Self-Adaptive Matrix Completion for Heart Rate Estimation from Face Videos under Realistic Conditions

Sergey Tulyakov[1], Xavier Alameda-Pineda[1], Elisa Ricci[2,3], Lijun Yin[4], Jeffrey F. Cohn[5,6], Nicu Sebe[1]

[1]University of Trento, Via Sommarive 9, 38123 Trento, Italy
[2]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
[3]University of Perugia, Via Duranti 93, 06123, Perugia, Italy
[4]State University of New York at Binghamton, Binghamton, NY 13902, USA
[5]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[6]Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, USA
{sergey.tulyakov,xavier.alamedapineda,niculae.sebe}@unitn.it,
eliricci@fbk.eu, lijun@cs.binghamton.edu, jeffcohn@pitt.edu

Abstract

Recent studies in computer vision have shown that, while practically invisible to a human observer, skin color changes due to blood flow can be captured on face videos and, surprisingly, be used to estimate the heart rate (HR). While considerable progress has been made in the last few years, still many issues remain open. In particular, state-of-the-art approaches are not robust enough to operate in natural conditions (e.g. in case of spontaneous movements, facial expressions, or illumination changes). Opposite to previous approaches that estimate the HR by processing all the skin pixels inside a fixed region of interest, we introduce a strategy to dynamically select face regions useful for robust HR estimation. Our approach, inspired by recent advances on matrix completion theory, allows us to predict the HR while simultaneously discover the best regions of the face to be used for estimation. Thorough experimental evaluation conducted on public benchmarks suggests that the proposed approach significantly outperforms state-of-the-art HR estimation methods in naturalistic conditions.

1. Introduction

After being shown in [23, 18] that changes invisible to the naked eye can be used to estimate the heart rate from a video of human skin, this topic has attracted a lot of attention in the computer vision community. These subtle changes encompass both color [27] and motion [4] and they are induced by the internal functioning of the heart. Since faces appear frequently in videos and due to recent and significant improvements in face tracking and alignment methods [3, 21, 13, 14, 29], facial-based remote heart rate estimation has recently become very popular [17, 30, 10, 25].

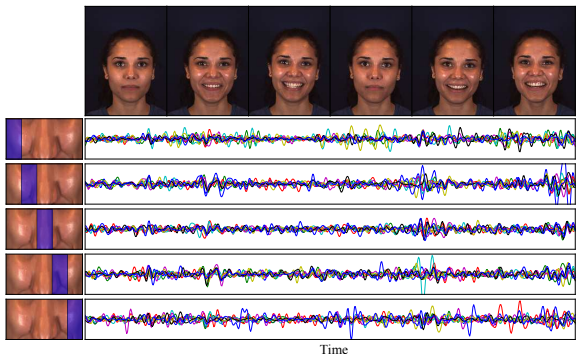

Figure 1. Motivation: Given a video sequence, automatic HR estimation from facial features is challenging due to target motion and facial expressions. Facial features extracted over time in different parts of the face (purple rectangles) show different temporal dynamics and are subject to noise, as they are heavily affected by movements and illumination changes. In this paper, we propose a novel approach to simultaneously estimate the HR signal and select the reliable face regions at each time for robust HR prediction.

Classical approaches successfully addressed this problem under laboratory-controlled conditions, *i.e.* imposing constraints on the subject's movements and requiring the absence of facial expressions and mimics [18, 27, 4]. Therefore, such methods may not be suitable for real world applications, such as monitoring drivers inside a vehicle or people exercising. Long-time analysis constitutes a further limitation of existing works [17, 18, 19]. Indeed, instead of es-

timating the *instantaneous* heart rate, they provide the *average* HR measurement over a long video sequence. The main disadvantage of using a long analysis window is the inability to capture interesting short-time phenomena, such as a sudden HR increase/decrease due to specific emotions [22].

In practice, another problem faced by researchers developing automatic HR measurement approaches, is the lack of publicly available datasets recorded under realistic conditions. A notable exception is the MAHNOB-HCI dataset [20], a multimodal dataset for research on emotion recognition and implicit tagging, which also contains HR annotations. Importantly, an extensive evaluation of existing HR measurement methods on MAHNOB-HCI have been performed by Li *et al.* [17]. However, the MAHNOB-HCI dataset suffers from some limitations, since the recording conditions are quite controlled: most of the video sequences do not contain spontaneous facial expressions, illumination changes or large target movements [17].

In this work, we tackle the aforementioned problems by introducing a novel approach for HR estimation from face videos and providing an extensive evaluation on two datasets: the MAHNOB-HCI, previously used for HR recognition research [17], and a spontaneous dataset with heart rate data and RGB videos (named MMSE-HR), which is a subset of the larger multimodal spontaneous emotion corpus (MMSE) [31] specifically targeted to challenge HR estimation methods.

Inspired by previous methods, we track the face in a given video sequence, so to follow rigid head movements [17], and extract chrominance features [10] to compensate for illumination variations. Importantly, most previous approaches preselect a face region of interest (ROI) that is kept constant through the entire HR estimation. However, the region containing useful features for HR estimation is a priori different for every frame since major appearance changes are spatially and temporally localized (Fig.1). Therefore, we propose a principled data-driven approach to automatically detect the face parts useful for HR measurement, that is to estimate the *time-varying mask* of useful observations, selecting at each frame the relevant face regions from the chrominance features themselves.

Recent advances on matrix completion (MC) theory [11] have shown the ability to recover missing entries of a matrix that is partially observed, *i.e. masked*. Up to the authors knowledge, we propose the first matrix completion-based learning algorithm able to *self-adapt*, that is to automatically select the useful observations, and call it *self-adaptive matrix completion* (SAMC). Intuitively, while learning the mask allows us to discard those face regions strongly affected by facial expressions or large movements, completing the matrix smooths out the smaller noise associated to the chrominance feature extraction procedure. The experiments we conducted on the MANHOB-HCI dataset clearly show that our method outperforms the state-of-the-art approaches for HR prediction. To further demonstrate the ability of our method to operate in challenging scenarios, we report a series of tests on the MMSE-HR dataset, where subjects show significant movements and facial expressions.

