# Exploring the Feasibility of Face video based Instantaneous Heart-rate for Micro-expression Spotting

Puneet Gupta, Brojeshwar Bhowmick, Arpan Pal
Embedded system and Robotics, TCS Research and Innovation, Kolkata-700106, India,
Email: {gupta.puneet5, b.bhowmick, arpan.pal}@tcs.com

## Abstract

*Facial micro-expressions (ME) are manifested by human reflexive behavior and thus they are useful to disclose the genuine human emotions. Their analysis plays a pivotal role in many real-world applications encompassing affective computing, biometrics and psychotherapy. The first and foremost step for ME analysis is ME spotting which refers to detection of ME affected frames from a video. ME spotting is a highly challenging research problem and even human experts cannot correctly perform it because MEs are manifested using subtle face deformations and that too for a short duration. It is well established that changes in the human emotions, not only manifest ME but it also introduces changes in instantaneous heart rate. Thus, the manifestation of ME and changes in the instantaneous heart rate are related to the change in human expressions and both of them are estimated using temporal deformations of the face. This provides the motivation of this paper that aims to explore the feasibility of variations in the instantaneous heart rate for performing the correct the ME spotting. Experimental results conducted on a publicly available spontaneous ME spotting dataset, reveal that the variations in instantaneous heart rate can be utilized to improve the ME spotting.*

## 1. Introduction

Perceiving the facial expression plays a pivotal role in affective computing, human psychology understanding and human communication [5]. Due to these applicabilities, correct understanding of human expression is an active research area. Human expression can be classified as: i) facial macro-expressions which are usually observed in day-to-day life and easily elucidated by humans; and ii) facial micro-expressions (ME) which are originated by human reflexive behavior. Since ME arises from the human reflexive behavior which is hard to control or suppress, they are hard to suppress and disclose the genuine human expressions [10]. Due to these reasons, ME is useful in numerous applications, some of the crucial amongst them are: i) business negotiation and security applications where lie detection is required to circumvent frauds [5]; ii) psychotherapy for monitoring suspicious intent or surveillance [10]; iii) applications in the realm of affective computing that require the understanding of human genuine emotions like commercial advertisement rating [16]; and iv) virtual/ augmented reality for face synthesis. Requirement of ME in such applications, has ignited the research in ME analysis.

ME analysis consists of two stages, viz., ME spotting which refers to the localization of ME frames in the face video and ME recognition where the spotted frames are analyzed to classify the expression. ME spotting is a highly challenging problem and even human experts who are trained for expression analysis cannot correctly perform the ME spotting. The reason behind this incapability is that ME are generated for short duration (usually 1/25 to 1/5 of a second) by subtle stretching or contraction of facial arteries located in the face areas belonging to like lips, eyes and mouth. Human eyes are unable to process such short duration subtle facial temporal movements [15]. Utilizing a human expert for ME spotting is not only error-prone, but highly expensive, as well. All these factors provide the motivation to propose an automatic ME spotting system in this paper, that can automatically and correctly localize the ME frames in the face video.

Prominent facial deformations are generated by the unavoidable pose variations and eye blinks. Unfortunately, subtle facial deformations generated by ME and these inevitable prominent facial deformations usually co-exist. Facial deformation due to ME can also produce with the deformations due to macro-expression such that ME and macro-expression belong to either similar or opposite attributes [22]. The facial deformations due to pose variations, eye blinks and macro-expression can be easily misinterpreted as deformations due to ME which eventually results in erroneous ME spotting. In the literature, ME spotting is performed by analyzing the temporal deformations that are produced by either variations in facial appearance or movements of discriminative facial points. Both these deformations are inadequate for correct ME spotting. The problem with such approaches are: i) the facial appearance is easily influenced by illumination, eye-blinking and macro-expressions; and ii) movements of the facial points are error-prone due to inaccurate localization.

Human emotions, not only results in the manifestation of ME but they also introduces changes in the instantaneous heart rate (HR) [4]. Various studies have been conducted to understand the correlation between variations in HR and human emotions, but it still requires rigorous analysis [14]. But it is well established that fluctuations in HR are observed when human emotion changes and this phenomenon is used in polygraph based lie detection [3]. The paper aims to explore the feasibility of variations in instantaneous HR for improving the ME spotting.

