A non-contact SpO_2 estimation using video magnification and infrared data

Thomas Stogiannopoulos, Grigorios-Aris Cheimariotis, Nikolaos Mitianoudis

Electrical & Computer Engineering Dep.
Democritus University of Thrace
Xanthi, Greece
{tstogian, gcheimar, nmitiano}@ee.duth.gr

***Abstract*—Peripheral oxygen saturation (SpO_2) is one important vital sign to be monitored in individuals, whose health is fragile, such as the elderly. Contactless SpO_2 monitoring using RGB cameras has been already developed with satisfactory results. This work explores the case of achieving an acceptable level of performance, when the lightning conditions are not optimal, particularly during night time, by processing solely infrared low-cost camera recordings. The Eulerian Video Magnification (EVM) technique was used to enhance the subtle differences in skin pixel intensity in the facial area. Two approaches were explored for performing regression: one using 12 novel features extracted from the amplified photoplethysmography (PPG) signal and Generalized Additive Models and a second using a 3D Convolution Neural Network (CNN) architecture on the raw amplified forehead video. The root mean square error in the estimated SpO_2 levels using both methods is minimal and in the accepted range for these applications.**

***Index Terms*—Eulerian Video Magnification (EVM), peripheral oxygen saturation (SpO_2), Infrared (IR), Regression Neural Networks, Generalized Additive Models (GAM), remote photoplethysmography (rPPG)**

I. Introduction

Oxygen saturation, a crucial physiological parameter, has been providing vital information about a patient's health status during treatment therapies, critical surgery operations or every day activity monitoring. The traditional contact pulse oximeter, which is most commonly placed on the fingertip, has been utilized frequently in daily tasks. However, it cannot be employed in cases of individuals, who have trembling hands or feet for a variety of reasons or strict health protocols are applied during a grand disease outbreak, such as the COVID-19 pandemic. In these cases, any medical equipment must not be shared among patients, if possible, to avoid any infections.

The peripheral oxygen saturation (SpO_2) level is expressed as the percentage of oxygenated hemoglobin over the total amount of hemoglobin (oxygenated and deoxygenated) in the blood [1], [3]. Most people lie within the range of 96%-98%. There are also cases, where it can reach up to 100% [1]. Its mathematical definition is given by:

$$SpO_2 = \frac{HbO_2}{HbO_2 + Hb} \times 100\% \quad (1)$$

where Hb is the deoxygenated hemoglobin and HbO_2 is the oxygenated hemoglobin. The SpO_2 must be continuously monitored both during surgical operations and in intensive care units (ICUs). The pulse oximeter is currently the instrument used most frequently for this purpose. Detecting variations in blood volume in the microvascular blood–tissue exchange network [15] is possible with the help of the straightforward, non-invasive and inexpensive optical technique, known as photoplethysmography (PPG). The concepts of photoplethysmography are used by oximeters to track vital signs. The pulse oximeter, which measures SpO_2, is a tool that typically relies on a clip fastened to the subject's index fingertip, earlobe or toe. Unfortunately, in order to measure, direct contact with the patient's skin is required, and this contact is not always feasible, e.g. in newborns. As a result, continuous monitoring SpO_2 without direct skin contact will improve the quality of treatment for these patients.

Contactless measuring techniques of vital signs perform effectively to address the aforementioned issues [2]. These investigations paved the way for innovations, such as contactless pulse oximetry, image photoplethysmography (iPPG), and remote photoplethysmography (rPPG). Video frames are used to analyze variations in the amount of light absorbed by tissue during contactless vital sign monitoring. To facilitate this, the Eulerian video magnification (EVM) technique was used to enhance the subtle differences in skin pixel intensity using imaging photoplethysmography (iPPG) [7].

There has been previous research on estimating peripheral oxygen saturation (SpO_2) remotely. Some studies processed signals recorded by RGB cameras using ambient light [3], [4]. Some other previous works have focused on video magnification techniques, such as the Eulerian Motion Magnification [5] or including additional signal transformations (e.g. Hermite Transform) [6]. Both of these approaches have implemented Motion Magnification using RGB footage, originated from a visual camera in a natural or ambient lighting environment and later applied linear regression to estimate the actual SpO_2 levels. One important element is that, in current literature, the SpO_2 is modelled as a function of K_a i.e. the ratio of ratios [5]. This ratio is obtained by analyzing the association between the red and green wavelengths [5], the red and IR wavelengths [4] or between the red and blue wavelengths of the PPG signal [6].

