Revisiting motion-based respiration measurement from videos

Qi Zhan#, Jingjing Hu#, Zitong Yu, Xiaobai Li and Wenjin Wang*

Abstract— **Video-based motion analysis gave rise to contactless respiration rate monitoring that measures subtle respiratory movement from a human chest or belly. In this paper, we revisit this technology via a large video benchmark that includes six categories of practical challenges. We analyze two video properties (i.e. pixel intensity variation and pixel movement) that are essential for respiratory motion analysis and various signal extraction approaches (i.e. from conventional to recent Convolutional Neural Network (CNN)-based methods). We find that pixel movement can better quantify respiratory motion than pixel intensity variation in various conditions. We also conclude that the simple conventional approach (e.g. Zero-phase Component Analysis) can achieve better performance than CNN that uses data training to define the extraction of respiration signal, which thus raises a more general question whether CNN can improve video-based physiological signal measurement.**

I. INTRODUCTION

Respiration rate (RR) is a critical vital sign for indicating the physiological status of a person, which has been applied in health monitoring to detect sleep disorders (e.g. apnea), cardiac arrest and stroke [1]. However, conventional respiration measurements that require contact-sensors attached to the human skin, such as electrodes, a strain gauge or respiratory effort belts, are uncomfortable and cumbersome to use for long-term continuous health monitoring.

In recent years, Video-based respiration measurement has been proposed and prototyped, which measures subtle chest/belly movement by tracking pixel movement [2]–[5]. Li [2] uses optical flow to track the motion trajectories of the features in four simulated sleep scenarios (i.e. left side, right side, supine and torso obscured). In [3], the motion features of the chest are tracked by using optical flow in four recording conditions (i.e. camera distances, illumination conditions, the clothing of subjects and the view of camera). In addition, pixel intensity variation has been exploited to measure respiratory motion, assuming that chest/belly motions change light reflections of the body surface [6], [7]. Carlo [7] used the variation of light intensity to measure respiration and studied the influence of the subject's clothing and gender on respiration measurement.

Various algorithmic approaches have been developed to extract the respiration signal from the aforementioned video properties (e.g. pixel intensity variation and pixel movement). Conventional approaches such as Principal Component Analysis (PCA) and signal averaging have been used to derive a respiration signal from motion signals. More recently, Convolutional Neural Network (CNN)-based methods have been introduced to measure the respiration signal from a video in an end-to-end fashion [8]–[10]. Qayyum [8] estimated the respiration rate on the pre-trained network from the videos, where the subjects perform breathing in the complex background. In [9], an SR-DNN model enhances the low-resolution thermal sequences to measure the respiration signal from the subjects without voluntary body movements. Deepphys [10] measures the respiration signal from the videos with different levels of head motion. The architectures of above CNN-based methods are 2D-CNN where the convolution operation is on the spatial contexts, which is different from 3D-CNN where the convolution operation is on the spatio-temporal contexts.

However, these methods are validated in different yet limited experimental conditions (e.g. different camera parameters and recording protocols), which is difficult to compare and conclude their performance. Therefore in this paper, we build a unified and fair benchmark to understand their performance and try to answer two specific questions:

1) *Which video property (pixel movement or pixel intensity variation) is better for respiratory motion measurement?*
2) *Which method (Zero-phase Component Analysis (ZCA), Averaging, 2D-CNN, 3D-CNN) can better measure a respiration signal from the given video property?*

To this end, we created a video benchmark dataset that includes six categories of practical challenges. We compared four signal extraction methods on two video properties. The results show that pixel movement is more robust than pixel intensity variation in capturing subtle respiratory motions in different conditions. But both video properties are not immune to non-respiratory motions. We also find that the simple conventional approach (e.g. ZCA) can achieve a better performance than 2D/3D CNN for respiration signal extraction (i.e. the best option is pixel movement based ZCA extraction). The benchmark and insights provided in this paper would be useful for designing new methods and applications in future.

This research is supported by the National Natural Science Foundation of China (Grant No. 61671204).

Wenjin Wang* is the corresponding author.

Qi Zhan# and Jingjing Hu# contributed equally to this work.

Qi Zhan and Jingjing Hu are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, People's Republic of China. E-mail: zhanqi@hnu.edu.cn, jingjingh@hnu.edu.cn

Zitong Yu and Xiaobai Li are with the Center for Machine Vision and Signal Analysis, University of Oulu, FI-90014, Finland. E-mail: Zitong.Yu@oulu.fi, Xiaobai.Li@oulu.fi

Wenjin Wang is with the Eindhoven University of Technology, The Netherlands. E-mail: wwang@tue.nl

II. METHODS

Fig. 1 shows the benchmark system in this study. Video properties and respiration signal extraction methods are specified in the following subsections.