HR estimation can be performed using either contact mechanisms that require user to place sensors (like electrocardiography (ECG) or photo-plethysmography (PPG)) on their body and non-contact mechanisms that do not require any user contact. The contact mechanisms require the fixed contact between the skin and sensor to be fixed [38]. It can introduce psychological bias which degrades the HR estimation [38]. In addition, long term HR monitoring is not feasible using contact mechanisms. In contrast, well known non-contact mechanisms based on Microwave Doppler and laser [11] need highly expensive and bulky sensors for HR estimation. Fortunately, another non-contact mechanism is proliferating that require non-contact face videos for HR estimation [7]. Such an acquisition can be performed using cheap and portable camera. In addition, these same videos are useful for ME analysis.

All these factors motivate us to employ non-contact face videos for estimating the variations in instantaneous HR and used it for improving the ME spotting. This is the main contribution of the paper. To the best of our knowledge, this is the first attempt towards exploring the face video based HR estimation for ME spotting.

The paper is organized in the following manner. The preliminaries required to properly understand the proposed system are discussed in the next section. It provides the brief overview of ME spotting and HR estimation using face videos. The next section describes the proposed ME spotting system using HR monitoring. The experimental results are analyzed in Section 4. Conclusions are given in the last section.

## 2. Preliminaries

This section provides the brief overview of existing ME spotting and HR estimation using face videos.

### 2.1. ME Spotting

Most of the existing work in the realm of ME analysis is mainly concentrated around ME recognition rather than ME spotting because of the unavailability of spontaneous ME spotting datasets. Recently few ME spotting datasets are made publicly available which ignites the research in ME spotting. Some such datasets are *SMIC-E-HS* [17] and *CASMEII* [37]. A ME spotting system usually consists of the following steps: i) face alignment where face present in each video frame is detected and aligned to a common reference so as to handle the geometric deformations; ii) feature encoding where subtle temporal deformations in the video frames are represented; and iii) spotting the ME frames by analyzing the temporal deformations.

The efficacy of ME spotting is highly dependent on the feature encoding. The ME spotting cannot be correct unless feature encoding does not correctly represent those subtle facial temporal deformations that are generated by ME and simultaneously mitigate the spurious temporal deformations generated by expressions and illumination [20]. One of the most extensively studied feature encoding for ME spotting is the appearance based feature encodings which are given by the variations introduced in the facial appearance or texture in the video frames [22, 34]. The facial appearance features in a video frame can be determined by applying main directional maximal difference [34] or histograms of Local Binary Patterns (LBP) [22]. In [19], optical flow is utilized to compute the local facial appearance features. The corresponding appearance based feature encoding can then be obtained by subtracting the facial appearance features in the subsequent frames.

The most valuable information related to ME usually resides in the following facial areas: lips, eyes and eye-brows. It is because expressions are manifested by the contraction or stretching of the facial arteries present in these facial areas. But most of the above mentioned appearance based feature encoding, consider full face region to extract the facial appearance rather than the facial regions. Hence, it can be inferred that appearance based feature encoding contain large redundant information. In addition, these facial appearance features are easily influenced by illumination, eye-blinking and macro-expressions which result in erroneous ME spotting. Hence, the utilization of only some discriminatory facial features known as facial landmarks is studied in [24] for feature encoding. The facial landmarks are chosen in such a manner that they can completely model the facial dynamics. The main prerequisite for the utilization of facial landmark in ME spotting is their correct localization and unfortunately, it is usually violated due to the following reasons: i) illumination significantly changes the appearance of facial regions; and ii) some landmark points are less discriminatory in the neighborhood like the landmarks lying on lip regions. Another reason for erroneous ME spotting is the inevitable eye-blinking that generates prominent temporal deformations near eye-areas which in-turn leads to spurious eye-brow appearance and landmark localization. It is observed in [22] that it is better to avoid the eye region during appearance based feature encoding to achieve better ME spotting.

The feature encoding is analyzed to detect the frames affected by ME. Generally, video frames can be classified [30] based on the facial expression as: (i) onset frames where expression comes into existence and thus the temporal deformations are escalating; (ii) apex frames where the expression is at its peak and thus the temporal deformation is the most prominent; (iii) offset frames where the expression diminishes and thus temporal deformations starts to fade away; and (iv) neutral frames where no expression is noticeable and thus temporal deformations are negligible. The apex frames constituting the frames affected by ME, are the frames corresponding to the peaks in the feature encoding because they contain large temporal deformations as compared to their neighborhoods [18]. It is quite possible that spurious peaks are generated in the feature encoding due to macro-expressions, eye-blinking and background noise. In such a case, the apex frame will be wrongly classified which in-turn results in wrong ME spotting.

