Abstract

We study the impact of prior knowledge about invariance for the task of heart rate estimation from face videos in the wild (e.g., in presence of disturbing factors like rigid head motion, talking, facial expressions and natural illumination conditions under different scenarios). We introduce features invariant with respect to the action of a differentiable local group of local transformations. As result, the energy of the blood volume signal is re-arranged in vector space with a more concentrated distribution. The uncertainty in the feature distribution is incorporated with a model that leverages the local invariance of the heart rate. During experiments the method achieved strong estimation performance of heart rate from face videos in the wild. To demonstrate the potential of the approach it is compared against recent algorithms on data collected to study the impact of the mentioned nuisance attributes. To facilitate future comparisons, we made the code and data for reproducing the results publicly available.

1. Introduction

In general, it can’t be expected to obtain clear signals from sensors. Signals are often affected by nuisance factors hiding the target function. This is a major drawback and makes the analysis of processes difficult, enforcing several constraints for real applications. One of these applications is the task of vital parameter estimation from face videos under arbitrary natural conditions. The signal is heavily influenced by endogenous as well as exogenous factors, like face and head motion, illumination changes as well as specific sensor and image properties [21] and, last but not least, human conditions itself like age and health [37].

The role of physiological states has a large impact on human state computing in computer vision, since it holds informations about the affective nature of the human interacting with the machine. During the last years, measuring blood volume changes and heart rate measurements from facial images became a part of top computer vision conferences [18] [19] [26] [38]. All these contributions focus on how to cope with motion like head pose variations and facial expressions since any kind of motion on a specific skin region of interest (ROI) will destroy the raw signal in a way that no reliable information can be extracted anymore. Besides being able to estimate vitality parameters like heart rate and respiration, the functional survey of wounds as well as quantification of allergic skin reaction [3] are further applications of camera-based blood perfu-
Figure 2. Example images of the 25 user data collection. From left to right: Face recordings during head resting conditions, during head rotations, during an exercise on a bicycle ergometer in a gym and during an urban conversation.

sion analysis. Recently, prediction of emotional states, stress [22, 30, 5], fatigue [35] and sickness [14] became interesting new achievements in this area, pushing the focus of this technology further towards human-machine interaction.

In contrast to the genuine medical use-case of the technology, in computer vision and human-machine interaction we can’t expect any cooperative behavior of the user without introducing lack of convenience and a reduction of the general acceptance. Thus, there’s a demand to produce features and models robust to nuisance factors, still preserving the desired target information. In a more philosophical sense, where everything seems to change just search for what is not changing.

The main contributions of this work are

- a feature representation for the problem of estimating heart rate using low-cost camera sensor technology invariant under rigid and non-rigid facial motions and varying illumination,

- a direct connection of the proposed feature presentation to a functional mathematical formulation for the quasi-periodic and non-stationary nature of heart rate and

- a set of uncompressed face video recordings with reference measurements collected under resting, head motion, a bicycle ergometer scenario and during an urban conversation.

The outline of this work is as follows. From the historical genuine up to the development of the state of the art in computer vision, the methodology of heart rate estimation from face videos will be reviewed. Followed by theoretical aspects, the feature and model space will be described. Based upon an evaluation on collected data the results will be presented and finally discussed.

