Nose Breathing or Mouth Breathing?
A Thermography-Based New Measurement for Sleep Monitoring

Huang, Zhengjie

Award date:
2020

Nose Breathing or Mouth Breathing? A Thermography-Based New Measurement for Sleep Monitoring

Zhengjie Huang, Wenjin Wang, and Gerard de Haan

Abstract

Nose breathing is preferred during sleep, although health issues may cause a subject to breathe through the mouth, and long-term mouth breathing may raise other health issues like sleep apnea. This paper proposes a first-ever classification of nose breathing and mouth breathing using the thermography of the subject. The measurement uses the relative temperature variations of different facial regions to classify mouth or nose breathing. This measurement is particularly health-/well-being relevant as it can be used as an early sign for sleep disorders or an indicator of sleep quality. An end-to-end processing flowchart has been provided for proof-of-concept validation on real-life recordings of thermal videos. Furthermore, we transfer knowledge from visible domain to thermal domain and augment existing data to address the lack of thermal training data to improve facial landmark accuracy. Two volunteers participated in our experiments and our proposed method achieved an overall classification accuracy of 96% in controlled lab conditions.

Index Terms

Thermography, Respiration.

Nose Breathing or Mouth Breathing? A Thermography-Based New Measurement for Sleep Monitoring

I. Introduction

HEALTHY people breathe with both nose and mouth. The nose can warm up and moisturize air from the environment. Also, the chemicals produced by the nose improve oxygen absorption in the lung. Breathing with mouth becomes necessary because of a blocked nose or high-intensity sports. Some people breathe with mouth occasionally while some breathe with mouth almost exclusively which in the long term can lead to a number of health issues like bad breath, periodontal disease, throat and ear infections [1], palatine, and pharyngeal tonsils hypertrophy [2]. It is even worse for children. A study consisting of 661 children participants aged from 6 to 12 years old shows that 26.8% of them are breathing with their mouth [3] and their facial growth can be affected that leads to unattractive facial features [4] if not treated in time. Furthermore, up to 42% of mouth breathers also have apnea according to a study [5]. Therefore, the mouth-or-nose breathing classification is important for the following reasons: early signs of mouth breathing can be captured by overnight monitoring for prevention purposes; the ratio of mouth breathing can be observed for evaluation of recovery from mouth breathing.

Due to the discomfort and inconvenience caused by the contact monitoring equipment, contact-less monitoring has been a popular topic over the past few years and has achieved great success. Most vital signs like respiratory rate, heart rate can be monitored remotely with cameras [6]–[10]. However, to the best of our knowledge, there is not even a contact-based method that monitors the subject's air passageway of breathing (e.g., nose, mouth, or both). Also, most of the respiratory rate monitoring methods based on RGB cameras count on the chest or abdominal motion [11], thorax motion [12] which can be inaccurate when apnea occurs since apnea sometimes is also accompanied with chest or abdominal motion. These drawbacks can be overcome with thermal cameras that capture the temperature information as airflow directly indicates a breath. Moreover, the price of thermal cameras has gone down dramatically thanks to the development of sensor technology that makes its application more affordable. Researchers [6], [13] have validated the use of the thermal camera for respiratory monitoring. Those thermal-based methods mostly assume nasal breathing which does not always hold. Therefore, it is also important to know whether the subject is breathing through nose or mouth while monitoring.

To allow more convenient and affordable monitoring of breathing air passageways, we present an end-to-end processing flowchart that can monitor the subject's breathing air passageways based on thermography in this paper. It is achieved simply with a thermal camera so that the monitoring is possible even at home and much of the inconvenience of in-hospital monitoring can be avoided. Our flowchart consists of two major components. First, face detection and facial landmark localization are done in the first frame to extract the nose and mouth; Then, we process signals (i.e., respiratory signal, motion signal) and compare the respiratory spectra from both regions to arrive at the nose/mouth breathing classification for every frame.

II. Methodology

In this section, we will give a detailed description of our processing flowchart as shown in Figure 1.

A. ROI extraction

To enable the measurement of the nose and mouth area, we first need to locate the nose and mouth, i.e., extract the ROI. Our ROI extraction is composed of three steps which are face detection, facial landmark localization, nose and mouth extraction.

1) Face detection: Face detection on RGB images has been advanced, by deep learning techniques and large annotated datasets, to an almost mature status. Nevertheless, due to the substantial differences between thermal images and RGB images, methods that work for RGB images do not necessarily work well on thermal images. Therefore, researchers are still working on robust face detection methods on thermal images. Pereira *et al.* [13] proposed to use the Otsu's multi-level threshold [14] to segment the image into multiple classes. It takes advantage of the fact that the face is usually the warmest object in the image. However, its performance degrades severely in scenes cluttered with objects of different temperatures and it includes some unwanted areas like neck area which is as warm as the face. The number of levels to segment is also dependent on the actual scene. Furthermore, their method relies on high-resolution thermal cameras that are many times more costly than those low-resolution ones. To obtain a Rectangular region of the face, Filipe *et al.* [15] project the image horizontally and vertically and calculate the maxima and minima of the projections to determine the start and end index of the face area. Marzec *et al.* [16] explored the characteristics of thermal facial images and assumed a few rules based on general temperature distribution on faces and facial anatomy. For example, the eyebrow area is usually less bright than the upper part of orbits area in thermography; the nose area is usually colder than the orbits area; the height of

