Post-print of': M. Szankin et al., "Long Distance Vital Signs Monitoring with Person Identification for Smart Home Solutions," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, 2018, pp. 1558-1561, doi: 10.1109/EMBC.2018.8512509.

Long Distance Vital Signs Monitoring with Person Identification for Smart Home Solutions

M. Szanki, *Member IEEE,* A. Kwasniewska, *Student Member IEEE, EMBS*, T. Sirlapu, M. Wang, J. Ruminski, *Member IEEE, EMBS*, R. Nicolas and M. Bartscherer

***Abstract*— Imaging photoplethysmography has already been proved to be successful in short distance (below 1m). However, most of the real-life use cases of measuring vital signs require the system to work at longer distances, to be both more reliable and convenient for the user. The possible scenarios that system designers must have in mind include monitoring of the vital signs of residents in nursing homes, disabled people, who can't move, constant support for people regardless of the performed activity (e.g. during sleeping), infants, etc. In this work we verified the possibility of remote pulse estimation at a distance above 5m. Additionally, we integrated the deep learning algorithm for person tracking and identification, even when facial features are not visible. In this way, we enabled the collection of user specific measurements to create personalized vital signs patterns and we provided the support for monitoring of multiple people using one video stream. The preliminary results showed that it is possible to accurately (RMSE < 2.8 beats per minute) extract pulse from visible light sequences acquired with a webcam at a distance of 6m after applying a proper image pre-processing algorithm.**

I. Introduction

In latest years smart home solutions became more popular due to the increased availability of affordable and reliable sensing infrastructure [1]. With further development of these solutions we can observe even larger need for expanding them to remote monitoring solutions that could be applied to telemedicine use cases [2]. More often we reach to our smart home systems not only for household monitoring, but also for inhabitants guarding. This development, however, creates a new set of problems that has to be addressed. Human safety requires not only saving a video stream of the accident – smart systems must be able to capture, process and respond to threats and accidents in real time. To increase their effectiveness, these solutions should be less prone to the distance between the subject and the sensor.

Recent advances in cloud computing, image processing and mobile technologies enables many solutions for remote vital signs monitoring that do not require users to wear any additional sensors/devices. Heart rate which is one of the most fundamental vital sign that can indicate potential health problems. In [3] it was proved that a heart rate can be estimated by analyzing the green channel of video sequences recorded with the RGB camera, as it contains the strongest *photoplethysmography* (PPG) signal. Other studies made use of an RGB video as well and applied the discriminative statistical model for Blood Volume Pulse estimation [4]. Poh et al. proposed to use a blind source separation of three color channels (R, G, B) into independent components to measure a cardiac pulse from automatically tracked facial area [5]. It has also been proved that the fusion of skin color variation and head motion is a promising approach for estimating a pulse using the video stream recorded with the webcam [6]. Later, it was shown that YUV color model can also be used for accurate remote heart rate estimation (mean squared error <2 beats per minute *bpm*) from short videos [7]. Recently, some attempts have been made to discover facial regions that produce most accurate results of pulse estimation by using of matrix completion theory [8]. However, most of the conducted studies assume that the face is placed at a short distance from the camera (<1m). Short distance provides better face visibility, greater influence of facial skin region on auto exposure and other automatic camera settings. As a result, the extracted PPG signal from a face region is usually characterized by higher dynamics and greater SNR.

Continuous acquisition of vital signs regardless of a distance from a camera allows for gathering data more frequently in various situations (e.g. sleeping, infants monitoring, support of people who are disabled or after injuries that affected their motor skills) to create a more detailed and precise health pattern profile. This can be especially useful in smart buildings with high density of occupancy (e.g. nursing homes, hospitals or even prisons), where person recognition and remote vital sign acquisition could be particularly useful. To achieve this, accurate person identification and tracking is an essential prerequisite. Existing solutions, though, are often based on detecting facial features [9] that are only visible at a short distance. Another approach considers body poses and/or movements as inputs to predict results [10]. However, if the person is still or imitates other person, this may lead to inaccurate results. Moreover, most of previous solutions are attribute based algorithms which may fail to tell the subtle difference between two identities when their clothing looks alike [11].

In this preliminary study, we want to investigate methods of person identification and pulse estimation for subjects at distances higher than 5m. We propose to apply a combination of the deeply-learned part-aligned body representation [12]

This work has been partially supported by Intel Corporation, USA. We thank all our colleagues from Intel Corp., who provided insight and expertise that greatly assisted the research. The research has been also partially supported by Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdansk University of Technology.