A. Pixel intensity variation-based methods

1) Light intensity variation: In [7], the intensity of reflected light measured by a camera contains two components: the illumination intensity and subject body reflection:

$$C(x, y, t) = I(x, y, t) * R(x, y, t), \tag{1}$$

where $C(x, y, t)$ is the reflected light intensity at $(x,\ y)$; $I(x, y, t)$ is the illumination intensity and $R(x, y, t)$ is the body reflection. If the light source remains stable, pixel intensity variation observed by the camera resembles the surface changes caused by the movement of chest/abdomen [7]. Therefore, respiration signal could be measured by tracking pixel intensity variation. According to [6], respiratory motion has strong energy on the vertical direction (y-axis) due to inhale and exhale. The intensity of the pixels is projected onto the y-axis to derive the respiration signal. Since the temporal standard deviation of the y-projected signal indicates the amount of motion, we select ten y-projected signals with the largest standard deviation to measure a respiration signal using different approaches (the Averaging approach and ZCA). For Averaging, the mean of ten y-projected signals is used as the respiration signal. For ZCA, we apply it on the ten y-projected signals to find the respiratory components. Top three periodic ZCA signals with the largest Signal-to-Noise Ratio (SNR) are selected and averaged. Note that ZCA is similar to PCA, with an extra step of back-projection of the de-mixed signals to eliminate the arbitrary sign problem of PCA [11].

2) Pixel intensity variation-based 2D/3D-CNN: To explore whether the pixel intensity variation can be used as the input of 2D/3D-CNN to measure respiration signal, we replace the input of DeepPhys (2D-CNN) in [10] with the image difference between two consecutive frames that measures intensity changes. The training label is the differentiated reference respiration signal that corresponds to the input of image difference. For 3D-CNN, we use the same architecture of PhysNet [12] that has been demonstrated for pulse extraction, which takes the input of the R-G-B channels of 128 frames of a video recorded at 20 frames per second (fps). Here we replace its training label by the reference respiration signal. Note that the mean square error is used as the loss function for both CNN methods.

B. Pixel movement-based methods

1) Optical flow: Optical flow measures the pixel displacement on a 2D plane [4]. Assuming $I(x, y, t)$ as the intensity of a pixel $(x,\ y)$ at t, it remains constant when moving to $x + \Delta x,\ y + \Delta$ at $t + \Delta t$:

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{2}.$$

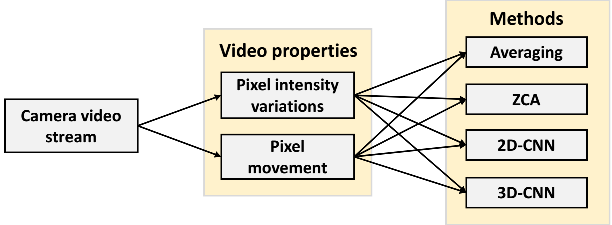


Fig. 1: Overview of the benchmark system, which compares the performance of four respiration signal extraction methods on two different video properties.

We use dense optical flow [13] to measure the pixel displacement caused by respiratory motion of chest/abdomen. Assuming respiratory motion has major energy on the vertical direction, we only use the pixel motion on the y-direction to generate motion traces, which are further used for respiratory component extraction (i.e. by Averaging or ZCA).

For the averaging method, it takes the mean of vertical motion traces as the respiration signal. For ZCA, we first down-scale the dense flow image into 10×10 pixels using the nearest-neighbor interpolation to reduce the computational complexity of ZCA. Then we use ZCA to decompose the 10×10 motion traces and select the respiratory component, in the same way as the ZCA used in pixel intensity variation.

2) Optical flow-based 2D/3D-CNN: To investigate the pixel movement-based 2D/3D-CNN, we use dense flow image as the input for DeepPhys and PhysNet, respectively. The only difference between 2D-CNN and 3D-CNN in this application is: 2D-CNN takes a single flow image between two consecutive frames as the input, while 3D-CNN takes a sequence of flow images (128 frames at 20 fps). The training labels for 2D/3D-CNN are differentiated respiration signals that associate with the same meaning of dense flow input (i.e. temporal changes of image pixels).