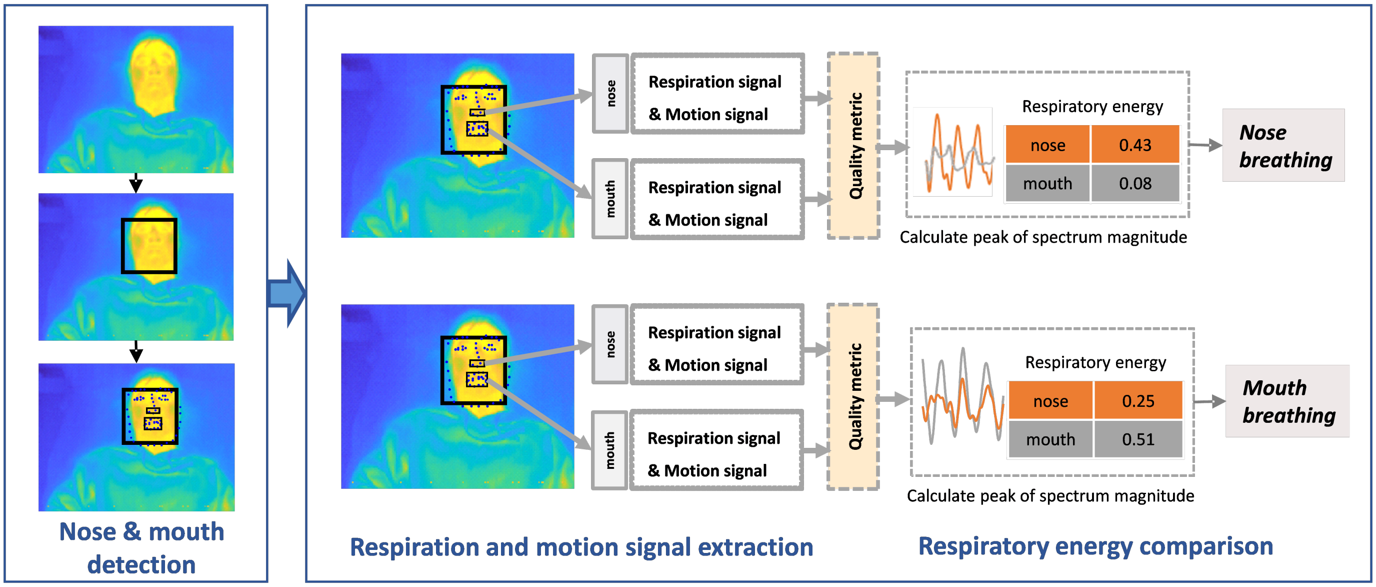

Nose & mouth detection
nose
mouth
Respiration signal & Motion signal
Respiration signal & Motion signal
Quality metric
Respiratory energy
nose 0.43
mouth 0.08
Calculate peak of spectrum magnitude
Nose breathing
nose
mouth
Respiration signal & Motion signal
Respiration signal & Motion signal
Quality metric
Respiratory energy
nose 0.25
mouth 0.51
Calculate peak of spectrum magnitude
Mouth breathing
Respiration and motion signal extraction
Respiratory energy comparison

Fig. 1. The flowchart for the proposed nose-or-mouth breathing classification consists of two main parts: (i) detect the face and facial landmarks to extract the nose and mouth region of interest; (ii) extract respiratory signals (temperature variations) from nose and mouth areas respectively and compare their respiratory spectra of both regions to arrive at the nose/mouth breathing classification.

a nose is similar to the length of the eyebrow, etc. Those rules may be applicable when detecting faces and selected points of the frontal view of a face but not for profile views. Histograms of Oriented Gradient (HOG) descriptors began to gain its popularity from 2005 after its success in human detection [17]. For classification purposes, the Support Vector Machine (SVM) [18] classifier is often used in combination with HOG descriptors. SVM is widely used as a binary or multi-class classifier due to its capability in separating data of high dimension into clusters. Kopaczka *et al.*conducted experiments on face detection and proved that machine learning-based methods are superior to specialized knowledge-based methods [19]. However, due to limited training data, it failed to detect faces of large poses.

Many heuristic face detection methods for thermal images assume a clear scene, their performance degrades severely with the presence of clutter and often include some unwanted area like the neck. Those limitations can be overcome with the utilization of facial features along with large training data. The face detector we used in this paper is a cascaded face detector. Firstly, we use a HOG-SVM detector trained on nearly 3000 thermal images [20]. Since the HOG descriptors are sensitive to the size of the image, we scale the image if no face can be detected in its original size. Secondly, if no face is detected in the previous stage, we use the heuristic method proposed by Filipe *et al.*since our input video always has a face in it.

2) Facial landmark localization: Facial landmarks can be helpful for tasks involving facial features like face recognition, emotion analysis, etc. In our case, instead of training two detectors for detecting nose and mouth separately, we use the landmarks to locate the ROI at once. Moreover, the facial anatomy can also increase confidence in ROI extraction such that things like the nose is under the mouth will not happen.