In this work, the main objective is to devise a method

that estimates the SpO_2 remotely using only infrared low-cost cameras, in order to operate during night-time. To the best of our knowledge, this is a novel effort that has many applications. To tackle the problem, we used Eulerian Motion Magnification to enhance the video and then experimented with two methodologies: one that uses 12 novel statistical features and traditional machine learning tools, such as Generalized Additive Models and a second that uses a CNN architecture on the raw magnified video. The presented results show that this task is possible with accurate predictions.

II. Proposed Methodology

A. Preprocessing

In this paper, the interest is to obtain an SpO_2 estimate without contact using an infrared camera. The facial area was chosen to be used for the estimation of SpO_2 and more specifically the forehead and the left/right cheek areas. The method uses 2-minute video clips from an infrared camera as inputs. At first, the face is detected in the video using the Viola-Jones algorithm [16], which seems to operate accurately in infrared videos. The next step was to identify the forehead and the left/right cheek region in the detected face. Based on the analysis in [8], [10], [11], it was possible to determine accurately and calculate the coordinates of the desired regions of interest (ROI), as shown in Fig. 1. It is ordinary calculus to use the presented ratios on the face detection bounding box, in order to estimate the coordinates of the three desired ROIs.

The next step is to use the Eulerian Video Magnification method, as proposed by Wu et al. [7] to amplify the facial blood-flow in the infrared stream. This method is of great importance, due to the fact that subtle variations on skin's surface, which are commonly unnoticeable to the naked eye, are now highlighted. The amplification factor was set to $\alpha = 120$, while the frequency range of amplification was set between 0.4 and $4Hz$. This is slightly larger than the range used in [3], because in cases of supraventricular tachycardia (SVT), the heart rate could peak at 240 bpm [9].

B. Feature Extraction

Inspired by [3], [5], a number of novel features that are intrinsically linked with the statistical properties of the iPPG signal can be extracted from the 3 desired regions. Extending those initially proposed by [3], [5], the following features are selected: a) the average of all frames' intensity averages, b) the standard deviation of all frames' intensity averages, c) the average of all frames' standard deviation of intensity and d) the standard deviation of all frames' standard deviation of intensity. Four features for three distinctive regions of interest sums up to twelve features in total. These features will be used as input to traditional machine learning regression algorithms in the experimental section.

C. Generalized Additive Models

In this paper, we examine the application of Generalized Additive Models (GAM) for regression [14]. The main distinction between a GAM and Generalized Linear Models

Fig. 1. Proportion analysis of the human face. The desired regions of interest (forehead, left-right cheek) are highlighted.

(e.g. Linear Regression) is that a GAM is permitted to learn non-linear relationships between dependent and independent variables. GAMs can offer regression results that can be represented by the sum of any number of flexible functions of each feature, known as splines. Splines are composite non-parametric functions that reveal non-linearities for each feature. Thus, a GAM can express the inference of a r.v. Y as the sum of a set of predictor r.v.s $X_1, X_2, \ldots, X_p$, as follows:

$$\mathcal{E}\{Y|X_1, X_2, \ldots, X_p\} = f_0 + \sum_{j=1}^{p} f_j(X_j) \qquad (2)$$

where $f_j(\cdot)$ are smooth nonparametric functions standardized, so that $\mathcal{E}\{f_j(X_j)\} = 0$ [14]. GAMs are particularly suitable for regression tasks, since the marginal effect of a single variable is independent of the other variables' values, and thus are able to capture nonlinear relationships and patterns. The GAM framework enables us to regulate the predictor functions' smoothness to avoid overfitting.

D. Deep Learning Regression

Due to their greater capability of collecting spatio-temporal features from video frames, convolutional neural networks with 3D kernels (3D CNNs) have recently gained a lot of popularity in the computer vision community. In this paper, we explore their efficiency in this regression task. More specifically, we decided to exploit only the motion amplified forehead video of size $64 \times 128 \times 300$, denoting $D = 300$ the number of frames and $(H = 64) \times (W = 128)$ the frame size. The input video is normalised to the $[0, 1]$ interval, whereas the SpO_2 output values also normalised to the $[0, 1]$ interval. The proposed network consists of two Conv3D layers, which are then followed by two Fully-Connected layers and a final output

layer, which features a linear activation function and yields the inferred value. The complete architecture of the proposed 3D CNN is shown in Table I. The loss function was the Mean Square Error (MSE), which was optimised using the Adam Optimizer with a learning rate of $\eta = 0.001$.