## 2.2. Face HR Estimation

Cardiovascular pulse propagate in the human body due to the contraction and expansion of the heart. This mechanism introduces the color variations which can be observed in the reflected light using a camera [25]. In addition, the cardiovascular pulse also introduces subtle movements in the face, which can be examined using camera [1]. Both these mechanisms are used for HR estimation using face video. Such HR estimation systems consist of three stages, viz. preprocessing, temporal signal extraction and HR estimation.

### 2.2.1 Preprocessing

In the first stage, face is detected from the input video by using the existing face detectors like Viola-Jones [31] or model based face detectors [2]. The detected face region can contain non-skin pixels belonging to the background or human hairs. Non-skin pixels do not contain enough information and hence they are eliminated using skin color discrimination. Unavoidable eye blinking can result in erroneous HR estimation hence the eye areas can also be eliminated to improve the HR estimation. The eye areas can be estimated by utilizing facial geometry heuristics or trained classifiers [33]. In some cases, facial boundary is also removed to improve the performance [7]. Region of interest (ROI) is defined using the remaining facial area. Full face, forehead region or cheek areas can be used as ROI [12]. ROI can also be adaptively defined by selecting few important rectangular face blocks [13].

### 2.2.2 Temporal Signal Extraction

Subtle motion or color variations introduced in the face video due to heart beats can be observed using Eulerian [25] or Lagrangian methodologies [1]. The variations introduced in subsequent frames are referred as temporal signals. In Lagrangian methodology, discriminating features are explicitly tracked and their variations in the subsequent frames provide the temporal signal. The efficacy of these temporal signals is dependent on correct localization, but unfortunately these are improperly localized due to illumination. In addition, the temporal signal are not useful when few discriminatory features are available. Another factor that restricts its utility is that extraction of such temporal signal is highly time-consuming. Hence, Eulerian methodology is introduced for temporal signal extraction [25]. The temporal signal in this methodology is given by the variations produced in the fixed region of interest (ROI). Such an implicit tracking is less time-consuming than the Lagrangian methodology. They are relevant only when subtle variations are present, as in case of HR estimation [36].

### 2.2.3 HR Estimation

Pulse signal is extracted from the temporal signal using blind source separation techniques [7]. The pulse signal is transformed into the frequency domain by applying Fast Fourier Transform (FFT) and the frequency containing the maximum amplitude in the spectrum corresponds to HR frequency [23]. It is possible that several spurious peaks are produced in the spectrum when noise is present in the pulse signal. In such a scenario, maximum amplitude peak might not corresponds to actual HR frequency, i.e., HR estimation is erroneous. The impact of such noise is mitigated by applying filtering techniques like Detrending filter is used to handle the non-stationary trend of a signal [28].

# 3. Proposed System

In this section, the ME spotting system is proposed which explore variations in the instantaneous HR so as to improve the spotting. It consists of the following five stages: i) Face alignment; ii) ROI extraction; iii) Evaluating instantaneous HR variations; iv) Determining plausible ME spots; and v) Post-processing of spots. In the first stage, face present in the face video frames is detected and rigid deformations present in them are removed by aligning them to a common reference. In the second stage, non-skin pixels in the facial area are removed and the remaining area is used to define the several ROIs. Subsequently, the ROIs are utilized to evaluate the instantaneous HR at multiple locations and they are followed by calculating the variations in instantaneous HR. In the fourth stage, the temporal deformation in the ROIs is encoded and the peaks corresponding