Existing facial landmark models can be roughly divided into three categories that are Constrained Local Model (CLM), holistic models, and machine learning-based models [21]. The development of facial landmark localization on RGB images has also been significant after deep learning came into play. Among all the machine learning-based models, deep architectures like ResNet [22] and Hourglass [23] are widely used due to the emerging of large annotated dataset and their proven success. Different from traditional Convolutional Neural Networks (CNN), the Hourglass architecture not only has a top-down, bottom up design but also intermediate supervision. Its distribution of capacity is more symmetric thanks to its residual module [22]. It is, therefore, able to capture features of different scales and retain them for later stages. Face Alignment Network (FAN) proposed by Bulat *et al.*is an Hourglass-based network. They trained the FAN on a 2D dataset and claimed to have achieved saturating performance on the dataset. Deep Alignment Network (DAN), proposed by Kowalski *et al.*, is a VGG-based network [24]. Its main contribution is that it makes use of landmark heatmaps to transfer the information of landmark estimates and refine the estimates iteratively across stages.

The research of facial landmark localization on thermal images is still at an early stage though. Due to substantial appearance of thermal images and RGB images, existing facial landmark models trained on RGB images do not work on thermal images. Moreover, thermal facial images are low in contrast and lack texture information compared to visible images due to the relatively uniform temperature distribution on the face. Kopaczka *et al.* [20] published a database of fully annotated thermal images, such that it is possible to train machine learning models for different purposes. They also trained an Active Appearance Model (AAM) on thermal images for facial landmark localization and then used the facial features extracted from facial landmarks for emotion classification. Furthermore, they evaluated the performance of an AAM (a holistic model) and a DAN [25] based model respectively. Their results showed that DAN outperforms AAM in many aspects including accuracy and speed. DAN also shows promise by outperforming the other two machine learning-based models - Multi-Task CNN and Patch-based fully convolutional neural network classifier (PBC) in experiments conducted by Poster *et al.* [26].

The advance of facial landmark model was not attributed to the prosperity of deep learning methods, but also large annotated dataset. However, when training data is not sufficient like in thermal domain, transfer learning can be of help. Transfer learning has been applied in many areas where machine learning is actively used like natural language processing (NLP) and Computer Vision. One notable example is sentiment classification where the task is to classify the reviews of a product into different categories. Training such a classifier is not hard nowadays. However, considering the amount of products, it is clearly not ideal to train a classifier for every product from scratch. Although the distribution of review data for different products can be different, they still have patterns in common which can be adapted to a new domain. This holds the same for our case where features like the face anatomy are shared for the thermal domain and visible domain.

To the best of our knowledge, the research of facial landmark localization on thermal images is quite limited. Considering that DAN has been proven feasible [20] and has its advantage over some traditional methods on thermal domain [26], we choose DAN as our facial landmark model. For comparison, we also trained FAN on thermal images since they performed similar on RGB images.

3) ROI validation: Given our ROI extraction methods as described before, there will always be a detected nose and mouth. However, there are cases that the detection results are not valid. For example, the landmarks may be incorrect; the landmarks are correct but the mouth or nose is covered by something leading to no temperature variations; large motion induced noises interfere heavily with respiratory signals. To exclude those cases, we propose a signal validation module that consists of two components: motion analyzer and signal validator. These two components are cascaded meaning that only if it passes one component will it go to the next component

Because we track the subject's face in the video on a frame by frame basis, the motion of the face will cause landmark variations in the temporal dimension. Since motion detection relies on temporal variations of landmarks, we define a window of 15 seconds to evaluation all motions within the window. The window size is determined as a result of a trade-off between response time (how long it takes for a window to perceive motion) and noise-tolerance (how much do noises affect our results). If the window is too large, a drastic motion may persist in many windows thus invalidate them. If the window is too short, noises are more likely to dominate our target signals either motion signals or respiratory signals. Such a window will also be consistently used in later stages for signal analysis and classification. By calculating the temporal variance of all landmarks as in Equation 1, we can know the amplitude of the motion. By rejecting all windows with amplitude higher than an empirical value, we can exclude windows that contain large motions from classification. Because the breathing air passageway is not likely to change frequently, we hold the last valid classification result until a new window is valid.

$$m = \frac{\sum_{i=1}^{n} \|p_i - \sigma(p_i)\|}{n} \tag{1}$$

where p_i is the position of the i-th landmark and $\sigma(p_i)$ refers to the variance of i-th landmark's location in a window.

Another assumption imposed by our method is that either nose or mouth will be used for breathing. Therefore, at least one of the two areas will have dominating respiratory signals given no motions involved. We use the energy percentage of signals within the respiratory band to represent the dominance of respiratory signal as in Equation 2. When the dominance value is larger than an empirical value which is 0.5 considering noises caused by sensor drift and slight motions, we say the signal is dominated by respiratory signals.

$$d = \frac{\sum_{f_{min} \leq f \leq f_{max}} Y_i(f)}{\sum_f Y_i(f)} \tag{2}$$

where f_{min} and f_{max} represents the lower bound and upper bound of the respiratory band respectively.

B. Mouth-or-nose breathing classification

The area used for air exchange has a higher temperature variation because the air inhaled from the environment is usually colder than the air exhaled through mouth or nose given the condition that the room temperature is stable and lower than the body temperature, which is usually the case. Therefore, the temperature variations caused by air exchange can be used for classification.

1) Respiratory signal extraction: Since the classification relies on temporal variations, it involves a video sequence. As aforementioned, we use a window of 15 seconds for signal analysis.

