Multispectral camera fusion increases robustness of ROI detection for biosignal estimation with nearables in real-world scenarios

Gaetano Scebba, *Student Member, IEEE*, Laura Tüshaus and Walter Karlen, *Senior Member, IEEE*

***Abstract*— Thermal cameras enable non-contact estimation of the respiratory rate (RR). Accurate estimation of RR is highly dependent on the reliable detection of the region of interest (ROI), especially when using cameras with low pixel resolution. We present a novel approach for the automatic detection of the human nose ROI, based on facial landmark detection from an RGB camera that is fused with the thermal image after tracking. We evaluated the detection rate and spatial accuracy of the novel algorithm on recordings obtained from 16 subjects under challenging detection scenarios. Results show a high detection rate (median: 100 %, 5^{th}–95^{th} percentile: 92 %–100 %) and very good spatial accuracy with an average root mean square error of 2 pixels in the detected ROI center when compared to manual labeling. Therefore, the implementation of a multispectral camera fusion algorithm is a valid strategy to improve the reliability of non-contact RR estimation with nearable devices featuring thermal cameras.**

I. Introduction

Respiratory rate (RR) is an essential vital sign, but with current, non-invasive methods cannot be measured objectively and reliably. Elevated RR is an early predictor of severe ilness, such as cardio-pulmonary arrest or respiratory tract infections [1]. While invasive RR monitoring devices used clinically frequently cause discomfort to the patient and may influence respiratory function [2], less invasive methods are readily available, but lack accuracy and reliability [3]. Recent designs of devices that measure biosignals with non-contact monitoring techniques, so called nearables, show that clinically needed accuracy can be achieved, but reliability remains a challenge [2].

Thermography is one non-invasive method to successfully measure RR [4], [5]. Thermal cameras sensitive to the radiation in the far-infrared (FIR) spectrum detect temperature fluctuations caused by the respiratory airflow visible in the regions of interest (ROIs) comprising the nostrils and/or the mouth. These changes in temperature are captured as oscillations of the pixel intensity within the ROI. Therefore, reliable and automatic detection of the ROI is a key factor for a robust RR monitoring system based on camera technology.

Despite the importance and clinical need for a reliable and automatic method to detect the ROI from FIR videos, recent approaches have proposed either the manual selection of the ROI [6], or developed and tested their methods under unrealistic experimental scenarios using bulky and expensive thermal cameras.

This work was supported by the Swiss National Science Foundation under Grant 150640.

Mobile Health Systems Lab, Institute of Robotics and Intelligent Systems, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland (email: gaetano.scebba@hest.ethz.ch)

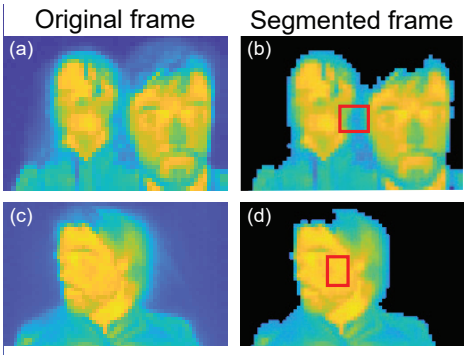

Fig. 1. Two examples where FIR image segmentation fails. When two subjects are visible in the camera's field of view (a) or the subject is in contact with a heat conductive surface (c), the face segmentation is erroneous (b, d) and consequently, the nose ROI incorrectly detected (red box).

The aim of this work was to develop a novel method for the automatic detection of the human nose ROI that is compatible with low-resolution thermal cameras, and improves the reliability of non-contact RR estimation with thermal imaging.

II. Related work

State-of-the-art methods for automated ROI detection rely on threshold-based methods, such as image segmentation and edge enhancement [7]. First, the image is segmented and the largest continuous area is labeled as the face region. Then, an edge enhancement mask is applied and the nose is identified in the central region of the segmented face area. This assumes an ideal scenario where only one subject is in the field of view and the face is the warmest region of the image.

In a more realistic scenario, this approach fails. If multiple subjects are visible in the camera's field of view (Fig. 1 a), the algorithm detects a single area that includes both faces (Fig. 1 b). If a subject is lying on a heat conductive surface, such as a bed (Fig. 1 c), the temperature of this surface increases, forming a thermal shadow. Again, the algorithm segments a larger area than desired (Fig. 1 d). In both cases, the image segmentation prevents the correct selection of the face, leading to an erroneous ROI identification.