Thus, the contribution of this paper is three-fold:

- We present a novel approach to address the problem of HR estimation from face videos in realistic conditions. To cope with large facial variations due to spontaneous facial expressions and movements, we propose a principled framework to automatically discard the face regions corresponding to noisy features and only use the reliable ones for HR prediction. The region selection is addressed within a novel matrix completion-based optimization framework, called *self-adaptive matrix completion*, for which an efficient solver is proposed.

- Our approach is demonstrated to be more accurate than previous methods for *average* HR estimation on publicly available benchmarks. In addition, we report short-term analysis results to show the ability of our method to detect *instantaneous* heart rate.

- We perform extensive evaluation on the commonly used MAHNOB-HCI dataset and a spontaneous MMSE-HR dataset including 102 sequences of 40 subjects, moving and performing spontaneous facial expressions. As we show, this dataset is valuable for instantaneous HR estimation.

2. Related Work

In this section, we briefly review previous works on remote heart rate measurement and on matrix completion.

2.1. HR Estimation from Face Videos

Cardiac activity measurement is an essential tool to control the subjects' health and is actively used by medical practitioners. Conventional contact methods offer high accuracy of cardiac cycle. However, they require specific sensors to be attached to the human skin, be it a set of electrocardiogram (ECG) leads, a pulse oximiter, or the more recent fitness tracker. To avoid the use of invasive sensors, non-contact *remote* HR measurement from visual data has been proposed recently by computer vision researchers.

Verkruysse *et al.* [23] showed that ambient light and a consumer camera can be used to reveal the cardio-vascular pulse wave and to remotely analyze the vital signs of a person. Poh *et al.* [18] proposed to use blind source separation on color changes caused by heart activity to extract the HR signal from a face video. In [27] an Eulerian magnification method is used to amplify subtle changes in a video stream

and to visualize temporal dynamics of the blood flow. Balakrishnan *et al.* [4] showed that subtle head motions are affected by cardiac activity, and these motions can be used to extract HR measurements from a video stream.

However, all these methods failed to address the problems of HR estimation in presence of facial expressions and subject's movements, despite their frequent presence in real-world applications. This limits the use of these approaches to laboratory settings. In [10, 25] a chrominance-based method to relax motion constraints was introduced. However, this approach was tested on a few not-publicly-available sequences, making it hard to compare with.

Li *et al.* [17] proposed an approach based on adaptive filtering to handle illumination and motion issues and they evaluated it on the publicly available MAHNOB-HCI dataset [20]. However, although this work represents a valuable step towards remote HR measurement from visual data, it also shares several major limitations with the previous methods. The output of the method is the *average* HR, whereas to capture short-term phenomena (*e.g.* HR variations due to instantaneous emotions) the processing of smaller time intervals is required. A further limitation of [17] is the MAHNOB-HCI dataset itself, since it is collected in a laboratory setting and the subjects are required to wear an invasive EEG measuring device on their head. Additionally, subjects perform neither large movements nor many spontaneous facial expressions.

In this work, we address the aforementioned limitations by proposing a novel method capable of predicting HR with higher accuracy than the state-of-the-art approaches and of robustly operating on short time sequences in order to detect the instantaneous HR. To our knowledge, while previous works [17, 25] have acknowledged the importance of selecting parts of the signal to cope with noise and provide robust HR estimates, this paper is the first to tackle this problem within a principled optimization framework.

2.2. Matrix completion

Matrix completion [11] approaches develop from the idea that an unknown low-rank matrix can be recovered from a small set of entries. This is done by solving an optimization problem, namely, a rank minimization problem subject to some data constraints arising from the small set of entries. Matrix completion has proved successful for many computer vision tasks, when data and labels are noisy or in the case of missing data, such as multi-label image classification [6], image retrieval and tagging [28, 9], manifold correspondence finding [16], head/body pose estimation [1] and emotion recognition from abstract paintings [2]. Most of these works extended the original MC framework by imposing task-specific constraints. For instance, in [9] a MC problem is formulated adding a specific regularizer to address the ambiguous labeling problem. Very importantly, even if most computer-vision papers based on matrix completion are addressing classification tasks, therefore splitting the matrix to be completed between features and labels, MC techniques can be used in general, without any structural splitting. Indeed, in [15] matrix completion is adopted to address the movie recommendation problem, where each column (row) represents a user (movie), and therefore each entry of the matrix shows the suitableness of a video for a user. In [16, 15], the MC problem is extended to take into account an underlying graph structure inducing a weighted relationship between the columns/rows of the matrix. In this paper, we were inspired by [16, 15, 1] in modeling the temporal smoothness of the HR signal. However, our method is essentially novel, since we are able to simultaneously recover the unknown low-rank matrix and the underlying data mask, corresponding to the most reliable observations.

3. HR Estimation using SAMC

In this section we describe the proposed approach for HR estimation from face videos, that has four main phases as shown in Figure 2. Phase 1 is devoted to process face images so to extract face regions, that are used in phase 2 to compute chrominance features. Phase 3 consists in the joint estimation of the underlying low-rank feature matrix and the mask using SAMC. Finally, phase 4 computes the heart rate from the signal estimate provided by SAMC.

3.1. Phases 1 & 2: From Face Videos to Chrominance Features

Inspired by previous methods on remote HR estimation, we use Intraface[1] to localize and track 66 facial landmarks. Many approaches have been employed for face frontalisation [24, 12]. However, in order to preserve the underlying blood flow signal, in the current study we define the facial region of interest (see Fig. 2-Phase 1), from which the HR will be estimated. The potential ROI is then warped to a rectangle using a piece-wise linear warping procedure, before dividing the potential ROI into a grid containing R regions.