First, we take the average value of the nose and mouth area from each frame to compose the time-series signal:

$$\begin{cases} \mu_{\mathrm{n},k} = \frac{1}{S(V_\mathrm{n})} \sum_{(p,q)\in V_\mathrm{n}} i(p,q,k) \\ \mu_{\mathrm{m},k} = \frac{1}{S(V_\mathrm{m})} \sum_{(p,q)\in V_\mathrm{m}} i(p,q,k) \end{cases} \tag{3}$$

where $i(p,q,k)$ represents the intensity of pixel at position (p,q) of video frame k; V_n and V_m represents of pixel collection of the nose area and mouth area respectively. $S(V)$ represents the number of pixels in V.

In addition, inspired by [27], we also take the standard deviation of the nose and mouth area from each frame for complementary use:

$$\begin{cases} \sigma_{\mathrm{n},k} = \sqrt{\frac{1}{S(V_\mathrm{n})-1} \sum_{(p,q)\in V_\mathrm{n}} \left|i(p,q,k) - \mu_{n,k}\right|^2} \\ \sigma_{\mathrm{m},k} = \sqrt{\frac{1}{S(V_\mathrm{m})-1} \sum_{(p,q)\in V_\mathrm{m}} \left|i(p,q,k) - \mu_{m,k}\right|^2} \end{cases} \tag{4}$$

Then, we construct windows by concatenating signals from 15 seconds of frames for further processing.

$$\begin{aligned} R_i &= s_i | s_{i+1} | ... | s_{i+N-1} \\ R_{i+1} &= s_{i+1} | s_{i+2} | ... | s_{i+N} \end{aligned} \tag{5}$$

where s_i refers to $\mu_{\mathrm{n},k}$, $\mu_{\mathrm{m},k}$, $\sigma_{\mathrm{n},k}$, or $\sigma_{\mathrm{m},k}$ and R_i is the i-th respiratory window; N refers to the number of frames in a window (450 in our case); Symbol | means signal concatenation.

Furthermore, each window is normalized such that they have a mean value of 0 and a standard deviation of 1.

$$\hat{s}_i = \frac{s_i - \mu_s}{\sigma_s} \tag{6}$$

where μ_s and σ_s are the average and standard deviation of all signals in a window, respectively.

Due to sensor drift and turbulence from the environment, there are inevitably noises that are outside the respiratory rate (RR) range which is 12 to 18 cycles per minute (cpm) [28] in original signals. To exclude these effects, we filter all signals outside the frequency range of possible RR. We also widen the range to 12 to 40 to include some abnormal cases. A second-order Butterworth filter is used such that signals of frequency lower than 0.167 Hz and higher than 0.667 Hz (corresponds to 12 cpm and 40 cpm) are filtered from windowed respiratory signal R_i.

2) Respiratory energy comparison: After filtering out the noises, the remaining signals consist of mainly breathing signals. By comparing the time-domain signals, specifically, respiratory signals of the nose area and mouth area, we can know which one contributes to air exchange or if both of them do.

As aforementioned, we use a window of 15 seconds (450 frames) and slide it through the whole video with a stride of 1 frame. Therefore, we get a classification result for each frame in the video except for the last 15 seconds. For each window R_i, we get its spectrum Y_i by 1D Fourier Transform.

$$Y_i = \mathcal{F}(R_i) \tag{7}$$

and choose the highest magnitude E within the respiration band as an energy level indicator..

$$E_i = \max_{f_{min} \leq f \leq f_{max}} (|Y_i(f)|) \tag{8}$$

where $|Y_i(f)|$ refers to the magnitude spectrum at frequency f; f_{min} and f_{max} refers to the minimum and maximum frequency of respiration band which are 0.167Hz and 0.667Hz respectively.

During nose breathing, the respiratory signal of the nose area will be stronger than that of the mouth area i.e., $E_{nose} > E_{mouth}$. However, this is not necessarily the case during mouth breathing. Because if the subject's nose is clear, there might still exist some involuntary air exchange through the nose even when the subjects try to breathe through the mouth. Since nose breathing and mouth breathing are not mutually exclusive, it might be better to give a ratio indicating how much mouth breathing contributes to the whole air exchange instead of having a binary classifier.

$$p = \frac{E_{\text{mouth}}}{E_{\text{nose}} + E_{\text{mouth}}} \tag{9}$$

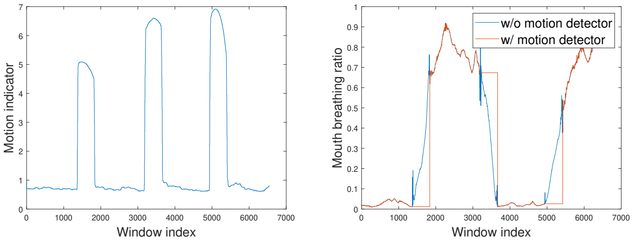


Fig. 2. Motion indicator and mouth breathing ratio of the first experiment. Motion occurs at the end of every minute accompanies with change of air passageways.

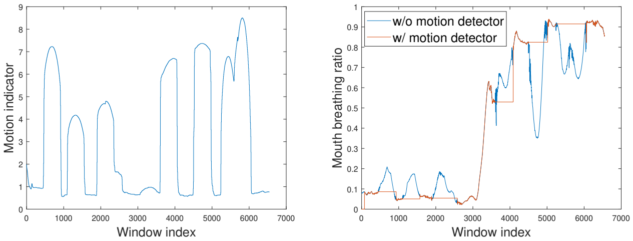


Fig. 3. Motion indicator and mouth breathing ratio of the second experiment.Motion occurs randomly. Subject was breathing through nose in the first two minute and mouth in the last two minute.