Facial landmark detection in visible images is widely used in computer vision and is a key method for a number of commercially available applications, such as facial 3D modeling, person identification, and emotion recognition [8]. Due to commercial requirements and the growing availability of large, annotated image datasets, numerous reliable algorithms based on machine learning techniques are now available to identify facial features in RGB images.

To overcome the above-mentioned limitations in ROI

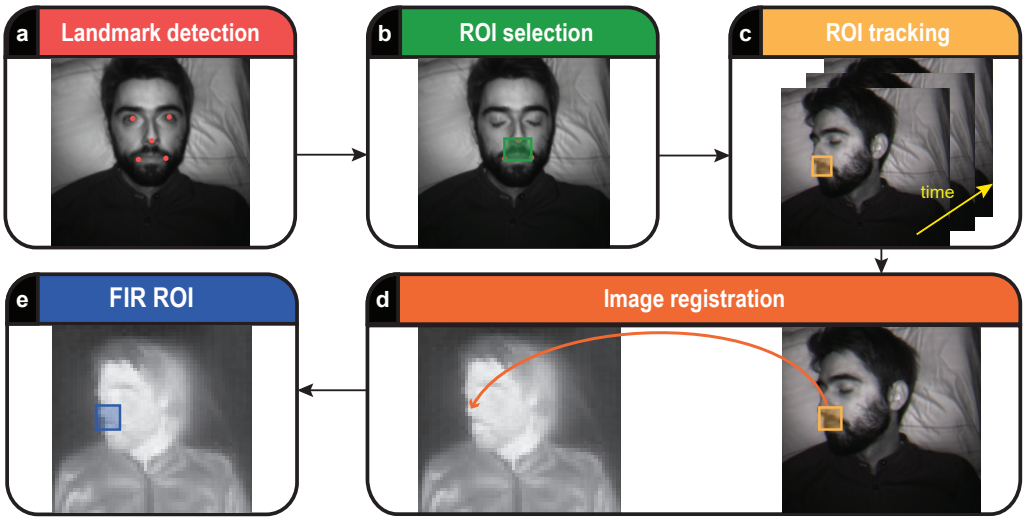

Fig. 2. Pipeline of the proposed multispectral ROI detection algorithm. (a) Facial landmarks are detected using neural networks, followed by (b) a ROI selection using the nose and the mouth corners. (c) The Kanade-Lucas-Tomasi algorithm tracks the position of the ROI over the RGB frames and compensates for head motion. (d) Coordinates of the tracked ROI are transformed into the (e) FIR image space.

identification, we investigated the benefits of a multispectral fusion approach that augments low-resolution thermal imaging with landmark detection performed in the visible spectrum, tested it with a previously unexplored protocol, and compared it to manual labels of nose ROI. This work provides a novel approach to accurately detect the ROI for non-contact estimation of the RR using thermography.

III. Algorithms

We propose a multispectral ROI detection algorithm pipeline that consists of facial landmark detection, ROI selection, tracking, and image registration using the combination of RGB and FIR images (Fig. 2). We use the RGB frames to identify and track the ROI, and define an image registration model to transform the ROI coordinates from the RGB to the FIR image space.

A. Facial landmark detection and ROI selection

We implemented the cascaded convolutional neural network (CCNN) model developed and trained by Zhang et al. to detect facial landmarks [9]. The model is able to identify five facial landmarks: two for the eyes, one for the nose, and two for the left and right corners of the mouth. Based on these points, we introduced the nose ROI as a rectangle with the distance between the mouth corners as width, and the distance between the nose and the lowest mouth corner as height. By selecting the entire nose/mouth area, the breathing signal can be extracted not only from the nostrils' airflow, but also from the mouth. In the cases where the CCNN model failed to detect the facial landmarks, the corresponding frame was discarded.

B. Tracking

We applied a tracker algorithm to the ROI in order to compensate for head motion. Once a valid ROI was detected, we extracted the feature points with a minimum eigenvalue algorithm developed by Shi and Tomasi [10]. The feature points were then tracked using the Kanade-Lucas-Tomasi (KLT) single points tracker algorithm [11], where a rigid affine transformation was computed based on the tracked points. To tackle potential errors in the tracking of the feature points, we adopted the method described by Kalal et al. to automatically detect the tracking failures [12] and exclude erroneously tracked points in the affine model estimation. This way, the motion of the ROI was modeled between consecutive frames. Detection was re-triggered every two seconds to further improve the robustness of the ROI identification.