The overall performance of the HR estimation method will strongly depend on the features extracted on each of the R sub-regions of the facial ROI. Ideally, we would select features that are robust to facial movements and expressions, while being discriminant enough to account for the subtle changes in skin color. Currently, the best features for HR estimation are the chrominance features, defined in [10]. The chrominance features for HR estimation are derived from the RGB channels, as follows. For each pixel the chrominance signal C is computed as the linear combination of two signals X_f and Y_f, *i.e.* $C = X_f - \alpha Y_f$, where $\alpha = \frac{\sigma(X_f)}{\sigma(Y_f)}$ and $\sigma(X_f)$, $\sigma(Y_f)$ denote the standard

[1] http://www.humansensing.cs.cmu.edu/intraface

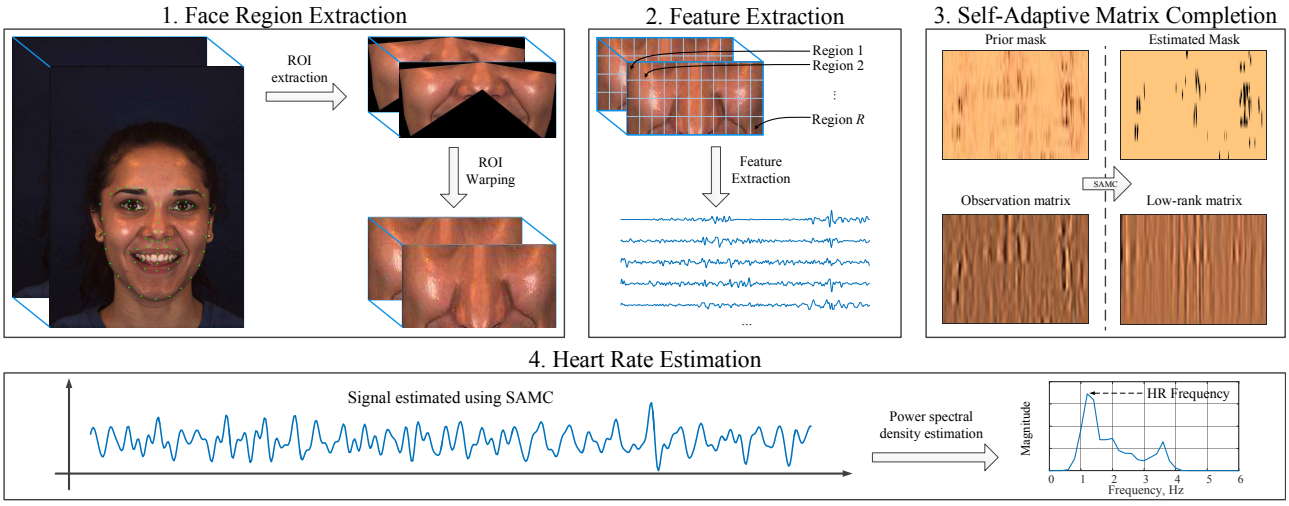

Figure 2. Overview of the proposed approach for HR estimation. During the first phase, we automatically detect a set of facial keypoints and use them to define a ROI. This region is then warped to a rectangular area and divided into a grid. For each small sub-region, chrominance features are computed (Phase 2). We then apply SAMC on the matrix of all feature observations to recover a smooth signal, while selecting from which sub-regions the signal is recovered (Phase 3). Welch's method [26] is used to estimate the power spectral density and thus the HR frequency (Phase 4).

deviations of X_f, Y_f. The signals X_f, Y_f are band-passed filtered signals obtained respectively from the signals X and Y, where $X = 3R_n - 2G_n$, $Y = 1.5R_n + G_n - 1.5B_n$ and R_n, G_n and B_n are the normalized values of the individual color channels. The color combination coefficients to derive X and Y are computed using a skin-tone standardization approach (see [10] for details). For each region $r = 1, \ldots, R$, the final chrominance features are computed averaging the values of the chrominance signals over all the pixels.

3.2. Phase 3: Self-Adaptive Matrix Completion

The estimation of HR from the chrominance features is challenging for mainly two reasons. Firstly, the chrominance features associated to different facial regions are not fully synchronized. In other words, even if the output signals of many regions are synchronized between them (mainstream underlying heart signal), the signal of many other regions may not be in phase with the mainstream. Secondly, face movements and facial expressions induce strong perturbations in the chrominance features. These perturbations are typically local in space and time while large in intensity (Fig.1). Therefore, we need to localize where these perturbations take place so not to use them in the HR estimation.

These two main difficulties are intuitively overcome by deriving a matrix completion technique embedding a *self-adaptation* strategy. On the one hand, since matrix completion problems are usually approached by reducing the matrix rank, the low-rank estimated matrix naturally groups the rows by their linear dependency. In our particular case, two rows are (near) linearly dependent if and only if the output signals they represent are synchronized. Therefore, the underlying HR signal is hypothesized to be in the vector subspace spanned by the largest group of linearly dependent rows of the estimated low-rank matrix.

On the other hand, the estimated low-rank matrix is enforced to resemble the observations. In previous MC approaches [6, 9, 1, 16], the non-observed part of the matrix consisted of the labels of the test set. Thus, the set of unknown matrix entries was fixed and known in advance. The HR estimation problem is slightly different since there are no missing observations, *i.e.* the matrix is fully observed. However, many of these observations are highly noisy, thus corrupting the estimation of the HR. Importantly, we do not know in advance which are the corrupted observations. This is why we believe that this problem naturally requires some form of adaptation, implying that the method selects the samples with which the learning is performed. Consequently, we name the proposed learning method *self-adaptive matrix completion* (SAMC).