2. Related Work

Photoplethysmography, short PPG, dates back to the late first half of the 20th century, when Molitor and Kniazak [25] recorded peripheral circulatory changes in animals. A year later, Hertzman [15] introduced the term Photoelectric Plethysmograph as “the amplitude of volume pulse as a measure of the blood supply of the skin”. Hertzman’s instrumentation comprised mainly of a tungsten arc lamp and a photomultiplier tube. An advancement to the classical PPG is the camera based PPGI (with I for Imaging) introduced by the work of Blazek [4]. The basic principle behind the measurement of blood volume changes in the skin by means of PPG is the fact that hemoglobin absorbs light much stronger at specific frequency bands than the remaining skin tissues. The first published visualization of pulsatile skin perfusion patterns in the time and frequency domain is given by Blazek [4]. Since classical signal processing is applied mostly to extract information out of the perfusion signals [17, 28, 40]. However later it is realized that motion of the skin ROI [17] and micro motion of the head due to cardiac activity [2, 23] inherently induces artifacts into the extracted signal, especially when lighting is neither uniform nor orthogonal, canceling motion artifacts during signal processing became an important aspect for reliable skin blood perfusion measurements [24]. A basic early idea of compensating the motion of the skin ROI by optical flow methods directly in the image plane [17] is followed by Poh et al. [28], who proposed to solve the problem by blind source separation using Independent Component Analysis (ICA) over the different color channels. However, Wedekind et al. [43] compared ICA in multiple setting and principal component analysis and showed limitations of either transform. In fact, in case the underlying signal basis is majoritarian Gaussian, ICA will not be able to determine a proper de-mixing matrix and the independent components cannot be obtained in a deterministic order [8]. A solution to this problem excluding scenarios that have periodic movements is discussed by Macwan et al. [20]. Tarassenko et al. [36] attempted to cope with light flicker by using an auto-regressive modeling and pole cancellation. Haan and Jeanne [9] and De Haan and Van Leest [10] proposed to map the PPGI-signals by linear combination of RGB data to a direction that is orthogonal to motion induced artifacts. An alternative, which does not require skin-tone or pulse-related priors in contrast to the channel mapping algorithms, determines the spatial subspace of skin-pixels and measure its temporal rotation for signal extraction [32]. Tulyakov et al. [38] proposed matrix completion to jointly estimate reli-
able regions and heart rate estimates whereby Li et al. [19] applied an adaptive least square approach to extract robust pulse frequencies. Both reported performance gains similar to De Haan and Jeanne [9], however they used compressed video data during their experiments, which raises some doubts on the validity of results. Wang et al. [41] reported an orthogonal behavior of skin color and motion artifacts but introduced a static operator for feature transformation representing results on private data. A new stochastic model formulation was introduced by Pilz et al. [27] outperforming Wang et al. [42], however the proposed features based upon vector quantization seem to be a rather time consuming heuristic approach. All these important contributions share the problem that there exists no consensus about an unique benchmarking criterion as well as an agreement on suitable open data sets for fair comparison of algorithms.

3. Methodology

In pattern recognition the invariance problem is anchored as traditional paradigm where the classification aims to be invariant with respect to action of a group that acts on set of features [32]. More generally, an invariant with respect to an equivalence relation is a property that is constant on each equivalence class [43]. Therefore, the invariance criterion is a necessary condition for the generalization ability of learning algorithms. Invariance can be regarded in the feature and model space. For heart rate from face videos the basic feature is usually computed over a set of pixel intensities out of aligned face regions and the model space over a set of suitable frequencies. In the following we describe features invariant with respect to the action of the Lie group $SE(3)$ of rigid transformations, the Special Euclidean group $SE(3)$ and a stochastic frequency representation invariant with respect to the quasi-periodic nature and non-stationarity of heart rate. The model space is based upon the previous works of Särkkä [31] and Pilz et al. [27].

The Feature Space

Regarding a common optical sensor signal

$$\vec{p} \in \mathbb{R}^n = \{R, G, B\}, n = 3$$

(1)

as spatial expectation over a skin operator $s$ and function of time $t$

$$\bar{x}(t) = \int_0^{\infty} \mathbb{E} [\{\vec{p} \mid s(\vec{p})\}] dt$$

(2)

we assume this multivariate observation is drawn by a normal distribution

$$\bar{x}(t) \sim \mathcal{N}(\vec{\mu}, \sigma^2).$$

(3)

Local invariance of blood volume changes as function of time for each input feature $\bar{x}(t)$ under transformations of a differentiable local group of local transformations $\mathcal{L}_T$ [33]

$$\frac{\partial}{\partial t} \bigg|_{t=0} = f(\mathcal{L}_T, \bar{x}(t)) = 0$$

(4)

can be approximately enforced by minimizing the regularizer

$$\frac{1}{T} \sum_{i=1}^{l} \left( \frac{\partial}{\partial t} \bigg|_{t=0} f(\mathcal{L}_T, \bar{x}_i) \right)^2.$$