A. Kwasniewska (alicja.kwasniewska@pg.edu.pl), J. Ruminski (jacek.ruminski@pg.edu.pl) are with Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Deptartment of Biomedical Engineering, Gdansk, Poland

M. Szankin (maciej.szankin@intel.com), A. Kwasniewska (alicja. kwasniewska@intel.com), T. Sirlapu (tejaswini.sirlapu@intel.com), M. Wang (mingshan.wang@intel.com), M. Bartscherer (marko.bartscherer @intel.com), R. Nicolas (rey.nicolas@intel.com) are with Intel Corp., San Diego, CA, USA

and facial features embedding [13] to enable analysis of personalized vital sign patterns. The accuracy of the applied approach was verified in various scenarios where a person is visible from different angles, perform various poses and have changed appearance features to address possible real-life smart home use cases. We also investigate if often-used pulse estimation method (like [7]) can also produce accurate results at longer distances (>5m). We additionally analyze if additional color magnification [14] applied to image sequences improves the accuracy of pulse estimation. Specifically, we verify the accuracy of (30s) video-based pulse estimation from signals obtained from forehead and face regions with the aggregation operation of pixels in YUV.

The paper is organized as follows: in Section II we describe methodology used for person identification and vital signs evaluation in a far field. Section III demonstrates preliminary results, further discussed in Section IV. The paper is concluded in Section V.

II. Methodology

The experiment of monitoring vital signs and person re-identification was conducted on a group of 12 healthy volunteers (6 males, 6 females, age: 31.6±5.9, representing all of skin color types from Fitzpatrick scale [15]) in an environment lit with incandescent downlight. At first, we wanted to validate if the system is able to identify and track a given person in various unconstrained scenarios:

1. *TC1 – Subject is walking naturally (~4-7m from camera), visible from different camera angles, visual appearance unchanged.*
2. *TC2 – Subject is performing various body poses (~4-7m from camera), visible from one camera angle, visual appearance unchanged.*
3. *TC3 – Different subjects are wearing the same clothes, walking naturally (~4-7m from camera), visible from different camera angles (3 volunteers).*

In second part of the experiment we wanted to verify the accuracy of long distance pulse estimation using imaging photoplethysmography:

4. *TC4 – Subject is sitting at a distance of 6m facing the camera, visual appearance unchanged. In this scenario, we used Zacurate 430-DL fingertip pulse oximeter for ground truth measurements.*

For all test cases a 30 seconds live video stream was captured using a Logitech C920 webcam with a 30 frames per second at 1920x1080 resolution and processed using a pipeline of three deep neural networks executed in parallel processes divided into two stages. In the first stage, we used the SSD network [16] trained on VOC 2007 dataset for body detection and FaceNet model [13] retrained on 1500 face pictures of each volunteer, extracted from previously recorded with the same camera 1-minute videos at the distance of 1m (divided into training, validation and test set in proportion 8:1:1). As a