In order to formalize the self-adaptive matrix completion problem let us assume the existence of R regions where chrominance features are computed during T video frames. This provides a chrominance observations matrix $\mathbf{C} \in \mathbb{R}^{R \times T}$. Ideally, in a scenario where we could trust all region features continuously, we would simply estimate the low-rank matrix that better approximates the matrix of observations $\mathbf{C}$, by solving: $\min_{\mathbf{E}} \nu \,\mathrm{rank}(\mathbf{E}) + \|\mathbf{E} - \mathbf{C}\|_{\mathcal{F}}^2$, where ν is a regularization parameter. Unfortunately, minimizing the rank is a NP-hard problem, and traditionally a

convex surrogate of the rank, the nuclear norm, is used [8]:

$$\min_{\mathbf{E}} \nu\|\mathbf{E}\|_* + \|\mathbf{E}-\mathbf{C}\|_{\mathcal{F}}^2. \quad (1)$$

Another intrinsic property of the chrominance features is that, since the underlying reason of their oscillation is the internal functioning of the heart, we should enforce the estimated chrominance features (those of the low-rank estimated matrix) to be within the heart-rate's frequency range. Inspired by [15, 16, 1] we add a temporal smoothing term by means of a Laplacian matrix $\mathbf{L}$:

$$\min_{\mathbf{E}} \nu\|\mathbf{E}\|_* + \|\mathbf{E}-\mathbf{C}\|_{\mathcal{F}}^2 + \gamma\,\mathrm{Tr}(\mathbf{E}\mathbf{L}\mathbf{E}^\top), \quad (2)$$

where γ measures the weight of the temporal smoothing within the learning process. $\mathbf{L}$ should encode the relational information between the observations acquired at different instants, thus acting like a relaxed band-pass filter. Indeed, imposing that $\mathbf{e}_r$ is band-pass filtered is equivalent to reduce $\|\mathbf{e}_r - \mathbf{e}_r\mathbf{T}\|^2 = \|\mathbf{e}_r\tilde{\mathbf{T}}\|^2$, where each column of $\mathbf{T}$ is a shifted replica of the band-pass normalized filter tap values so that the product $\mathbf{e}_r\mathbf{T}$ boils down to a convolution and $\tilde{\mathbf{T}}$ is a copy of $\mathbf{T}$ with zeros in the diagonal, since the band-pass filter is normalized. Imposing this for all R regions at once writes: $\mathrm{Tr}(\mathbf{E}\tilde{\mathbf{T}}\tilde{\mathbf{T}}^\top\mathbf{E}^\top)$, and therefore $\mathbf{L} = \tilde{\mathbf{T}}\tilde{\mathbf{T}}^\top$.

As previously discussed, the estimated matrix should not take into account the observed entries associated to large movements or spontaneous facial expressions. We model this by including a masking binary matrix $\mathbf{M} \in \{0,1\}^{R\times T}$ in the previous equation as [6]:

$$\min_{\mathbf{E}} \nu\|\mathbf{E}\|_* + \|\mathbf{M}\circ(\mathbf{E}-\mathbf{C})\|_{\mathcal{F}}^2 + \gamma\,\mathrm{Tr}(\mathbf{E}\mathbf{L}\mathbf{E}^\top), \quad (3)$$

where $\circ$ stands for the element-wise (Hadamard) product and the entries of the matrix $\mathbf{M}$ are 1 if the corresponding entry in $\mathbf{C}$ has to be taken into account for the HR estimation and 0 otherwise.

Importantly, while in the previous studies $\mathbf{M}$ was known in advance, in the present study we have to estimate it. We naturally interpret this as a form of adaptation since $\mathbf{M}$ is a observation-selection variable indicating from which observations should the method learn at each iteration. The masking matrix $\mathbf{M}$ should select the largest possible amount of samples that provide useful information for the estimation of the HR. Moreover, when available, it would be desirable to use a prior for the mask $\mathbf{M}$, taking real values between 0 and 1, $\widetilde{\mathbf{M}} \in [0,1]^{R\times T}$. The complete SAMC optimization problem writes:

$$\min_{\mathbf{E},\mathbf{M}} \nu\|\mathbf{E}\|_* + \|\mathbf{M}\circ(\mathbf{E}-\mathbf{C})\|_{\mathcal{F}}^2 + \gamma\,\mathrm{Tr}(\mathbf{E}\mathbf{L}\mathbf{E}^\top) - \beta\|\mathbf{M}\|_1 + \mu\|\mathbf{M}-\widetilde{\mathbf{M}}\|_{\mathcal{F}}^2, \quad (4)$$

The parameters β and μ regulate respectively the number of selected observations and the importance of prior information. In this paper the prior mask $\widetilde{\mathbf{M}}$ is defined as the negative exponential of the local standard deviation of the signal. Our intuition is that, if the signal has small local standard deviation, the chrominance variation within the region is due to the heart-rate and not to head movements or facial expressions, and therefore that matrix entry should be used to estimate the HR.

3.2.1 Solving SAMC

The SAMC optimization problem in (4) is not jointly convex in $\mathbf{E}$ and $\mathbf{M}$. Moreover, even in the case the masking matrix $\mathbf{M}$ was fixed, (4) would contain non-differential and differential terms and a direct optimization would be challenging. Instead, alternating methods have proven to be successful in solving (i) convex problems with non-differential terms and (ii) marginally convex problems that are not jointly convex. More precisely, we derive an optimisation solver based on the alternating direction method of multipliers (ADMM) [5]. In order to derive the associated ADMM method, we first define the augmented Lagrangian problem associated to (4):