III. Experiments

This section contains three parts. The first part is about a motion detector. We will show an example of two videos with motions and how we exclude the invalid signals. The second part is about the facial localization methods. We will show how transfer learning and data augmentation techniques improve the training process and the final localization accuracy. The last part is about the classifier in which we describe the experimental setup and protocol for evaluating our proposed classification method as well as the results of our method.

A. Motion detector

In the experiments, we test our motion detector on different motions. In the first video, the subject was instructed to change air passageways with large motion every minute. In the second video, the subject was asked to have random and frequent motions while breathing through nose in the first two minutes and breathing through mouth in the last two minutes. The motion indicator and mouth breathing ratio for these two videos are shown in Figure 2 and 3. We can see that the appearance of the motion does increase the value of the motion indicator and after we exclude all the windows with unacceptable motions (motion indicator > 1), the remaining windows can be easily and correctly classified.

B. Domain Transfer and Data augmentation

We compare the performance of facial landmark models before and after applying domain transfer and data augmentation techniques in this section. As aforementioned, two state-of-art models will be used.

1) Model Training: We trained both networks as per their creator's setting to maximize their performance. FAN was implemented in Torch7 [29]. The initial learning rate was set to 10^{-4} which was dropped to 10^{-5} after 15 epochs and to 10^{-6} after another 15 epochs. We trained a total of 40 epochs and the accuracy indeed stopped increasing before it ended. DAN was trained end-to-end. The initial learning rate was set to 10^{-3}. An Adam optimizer [30] was used to adaptively adjust the learning rate and add momentum.

Initially, we trained both models only on the thermal dataset [20] as a baseline for both model. Then we trained both models on the augmented thermal dataset as an augmented version. The domain transfer was achieved by fine-tuning a model that had been trained on RGB dataset with the thermal dataset. Lastly, we combine the domain transfer and data augmentation by fine-tuning a RGB pre-trained model on augmented thermal dataset. The pre-trained version of DAN was trained on the 300W dataset [31] with a total of 3148 images and the Menpo dataset [32] with 6679 training image. FAN was pre-trained on 300-W-LP [33] which was augmented from 300-W dataset. To address the issue that there is a lack of annotated data and it is almost impossible to annotate the hidden landmarks of the face, Zhu *et al.*employed 3D models to augment annotated frontal faces to semi-frontal faces and even side-view faces.

2) Model evaluation: We first evaluate the performance of the training process. The training error of two networks with respect to the training epochs can be found in Figure 4. For both networks, we can see that the training error of those with domain transfer after the first epoch is a lot lower than those without. Only in a few epochs (2 to 3), the models reached an almost saturate status when domain transfer is applied. Also, the data augmentation, which increases the diversity of the dataset, improves accuracy. The standard way of evaluating the performance of a facial landmark model is to calculate the Normalized Mean Error (NME) between its predictions and the ground truth as in Equation 10. The mean error for an image is simply the average error of all landmarks.

$$d(\hat{s}, s_{gt}) = \frac{\|\hat{s} - s_{gt}\|}{l_{norm}} \tag{10}$$

where $\hat{s}$ and s_{gt} are the prediction and the ground truth respectively and $\|.\|$ denotes calculating the L2 norm of two landmarks. l_{norm} is a normalization value so that image size does not scale the error. In line with prior art [25] [31] [34], we take the inter-ocular distance as the normalized distance.

The mean error can indicate how well the prediction fits the face for one image. However, when it comes to a large dataset, only taking the mean is not very informative since it can be biased by a small portion of data. Therefore, we use the average error in combination with the Cumulative Error Distribution (CED) of the whole dataset. We also used the Area-Under-the-Curve (AUC) which is calculated as the area under the CED curve up to a error threshold and then divided by that threshold as in Equation 11.

$$\text{AUC}_\alpha = \frac{\int_0^\alpha \text{CED}(x)dx}{\alpha} \tag{11}$$

3) Performance Comparison: In this section, the performance of the final models will be presented to show the difference of models before and after applying techniques. The training error of the whole training process will be presented to showcase how domain transfer boost the training.

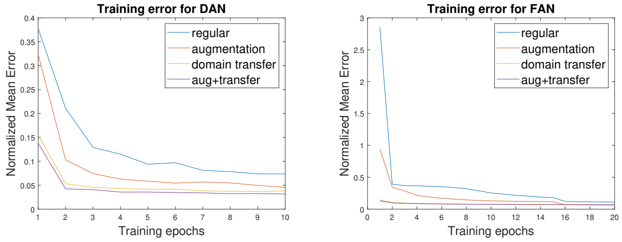


Fig. 4. Training error of DAN and FAN during the whole training process.

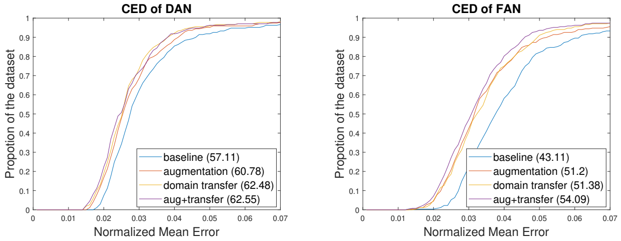


Fig. 5. CED of the final model of DAN and FAN. The corresponding $\text{AUC}_{0.07}$ (in %) follows the model name in the legend.