Figure 1. Visualization of 2 consecutive frames with a) and without b) EVM

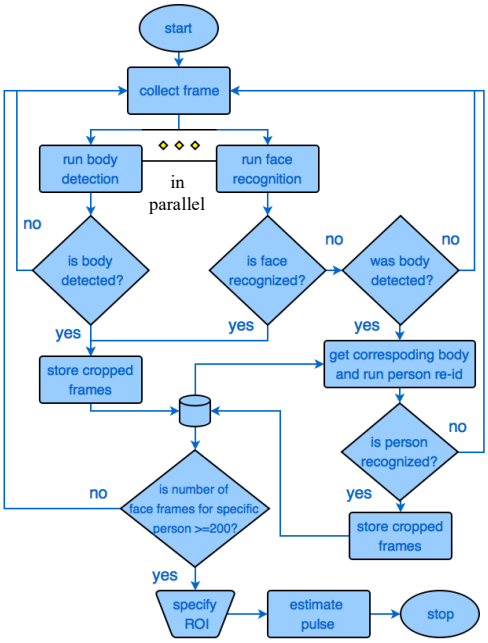

Figure 2. Flow of the proposed methodology

result of training the FaceNet model on our data, for each volunteer we stored a facial feature vector of a size 128. This size of vector has been previously verified to produce sufficient number of features that uniquely identify a person [13]. During the inference, for each detected face in the video stream, a feature vector of the same size was extracted and compared against existing facial features vectors. If a face was properly recognized in a video stream (a cropped sub-image of a body detected with the SSD network was saved with a label specifying the person ID. In this case, the second stage was not activated. If the first stage was not able to detect face but detect body (e.g. subject standing with back towards the camera), the bounding box with body location was sent to the second layer, in which Re-identification [12] (ReID) model trained on CUHK03 dataset was run. The output from this model was represented as the vector of 512 features and was compared with all other stored feature vectors. Each of the stored feature vectors corresponded to the output from re-id model extracted for al previously cropped body areas. In order to preserve constrained resources available on the edge device, we only saved 4 most recent feature vectors for each person, that were later used for person re-identification.

Additionally, to verify the possibility of pulse rate estimation at a long distance in TC4, the consecutive 30 seconds of captured video stream were cropped to the detected facial area and converted to the YUV420P color space while reducing the frame rate to 15Hz. In this part of the experiment, we also tested if applying Eulerian Video Magnification [14] (EVM) algorithm can improve the accuracy of pulse estimation. Since the heart rate during rest for an adult typically varies from 60 to 100 beats per minute, in EVM color magnification algorithm we applied filtering with a range from 1Hz to 1.67Hz. The resulting signal was then amplified by 20, as indicated in [14] for face motion. Both magnified and non-magnified sequences were analyzed (examples of frames presented in Fig. 1.) to obtain the pulse rate. For this, two regions of interests (ROI) were manually

selected – on the forehead and on the whole face. Values of pixels inside these areas were averaged for each frame in a sequence of last ~200 samples, producing raw signals further filtered with the bandpass Butterworth filter (frequency between 0.67-4Hz), because according to American Heart Association [17] heart rate for an adult varies from 40 for athletes to higher values up to 200 during exhaustive activities. Then, as verified in [7], *ePR_sp* estimator was applied for heart rate estimation. The estimated signals were compared against readouts from the pulse oximeter. Flow of the whole execution is presented in Fig. 2.

III. Results

The results of pulse rate estimation from a forehead and a face ROI, both for magnified and non-magnified sequences, together with ground truth measurements are presented in Table I and Table II, respectively. Table III presents calculated Root Mean Square Error between estimated and ground truth heart rates. All measurements presented in tables I, II and III were based on TC4, as only in this case subjects were uninterruptedly exposing their forehead (ROI used for vital signs evaluation) for extended period of time.

TABLE I. Pulse rates evaluated from forehead region

	Ground	EVM		No EVM	
Person	**Truth**	***Y***	***V***	***Y***	***V***
1	65.0	65.769	65.769	69.231	69.231
2	70.0	72.764	72.764	7.347	7.347
3	85.0	83.191	83.191	7.258	7.258
4	84.0	85.537	85.537	21.600	21.600
5	77.0	72.764	72.764	18.367	18.367
6	52.0	51.851	51.851	7.200	7.200
7	86.0	86.611	86.611	72.289	72.289
8	85.0	89.990	89.990	89.990	89.990
9	59.0	59.504	59.504	14.400	14.400
10	54.0	59.990	59.990	57.599	57.599
1	72.0	71.285	71.285	72.874	72.874
12	75.0	74.687	74.687	36.000	36.000

TABLE II. Pulse rates evaluated from facial region

	Ground	EVM		No EVM	
Person	**Truth**	***Y***	***V***	***Y***	***V***
1	65.0	64.800	64.800	64.800	64.800
2	70.0	72.764	72.764	7.347	7.347
3	85.0	75.639	75.639	7.258	7.258
4	84.0	70.661	70.661	14.400	14.400
5	77.0	91.913	91.913	91.835	91.835
6	52.0	51.851	51.851	7.200	7.200
7	86.0	82.845	82.845	50.602	50.602
8	85.0	74.990	74.990	7.200	7.200
9	59.0	48.347	48.347	10.800	10.800
10	54.0	56.249	56.249	50.399	50.399
11	72.0	71.285	71.285	72.874	72.874
12	75.0	78.421	78.421	31.500	31.500

TABLE III. Root Mean Square Error of estimated heart rate

	EVM		No EVM	
Region	***Y***	***V***	***Y***	***V***
Forehead	2.796	2.796	43.839	43.839
Face	7.834	7.835	48.798	48.798