III. Experiments

A. Dataset - Implementation

A dataset of infrared facial videos with SpO_2 measurements had to be constructed, in order to evaluate the performance of the proposed methods. In total, 16 subjects participated in two experiments each. It was determined that 2 minutes would be an adequate monitoring duration for each subject, in order to remain calm and relaxed. A commercial pulse oximeter (JPD-500D ControlBios Oxicore Pulse Oximeter) was employed as the reference point that continuously tracked each participant's oxygen saturation during the two 2-minute length trials. Half of the subjects were captured in a dark room and the other half were captured in a room with natural sunlight, using the infrared camera in either case. Each participant was asked to remain as still as possible for 2 minutes twice. During the first video recording, all participants were asked to breathe normally. During the second video recording, the subjects were asked to hold their breath, as many times and as long as they felt comfortable, in order to capture lower SpO_2 measurements. A wired Google Nest Cam (1920×1080/Full HD, 15 fps@night time & 30 fps@day time) was used for capturing, which was placed at eye-level and switched to the "Infrared Always" mode. The participants were seated at 75 cm away from the camera. This distance was chosen in order to avoid excessive distortion, caused by the wide-angle camera lenses. All data and SpO_2 reference values are publicly available and can be found on the paper's GitHub page .

The motion magnification step was run on MATLAB R2018b, based on the code, available by [7]. The machine learning regression models and GAM were available in Python from the scikit-learn and pygam libraries. The proposed 3D CNN architecture was implemented in Tensorflow v.2.9.1 on a machine running Ubuntu 22.04, with an Intel i9-11900KF@3.5GHz with 64GB RAM and an NVidia RTX A6000 with 48GB RAM[1].

B. Comparison

The ISO 80601-2-61:2019 [13] defines the accuracy as the root mean square difference between the estimated values SpO_{2i} and reference values S_{Ri} and is given by:

$$A_{rms} = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(SpO_{2i} - S_{Ri})^2}{n}} \qquad (3)$$

where n is the number of samples. According to BS EN ISO 80601-2-61:2019, the root mean square accuracy must not exceed 2% of SpO_2 range [13] or alternatively the mean square error of the testing pair values must not exceed 4%.

[1]The developed code can be found at [20]

TABLE I
The Proposed 3D CNN Architecture for video sequences

Layer/Stride	Contents	Output Size (HxWxDxC)
Input Clip	-	64x128x300x1
Conv3D	$\begin{bmatrix} Conv3D(16, kernel = (5,5,5)) \\ MaxPooling3D(pool = (3,3,3)) \\ Dropout(p = 0.5) \\ Activation = ReLU \\ initializer = he\_uniform \end{bmatrix}$	20x41x98x16
Conv3D	$\begin{bmatrix} Conv3D(16, kernel = (5,5,5)) \\ MaxPooling3D(pool = (3,3,3)) \\ Dropout(p = 0.5) \\ Activation = ReLU \\ initializer = he\_uniform \end{bmatrix}$	5x12x31x32
Flatten	-	59520
FC1	$\begin{bmatrix} Dense(128) \\ Activation = ReLU \\ initializer = he\_uniform \end{bmatrix}$	128
FC2	$\begin{bmatrix} Dense(128) \\ Activation = ReLU \\ initializer = he\_uniform \end{bmatrix}$	128
Output	$\begin{bmatrix} Dense(1) \\ Activation = Linear \end{bmatrix}$	1

The aim of this paper was to verify whether there is any correlation between the collected infrared data, i.e. grayscale video, and the peripheral oxygen saturation (SpO_2). We tested machine learning and deep learning regression algorithms to demonstrate that features of the iPPG signals, extracted from the forehead and left and right cheeks, are adequate to fulfil medical standards and accuracy criteria, such as those stated from USA [12] and EU/UK (BS EN ISO 80601-2-61:2019) [13]. For the machine learning regression, apart from the proposed GAM models, we also tested linear regression [17], Ridge regression [18] and Lasso regression [19] for alpha = 0.0001 and alpha = 0.001.

The developed dataset of 190 video clips was shuffled and divided into 152 clips for training and 38 clips for testing, while applying 5-fold cross validation and multiple runs (20 runs per algorithm). No division between different light conditions was made. The accuracy of each algorithm is estimated by the mean of the 5 folds and 20 runs using the definition in (3).