$$\min_{\mathbf{E},\mathbf{F},\mathbf{M},\mathbf{Z}} \nu\|\mathbf{E}\|_* + \|\mathbf{M}\circ(\mathbf{F}-\mathbf{C})\|_{\mathcal{F}}^2 + \gamma\,\mathrm{Tr}(\mathbf{F}\mathbf{L}\mathbf{F}^\top) - \beta\|\mathbf{M}\|_1 + \mu\|\mathbf{M}-\widetilde{\mathbf{M}}\|_{\mathcal{F}}^2 + \langle\mathbf{Z},\mathbf{E}-\mathbf{F}\rangle + \frac{\rho}{2}\|\mathbf{E}-\mathbf{F}\|_{\mathcal{F}}^2, \quad (5)$$

where $\mathbf{F}$ is defined to split the terms of (4) that depend on $\mathbf{E}$ into those that are differential and those that are not. The variable $\mathbf{Z}$ represents the Lagrange multipliers constraining $\mathbf{E}$ to be equal to $\mathbf{F}$, further regularized by the term $\|\mathbf{E}-\mathbf{F}\|_{\mathcal{F}}^2$. The ADMM solves the optimisation problem by alternating the direction of the optimisation while keeping the other directions fixed. Specifically, solving (5) requires alternating the following three steps until convergence:

E/M-step With fixed $\mathbf{F}$ and $\mathbf{Z}$ the optimal value of $\mathbf{E}$ is obtained by solving:

$$\min_{\mathbf{E}} \nu\|\mathbf{E}\|_* + \frac{\rho}{2}\|\mathbf{E}-\mathbf{F}+\rho^{-1}\mathbf{Z}\|_{\mathcal{F}}^2. \quad (6)$$

The solution of such problem is given by the shrinkage operator applied to $\mathbf{F}-\rho^{-1}\mathbf{Z}$, see [7]. Formally, if we write the singular value decomposition of $\mathbf{F}-\rho^{-1}\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, the optimal value for $\mathbf{E}$ is:

$$\mathbf{E}^* = \mathbf{U}\mathcal{S}_{\frac{\nu}{\rho}}(\mathbf{D})\mathbf{V}^\top, \quad (7)$$

where $\mathcal{S}_\lambda(x) = \max(0, x-\lambda)$ is the soft-thresholding operator, applied element-wise to $\mathbf{D}$ in (7).

The optimal value for $\mathbf{M}$ is obtained from the following optimisation problem:

$$\min_{\mathbf{M}} \|\mathbf{M}\circ(\mathbf{F}-\mathbf{C})\|_{\mathcal{F}}^2 - \beta\|\mathbf{M}\|_1 + \mu\|\mathbf{M}-\widetilde{\mathbf{M}}\|_{\mathcal{F}}^2, \quad (8)$$

which can be rewritten independently for each entry of $\mathbf{M}$:

$$\min_{m_{rt}\in\{0,1\}} (f_{rt} - o_{rt})^2 m_{rt} + \mu(m_{rt} - \widetilde{m}_{rt})^2 - \beta m_{rt}. \quad (9)$$

The solution is straightforward:

$$m^*_{rt} = \begin{cases} 1 & (f_{rt} - o_{rt})^2 + \mu(1 - 2\widetilde{m}_{rt}) < \beta, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Intuitively, this means that a chrominance feature is selected for learning if (i) the entry of the smoothed low-rank estimation $\mathbf{F}$ is close to the corresponding entry in $\mathbf{C}$ and (ii) that chrominance feature should be selected a priori. Remarkably, this criterion is a mixture of the a posteriori representation power and the a priori knowledge.

F-step With fixed $\mathbf{E}$, $\mathbf{Z}$ and $\mathbf{M}$, the optimal value of $\mathbf{F}$ is obtained by solving the following optimisation problem:

$$\min_{\mathbf{F}} \|\mathbf{M}\circ(\mathbf{F}-\mathbf{C})\|^2_{\mathcal{F}} + \gamma \operatorname{Tr}(\mathbf{F}\mathbf{L}\mathbf{F}^\top) + \frac{\rho}{2}\|\mathbf{F}-\mathbf{E}-\rho^{-1}\mathbf{Z}\|^2_{\mathcal{F}}. \quad (11)$$

Eq. 11 is a particular case of the problem solved in [15, 16]. Importantly, in our case there is no need to solve a linear system of dimension RT as in [15, 16], but we require to solve R linear systems of dimension T as in [1]. From a numerical point of view this is quite advantageous, since larger linear systems tend to be numerically more unstable. More precisely, (11) can be rewritten independently for each of the R rows of $\mathbf{F}$:

$$\min_{\mathbf{f}_r} \|\mathbf{M}_r(\mathbf{f}_r - \mathbf{o}_r)\|^2 + \gamma \mathbf{f}_r \mathbf{L} \mathbf{f}_r^\top + \frac{\rho}{2}\|\mathbf{f}_r - \mathbf{e}_r - \rho^{-1}\mathbf{z}_r\|^2, \quad (12)$$

where lower-case bold letters denote rows of the respective matrices and $\mathbf{M}_r = \operatorname{diag}(\mathbf{m}_r)$. The solution of the previous system is straightforward:

$$\mathbf{f}^*_r = (2\mathbf{M}_r + 2\gamma\mathbf{L} + \rho\mathbf{I}_T)^{-1}(2\mathbf{M}_r\mathbf{o}_r + \rho\mathbf{e}_r + \mathbf{z}_s), \quad (13)$$

where $\mathbf{I}_T$ is the T-dimensional identity matrix.

Z-step The optimal value of $\mathbf{Z}$ is taken from [5]:

$$\mathbf{Z}^* = \mathbf{Z} + \rho(\mathbf{E} - \mathbf{F}), \quad (14)$$

where the right-hand side represent the current values.