C. Image registration

We registered images to accurately share the location of the ROI between RGB and FIR frames. The coordinates of the ROI identified in the RGB frame X_{RGB} and Y_{RGB} were transformed such as

$$\begin{bmatrix} X_{FIR} \\ Y_{FIR} \end{bmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \begin{bmatrix} X_{RGB} \\ Y_{RGB} \\ 1 \end{bmatrix},$$

where s_x and s_y specify scaling factors, and t_x and t_y specify the translation factors. Such factors are dependent on the distance between the subject and the camera. For this reason, the transformation model was defined during the calibration phase of the camera and the same parameters (scaling and translation factors) were applied to analyze all the recordings with the matching subject-to-camera distance.

IV. Methods

A. Experimental setup

We connected two separate cameras to a microcomputer (Raspberry Pi 3 B, Raspberry Pi Foundation, UK) to allow the synchronized recording of the RGB and FIR video streams. Custom software was developed for video data collection. It simultaneously captured the videos from the

cameras, compressed, and stored them on an SD card. RGB videos were recorded with a See3cam_CU40 camera (econ-Systems, Chennai, India) with a frame rate of 15 Hz and a resolution of 336x190 pixels. The FIR videos were recorded with a FLIR Lepton camera (FLIR Systems Inc., California, USA) with a frame rate of 8.7 Hz and a resolution of 80x60 pixels. A near-infrared LED array enabled recordings in environments with insufficient lighting.

B. Experimental protocol

After institutional ethics board approval (ETH Zurich 2017-N-60) and informed written consent, healthy volunteers were enrolled in the study. The experimental protocol contained three scenarios that were specifically designed to challenge the accurate identification of a ROI.

a) Standing: Subjects were asked to stand upright in front of the camera at 1 m for 2 minutes and breathe following a metronome at 3 different breathing rates (10, 20 and 40 breaths/min). This way ideal conditions with spontaneous motion artifacts were achieved.

b) Multiple: Two subjects were standing at approximately 1 m distance from the camera, partly behind each other. After 90 s, subject # 2 was asked to move in front of the camera partially covering subject # 1, stand there for 90 s, and to move back to their original position passing between subject # 1 and the camera. The total duration of the recording was 3 minutes.

c) Supine: Subjects were asked to adopt non-frontal head positions while lying on a bed with the camera placed at 80 cm above the bed. After 30 s of frontal head view, subjects rotated their head to $45°$ left and right. Each position was maintained for 2 minutes (total 4.5 minutes). The videos were recorded in a dark environment with no artificial lighting.

C. Analysis

We transferred both RGB and FIR videos from the SD card to a PC and performed the video processing with Matlab R2017b (MathWorks Inc., Natick, USA).

a) ROI detection rate: We used the ROI detection rate metric to quantify the success rate of the identification of the ROI in each frame [13] such as

$$ROI\ detection\ rate = \frac{\#ROI\ detected}{\#frames} \times 100.$$

b) ROI spatial accuracy: In order to evaluate the spatial accuracy of our multispectral ROI detection algorithm, the lead author manually labeled the nose area with a 16 pixel ^{2} rectangle, placing its top edge over the nose tip. The root mean square error (RMSE) between the central coordinates was computed such as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}^{n}(X_i^{aut} - X_i^{man})^2 + (Y_i^{aut} - Y_i^{man})^2},$$

where n is the number of automatically detected ROIs, and $X_i^{aut}, Y_i^{aut}, X_i^{man}, Y_i^{man}$ are the coordinates of the ROI central point.

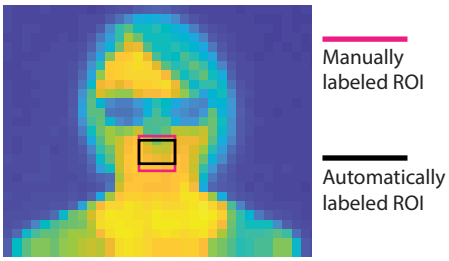

Fig. 3. Comparison between the automatically and manually detected ROIs.