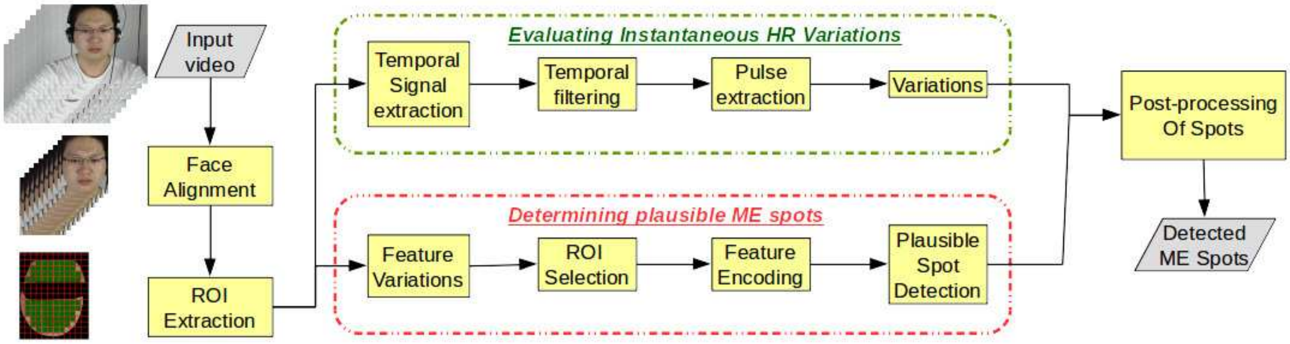

Figure 1. Flow-graph of the proposed system

to the encoding is used to locate the plausible ME spots. Eventually, variations in instantaneous HR are utilized to select the genuine ME spots from the plausible ME spots. The flow-graph of the proposed system is shown in Figure 1.

## 3.1. Face Alignment

The input video contains human face along with large background. But instantaneous HR estimation and ME spotting require only the facial region, thus the facial region is detected and the remaining area (i.e., background) is removed. For this purpose, accurate localization of discriminating facial landmarks is required. We employed Constrained Local Neural Field model (CLNF) [2] for the facial landmark detection. These landmarks represent the contour of eyes, lips, nose and facial boundary. CLNF first applies Viola-Jones face detector [32] to detect plausible faces areas. Subsequently, the actual face is determined from the plausible face areas and facial landmarks are localized using the global and local facial models. The face area is determined using the facial landmarks and the background (or remaining area) is removed. The detected face contains rigid deformations like translation and rotations which degrade the performance of HR estimation and ME spotting. Thus, the detected face region is aligned to a common reference. We normalize the detected faces such that their eye distance is fixed, as proposed in [21].

## 3.2. ROI Extraction

The efficacies of HR estimation and ME spotting are restricted by the unavoidable eye-blinking. We handle this issue by removing the eye areas for the processing. The eye areas are given by the convex hull of only those facial landmarks that correspond to the eyes. In addition, it is observed in [7] that even slight movement in face boundaries introduces large color variations and thereby significantly degraded the HR estimation. Thus, we apply morphological erosion to remove the facial boundaries [9]. Moreover, it is observed in [8] that if full face is considered as one ROI, then expression variations localized in small face area can tremendously deteriorate the HR estimation. Hence, we extract several non-overlapping square blocks from the resultant image and mark the blocks containing only skin pixels as ROIs.

## 3.3. Evaluating Instantaneous HR Variations

In this stage, instantaneous HR and their variation are evaluated. It consists of the following steps: i) Temporal signal extraction; ii) Temporal filtering; iii) Pulse extraction; and iv) Variations.

### 3.3.1 Temporal Signal Extraction

Each ROI consisting of skin pixels, contains the cardiovascular pulse information in terms of color variations. It is noticed in [29] that green channel intensities provide the most useful information regarding cardiovascular pulse as compared to red and blue channel intensities. Reason for this phenomenon are: i) green light better penetrates inside the skin as compared to blue light; and ii) green light provides better hemoglobin absorption than red light [29]. Thus, temporal signal of an ROI is provided by its mean green value of pixels [7]. That is, the temporal signal of $i^{th}$ ROI, $T^i$ is:

$$T^i = \left[ \sum_{(x,y)\in R^i} I_g^1\left(x,y\right), \cdots, \sum_{(x,y)\in R^i} I_g^n\left(x,y\right) \right] \quad (1)$$

where $n$ is the total number of frames; $(x, y)$ represents the pixel location; $R^i$ is the $i^{th}$ ROI; and $I_g^k$ contains the green channel intensities of $k^{th}$ frame.

### 3.3.2 Temporal Filtering

Band-pass filter is employed to minimize the noise in temporal signals. It is observed in [8] that human heart beat within the range of 42 to 240 beat-per-minute (bpm). Thus, the frequency range for the band-pass filter is set from

0.7 to 4 Hz. Moreover, the temporal signals contain non-stationary trend due to focus or illumination changes. Thus, Detrending filter is also applied to remove the noise from temporal signals [28].