3.3. Phase 4: HR Estimation

Once the SAMC solver converges to an optimal solution for $\mathbf{E}$, we can simply hypothesize that, since the main underlying signal is the one associated to the heart rate, the largest singular value of $\mathbf{E}$, would encode the information associated to the sought signal. Therefore, we write the singular value decomposition of $\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, it is reasonable to take the first column of $\mathbf{V}$, $\mathbf{V}_1$ as the estimated underlying HR signal. Finally, the Welch's power spectral density estimation method [26] is employed to obtain the frequency in $\mathbf{V}_1$ with the largest energy f_{HR}. For the instantaneous HR measurement to get f_{HR} we follow [10] and simply detect the highest peak in the Fourier domain of the estimated signal. The HR measured from the input video is then computed as $H = 60 f_{HR}$.

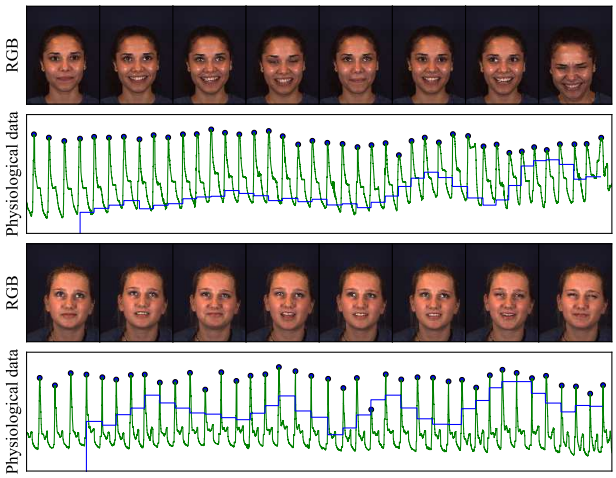

Figure 3. Two examples of video sequences from the MMSE-HR dataset where the subjects experience fear. For each subject two rows are shown. Top: the recorded RGB-video frames. Bottom: physiological data. Note how the heart rate (the blue line) increases when each subject experiences fear.

4. Experimental Evaluation

4.1. Datasets

We conducted experiments on two datasets: the publicly available MAHNOB-HCI dataset [20] and the MMSE-HR dataset. As demonstrated by our experimental results, the latter dataset contains more challenging sequences, due to subjects' movements and facial expressions.

The MAHNOB-HCI dataset is a multimodal dataset with 20 high resolution videos per subject. It contains 27 subjects (12 males and 15 females) in total, and each subject participated in two experiments: (i) emotion elicitation and (ii) implicit tagging. Following [17], in our experiments we used a 30 second interval (frames from 306 through 2135) of 527 sequences. To compute the ground truth heart rate for each video sequence we used the second channel (EXG2) of the corresponding ECG waveforms (see [20]).

The MMSE-HR dataset[2] is a subset of the MMSE database [31] specifically targeted to challenge heart rate estimation algorithms. The MMSE-HR dataset includes 102

[2]The MMSE-HR is included in the full dataset (MMSE) [31] which will be made available to the research community through the Binghamton University

RGB videos and heart-rate data of 40 participants with diverse ethnic/racial ancestries. Two examples are given in Fig. 3 (Note how the HR changes during the recording when each person experiences fear. This supports the value of the dataset for research on instantaneous HR estimation). The physiological data were collected by Biopac Mp150 data acquisition system[3], including heart-rate, mean blood pressure, and other physiological signals, working at 1 kHz. All sensors were synchronized. More details regarding data collection and recording setup can be found in [31].

To compute the ground truth HR signal for both datasets we used a peak detection method from the MNE package[4].

4.2. Settings

To evaluate the performance of the proposed approach and compare it with previous methods, we consider five commonly used metrics in the literature on remote HR analysis [17]. Specifically, we define $H_e(i) = H_p(i) - H_{gt}(i)$, *i.e.* the difference between the predicted heart rate $H_p(i)$ and the ground truth heart rate $H_{gt}(i)$ for the i-th video sequence. We report the mean M_e and the standard deviation SD_e of H_e over all sequences. We also adopt the Root Mean Squared Error ($RMSE$), the mean of error-rate percentage $M_{eRate} = \sum_{i=1}^{N} \frac{|H_e(i)|}{H_{gt}(i)}$ and the Pearson's correlation ρ between signals $H_p = \{H_p(1), ..., H_p(N)\}$ and $H_{gt} = \{H_{gt}(1), ..., H_{gt}(N)\}$, being N is the number of video sequences. In all our experiments the parameters of the proposed method have been selected by cross-validation on a subset of MMSE-HR and set to $\nu = 0.0357$, $\gamma = 0.01$, $\mu = 0.0011$ and $\beta = 0.0005$. Importantly, these parameters were used throughout all our experiments for the two datasets, supporting the generalization ability of SAMC.

4.3. Results

Average HR prediction. In the first series of experiments we compare the proposed approach with several state-of-the art methods for average HR prediction on the MAHNOB-HCI dataset. Specifically we consider the approaches described in [18, 19, 4, 17, 10]. Performance on MAHNOB-HCI is given in Table 1. To perform a quantitative comparison, we have implemented the methods in [17] and [10][5], since their code is not available, while the performance measures for [18, 19, 4] are taken from [17]. It is evident that, while HR estimation on MAHNOB-HCI represents a challenging task for early methods, the more recent approaches, [17] and [10], achieve high accuracy. Moreover, our approach outperforms competing methods by a small margin. This can be explained by the fact that MAHNOB-HCI does not contain many sequences with subject's movements and facial expression changes, while SAMC has been designed to explicitly cope with the spatially localized and intense noise they generate.

Table 1. Average HR prediction: comparison among different methods on MAHNOB-HCI dataset (best performance in bold).

Method	$M_e(SD_e)$	$RMSE$	M_{eRate}	ρ
Poh *et al.* [18]	-8.95 (24.3)	25.9	25.0%	0.08
Poh *et al.* [19]	2.04 (13.5)	13.6	13.2%	0.36
Balakrishnan *et al.* [4]	-14.4(15.2)	21.0	20.7%	0.11
Li *et al.* [17]	-3.30 (6.88)	7.62	6.87%	0.81
De Haan *et al.* [10]	4.62 (6.50)	6.52	6.39%	0.82
SAMC	**3.19 (5.81)**	**6.23**	**5.93%**	**0.83**

Table 2. Average HR prediction: comparison among different methods on MMSE-HR (best performance in bold).