C. Results

The results from the conducted experiments are summarised in Table II, where the best performance is indicated in bold. It is evident that all models have not exceeded the 2% tolerance of RMSE that is dictated by the aforementioned directive. From all the tested frameworks the GAM seems to predict the correct SpO_2 more accurately. The 3D CNN, although it performs favourably, falls behind many traditional machine learning regression techniques. This might be due to the relative small number of data of the dataset and the fact that only the forehead data are used.

TABLE II
COMPARISON IN TERMS OF MEAN ABSOLUTE ERROR, MEAN SQUARE ERROR AND ROOT MEAN SQUARE ERROR

Model	MAE	MSE	RMSE
Linear Regression	1.266	2.982	1.727
Ridge Regression	1.410	2.903	1.704
Lasso (alpha = 0.0001)	1.396	2.874	1.695
Lasso (alpha = 0.001)	1.445	3.019	1.738
GAM	**1.123**	**2.662**	**1.632**
3D CNN	1.417	2.935	1.713

TABLE III
NUMBER OF FEATURES INVESTIGATION AND BEST PERFORMANCE

Model	Min RMSE	No. of features
Linear Regression	1.517	4
Ridge Regression	1.671	8
Lasso (alpha = 0.0001)	1.659	9
Lasso (alpha = 0.001)	1.727	5
GAM	**1.440**	12
3D CNN	1.687	-

In the previous study, all 12 features are presented as input to the machine learning approaches, except for the 3D CNN, which uses the amplified video stream. As an ablation study, we wanted to check whether all these 12 features are useful for each algorithm. As a result, all possible combinations of subsets of these features were tested with each regression algorithm. In Table III, we depict the best attained RMSE score for each method and the number of features that produced this score. It appears that only the GAM produces the best score using all 12 features, whereas the other methods, rely on subsets of the original features. It is noteworthy that Linear Regression uses only 4 features to produce its best result, which include statistical measurements from all examined facial areas. This shows that all the selected regions are important for this task. Finally, Table III shows the best attainable result (RMSE) from each method (including 3D CNN) and not the average that is depicted in Table II.

One limitation of the current research can be seen in Fig. 2 and Fig. 3, where the distribution of actual SpO_2 measurements and the age distribution of the participants are visualised. The vast majority of the participants were undergraduate and postgraduate students without any health complications, even though heavy smokers were included. Even when the students were asked to hold the breath in order to cause controlled drop in oxygen concentration, as suggested in previous studies, in most cases the measurement drop was small or negligible. Consequently, most measurements lie within the 96%-100% range and only a few below 95%, which demonstrates a situation that requires possible medical assistance. Hence, the experiments would be more complete, if more people with low levels of SpO_2 could be captured.

IV. CONCLUSIONS

In this work, we attempt to measure SpO_2 levels from human subjects in a contactless low-light scenario. The only required equipment is a commercial low-cost infrared camera, such as those available for home surveillance. Using motion-amplified infrared video and 12 novel statistical features, we have shown that it is possible to estimate the actual levels of SpO_2 with very small tolerance that satisfies the world's standards. We have tested a variety of machine learning regression techniques, as well as a 3D CNN, with the one based on GAM producing the most accurate estimates. The produced dataset and methods are available online. The proposed method's key advantage is the potential of full-day use in cases, such as elderly care facilities, hospitalized patients where a medical emergency, serious or not, may occur at any time. For future work, we would like to expand the current dataset with more people that exhibit lower (pathological) levels of SpO_2.

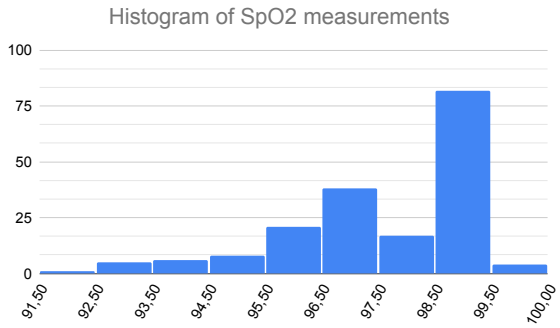

Fig. 2. Distribution of SpO_2 measurements.

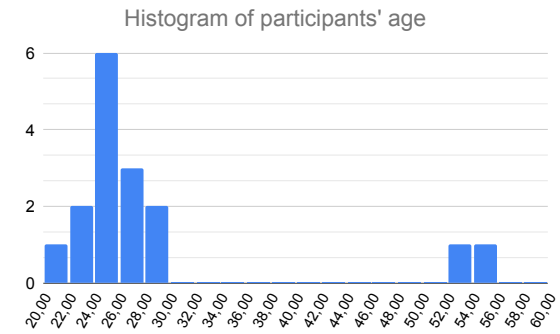

Fig. 3. The age distribution of subjects.