### 3.3.3 Pulse Extraction

Each temporal signal contains pulse signal along with the noise. In case of multiple temporal signals, the pulse signal is extracted using blind source separation by estimating the individual source components [25]. Amplitudes of pulse signal and noise in the temporal signals depend on the facial structure, user characteristics (like skin color) and environmental settings (like illumination). Hence, z-score normalization [27] is applied to normalize the temporal signals. The pulse signal is extracted from these temporal signals by applying the kurtosis based optimization, as utilized in [8] because it quickly provides the global convergence and it is proved in [7] that such a pulse estimation provides correct pulse extraction.

### 3.3.4 Variations

The pulse signal is divided into overlapping windows and HR is estimated from each window. Such HR estimates are referred as instantaneous HR. The size of each overlapping window is 60 frames and there is an overlap of 30 frames in the subsequent windows. The variations in instantaneous HR at $m^{th}$ frame, $v_m$ is given by:

$$v_m = \sum_{a=-p}^{p} \left| h_i - h_{(i+a)} \right| \tag{2}$$

where $p$ is the number of neighbors; $|\bullet|$ denotes the absolute operation; $i$ is the fragment containing $m^{th}$ frame; and $h_l$ denotes the instantaneous HR in $l^{th}$ fragment. The value of $p$ is set to 2.

## 3.4. Determining plausible ME spots

In this section, the extracted ROIs are used to determine the plausible ME spots. Initially, the feature variations in each ROI are evaluated. Based on the feature variations, only some ROIs are selected that might contain the temporal deformations due to ME. Appearance based features are encoded using these selected ROI and the encoding is used to extract the plausible ME spots.

### 3.4.1 Feature Variations

Variations in the features are required for ME spotting. In this section, such variations are evaluated using changes in facial appearance. We calculate the feature variations in each ROI by utilizing Eulerian methodology described in [8]. That is, the intensity variations in a particular ROI provide the variations in appearance feature. More clearly, feature variation corresponding to $i^{th}$ ROI in $\alpha^{th}$ frame, $F_i(\alpha)$ is given by:

$$F_i(\alpha) = \sum_{(a,b)\in R^i} \left( G^{\alpha}_{(a,b)} - \frac{\left( G^{(\alpha+k)}_{(a,b)} + G^{(\alpha-k)}_{(a,b)} \right)}{2} \right)^2 \tag{3}$$

where $R^i$ denotes the $i^{th}$ ROI and $G^w_{(a,b)}$ is the gray-scale intensity at pixel location $(a, b)$ of the $w^{th}$ frame. It can be observed that in the equation, the feature difference is evaluated by subtracting the features of sequential frames within a specified interval (defined by $K$) rather than subtracting the features of alternating frames. It provides a better features representation than considering alternating frame difference [18].

### 3.4.2 ROI Selection

Usually, small facial regions are affected by the ME and they are sufficient for correct ME spotting [18]. Hence we choose only few ROI that contain significant temporal deformations. The total variation in an ROI, $d\,(i)$ is given by adding its features variation, i.e.,

$$d\,(i) = \sum_{i=q+1}^{n-q} F_i\,(\alpha) \tag{4}$$

where $F_i$ is the feature variation for $i^{th}$ block; $n$ is the number of video frames; and $q$ is the interval used for frame feature difference. We selected 40% of the ROI containing the largest total variations.

### 3.4.3 Feature Encoding

The appearance based feature encoding is obtained by adding the feature variations of the selected ROI, i.e.,

$$A\,(\alpha) = \sum_{i\in B} F_i\,(\alpha) \tag{5}$$

where $A$ is the appearance based feature encoding; $F_i$ is feature variation for $i^{th}$ block; $\alpha$ is the frame number; and $B$ stores the index of the selected ROI.

### 3.4.4 Plausible Spot Detection

ME are given by the peaks in the feature encoding. It is observed that spurious peaks can be generated by the noise. Such spurious peaks usually contain low amplitude as compared to the genuine ME peaks. Hence a threshold, $T$ is defined to remove some spurious peaks [18]. It is set according to the algorithm proposed in [18], i.e.,

$$T = A_{mean} + \tau \times \left( A_{max} - A_{mean} \right) \tag{6}$$

where $A_{mean}$ and $A_{max}$ indicate the mean and maximum value in $A$ respectively; and $\tau$ is a predefined parameter. All the peaks whose magnitude are greater than $T$ are marked as plausible ME peaks.