The training error of the training process for both models can be found in Figure 4. It is obvious for both models that only after one epoch of training, the models that applied domain transfer had an error that was a lot lower than those without and reached an almost saturate state in just a few epochs.

The CED curve of the best version (performs the best on validation set) of both models are in Figure 5. It also shows the benefits provided by domain transfer and data augmentation. Overall, either domain transfer or data augmentation have improved the model performance on the test set. The quantitative results - $\text{AUC}_{0.07}$ (in $\%$) of all models, can be found in the legends of Figure 5. With both techniques combined, a performance gain of 25% and 9.5% was achieved by FAN and DAN respectively.

Considering the superior performance of DAN, we chose DAN for facial landmark localization in later experiments.

C. Experimental setup and protocol

Two volunteers (2 males) participated in our experiments. Thermography videos were recorded using a FLIR E50 camera[1]. It is a Long Wave Infrared (LWIR) camera featured with thermal sensitivity of better than 0.05K and a spatial resolution of 240×180 pixels. The videos were recorded at 30 frames per second (fps). Furthermore, this camera does calibration/non-uniformity compensation (NUC) every three to four minutes by default in real-time to adapt the sensing range to the full span of the temperature range of the scene.

[1] www.flir.com

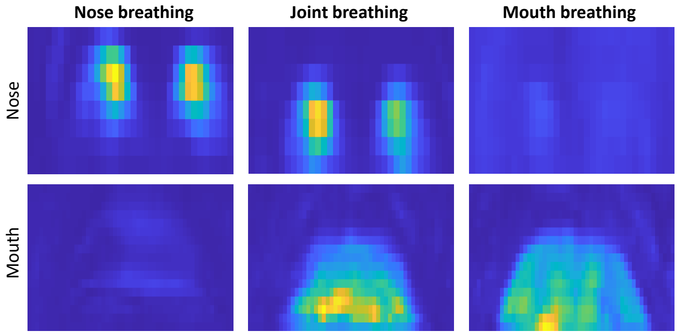


Fig. 6. Temporal temperature variations of the nose area and mouth area during nose breathing, joint breathing, and mouth breathing. Only the nose area has large temporal variations (airflow) during nose breathing and only mouth area has large temporal variations during mouth breathing while both areas have large temporal variation in joint breathing.

The calibration/NUC normally takes around 0.5s, after which the whole image is adjusted by an offset.

All videos were recorded with the subjects lying in the bed in order to simulate a sleeping condition, faces facing right towards the camera. The camera was placed at around 50 cm from the subject's face to cover the view of the whole pillow area such that the face was in the camera view, even with slight body motions. Each video is four-minute-long. Subjects were asked to breathe with their nose in the first minute and the third minute, mouth in the second and fourth minute while keep stationary and they strictly followed the time protocol. This protocol was used to generate the reference for the benchmark.

During the experiments, we found that some subjects had involuntary air exchange through the nose during mouth breathing which is neither purely nose breathing nor mouth breathing. Therefore, we define a third class of breathing called joint breathing in addition to nose breathing and mouth breathing in which both nose and mouth are used for air exchange. Temporal temperature variations of the nose and mouth area of the three breathing classes are shown in Figure 6. Theoretically, p (as defined in Equation 9) should be 0 during nose breathing and 1 during mouth breathing but considering the environment turbulence and sensor noises within the respiratory band, it is impractical to expect p to be exactly 0 or 1. Therefore, we empirically specify a value range for different classes: (1) $p < 0.2$: nose breathing; (2) $p > 0.8$: mouth breathing; (3) $0.2 \leq p \leq 0.8$: joint breathing.

We also manually annotate the windows with three breathing classes based on the temporal variation of the nose area and mouth area in the thermography sequence. And the major difference between the specified protocol and the annotated labels is that some subjects were actually breathing through both nose and mouth when asked to breathe through their mouth.

The accuracy of our method is measured by:

$$acc = \frac{n_{\text{success}}}{n_{\text{success}} + n_{\text{failure}}} \tag{12}$$

where n_{success} and n_{failure} refer to the number of successful classifications and wrong classifications.

D. Results and discussion

This section describes the experimental results of two test subjects. It also discusses the performance difference between mean traces and std traces. The evaluation was implemented and performed using MATLAB[2] (MATLAB R2019b, The MathWorks Inc., Natick, MA, USA).

1) Spectrogram analysis: Figure 7 compares the results obtained by using mean traces and standard deviation traces. Compared to the mean traces of the nose or mouth area which represents the average intensity of the area and captures temporal variations, the standard deviation traces are invariant to the area of ROI due to its characteristics. Moreover, the standard deviation is also immune to the self-calibration of thermal camera which introduces a global shift of the temperature value, as the local standard deviation does not change.

These two experiments show clear nose or mouth breathing which is quite distinguishable as shown in Figure 7.

2) Classification accuracy: The mouth breathing ratio traces of two test subjects can be found in Figure 8. As specified by the experimental protocol, the ratio should be low in the first and third minutes and high in the second and fourth minutes. There is also a transition period (in grey shadings) in which windows have both nose breathing and mouth breathing.

	s1	**s2**	avg
nose	1.00	0.73	0.87
joint	N/A	N/A	N/A
mouth	0.05	0.11	0.08
avg	0.53	0.42	0.48

TABLE I
CLASSIFICATION ACCURACY FOR TWO TEST SUBJECTS (MEAN-METHOD).

	s1	**s2**	avg
nose	1.00	1.00	1.00
joint	N/A	N/A	N/A
mouth	0.92	1.00	0.96
avg	0.98	1.00	0.98

TABLE II
CLASSIFICATION ACCURACY FOR TWO TEST SUBJECTS (STD-METHOD).