III. EXPERIMENT AND RESULT

A. Experimental setup

HNU respiration dataset We created a total of 66 recordings for six challenge categories: (1) Breathing patterns (Bre: deep breath, shallow breath and normal breath); (2) Illumination conditions (Ill: bright, dark and varying); (3) Postures (Pos: sitting-front, sitting-side, sitting-back, lying-side and lying-front); (4) Camera distances (Cam: 1.5 m and 2.5 m); (5) Backgrounds (Bac: simple and complex); (6) Non-respiration motion (Nrm: intentional body movement).

Six healthy adult subjects (4 females and 2 males) were guided to mimic the sinusoidal breathing pattern displayed on a frontal screen during the recording. The reference was the sinusoidal signal with the frequency between 0.167-0.33 Hz shown on the screen. Videos were recorded by a regular RGB camera (Global shutter RGB CMOS camera USB M2ST036-H from Shenzhen city Shen Technology Co. Ltd.) for one minute duration. Each video was saved in the lossless BMP format (640×480 pixels, 8-bit depth) and constant frame rate (20 fps). This study was approved by Hunan University and written consent forms were obtained from the participants.

TABLE I: Average RMSE obtained by four methods on two video properties in different challenges. **Boldface** character denotes the best result per row (challenge category).

Challenges	Pixel intensity variation-based (frames)				Pixel movement-based (frames)			
	AVG	ZCA	2D-CNN	3D-CNN	AVG	ZCA	2D-CNN	3D-CNN
Bre	35.7	30.9	53.2	30.6	4.8	**4.6**	14.4	28.4
Ill	37.5	32.6	60.8	27.8	19.3	**4.9**	35.5	18.3
Pos	42.7	44.83	42.5	31.3	10.6	**6.3**	23.5	23.3
Cam	72.4	33.9	53.9	46.3	13.9	**4.7**	35.3	21.5
Bac	81.3	34.1	44.2	44.3	**5.0**	6.9	5.7	18.6
Nrm	63.3	95.8	90.8	41.6	51.2	**10.4**	46.0	30.7
Total	55.5	45.4	57.6	37.0	17.5	**6.3**	26.7	23.5

TABLE II: Average accuracy obtained by four methods on two video properties in different challenges.

Challenges	Pixel intensity variation-based (%)				Pixel movement-based (%)			
	AVG	ZCA	2D-CNN	3D-CNN	AVG	ZCA	2D-CNN	3D-CNN
Bre	18.3	35.1	26.8	25.1	**75.8**	73.0	53.5	23.7
Ill	35.3	55.6	23.4	16.9	74.9	**76.8**	32.2	23.5
Pos	15.5	22.9	33.01	14.5	65.9	**67.3**	45.5	15.6
Cam	11.3	33.9	17.7	12.9	58.0	**70.6**	30.7	20.1
Bac	7.8	29.1	20.3	25.1	58.7	59.7	**66.0**	21.1
Nrm	4.0	0.0	23.4	12.0	40.1	**45.3**	36.9	21.2
Total	15.4	29.4	24.1	17.8	62.2	**65.5**	44.1	20.9

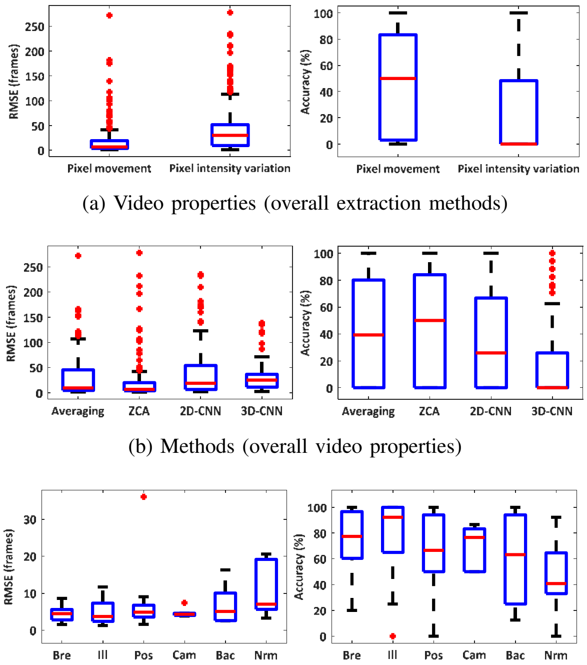
Each subject had a default recording condition: the subject performs normal breathing, sitting on a chair that is 1.5 m away from the camera, with bright illumination and simple background. For the rest recordings of this subject, we changed one challenge factor for each recording. We mention that for the CNN training, the HNU dataset is divided into six subject-independent groups for five-fold cross validation.