(5)

For the covariance matrix of the observation

$$\{\bar{x}_i : i = 1, ..., l\}$$

(6)

with respect to the transformations $\mathcal{L}_T$

$$C := \frac{1}{T} \sum_{i=1}^{l} \left( \frac{\partial}{\partial t} \bigg|_{t=0} \mathcal{L}_T, x_i \right) \left( \frac{\partial}{\partial t} \bigg|_{t=0} \mathcal{L}_T, x_i \right)^\top$$

(7)

and the corresponding symmetric eigenvalue problem

$$CV = \Lambda$$

(8)

we find an operator $P$ with corank $k = 1$ [34] for

$$\lim_{l \to \infty} P = I - VV^\top$$

(9)

and the corresponding feature vector

$$\tilde{x} = P \cdot \bar{x}.$$ 

(10)

The observation $\{\bar{x}_i : i = 1, ..., l\}$ is located on the null space defined by the projection operator $P$

$$\mathcal{H} = \mathcal{N}(P)$$

(11)

The hyperplane $\mathcal{H}$ is a linear subspace of $\mathbb{R}^n$.

Fig. 3 illustrates the feature transformation step.
For each resulting input feature $\tilde{x}$ under transformations of a differentiable local group of local transformations $L_{R}$ minimizing the corresponding regularizer, according to the principles in [4] and [5], yields to [29]

$$\lim_{l \to \infty} C_{\tilde{x}} - A \cdot \lim_{l \to 0} C_{\tilde{x}} \cdot A^{T} = 0$$

with

$$A = \lim_{l \to \infty} C_{\tilde{x}}^{1/2} \cdot \lim_{l \to 0} C_{\tilde{x}}^{-1/2}$$

and

$$b = \mathbb{E}[[\tilde{x}_{i}]] - \mathbb{E}[[A \cdot \tilde{x}_{1}]]$$

resulting to feature representation obtained by the linear transform

$$\tilde{x}_{SE(3)} = A \cdot \tilde{x} + b$$

Other common convenient solutions for the estimation of rigid transforms can be found following the works of Arun et al. [1] and Umeyama [39].

The choice of the interval $\lim_{l \to 0}$ plays a crucial role. We did experiments by computing the argument of the derivatives of the largest eigenvalue over the integral

$$\text{arg}(t) = \int_{0}^{2\pi} d\lambda_{1}(t) dt$$

with respect to the symmetric eigenvalue problem

$$\lim_{l \to \pi f t} C_{\tilde{x}}, V = VA.$$ (17)

Fig. 4 illustrates the feature rotation.

The observation $\{\tilde{x}_{SE(3)_{i}} : i = 1, ..., l\}$ is be distributed on the topological manifold

$$S^{1} = \{\tilde{x}_{SE(3)} \in \mathbb{C} | C \in \mathbb{H} : \|\tilde{x}_{SE(3)}\| = 1\},$$

for any real number $0 \leq c < 1$ with

$$\tilde{x}_{SE(3)_{c}} \in \mathbb{C}$$

for each harmonic component. Even when the frequency is discontinuous the signal remains continuous. Fig. 5 shows such a stochastic oscillator with time-varying frequency and amplitude. The stochastic state space for the resonator signal yields to

$$\frac{dx(t)}{dt} = F_{0}(f(t))x(t) + L(t),$$

$$c(t) = Hx(t).$$

Interestingly, this corresponds to the kernel formalism of a Gaussian process, where the Gaussian process is constructed as a solution to a $m$th order linear stochastic differential equation [13]. Here, the computational complexity
is linear, whereby the genuine kernel formalism is of cubic nature.

![Stochastic Oscillator](image)

Figure 5. A simulated trajectory of a stochastic oscillator with frequency trace in a possible range typical for a human in resting state.