V. Results

Videos of sixteen healthy subjects (7 female and 9 male, mean age: 25 ± 2 years) were recorded. Two recordings (subjects # 7 and # 11) of the *Supine* scenario were excluded due to a failure in storing the data and a misalignment of the camera system, respectively. Seven subjects had their face partially covered by a beard and six were wearing glasses. A total of 143 minutes of video were recorded. An example of the automatically and manually detected ROIs is depicted in Fig. 3.

A. ROI detection rate

There was a high success rate in the detection of the ROI (Fig. 4). The median ROI detection rate obtained across the three investigated scenarios was 100 % and the 5th-95th percentile range was 92 %-100 %. Only three subjects (# 12 in *Multiple*, # 16 and # 8 in *Supine* scenarios) showed lower ROI detection rate (94 %, 84 % and 73 %, respectively).

B. ROI spatial accuracy

Our multispectral ROI detection algorithm identified the ROI with a mean RMSE of 2 ±1 pixels over the three experimental scenarios (mean RMSE *Standing*: 2 pixels, *Multiple*: 2 pixels, *Supine*: 3 pixels). The mean ROI area of each scenario (12, 11, and 23 pixel^2, respectively) was comparable to the mean area of the manually labelled ROI (16 pixel^2).

VI. Discussion

We present a new method for the automatic detection of the nose ROI for low-spatial resolution FIR cameras using the sensor fusion of RGB and FIR images. Our multispectral nose ROI detection algorithm identifies the ROI solely in the RGB images using a deep-learning algorithm. The data fusion is realized by a geometric transformation of the ROI coordinates between RGB and FIR images. We evaluated the performance of the algorithm using videos of a diverse group of subjects exposed to real-world scenarios with challenging setups for ROI detection, such as multiple faces in the camera's field of view and different face orientations during a supine position.

Our findings demonstrate that the combination of different spectral image sources provides a reliable automatic ROI detection in low-resolution thermal images. This is supported not only by the high success rate of the detection, but also by the accuracy of the spatial location of the detected ROI,

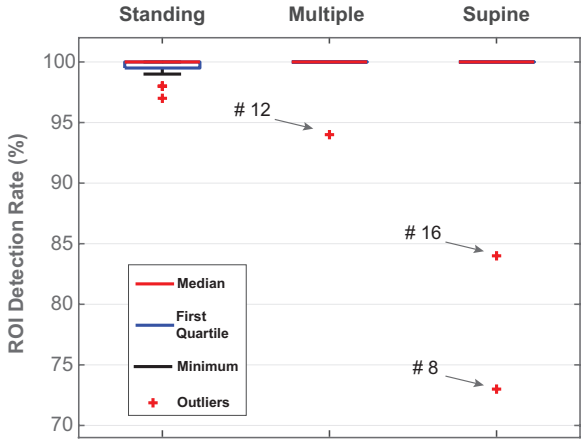

Fig. 4. ROI detection rates (%) computed for the *Standing* (n = 16), *Multiple* (n = 16), and *Supine* (n = 14) scenario. The boxplots illustrate the distribution across subjects. For *Multiple* and *Supine* scenarios, the median (red), the first quartile (blue), and minimum (black) overlap. Subjects indicated by arrows are outliers because subject # 12 sneezed during the recording, and subjects # 8 and # 16 had the face partially covered by a beard.

and the size of the mean area of the detected ROI compared to the manually labeled one.

We included subjects with a broad range of facial characteristics, such as facial hair and glasses, to specifically challenge our algorithm. Although our algorithm showed promising results for the majority of the tested subjects, we observed poor ROI detection rate during the *Supine* scenario for two subjects. These subjects' faces both were covered by a beard, which might have caused ROI detection failure during non-frontal views of the face. In these challenging situations, the FIR frame could be used as fallback source to detect the ROI, as we have shown previously [13].

In the case of the lower detection rate for subject # 12 in the *Multiple* scenario, the subject sneezed during the recording, resulting in large head movements that led to this decreased detection rate.

The challenge of reliably identifying a nose ROI in low-resolution thermal imaging was recently tackled by Cho et al., who proposed a novel algorithm based on the computation of the gradient map of only the FIR frame in the tracking step [6]. Although the experimental protocol included non-frontal views of the face, the authors designed an experimental setup where the camera was placed so close to the subject's face that it can no longer be considered non-invasive.

