–11. DOI: https://doi.org/10.1109/JTEHM.2018.2818687

[139] Fabian Schrumpf, Christoph Mönch, Gerold Bausch, and Mirco Fuchs. 2019. Exploiting weak head movements for camera-based respiration detection. In *Proceedings of the 2022 41th Annual International Conference of the IEEE Engineering in Medicine & Biology Society*.

[140] Jay Schulkin and Peter Sterling. 2019. Allostasis: A brain-centered, predictive mode of physiological regulation. *Trends in Neurosciences* 42, 10 (2019), 740–752.

[141] Christopher G. Scully, Jinseok Lee, Joseph Meyer, Alexander M. Gorbach, Domhnull Granquist-Fraser, Yitzhak Mendelson, and Ki H. Chon. 2011. Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Transactions on Biomedical Engineering* 59, 2 (2011), 303–306.

[142] Eli Sennesh, Jordan Theriault, Dana Brooks, Jan-Willem van de Meent, Lisa Feldman Barrett, and Karen S. Quigley. 2022. Interoception as modeling, allostasis as control. *Biological Psychology* 167 (2022), 108242.

[143] Dangdang Shao, Yuting Yang, Chenbin Liu, Francis Tsow, Hui Yu, and Nongjian Tao. 2014. Noncontact monitoring breathing pattern, exhalation flow rate and pulse transit time. *IEEE Trans. Biomedical Engineering* 61, 11 (2014), 2760–2767.

[144] Hang Shao, Lei Luo, Jianjun Qian, Mengkai Yan, Shangbing Gao, and Jian Yang. 2025. Video-based multiphysiological disentanglement and remote robust estimation for respiration. *IEEE Transactions on Neural Networks and Learning Systems* 36, 5 (2025), 8360–8371. DOI: https://doi.org/10.1109/TNNLS.2024.3424772

[145] Shashank Sharma, Sourya Bhattacharyya, Jayanta Mukherjee, Parimal Kumar Purkait, Arunava Biswas, and Alok Kanti Deb. 2015. Automated detection of newborn sleep apnea using video monitoring system. In *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*. 1–6. DOI: https://doi.org/10.1109/ICAPR.2015.7050675

[146] Ali I. Siam, Nirmeen A. El-Bahnasawy, Ghada M. El Banby, Atef Abou Elazm, and Fathi E. Abd El-Samie. 2020. Efficient video-based breathing pattern and respiration rate monitoring for remote health monitoring. *JOSA A* 37, 11 (2020), C118–C124.

[147] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2012), 42–55. DOI: https://doi.org/10.1109/T-AFFC.2011.25

[148] Jeremy Speth, Nathan Vance, Patrick Flynn, and Adam Czajka. 2023. Non-contrastive unsupervised learning of physiological signals from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[149] Jeremy Speth, Nathan R. Vance, Benjamin Sporrer, Lu Niu, Patrick Flynn, and Adam Czajka. 2024. MSPM: A multisite physiological monitoring dataset for remote pulse, respiration, and blood pressure estimation. *IEEE Transactions on Instrumentation and Measurement* 73 (2024), 1–14. DOI: https://doi.org/10.1109/TIM.2024.3476556

[150] Ronny Stricker, Steffen Müller, and Horst-Michael Gross. 2014. Non-contact video-based pulse rate measurement on a mobile service robot. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 1056–1062.

[151] Xiuchao Sui, Shaohua Li, Xue Geng, Yan Wu, Xinxing Xu, Yong Liu, Rick Siow Mong Goh, and Hongyuan Zhu. 2022. CRAFT: Cross-Attentional Flow Transformers for Robust Optical Flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[152] Zhaodong Sun and Xiaobai Li. 2022. Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast. In *Proceedings of the European Conference on Computer Vision*. Springer, 492–510.

[153] Zhaodong Sun and Xiaobai Li. 2024. Contrast-Phys+: Unsupervised and weakly-supervised video-based remote physiological measurement via spatiotemporal contrast. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 46, 08 (August 2024), 5835–5851. DOI: https://doi.org/10.1109/TPAMI.2024.3367910

[154] Nor Surayahani Suriani, Nur Syahida Shahdan, Nan Md. Sahar, and Nik Shahidah Afifi Md. Taujuddin. 2022. Non-contact facial based vital sign estimation using convolutional neural network approach. *International Journal of Advanced Computer Science and Applications* 13, 5 (2022). DOI: https://doi.org/10.14569/IJACSA.2022.0130546

[155] Chihiro Takano and Yuji Ohta. 2007. Heart rate measurement based on a time-lapse image. *Medical Engineering and Physics* 29, 8 (2007), 853–857.

[156] K. Song Tan, Reza Saatchi, Heather Elphick, and Derek Burke. 2010. Real-time vision based respiration monitoring system. In *Proceedings of the CSNDSP*.

[157] Lionel Tarassenko, Mauricio Villarroel, Alessandro Guazzi, Joao Jorge, DA Clifton, and Chris Pugh. 2014. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. *Physics and Measurement* 35, 5 (2014), 807.

[158] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*.

[159] Daniel Myklatun Tveit, Kjersti Engan, Ivar Austvoll, and Øyvind Meinich-Bache. 2016. Motion based detection of respiration rate in infants using video. In *Proceedings of the ICIP*.

[160] Anton M. Unakafov. 2018. Pulse rate estimation using imaging photoplethysmography: Generic framework and comparison of methods on a publicly available dataset. *Biomedical Physics and Engineering Express* 4, 4 (2018), 045001.

[161] Vidyadhar Upadhya, Avishek Chatterjee, A. P. Prathosh, and Pragathi Praveena. 2016. Respiration monitoring through thoraco-abdominal video with an LSTM. In *Proceedings of the 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering*. IEEE, 165–171.