B. Evaluation metrics

- **Root Mean Square Error (RMSE)** The RMSE is used to measure the difference between the peak location of the measurement and reference.
- **Accuracy** It refers to the percentage where the difference between the peak location of the measurement and reference is smaller than 6 frames (20 fps).

C. Results and discussion

Tables I-II summarize the averaged evaluation results of the benchmark. It is clear that the methods using pixel movement features (e.g. optical flow) are, in general, better than using pixel intensity variation. For instance, pixel movement is more robust to the challenges such as illumination and camera distance. We expect the benefits of using pixel movement are twofold: (i) pixel intensity is affected by both the illumination changes and body motions, whereas pixel movement (measured by optical flow) that only measures body motion is robust to illumination changes; (ii) optical flow separates motions into vertical and horizontal directions. Only the vertical motion (with strong respiratory component) is used for measurement, which is in principle robust to motion disturbances on the horizontal direction. In contrast, pixel intensity variation cannot differentiate reflection changes on different directions.



(a) Video properties (overall extraction methods)

(b) Methods (overall video properties)

(c) Challenges (pixel movement + ZCA)

Fig. 2: Box-plots of RMSE and accuracy in terms of (a) video properties; (b) methods; and (c) challenge categories. In each panel, the median values are indicated by red bars inside the blue boxes, the quartile range by boxes, the full range by whiskers, the outliers by red cross.

Fig. 2 shows the statistical comparison among the benchmark in terms of video properties, extraction methods and challenge categories. Fig. 2 (a) confirms our observation that pixel movement is indeed more robust than pixel intensity variation in quantifying subtle respiratory motions. Fig. 2 (b) shows that ZCA has the overall best performance in both video properties, second by Averaging that has no respiratory component de-mixing and selection. We also see that simple and conventional solutions of Averaging and ZCA outperform 2D/3D-CNN. More specifically, 2D-CNN has larger RMSE than 3D-CNN. The reason could be that the convolution of 2D-CNN is performed on the pixel data (either intensity variations or movements) with limited temporal information (i.e. only between two consecutive frames), whereas 3D-CNN can perceive longer spatio-temporal context (with more information related to respiratory activity) by its 3D convolution. However, the accuracy of 3D-CNN is lower because of the phase shift of the produced respiration signal, which is essentially due to its internal 3D processing (i.e. artifact of the default network). In Fig. 2 (c), we can see that non-respiratory motion is the most challenging factor for the best benchmarked approach that uses pixel movement with ZCA, i.e. motion-based respiration measurement is very difficult to be motion robust.

Since simple approaches like Averaging achieve better performance than CNN for respiration signal extraction, we somehow doubt whether CNN is suitable for this assignment. CNN has shown impressive performance in computer vision tasks such as recognition and classification that use

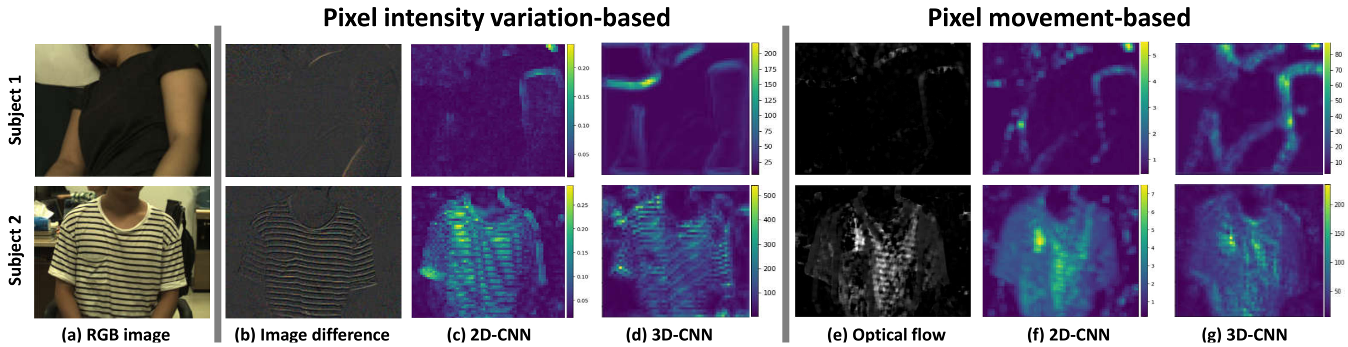


Fig. 3: Visualization of activation maps measured by 2D-CNN and 3D-CNN based on two video properties. Two subjects are exemplified: top row — the subject wears textureless cloth; bottom row — the subject breaths in a complex background.