Since the frequency is unknown, the state space depends on an additional latent variable

\[
\frac{dx(t)}{dt} = F_0(\theta) x(t) + L e(t),
\]

(28)

\[
c(t) = H(\theta) x(t)
\]

(29)

such that

\[
\theta \in \Omega = \{ \theta^1, \ldots, \theta^S \}
\]

(30)

forming a Markov chain with transition matrix \( \Pi \) with transition probabilities

\[
P(\theta_i | \theta_{i-1}) = \Pi_{ij}.
\]

(31)

The solution is given by computing the Gaussian mixture approximation to the joint posterior distribution of the latent variables and states [6]. Pilz et al. [27] stated that slow varying drifts can be modeled by a Wiener process, whereby any kind of violation of the smoothness criterion yields to a Poisson process. We didn’t consider the drift and jump case since the feature representation based upon local group invariance already accounts for such behavior.

4. Experiments

We divided the experimental procedure into two phases, a data collection and an algorithmic benchmarking.

Data

To justify the data collection effort, we state that to the best of our knowledge other data is either private therefore not accessible, recorded or distributed using image compression techniques or doesn’t reflect multiple recordings in order to study the specific questions concerning the mentioned problems that comes with the task of heart rate measurement from face videos. Although Heusch et al. [16] and Bobbia et al. [7] introduced public data for evaluating remote heart rate measurements along with reference implementation of recent algorithms, we received the video recordings done using image compression techniques or didn’t receive access privileges yet. Therefore we designed scenarios ranging from controlled and easy to uncontrolled and more difficult. We built up on the concept and user pool of Pilz et al. [27] but re-recorded the users to archive a better time synchronization between the camera and reference device. As result we recorded four different sessions. The first session consists of a resting scenario where no head or facial motion is performed and the illumination is more or less static. In the second session the users are asked to perform head as well as facial motions but the illumination remains static. The third session is performed during an exercise on a bicycle ergometer in a gym where no further instructions are given to the user. The fourth sessions is recorded during an urban conversation including head and facial motions as well as natural varying illumination conditions. Fig. 6 shows some example images taken from the recordings during the different sessions. Every session is recorded over an one minute time span, except the ergometer session which is recorded over a 5 minute period. In total 25 users were ask to participate resulting in an amount of 100 video recordings with approximately 200 min total duration; more than three hours. The data collection consists of 20 male and 5 female in the range of 25–42 years. The majority ethnicity is Caucasian. The camera device is selected as Logitech HD 270 webcam and as reference ground truth measurements we synchronized a common finger pulseoximeter, a CMS50E PPG device, over its serial port communication protocol. The average frame rate of the camera is set to 25 FPS and for the pulseoximeter 60 FPS. The camera video stream is captured uncompressed with auto-exposure and stored into an AVI container. For every captured image the time stamp is stored too. The pulseoximeter signal is stored together with the device’s pre-computed heart rate information.

Evaluation

The benchmarking of the described feature and model space is conducted against established methods. To this we count the ICA [28] approach as source separation method, the Spatial Subspace Rotation (SSR) [42] and the Projection Orthogonal to Skin (POS) [41] as feature transform methods. We performed tests for each session respectively as well as a separate run for the POS and the Local Group Invariance (LGI) on the entire data set. The signal processing procedure is selected to be equal for every approach. For each video frame a common face finder is used to pre-select the region of interest. A skin operator is applied onto the region by thresholding the blue- and red-difference chroma components. For the set of obtained RGB-pixels the expectation is computed and stored as three-dimensional time

\[1\text{We also re-implemented other methods [9, 19, 38, 20], since their code is not available. However, we obtained worse results.}\]
Figure 6. Comparison of a users spectrograms for the different data sessions computed with different algorithms. The rows (from top to down) reflect a head resting session, a head rotation session, a bicycle ergometer exercise and an urban conversation. The columns (from left to right) correspond to the reference PPG device, the green channel, the ICA source separation, the SSR, the POS and the Local Group Invariance (LGI) method. During the resting scenario each algorithm is able to extract reasonable heart rate information. When moving the head nearly every algorithm starts to fail. During the ergometer exercise the interference of the pedal frequency is visible over all algorithms. Here the LGI method benefits from its model space. The urban conversation as fully uncontrolled scenario with rigid and non-rigid facial motion along with illumination changes makes it very difficult to extract robust signals