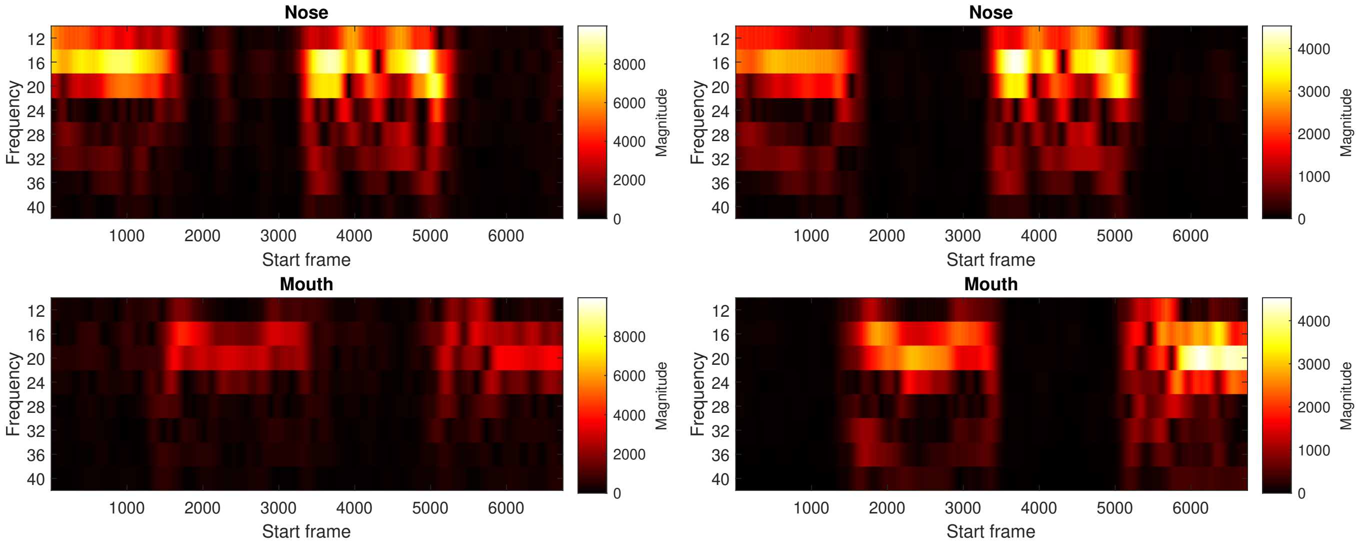
These two video recordings are four minutes long (240 seconds. 7200 frames), some of which are not available for classification because they contain both nose breathing and mouth breathing frames. Therefore, we have classification results for around 180 seconds (5400 windows).

The classification accuracy is shown in Table I and II. Overall, our proposed method worked descent on these video recordings and obtained an average accuracy of 98% with std and 48% with mean.

IV. CONCLUSION

In this paper, a conceptually new measurement, nose-or-mouth breathing classification, has been proposed using thermography, which has clinical/well-being relevance, and can be used as a new feature for sleep monitoring (e.g. early sign for sleep disorder). We also demonstrated this new measurement with an end-to-end image/signal processing flowchart. The

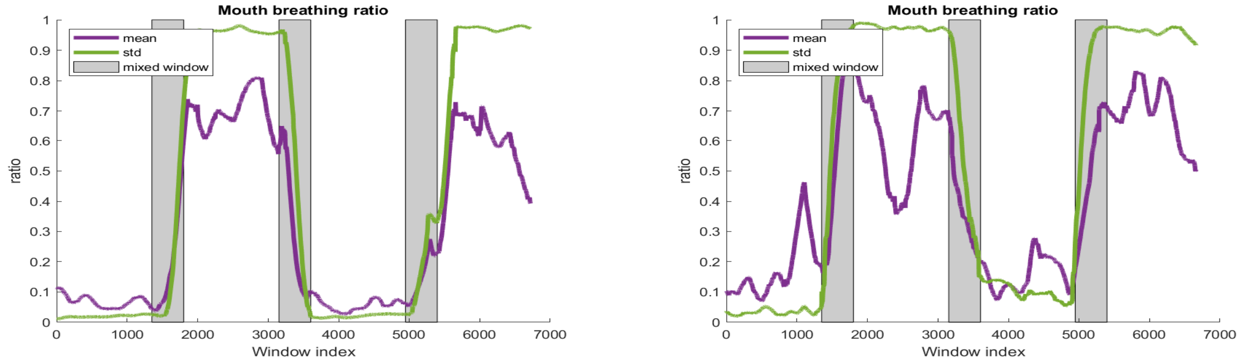
[2] www.mathworks.com



(a) Clear mouth/nose breathing (mean).

(b) Clear mouth/nose breathing (std).

Fig. 7. Spectrogram of respiratory signals from nose and mouth areas. Spectrograms on the left side are obtained by taking the mean value of the nose or mouth area while spectrograms on the right side are obtained by taking the standard deviation (std).