Method	$M_e(SD_e)$	$RMSE$	M_{eRate}	ρ
Li *et al.* [17]	11.56 (20.02)	19.95	14.64%	0.38
De Haan *et al.* [10]	9.41 (14.08)	13.97	12.22%	0.55
SAMC	**7.61 (12.24)**	**11.37**	**10.84%**	**0.71**

Table 3. Self-adapting (SA) vs. non-adapting (NA) MC.

	p	$M_e(SD_e)$	$RMSE$	M_{eRate}	ρ
SA	20	8.13 (12.08)	12.13	10.74	0.68
	40-100	8.22 (12.24)	12.23	10.84	0.67
NA	20	55.39 (36.86)	65.99	68.21	0.08
	40	35.90 (41.29)	51.47	44.76	0.16
	60	22.40 (33.79)	37.06	27.91	0.17
	80	9.41 (14.53)	14.63	11.91	0.49
	100	10.05 (15.23)	15.13	12.98	0.47

To demonstrate the advantages of our method, we perform similar experiments on the more challenging sequences of the MMSE-HR dataset. Here, we only compare our method against the best-performing methods from Table 1. Table 2 reports the results of our evaluation. On this difficult dataset, due to its capacity to select the most reliable chrominance features and ignore the noisy ones, the proposed SAMC achieves significantly higher accuracy than the state-of-the-art.

Effect of self-adaptation. In order to show the benefits of adopting the proposed self-adaptation strategy, we provide results with a fixed binary mask $\mathbf{M}$ (*i.e.* without self-adaptation) and compare them to those obtained with self-adaptation in Table 3. The first column corresponds to the percentile of the values of the prior $\widetilde{\mathbf{M}}$ used to construct the initial mask. More precisely, for a value p, the initial mask is 1 only in the entries corresponding to the $p\%$ regions with the lowest standard deviation. Therefore, $p = 100\%$ corresponds to an (initial) mask matrix of all 1's. Clearly, the

[3] http://www.biopac.com/

[4] http://martinos.org/mne/stable/index.html

[5] We also reimplemented the more recent method based on chrominance features in [25]. Unfortunately, perhaps due to the fact that the method is exhaustively described, we obtained worse results than those we obtained with [10]. Therefore we choose to report our results using [10].

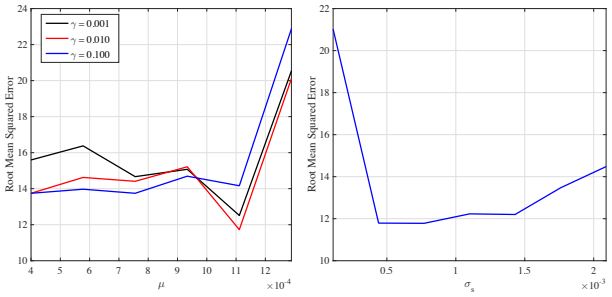

Figure 4. Left: performance at varying values of the γ and μ. Right: RMSE dependency on σ_s.

choice of p is crucial when the matrix is fixed, but almost irrelevant when there is self-adaptation. Also, self-adaptation systematically outperforms the fixed mask case.

Finally, Fig. 4 (left) shows the performance of the proposed approach at different values of parameters μ and γ for the experiments on the MMSE-HR dataset. As shown in the figure, very small and very large values of μ (indicating an increase and a reduction of the influence of the prior mask), correspond to a decrease of performance. Similarly, for the parameter γ, weighting the influence of the Laplacian term, a local optimum can be obtained for $\gamma = 0.01$. Fig. 4 (right) shows similar behavior for σ, used to compute the prior mask as the negative entry-wise exponential of the matrix of standard deviations normalized by σ: $\widetilde{\mathbf{M}} = \mathrm{e}^{(-\mathbf{S}/\sigma)}$.

Short-time HR estimation. To demonstrate the ability of our method to recognize instantaneous HR, we selected 20% of the recorded sequences where there is a very strong heart-rate variation. We split each sequence into non-overlapping windows of length 4, 6, and 8 seconds and process each window independently with [10] and SAMC, since the approach in [17] is not suitable for instantaneous HR prediction. Table 4 shows the results of our short-time window analysis. The table supports the intuition that, the smaller the window, the more difficult is for a method to reliably estimate the HR. Importantly, SAMC consistently outperforms [10] for all window lengths and produces reliable estimates starting from the 4-second windows.

To show that our method is able to follow the changes in subject's HR, we additionally report the predicted heart rate for three sequences of different length. Figure 5 shows the results of three selected video sequences processed by our method. Note that although the method is not able to predict the exact HR for every window, providing the value close to the ground truth, a sudden increase/decrease is well localized in time.

Running time. The proposed approach is fast, enabling real-time HR analysis. On average, phase 1 runs at 50 fps, while phase 2 runs at around 30 fps. Phase 3 and 4 have the smallest execution time, reaching 550 fps. Running times were measured using a single core implementation on a conventional laptop with an Intel Core i7-4702HQ processor.

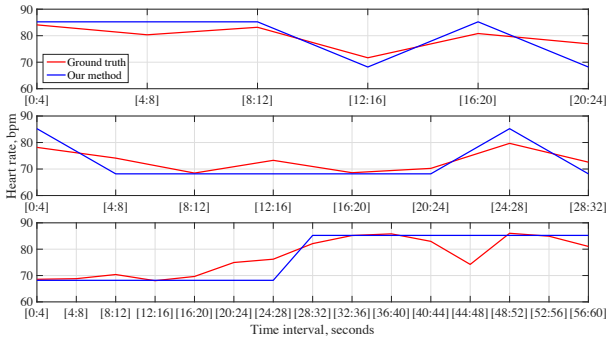

Figure 5. Heart rate recognition results for three sequences, using window size of 4 seconds. Y-axis shows the interval over which the heart rate was computed.