After running all test cases, we created a gallery of persons' profiles (4 feature vectors per volunteer per test case) that was later used for testing applied algorithms. Only 3 subjects participated in TC3, so the total number of feature vectors in the gallery was 156. If a face was detected in these frames, we also stored the extracted facial features vectors. Table IV presents accuracy of person detection and re-identification in terms of a percentage of correctly identified subjects based on combined results from ReID and FaceNet. In each scenario, 4 feature vectors for each volunteer were compared with all other feature files stored in the gallery. Also, if the face was detected, the extracted facial features vector was compared against stored facial users' profiles. The final decision about the identification was made as the logical disjunction of ReID and FaceNet decisions (correct if any of these models produced a correct prediction). Examples of query images (left in each pair) and identified results (right in each pair) for TC1, TC2 and TC3 are presented in Fig. 3, 4, 5

Figure 3. Query image (blue) and identified result (green) for TC1

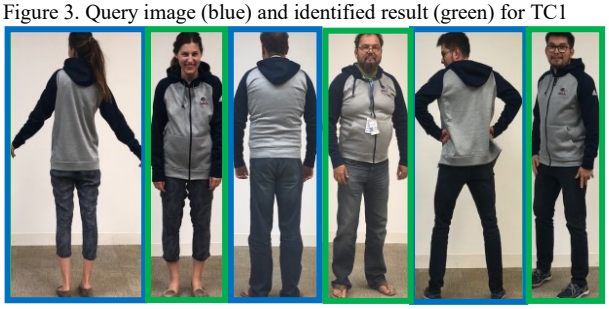

Figure 4. Query image (blue) and identified result (green) for TC2

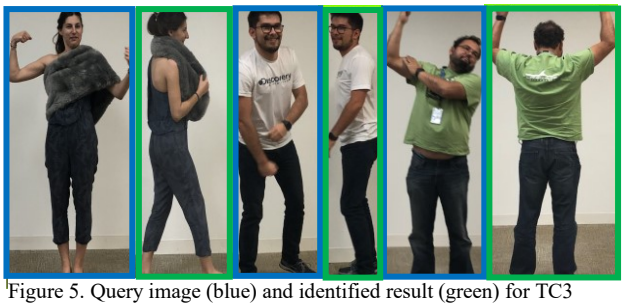

Figure 5. Query image (blue) and identified result (green) for TC3

TABLE IV. Test results for Person Detecion and Re-ID

TC	**Accuracy SSD [%]**	**Accuracy ReID + FaceNet [%]**
1	100±0.0	100±0.0
2	100±0.0	89.5±12.8
3	100±0.0	91.7±14.4
4	100±0.0	100±0.0

IV. Discussion

In this work a possibility of extracting heart rate from video sequences captured at a long distance (~6m) was evaluated using estimator applied to signals extracted by averaging pixel values inside face and forehead areas. Additionally, a combination of face recognition and a person re-identification models was employed to collect measurements for each user separately. The preliminary

results proved that it is possible to accurately estimate a heart rate from much further away than achieved by known state of the art methods, if a proper preprocessing method is applied. The RMSE of pulse computation after magnifying color changes is much smaller (<2.8 beats per minute for a forehead area) when compared to non-enhanced videos (~44 beats per minute for a forehead area). It was also observed that there is no difference between pulse rate estimation from signals acquired for Y and V channels. Analysis performed for different ROI showed that forehead area allows for more reliable parameter computations. This could be caused by appearance of artifacts in the facial regions, e.g. blinking, additionally amplified by EVM algorithm. For non-EVM measurements the difference between these two regions was smaller (EVM: 64%; non-EVM: 10%). Although average error was low, for some subjects we observed higher differences between ground truth and estimations. This could be caused by small involuntary movements and should be further investigated. In this work we applied *ePR_sp* estimatator, as in our previous studies [7] it proved to produce most reliable results, however in future work we would also like to investigate other estimators for long distance vital signs evaluation, such as periodicity of peaks estimator [7].

We discovered that combination of deep learning models, with minimal effort required for retraining, allows for precise body detection and person identification, regardless of visual features, performed poses and camera angles. In all cases accuracy of person recognition was higher than 89%. The proposed algorithm can be applied to various applications in smart buildings, especially where more than one person is visible in the video stream, e.g. nursing homes, hospitals or even office buildings. In these cases, person identification and tracking are essential to distinguish measurements acquired at a distance for different subject (the possible scenario includes continuous tracking of a person, e.g. in the hospital and estimating a pulse when he/she sits down).