(a) Mouth breathing ratio of subject 1.

(b) Mouth breathing ratio of subject 2.

Fig. 8. Mouth breathing ratio traces of two subjects. It is pure nose breathing frames for windows whose start frame is less than 1350, between 3600 and 4950, pure mouth breathing frames for windows whose start frame is between 1800 and 3150, between 5400 and 6850. Others (shaded area) are mixed windows which consist of both nose breathing frames and mouth breathing frames.

results showed that our proposed method achieved a classification accuracy of 98% in controlled lab conditions. And the performance gain provided by domain transfer and data augmentation potentially unleash the possibilities of applications in thermal domain.

However, it is also worth noting that his study is only a lab-based feasibility study that assumes no motion, a frontal face, and limited environment turbulence. In real-life scenarios, some assumptions no longer hold. For instance, subjects are not likely to keep the same sleep position all night. They may sleep on the back or on the side alternately. Also, when the air circulation is fast (e.g., strong wind from natural or from electrical fans), the temperature variation of the nose and the mouth area can be disrupted. Furthermore, when it comes to side view, our proposed method may no longer work since the nostrils and mouth are not visible in the image.

In order to increase relevance in realistic sleeping conditions and improve classification accuracy, efforts will be necessary to improve nose/mouth localization in non-frontal sleeping poses, detect and exclude abnormal cases (e.g., motion, false localization). More real-life recordings should be made to identify issues that may arise in real-life scenarios for further revising our method.

References

[1] Healthline, "Mouth breathing: Symptoms, complications, and treatments," https://www.healthline.com/health/mouth-breathing, accessed: 2020-01-13.

[2] H. E. Montgomery-Downs and D. Gozal, "Sleep habits and risk factors for sleep-disordered breathing in infants and young toddlers in louisville, kentucky," *Sleep medicine*, vol. 7, no. 3, pp. 211–219, 2006.

[3] D. C. L. d. Santos *et al.*, "Study of the prevalence of predominantly oral breathing and possible implications for breastfeeding in schoolchildren from são caetano do sul - sp - brazil (original article is in portuguese)," *master's dissertation*, 2004.

[4] Y. Jefferson, "Mouth breathing: adverse effects on facial growth, health, academics, and behavior," *Gen Dent*, vol. 58, no. 1, pp. 18–25, 2010.

[5] S. C. Izu, C. H. Itamoto, M. Pradella-Hallinan, G. U. Pizarro, S. Tufik, S. Pignatari, and R. R. Fujita, "Obstructive sleep apnea syndrome (osas) in mouth breathing children," *Brazilian journal of otorhinolaryngology*, vol. 76, no. 5, pp. 552–556, 2010.

[6] J. Fei and I. Pavlidis, "Virtual thermistor," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 250–253.

[7] R. Janssen, W. Wang, A. Moço, and G. de Haan, "Video-based respiration monitoring with automatic region of interest detection," *Physiological measurement*, vol. 37, no. 1, pp. 100–104, 2015.

[8] B. G. Vainer, "A novel high-resolution method for the respiration rate and breathing waveforms remote monitoring," *Annals of biomedical engineering*, vol. 46, no. 7, pp. 960–971, 2018.

[9] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote ppg," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, 2016.

[10] Z. Zhu, J. Fei, and I. Pavlidis, "Tracking human breath in infrared imaging," in *Fifth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'05)*. IEEE, 2005, pp. 227–231.

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

[12] 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. 523–529, 2011.

[13] C. B. Pereira, X. Yu, M. Czaplik, R. Rossaint, V. Blazek, and S. Leonhardt, "Remote monitoring of breathing dynamics using infrared thermography," *Biomedical optics express*, vol. 6, no. 11, pp. 4378–4394, 2015.

[14] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[15] S. Filipe and L. A. Alexandre, "Thermal infrared face segmentation: A new pose invariant method," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2013, pp. 632–639.

[16] M. Marzec, R. Koprowski, and Z. Wróbel, "Detection of selected face areas on thermograms with elimination of typical problems," *Journal of medical informatics & technologies*, vol. 16, pp. 151–160, 2010.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] M. Kopaczka, J. Nestler, and D. Merhof, "Face detection in thermal infrared images: A comparison of algorithm-and machine-learning-based approaches," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2017, pp. 518–529.

[20] M. Kopaczka, R. Kolk, J. Schock, F. Burkhard, and D. Merhof, "A thermal infrared face database with facial landmarks and emotion labels," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 5, pp. 1389–1401, 2018.

[21] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] M. Kowalski, J. Naruniec, and T. Trzcinski, "Deep alignment network: A convolutional neural network for robust face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 88–97.

[26] D. Poster, S. Hu, N. Nasrabadi, and B. Riggan, "An examination of deep-learning based landmark detection methods on thermal face imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[27] W. Wang, A. C. den Brinker, and G. De Haan, "Full video pulse extraction," *Biomedical Optics Express*, vol. 9, no. 8, pp. 3898–3914, 2018.

[28] D. C. Rizzo, *Fundamentals of anatomy and physiology*. Cengage Learning, 2015.

[29] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS workshop*, no. CONF, 2011.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 397–403.

[32] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 170–179.

[33] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 146–155.

[34] Z.-H. Feng, G. Hu, J. Kittler, W. Christmas, and X.-J. Wu, "Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3425–3440, 2015.