image contexts/features to differentiate objects or activities. However, human physiology is not appearance dependent. A subject's respiration rate and heart rate is determined by its cardio-respiratory system, not by facial appearance or clothing. Different subjects may have very different facial features (exploited by face recognition) but their vital signs could be very similar. Therefore, a rationale for signal extraction would be first eliminating the interfering appearance factor in images to reduce intra-subject variance and then retrieving the bio-signals. Another example is the camera-based photoplethysmography that measures blood perfusion beneath the skin but not outside the skin (not associated with appearance features). In addition, CNN-based physiological measurement still resembles a black-box. It is unclear how the image data is mapped to the physiological variables and the principles for such mapping/training are yet unknown, i.e. a fully transparent and explainable system is essential for making medical claims for healthcare technology.

To gain more insights, we show the activation maps obtained by 2D-CNN and 3D-CNN based on different video properties in Fig. 3. The activation maps suggest that CNN has been focused on the spatial contexts that are not relevant for respiration (e.g. pillow boundaries, subject hair, background edge) and these may deteriorate the performance as they are not the source of respiration. In the end, we mention that the challenges (e.g. motion robustness) that cannot be addressed by conventional approaches can nether be resolved by CNN. The role of CNN in vital signs extraction needs to be justified. Currently we consider it to be more suitable to be used as front-end steps in a camera vital signs monitoring system, such as region of interest (e.g. face or chest) detection and tracking where we have sufficient understanding and knowledge.

IV. CONCLUSIONS

In this paper, we revisit the video-based respiration monitoring that measures the respiratory motion. We use a large benchmark dataset with six categories of challenges to validate four signal extraction methods (from conventional to CNN-based approaches) on two video properties (i.e. pixel intensity variation and pixel movement). We conclude that pixel motion features are more robust for respiration signal extraction. The major challenge for both video properties is the disturbance of non-respiratory motion. Simple conventional approach (e.g. ZCA) outperforms CNN for signal extraction. We hope that the benchmark and insights gained in this study can help the video health monitoring community to improve and apply the techniques for video-based respiration monitoring.

REFERENCES

[1] A. Steinschneider, "Prolonged apnea and the sudden infant death syndrome: clinical and laboratory observations," *Pediatrics*, vol. 50, no. 4, pp. 646–654, 1972.

[2] M. H. Li, A. Yadollahi, and B. Taati, "A non-contact vision-based system for respiratory rate estimation," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2014, pp. 2119–2122.

[3] C. Wiede *et al.*, "Remote respiration rate determination in video data-vital parameter extraction based on optical flow and principal component analysis," in *International Conference on Computer Vision Theory and Applications*, vol. 5. SciTePress, 2017, pp. 326–333.

[4] T. Lukáč, J. Púčik, and L. Chrenko, "Contactless recognition of respiration phases using web camera," in *24th International Conference Radioelektronika*. IEEE, 2014, pp. 1–4.

[5] R. Janssen *et al.*, "Video-based respiration monitoring with automatic region of interest detection," *Physiol. Meas*, vol. 37, no. 1, pp. 100–114, 2016.

[6] M. Bartula, T. Tigges, and J. Muehlsteff, "Camera-based system for contactless monitoring of respiration," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 2672–2675.

[7] C. Massaroni *et al.*, "Non-contact monitoring of breathing pattern and respiratory rate via rgb signal measurement," *Sensors*, vol. 19, no. 12, p. 2758, 2019.

[8] A. Qayyum *et al.*, "Convolutional neural network approach for estimating physiological states involving face analytics," in *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*. IEEE, 2019, pp. 68–72.

[9] A. Kwasniewska *et al.*, "Evaluating accuracy of respiratory rate estimation from super resolved thermal imagery," in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2744–2747.

[10] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 349–365.

[11] L. Iozzia, L. Cerina, and L. Mainardi, "Relationships between heart-rate variability and pulse-rate variability obtained from video-ppg signal using zca," *Physiol. Meas*, vol. 37, no. 11, p. 1934, 2016.

[12] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2019, pp. 1–12.

[13] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian conference on Image analysis*. Springer, 2003, pp. 363–370.