[162] Mark Van Gastel, Sander Stuijk, and Gerard de Haan. 2016. Robust respiration detection from remote photoplethysmography. *Biomedical Optics Express* 7, 12 (2016), 4941–4957.

[163] Wim Verkruysse, Lars O. Svaasand, and J. Stuart Nelson. 2008. Remote plethysmographic imaging using ambient light. *Optics Express* 16, 26 (2008), 21434–21445.

[164] Mauricio Villarroel, Sitthichok Chaichulee, João Jorge, Sara Davis, Gabrielle Green, Carlos Arteta, Andrew Zisserman, Kenny McCormick, Peter Watkinson, and Lionel Tarassenko. 2019. Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit. *NPJ Digital Medicine* 2, 1 (2019), 1–18.

[165] Ching-Wei Wang, Andrew Hunter, Neil Gravill, and Simon Matusiewicz. 2013. Unconstrained video monitoring of breathing behavior and application to diagnosis of sleep apnea. *IEEE Transactions on Biomedical Engineering* 61, 2 (2013), 396–404.

[166] Hao Wang, Euijoon Ahn, and Jinman Kim. 2022. Self-supervised representation learning framework for remote physiological measurement using spatiotemporal augmentation loss. In *Proceedings of the AAAI*.

[167] Hengliang Wang, Kin Siu, Kihwan Ju, and Ki H Chon. 2006. A high resolution approach to estimating time-frequency spectra and their amplitudes. *Annals of Biomedical Engineering* 34, 2 (2006), 326–338.

[168] Haopeng Wang, Yufan Zhou, and Abdulmotaleb El Saddik. 2021. VitaSi: A real-time contactless vital signs estimation system. *Computers and Electrical Engineering* 95 (2021), 107392. DOI: https://doi.org/10.1016/j.compeleceng.2021.107392

[169] Jieying Wang, Caifeng Shan, Zhaoyang Liu, Shuwang Zhou, and Minglei Shu. 2025. Physiological information preserving video compression for rPPG. *IEEE Journal of Biomedical and Health Informatics* 29, 5 (2025), 3563–3575. DOI: https://doi.org/10.1109/JBHI.2025.3526837

[170] Wenjin Wang and Albertus C. den Brinker. 2022. Algorithmic insights of camera-based respiratory motion extraction. *Phys. measurement* 43, 7 (2022), 075004.

[171] Wenjin Wang, Albertus C. den Brinker, Sander Stuijk, and Gerard de Haan. 2016. Algorithmic principles of remote PPG. *IEEE Trans. Biomedical Engineering* 64, 7 (2016), 1479–1491.

[172] Bing Wei, Xuan He, Chao Zhang, and Xiaopei Wu. 2017. Non-contact, synchronous dynamic measurement of respiratory rate and heart rate based on dual sensitive regions. *Biomedical Engineering Online* 16, 1 (2017), 1–21.
[173] Stefan Wiesner and Ziv Yaniv. 2007. Monitoring patient respiration using a single optical camera. (2007), 2740–2743.
[174] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. 2012. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics* 31, 4 (2012), 1–8.
[175] Taihong Xiao, Jinwei Yuan, Deqing Sun, Qifei Wang, Xin-Yu Zhang, Kehan Xu, and Ming-Hsuan Yang. 2020. Learnable cost volume using the cayley representation. In *Proceedings of the European Conference on Computer Vision*.
[176] Yuzhe Yang, Xin Liu, Jiang Wu, Silviu Borac, Dina Katabi, Ming-Zher Poh, and Daniel McDuff. 2022. SimPer: Simple self-supervised learning of periodic targets. In *The Eleventh International Conference on Learning Representations*.
[177] Zitong Yu, Wei Peng, Xiaobai Li, Xiaopeng Hong, and Guoying Zhao. 2019. Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
[178] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Yawen Cui, Jiehua Zhang, Philip Torr, and Guoying Zhao. 2023. Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer. *International Journal of Computer Vision* 131, 6 (2023), 1307–1330.
[179] Zitong Yu, Yuming Shen, Jingang Shi, Hengshuang Zhao, Philip Torr, and Guoying Zhao. 2022. PhysFormer: Facial Video-based physiological measurement with temporal difference transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Los Alamitos, CA, USA, 4176–4186. DOI: https://doi.org/10.1109/CVPR52688.2022.00415
[180] Zijie Yue, Miaojing Shi, and Shuai Ding. 2023. Facial video-based remote physiological measurement via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 13844–13859. DOI: https://doi.org/10.1109/TPAMI.2023.3298650
[181] Mingliang Zhai, Xuezhi Xiang, Ning Lv, and Xiangdong Kong. 2021. Optical flow and scene flow estimation: A survey. *Pattern Recognition* 114 (2021), 107861. DOI: https://doi.org/10.1016/j.patcog.2021.107861
[182] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[183] Changchen Zhao, Chun-Liang Lin, Weihai Chen, and Zhengguo Li. 2018. A novel framework for remote photoplethysmography pulse extraction on compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
[184] Xiujuan Zheng, Wenqin Yan, Boxiang Liu, Yue Ivan Wu, and Haiyan Tu. 2025. Estimation of heart rate and respiratory rate by fore-background spatiotemporal modeling of videos. *Biomedical Optics Express* 16, 2 (2025), 760–777.
[185] Sayyedjavad Ziaratnia, Tipporn Laohakangvalvit, Midori Sugaya, and Peeraya Sripian. 2024. Multimodal deep learning for remote stress estimation using CCT-LSTM. In *Proceedings of the WACV*.

Appendix

A Online Resources

Code for reproducing the experiments can be accessed freely at https://github.com/phuselab/resPyre.

Received 19 September 2023; revised 25 February 2025; accepted 8 October 2025