Table 4. Short-time window analysis. Results for three windows sizes are reported: 4, 6, and 8 seconds.

	Method	$M_e(SD_e)$	$RMSE$	M_{eRate}	ρ
4 s	De Haan *et al.* [10]	**-1.85** (15.77)	15.83	9.92%	0.67
	SAMC	2.12 (**11.51**)	**11.66**	**9.15%**	**0.78**
6 s	De Haan *et al.* [10]	-2.21 (19.21)	19.27	11.81%	0.33
	SAMC	**0.32** (**8.29**)	**8.27**	**7.30%**	**0.80**
8 s	De Haan *et al.* [10]	**0.81** (11.49)	11.46	8.60%	0.63
	SAMC	1.62 (**9.67**)	**9.76**	**7.52%**	**0.71**

5. Conclusions

We presented a novel framework for remote HR estimation from visual data. At the core of our approach, there is a novel optimization framework, named self-adaptive matrix completion, which outputs the HR measurement while simultaneously selecting the most reliable face regions for robust HR estimation. This strategy permits to discard noisy features, due to spontaneous target's movements and facial expressions. As demonstrated by our experimental evaluation, the proposed approach provides accurate HR estimates and outperforms state-of-the-art methods not only in the case of long-time windows, but also for short-time analysis. Extensive experiments conducted on the MMSE-HR dataset support the value of the adopted self-adaption strategy for HR estimation. Future work guidelines include devising novel feature representations, in alternative to chrominance signals, to further improve the robustness to varying illumination conditions as well as exploiting the feasibility of combining the predicted HR measurements with visual features for spontaneous emotion classification.

6. Acknowledgments

The material (dataset part) is based upon the work supported in part by the NSF under grants CNS-1205664 and CNS-1205195.

References

[1] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe. Analyzing free-standing conversational groups: A multimodal approach. In *ACM Multimedia*, 2015. 3, 4, 5, 6

[2] X. Alameda-Pineda, Y. Yan, E. Ricci, and N. Sebe. Recognizing emotions from abstract paintings using non-linear matrix completion. In *CVPR*, 2016. 3

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013. 1

[4] G. Balakrishnan, F. Durand, and J. Guttag. Detecting pulse from head motions in video. In *CVPR*, 2013. 1, 3, 7

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. 5, 6

[6] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE TPAMI*, 37(1):121–135, 2015. 3, 4, 5

[7] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. 5

[8] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009. 5

[9] C.-H. Chen, V. M. Patel, and R. Chellappa. Matrix completion for resolving label ambiguity. In *CVPR*, 2015. 3, 4

[10] G. De Haan and V. Jeanne. Robust pulse rate from chrominance-based rPPG. *IEEE Transaction on Biomedical Engineering*, 60(10):2878–2886, 2013. 1, 2, 3, 4, 6, 7, 8

[11] A. Goldberg, B. Recht, J. Xu, R. Nowak, and X. Zhu. Transduction with matrix completion: Three birds with one stone. In *NIPS*, 2010. 2, 3

[12] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015. 3

[13] A. Jeni, J. F. Cohn, and T. Kanade. Dense 3D Face Alignment from 2D Videos in Real-Time. In *FG*, 2015. 1

[14] A. Jourabloo and X. Liu. Pose-Invariant 3D Face Alignment. In *ICCV*, 2015. 1

[15] V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst. Matrix completion on graphs. In *NIPS Workshops*, 2014. 3, 5, 6

[16] A. Kovnatsky, M. M. Bronstein, X. Bresson, and P. Vandergheynst. Functional correspondence by matrix completion. *CVPR*, 2015. 3, 4, 5, 6

[17] X. Li, J. Chen, G. Zhao, and M. Pietikainen. Remote Heart Rate Measurement From Face Videos Under Realistic Situations. In *CVPR*, 2014. 1, 2, 3, 6, 7, 8

[18] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics express*, 18(10):10762–10774, 2010. 1, 2, 7

[19] M. Z. Poh, D. J. McDuff, and R. W. Picard. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering*, 58, 2011. 1, 7

[20] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transaction on Affective Computing*, 3, 2012. 2, 3, 6

[21] S. Tulyakov and N. Sebe. Regressing a 3D face shape from a single image. In *ICCV*, 2015. 1

[22] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Scientific reports*, 4, 2014. 2

[23] W. Verkruysse, L. O. Svaasand, and J. S. Nelson. Remote plethysmographic imaging using ambient light. *Optics Express*, 16(26):21434, 2008. 1, 2

[24] W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, and N. Sebe. Recurrent face aging. In *CVPR*, 2016. 3

[25] W. Wang, S. Stuijk, and G. D. Haan. Exploiting Spatial Redundancy of Image Sensor for Motion Robust rPPG. *IEEE Transactions on Biomedical Engineering*, 62(2):415–425, 2015. 1, 3, 7

[26] P. D. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967. 4, 6

[27] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman. Eulerian video magnification for revealing subtle changes in the world. In *ACM Transactions on Graphics*, 2012. 1, 2

[28] L. Wu, R. Jin, and A. K. Jain. Tag completion for image retrieval. *IEEE TPAMI*, 35(3):716–727, 2013. 3

[29] X. Xiong and F. De La Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 1

[30] S. Xu, L. Sun, and G. K. Rohde. Robust efficient estimation of heart rate pulse from video. *Biomedical optics express*, 5:1124–35, 2014. 1

[31] Z. Zhang, J. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, and L. Yin. Multimodal spontaneous emotion corpus for human behavior analysis. In *CVPR*, 2016. 2, 6, 7