5. Conclusions

In this work we have presented a functional approach for the task of heart rate estimation from face videos under the load of nuisance factors. We performed evaluation on data collected under everyday facial motions and environment-

series for further processing. Each signal obtained by the different algorithms is band-filtered in the range between 0.5 and 2.0 Hz. We increased the upper ranger to 2.5 Hz for the processing of the ergometer sessions. All filtered signals are then analyzed by standard Fourier based spectral method with window size of 256 samples and overlap of 90 percent. A maximum peak energy criterion is applied over the spectral traces to determine the heart rate candidates. The PPG signals are analyzed in the same way but initially resampled to the camera frame rate. Correlation coefficients are computed against the PPG reference heart rate together with the root-mean-square error (RMSE) for each session and algorithm respectively. For the two full database runs correlation plots and Bland-Altman plots are computed additionally.

Fig. 6 compares the spectrograms for a single users over all sessions against the different algorithms given the reference measurements. Under controlled motion free conditions stable performance is obtained by all algorithms. Under motion scenarios it becomes more and more challenging to robustly extract the heart rate whereby under the fully uncontrolled urban conversation most of the algorithms to fail completely. We observed this behavior during the entire evaluation. The results for each session are presented in Table 1. During the resting scenario the LGI method performs slightly worse. For all other sessions the LGI method archives quite robust results where the others mostly start to fail. Fig. 7 compares the estimation performance between the POS and the LGI approach. The correlation for the POS method is heavily affected by outliers. Although the LGI approach results in a better statistical performance, it shows an estimation bias of approximately 4 BPM. This also explains why the LGI method performs slightly worse during the resting scenario.

5. Conclusions

In this work we have presented a functional approach for the task of heart rate estimation from face videos under the load of nuisance factors. We performed evaluation on data collected under everyday facial motions and environmen-
Figures 7. Correlation and Bland-Altman plots for the POS [41] (upper plots) and LGI (lower plots) method over the entire data collection.

POS archives a correlation of 0.35 with a RMSE of 21 BPM and LGI a correlation of 0.87 with a RMSE of 11 BPM. Many outliers can be attributed to false predictions during the gym session where the heart rate is confused with the pedal frequency.

<table>
<thead>
<tr>
<th>Session</th>
<th>ICA</th>
<th>SSR</th>
<th>POS</th>
<th>LGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resting</td>
<td>0.97/1.4</td>
<td>0.97/2.0</td>
<td>0.96/2.1</td>
<td>0.96/3.3</td>
</tr>
<tr>
<td>Rotation</td>
<td>0.16/10.8</td>
<td>0.51/7.6</td>
<td>0.56/5.3</td>
<td>0.97/2.9</td>
</tr>
<tr>
<td>Gym</td>
<td>0.41/16.6</td>
<td>0.08/18.6</td>
<td>0.09/23.1</td>
<td>0.63/13.1</td>
</tr>
<tr>
<td>Talk</td>
<td>0.13/23.1</td>
<td>0.14/15.4</td>
<td>0.3/12.5</td>
<td>0.72/4.3</td>
</tr>
</tbody>
</table>

Table 1. Pearson’s correlation coefficient and RMSE of prediction for the ICA [28], the SSR [42], the POS [41] and the LGI method.

6. Acknowledgments

We would like to thank the MedAix fitness center in Aachen for providing the bicycle ergometers. This work was funded, in part, by the German Federal Ministry of Education and Research (BMBF) under grant agreement VIVID 01SF15024 and by CanControls GmbH Aachen.

References