However, to further improve proposed implementation, we are considering the use of additional deep learning algorithms for facial areas detection (e.g. forehead region), as described in [18], to fully automate vital signs calculation. Additional challenges for future work also include taking measurements for a person after physical exercise or during illness (in both cases the pulse rate is characterized by a higher dynamic range). In the performed studies it was proved that person tracking and vital signs analysis at a long distance is possible. Therefore, similar measurements can be performed almost everywhere with the use of standard security camera, e.g. determining the stress level or state of the health. This leads to some security concerns and should be further discussed and address in future implementations.

V. Conclusion

In this work a possibility of vital signs estimation at far distance was evaluated. The color magnification algorithm applied to image sequences acquired at 6m with a standard RGB webcam allowed for reducing the RMSE of pulse evaluation with *ePR_sp* estimator from ~43.8 to 2.8 bpm for a forehead region. The proposed deep learning multi model pipeline allowed for reliable person identification, yet to confirm suitability of this solution for smart home devices computational performance should be measured in future work.

References

[1] G. Civitarese and C. Bettini, "Monitoring objects manipulations to detect abnormal behaviors," *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, Kona, HI, 2017, pp. 388-393.

[2] L. Walsh, A. Kealy, J. Loane, J. Doyle and R. Bond, "Inferring health metrics from ambient smart home data," *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Belfast, 2014, pp. 27-32.

[3] W. Verkruysse, L.O. Svaasand, J.S. Nelson, (2008). "Remote plethysmographic imaging using ambient light," *Optics Express*, 16(26), 21434–21445.

[4] A. Osman, J. Turcot and R. E. Kaliouby, "Supervised learning approach to remote heart rate estimation from facial videos," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Ljubljana, 2015, pp. 1-6. doi: 10.1109/FG.2015.7163150

[5] M.Z. Poh, D.J. McDuff, R.W. Picard, (2010). "Non- contact, automated cardiac pulse measurements using video imaging and blind source separation, " *Optics Express*, 18,10762–10774.

[6] C.H. Antink, H. Gao, C. Brüser, and S. Leonhardt, 2015. "Beat-to-beat heart rate estimation fusing multimodal video and sensor data," *Biomedical optics express*, 6(8), pp.2895-2907.

[7] J. Rumiński, 2016. "Reliability of pulse measurements in videoplethysmography," *Metrology and Measurement Systems*, 23(3), pp.359-371.

[8] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J.F. Cohn and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions" *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* pp. 2396-2404, 2016

[9] A. L. Nasution, D. B. Sena Bayu and J. Miura, "Person identification by face recognition on portable device for teaching-aid system: Preliminary report," *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*, Bandung, 2014, pp. 171-176. doi: 10.1109/ICAICTA.2014.7005935

[10] B.C. Munsell, A. Temlyakov, C. Qu, and S. Wang, 2012, "Person identification using full-body motion and anthropometric biometrics from kinect videos," *In European Conference on Computer Vision (pp. 91-100). Springer,* Berlin, Heidelberg.

[11] E. Ahmed, M. Jones, and T.K. Marks. "An improved deep learning architecture for person re-identification," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* pp.3908-3916, 2015

[12] L. Zhao, X. Li, J. Wang and Y. Zhuang, 2017. "Deeply-learned part-aligned representations for person re-identification," arXiv preprint arXiv:1707.07256.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, 2015. "Facenet: A unified embedding for face recognition and clustering, " *In Proc. of the IEEE conference on computer vision and pattern recognition* (pp. 815-823).

[14] H.Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, W. T. Freeman "Eulerian Video Magnification for Revealing Subtle Changes in the World" *ACM Transactions on Graphics, Volume 31, Number 4 (Proc. SIGGRAPH),* 2012

[15] T.B. Fitzpatrick, 1988. "The validity and practicality of sun-reactive skin types i through vi", *Archives of Dermatology, 124 (6): 869–871,* doi:10.1001/archderm.1988.01670060015008

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, 2016, "Ssd: Single shot multibox detector, " *In European conference on computer vision* (pp. 21-37). Springer, Cham.

[17] American Heart Association, accessed: 4/16/2018, available: https://bit.ly/2ESbNbX

[18] A. Kwasniewska, J. Ruminski, K. Czuszynski, M. Szankin, "Real-time Facial Features Detection from Low Resolution Thermal Images with Deep Classification Models", *Journal of Medical Imaging and Health Informatics 2018*, in print