### 3.5. Post-processing of spots

Amongst all the detected plausible peaks, only some peaks corresponds to the genuine ME spot and the remaining peaks are falsely generated due to noise. We employ the variations in instantaneous HR to classify the plausible peaks into genuine and false peaks. It is based on the intuition that HR fluctuates when expression changes, thus variations in instantaneous HR should be higher at the frames affected by ME. Hence, we classify the peaks using:

$$Classify\,(s) = \begin{cases} Genuine, & \text{if } v_s > t_f \\ Spurious, & \text{otherwise} \end{cases} \qquad (7)$$

where $s$ is a plausible location of ME; $v_s$ is the variations in instantaneous HR at $s^{th}$ location frame; and $t_f$ is a predefined threshold which is set to 10 bpm.

## 4. Experimental Results

### 4.1. Dataset Description

The efficacy of the proposed ME spotting using variations in instantaneous HR is tested using $CAS(ME)^2$ dataset [26]. It is important to note that HR evaluation requires large number of video frames. Since other ME spotting datasets (like SMIC datasets) provide small duration video frames that are unsuitable for HR estimation, they are avoided in the performance evaluation. The dataset contains 57 micro-expressions acquired from 22 subjects. Logitech Pro C920 camera is utilized to acquire its face videos at 30 fps and its resolution is set to 640 × 480 pixels. Moreover, uniform illumination is maintained during the acquisition. The dataset is provided with the onsets and offsets of ME. These are used as the ground-truth.

### 4.2. Performance Metrics

The performance is evaluated by verifying the genuine peak denoting the apex of ME with ground truth. If the genuine peak lies within the range of [onset-20, offset+20], then the spotted ME is considered as true positive; otherwise false positive. Behavior of true positive rate (TPR) and false positive rate (FPR) is analyzed for the performance evaluation. TPR is given by the percentage of true positives, divided by the total number of ME; while FPR is given by the percentage of false positives, divided by the total number of false peaks. Performance of ME spotting is evaluated using receiver operating characteristic (ROC) curves whose x and y axis denote FPR and TPR respectively. The thresholds required for plotting the ROC are obtained by varying the predefined parameter $\tau$ in Equation (6) from 0 to 1.

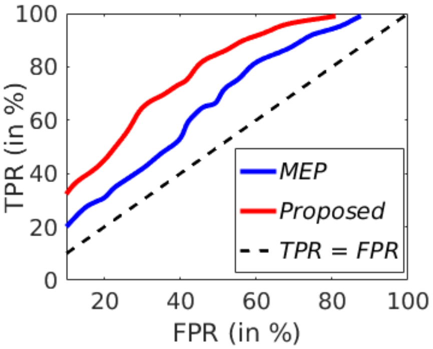

Figure 2. ROC curves of ME spotting systems.

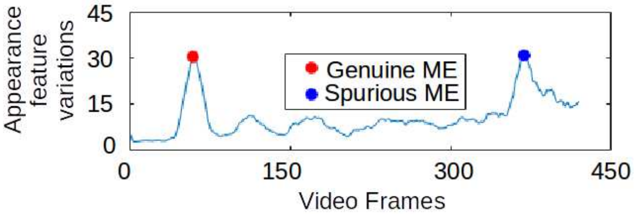

Figure 3. An example where variations in instantaneous HR has successfully classified plausible ME spots as either genuine or spurious ME spots. The variations in instantaneous HR at the depicted genuine and spurious ME spots are 16 bpm and 3 bpm respectively.

### 4.3. Analysis

For the performance analysis, the proposed system is compared with system $MEP$. $MEP$ is obtained by utilizing only the proposed ME spotting while excluding the utilization of instantaneous HR. Hence it provides only the plausible ME spots. ROC curves of the proposed system and $MEP$ are depicted in Figure 2. It can be seen that a large number of genuine MEs is spotted at low thresholds, but it also provides a large number of spurious ME spots. As the threshold value, $T$ increases both genuine and spurious ME spots start decreasing. In addition, it can be observed from Figure 2 that the proposed system performs better that $MEP$ which indicate that variations in instantaneous HR can successfully classify some plausible ME spots as either genuine or spurious. One such example where the variations in instantaneous HR has successfully classify the ME spots is shown in Figure 3.