ACKNOWLEDGEMENTS

This work was supported by the project "Study, Design, Development and Implementation of a Holistic System for Upgrading the Quality of Life and Activity of the Elderly" (MIS 5047294) which is implemented under the Action "Support for Regional Excellence", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). We would like to thank all volunteers for their participation.

REFERENCES

[1] B. R. O'Driscoll, L. S. Howard, J. Earis, and V. Mak, "British Thoracic Society Guideline for oxygen use in adults in healthcare and emergency settings", BMJ Open Respiratory Research, Vol. 4, No. 1, p. e000170, May 2017.

[2] W. Verkruysse, M. Bartula, E. Bresch, M. Rocque, M. Meftah, and I. Kirenko, "Calibration of Contactless Pulse Oximetry", Anesthesia and Analgesia, Vol. 124, No. 1, pp. 136–145, Jan. 2017.

[3] L. Kong et al., "Non-contact detection of oxygen saturation based on visible light imaging device using ambient light", Optics Express, Vol. 21, No. 15, pp. 17464, Jul. 15, 2013.

[4] A. Al-Naji, G. A. Khalid, J. F. Mahdi, and J. Chahl, "Non-Contact SpO2 Prediction System Based on a Digital Camera", Applied Sciences, Vol. 11, No. 9, pp. 4255, May 07, 2021.

[5] A. de Fatima Galvao Rosa and R. C. Betini, "Noncontact SpO2 Measurement Using Eulerian Video Magnification", IEEE Transactions on Instrumentation and Measurement, Vol. 69, No. 5, pp. 2120–2130, May 2020.

[6] J. Brieva, E. Moya-Albor, and H. Ponce, "A non-contact SpO2 estimation using a video magnification technique", 17th International Symposium on Medical Information Processing and Analysis, Dec. 10, 2021.

[7] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world", ACM Transactions on Graphics, Vol. 31, No. 4, pp. 1–8, Jul. 2012.

[8] J. Milutinovic, K. Zelic, and N. Nedeljkovic, "Evaluation of Facial Beauty Using Anthropometric Proportions", The Scientific World Journal, Vol. 2014, pp. 1–8, 2014.

[9] C. Garratt, D. Ward, A. Antoniou, and A. John Camm, "Misuse of Verapamil in pre-excited atrial fibrillation", The Lancet, Vol. 333, No. 8634., pp. 367–369, Feb. 1989.

[10] K. S. Kaya, B. Türk, M. Cankaya, N. Seyhun, and B. U. Coşkun, "Assessment of facial analysis measurements by golden proportion", Brazilian Journal of Otorhinolaryngology, Vol. 85, No. 4. pp. 494–501, Jul. 2019.

[11] J. W. Fernandes, "The Legacy of Art in Plastic Surgery", Plastic and Reconstructive Surgery - Global Open, Vol. 9, No. 4, pp. e3519, Apr. 2021.

[12] Center for Devices and Radiological Health, "Pulse oximeter accuracy and limitations", U.S. Food and Drug Administration, 11-Jul-2022. [Online]. Available: https://www.fda.gov/medical-devices/safety-communications/pulse-oximeter-accuracy-and-limitations-fda-safety-communication. [Accessed: 18-Jan-2023].

[13] "ISO 80601-2-61:2017", ISO, 01-Feb-2018. [Online]. Available: https://www.iso.org/standard/67963.html. [Accessed: 18-Jan-2023].

[14] T. Hastie and R. Tibshirani, "Generalized Additive Models", Statistical Science, Vol. 1, No. 3, Aug. 01, 1986.

[15] J. R. Levick, An introduction to cardiovascular physiology. London: Hodder Arnold, 2011.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), 8-14 December 2001, Kauai, HI, USA.

[17] D. Freedman, "Statistical Models: Theory and Practice", Cambridge University Press, 2005.

[18] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems", Technometrics, Vol. 42, No. 1, pp. 80–86, Feb. 2000.

[19] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso", Journal of the Royal Statistical Society: Series B (Methodological), Vol. 58, No. 1, pp. 267–288, Jan. 1996.

[20] T. Stogiannopoulos, "Tomstog/infrared-SPO2: The dataset used for the 'A non-contact SpO2 estimation using video magnification and Infrared Data' publication", GitHub, 23-Oct-2022. [Online]. Available: https://github.com/TomStog/Infrared-SpO2. [Accessed: 18-Jan-2023].