The camera-to-subject distance unavoidably influences the shape of the human face represented in the FIR image, especially when low-resolution FIR cameras are adopted. A set of standardized guidelines with detailed instructions for the subjects' facial characteristics, camera-to-subject distances, scene illumination, and realistic scenarios are needed for the advancement of non-contact RR estimation methods. In our case, the integration of two cameras increased the complexity of the experimental setup. Recent advances in thermal imaging technologies consist of new mobile imaging solutions that integrate both visible and thermal imaging modalities, making the integration of such a complex system effortless for non-contact camera-based methods used for RR estimation.

The reliability of our proposed approach can be further improved. The image registration model between the two cameras led to a dependence of our algorithm on the camera-to-subject distance. This dependency could be overcome by defining a spatial range, wherein the frames of two cameras would be accurately registered. Furthermore, a quantitative measurement of the signal quality extracted from the ROIs could further improve the reliability.

VII. Conclusion

We demonstrate that the adoption of RGB cameras provides acceptable reliability and accuracy in the ROI detection with low-resolution thermal cameras. This study addresses the challenge of detecting the nose ROI, which is a fundamental step for the non-contact estimation of the RR with nearable devices featuring thermal cameras.

Acknowledgements

We thank T. Meyer for the technical development and assistance with data collection, and J. Lim for commenting the manuscript. We are grateful to all the participants for volunteering in the study.

References

[1] P. B. Lovett, J. M. Buchwald, K. Stürmann, and P. Bijur, "The vexatious vital: Neither clinical measurements by nurses nor an electronic monitor provides accurate measurements of respiratory rate in triage," *Ann. Emerg. Med.*, vol. 45, no. 1, pp. 68–76, 2005.

[2] F. Q. Al-Khalidi, R. Saatchi, D. Burke, H. Elphick, and S. Tan, "Respiration rate monitoring methods: A review," *Pediatric Pulmonology*, vol. 46, no. 6, pp. 523–9, 2011.

[3] M. Elliott, "Why is Respiratory Rate the Neglected Vital Sign? A Narrative Review," *Int Arch Nurs Heal. Care*, vol. 2, no. 3, pp. 2–5, 2016.

[4] F. Q. Al-Khalidi, R. Saatchi, D. Burke, and H. Elphick, "Tracking human face features in thermal images for respiration monitoring," *Int. Conf. Comput. Syst. Appl. AICCSA 2010*, pp. 1–6, 2010.

[5] A. K. Abbas, K. Heimann, K. Jergus, T. Orlikowsky, and S. Leonhardt, "Neonatal non-contact respiratory monitoring based on real-time infrared thermography," *Biomed. Eng. Online*, vol. 10, no. 1, pp. 93–109, 2011.

[6] Y. Cho, S. J. Julier, N. Marquardt, and N. Bianchi-Berthouze, "Robust tracking of respiratory rate in high-dynamic range scenes using mobile thermal imaging," *Biomed. Opt. Express*, vol. 8, no. 10, pp. 4480–503, 2017.

[7] C. B. Pereira, X. Yu, M. Czaplik, R. Rossaint, V. Blazek, and S. Leonhardt, "Remote monitoring of breathing dynamics using infrared thermography," *Biomed. Opt. Express*, vol. 6, no. 11, pp. 4378–94, 2015.

[8] A. Zadeh, T. Baltrušaitis, and L.-P. Morency, "Convolutional experts network for facial landmark detection," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. - CVPR2017*, pp. 6–14, 2017.

[9] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–503, 2016.

[10] J. Shi and C. Tomasi, "Good features to track," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR1994*. IEEE Comput. Soc. Press, 1994, pp. 593–600.

[11] C. Tomasi and T. Kanade, "Detection and Tracking of Point Features," *Int. J. Comput. Vis.*, vol. 9, pp. 137–54, 1991.

[12] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-Backward Error: Automatic Detection of Tracking Failures," in *20th Int. Conf. Pattern Recognit. 2010*, 2010, pp. 2756–9.

[13] G. Scebba, J. Dragas, S. Hu, and W. Karlen, "Improving ROI detection in photoplethysmographic imaging with thermal cameras," in *39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 2017*, 2017, pp. 4285–8.