### 4.4. Failure Cases and Future Work

After rigorous experimentation, we observe that most of the false ME spots are reported due to eye-blinking even though we have excluded the eye areas from the ROI. It indicates that eye-blinking induces subtle temporal deformations in other facial regions. Another reason behind the poor performance is that the instantaneous HR fails to detect false ME spots due to the following reasons:

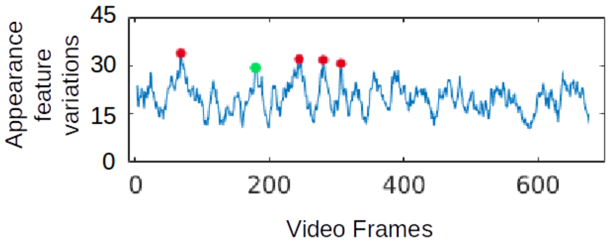

Figure 4. Failure case of variations in instantaneous HR for classifying the plausible ME spots. It shows true positive in green color while false positives in red color.

1. The instantaneous HR is spurious because the face video contains large facial movements due to noise.
2. The detected plausible ME peaks are close to each other and only some of them are genuine peaks. In such a case, all the plausible peaks have large variations of instantaneous HR. One such example is shown in Figure 4.
3. The threshold $t_f$ on variation of instantaneous HR is fixed. It is better to adaptively select this threshold because the variation in human HR depends on human psychological and health parameters.

In the future, we will try to improve the ME spotting using instantaneous HR by: i) providing better face video based HR estimation; ii) providing better feature encoding for ME spotting; and iii) adaptively selecting the threshold on variation of instantaneous HR. Better face video based HR estimation can be performed by utilizing all the color channels instead of just employing the green channel [35]. Similarly, ME analysis can be improved by considering the inter-beat intervals instead of instantaneous heart-rate because it is more closely related to human emotions than instantaneous HR [6]. In addition, ME analysis can be improved by analyzing the impact of different ME on variations in instantaneous HR because positive and negative emotions impact the HR in a different manner [6].

## 5. Conclusions

An automatic ME spotting system based on variations in instantaneous HR has been proposed in this paper. ME spotting is required in many real-world applications encompassing affective computing, biometrics and psychotherapy. But it is a highly challenging problem because MEs are manifested using subtle face deformations and that too for a short duration. Moreover, prominent inevitable facial deformations generated by pose variations, macro-expression and eye blinking can be easily misinterpreted as ME. We have explored the feasibility of variations in the instantaneous HR for correctly verifying the ME spots. The motivation behind the exploration is this that both manifestation of ME and changes in instantaneous HR are related to the change in human emotions.

Experimental results conducted on a publicly available spontaneous ME spotting dataset, have demonstrated that the variations in instantaneous heart rate can be utilized to verify the ME spots and eventually improve the spotting. We will improve the proposed system in the future by providing better face video based HR estimation; incorporating better feature encoding for ME spotting; and adaptively selecting the threshold on variation of instantaneous HR.

## ACKNOWLEDGMENT

Authors are thankful to the researchers, who have provided us the access of $CAS(ME)^2$ dataset.

## References

[1] G. Balakrishnan, F. Durand, and J. Guttag. Detecting pulse from head motions in video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3430–3437, 2013.

[2] T. Baltrusaitis, P. Robinson, and L.-P. Morency. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[3] N. R. Council et al. *The polygraph and lie detection*. National Academies Press, 2003.

[4] H. D. Critchley, P. Rotshtein, Y. Nagai, J. O'doherty, C. J. Mathias, and R. J. Dolan. Activity in the human brain predicting differential heart rate responses to emotional facial expressions. *Neuroimage*, 24(3):751–762, 2005.

[5] P. Ekman. Lie catching and microexpressions. *The philosophy of deception*, pages 118–133, 2009.

[6] J. J. Gross and R. W. Levenson. Hiding feelings: The acute effects of inhibiting negative and positive emotion. *Journal of abnormal psychology*, 106(1):95, 1997.

[7] P. Gupta, B. Bhowmick, and A. Pal. Accurate heart-rate estimation from face videos using quality-based fusion. In *IEEE International Conference on Image Processing, (ICIP)*, pages 4132–4136, 2017.

[8] P. Gupta, B. Bhowmick, and A. Pal. Serial fusion of eulerian and lagrangian approaches for accurate heart-rate estimation using face videos. In *IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2834–2837. IEEE, 2017.

[9] P. Gupta and P. Gupta. An efficient slap fingerprint segmentation and hand classification algorithm. *Neurocomputing*, 142:464–477, 2014.

[10] E. A. Haggard and K. S. Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of research in psychotherapy*, pages 154–165. Springer, 1966.

[11] M.-C. Huang, J. J. Liu, W. Xu, C. Gu, C. Li, and M. Sarrafzadeh. A self-calibrating radar sensor system for measuring vital signs. *IEEE transactions on biomedical circuits and systems*, 10(2):352–363, 2016.

[12] S. Kwon, J. Kim, D. Lee, and K. Park. Roi analysis for remote photoplethysmography on facial video. In *IEEE International Conference of the Engineering in Medicine and Biology Society (EMBC)*, pages 4938–4941. IEEE, 2015.

[13] A. Lam and Y. Kuno. Robust heart rate measurement from video using select random patches. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3640–3648, 2015.

[14] R. D. Lane, K. McRae, E. M. Reiman, K. Chen, G. L. Ahern, and J. F. Thayer. Neural correlates of heart rate variability during emotion. *Neuroimage*, 44(1):213–222, 2009.

[15] A. C. Le Ngo, Y.-H. Oh, R. C.-W. Phan, and J. See. Eulerian emotion magnification for subtle expression recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1243–1247. IEEE, 2016.

[16] P. Lewinski, M. L. Fransen, and E. S. Tan. Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. *Journal of Neuroscience, Psychology, and Economics*, 7(1):1, 2014.

[17] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[18] X. Li, H. Xiaopeng, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*, 2017.

[19] S.-T. Liong, J. See, R. C.-W. Phan, A. C. Le Ngo, Y.-H. Oh, and K. Wong. Subtle expression recognition using optical strain weighted features. In *Asian Conference on Computer Vision*, pages 644–657. Springer, 2014.

[20] S.-T. Liong, J. See, R. C.-W. Phan, Y.-H. Oh, A. C. Le Ngo, K. Wong, and S.-W. Tan. Spontaneous subtle expression detection and recognition based on facial strain. *Signal Processing: Image Communication*, 47:170–182, 2016.

[21] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[22] A. Moilanen, G. Zhao, and M. Pietikäinen. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *International Conference on Pattern Recognition (ICPR)*, pages 1722–1727. IEEE, 2014.

[23] T. F. of the European Society of Cardiology et al. Heart rate variability standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17:354–381, 1996.

[24] D. Patel, G. Zhao, and M. Pietikäinen. Spatiotemporal integration of optical flow vectors for micro-expression detection. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 369–380. Springer, 2015.

[25] 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(1):7–11, 2011.

[26] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu. Cas (me)ˆ 2: a database for spontaneous macro-expression and micro-expression spotting and recognition. *IEEE Transactions on Affective Computing*, 2017.

[27] A. A. Ross, K. Nandakumar, and A. Jain. *Handbook of multibiometrics*, volume 6. Springer Science & Business Media, 2006.

[28] M. P. Tarvainen, P. O. Ranta-Aho, P. A. Karjalainen, et al. An advanced detrending method with application to hrv analysis. *IEEE Transactions on Biomedical Engineering*, 49(2):172–175, 2002.

[29] H. E. Tasli, A. Gudi, and M. den Uyl. Remote ppg based vital sign measurement using adaptive facial regions. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 1410–1414. IEEE, 2014.

[30] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 149–149. IEEE, 2006.

[31] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 511–518. IEEE, 2001.

[32] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[33] P. Wang, M. B. Green, Q. Ji, and J. Wayman. Automatic eye detection and its validation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops (CVPR Workshops)*, pages 164–164. IEEE, 2005.

[34] S.-J. Wang, S. Wu, X. Qian, J. Li, and X. Fu. A main directional maximal difference analysis for spotting facial movements from long-term videos. *Neurocomputing*, 230:382–389, 2017.

[35] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2017.

[36] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31(4), 2012.

[37] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014.

[38] Z. Zhang, Z. Pi, and B. Liu. TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on biomedical engineering*, 62(2):522–531, 2015.