# Emotional State Recognition with Micro-expressions and Pulse Rate Variability

Reda Belaiche[1(✉)], Rita Meziati Sabour[1], Cyrille Migniot[1], Yannick Benezeth[1], Dominique Ginhac[1], Keisuke Nakamura[2], Randy Gomez[2], and Fan Yang[1]

[1] ImViA EA 7535, Univ. Bourgogne Franche-Comté, Dijon, France
{reda.belaiche,rita.meziatisabour,cyrille.migniot,
yannick.benezeth,dominique.ginhac,fan.yang}@u-bourgogne.fr
[2] Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako-shi, Saitama, Japan
{keisuke,r.gomez}@jp.honda-ri.com

**Abstract.** Machine learning has known a tremendous growth within the last years, and lately, thanks to that, some computer vision algorithms started to access what is difficult or even impossible to perceive by the human eye. It is then natural that scientists began looking for ways to probe humans' emotions and their psyche with this technology. In this paper, we study the feasibility of recognizing and classifying the abstract concept of *emotional states* from videos of people facing a regular RGB camera. We do so by using the barely perceptible micro facial expressions humans cannot control, as well as the spontaneous variations of the pulse rate that we estimated using remote photoplethysmography. We compare these two modalities and our experimental results show that it is possible to classify emotional states from these implicit information gathered from regular cameras with encouraging performances.

**Keywords:** Affective computing · Facial expressions · LBP · Pulse rate variability · Remote photoplethysmography

## 1 Introduction

People's general emotional state and mood have been much studied topics in the fields of psychology and medicine. It has been proved that a person's emotional state can impact their reaction time and learning ability in the short term, and even their health in the long run [1,2]. Being able to automatically predict a person's emotional state offers various real-world applications [3,4] such as neuromarketing [5] or automobile drivers' monitoring [6]. Contrary to simple emotions, emotional states are complex states of the human mind that are provoked by their surroundings or internal thoughts over a certain period of time. To recognize a person's emotional state, computer scientists researched many cues: gestures, voice intonations, and also macro facial expressions variations


E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11751, pp. 26–35, 2019.
https://doi.org/10.1007/978-3-030-30642-7_3

over time. More recently, the scientific community started to gain interest in the exploration of micro-expressions [7].

Facial expressions offer very important benchmarks in every day's social interactions. Most people are familiar with macro facial expressions, however, few people are aware of the existence of micro facial expressions [7,8], and even fewer know how to detect and recognize said micro-expressions. Initially discovered by Haggard and Isaacs [9], micro-expressions are a type of involuntary facial expressions that are extremely fast and of very low intensity. Their duration is within 1/4 s, which makes their localization and analysis rather complicated tasks. Micro-expressions can occur in two situations: conscious suppression and unconscious repression. Conscious suppression happens when a person intentionally tries to stop themselves from showing their true emotions or tries to hide them. Unconscious repression occurs when the subject himself does not realize their true emotions. In both cases, micro-expressions betray the subject's real emotions independently from his awareness of their existence.

Another avenue of research in emotion recognition is based on the analysis of physiological signals such as skin temperature, electrodermal activity or electromyography. Among many physiological features, the variations in the cardiac rhythms are an interesting indicator of the autonomic function [10]. In fact, the heart beating rate continuously fluctuates. The *Heart Rate Variability* (HRV) is conventionally defined as the time intervals between successive beats and is usually estimated from Electrocardiogram recordings. In the last decades, non-contact methods to evaluate the cardiac activity have been developed. One particular method is *Remote Photopletysmography* (RPPG) [11,12], which enables to estimate the pulse rate from a video. The basic principle of RPPG stems from reflective photoplethysmography where the light reaching a camera is modulated by the blood pulsations of the skin. The rhythmic beating of the heart results in the pulsating blood volume alterations, which in turn lead to minute changes in the skin color that can be quantified using different signal processing techniques to generate a cardiac signal. From an RPPG signal, the *Pulse Rate Variability* (PRV) can be estimated by calculating the pulse-to-pulse time intervals. Both HRV and PRV describe changing heart beat rhythms, and multiple researches have demonstrated similarity between PRV and HRV [13,14]. Therefore, the PRV can also be an indicator of the autonomic activity [15].

In this paper, we explore the feasibility of classifying and predicting a person's emotional state, based on two kinds of hidden information only perceptible to computer vision algorithms. This is realized through facial expression recognition and its recent extension to micro-expression recognition on one hand, and the analysis of the remotely measurable PRV on the other hand. These two approaches are quite complementary since *Micro/Macro-Facial Expressions* (M/M-FEs) describe the pixel content of the whole image at small intervals of time, while PRV processes the signal given by the color changes in pixels over a longer duration. In this work, we use the $CAS(ME)^2$ dataset [16], in which the subjects were exposed to different kinds of excitation videos to induce different reactions while they themselves were filmed. The main objective of this dataset is

to facilitate the development of algorithms for micro-expressions spotting from long video streams and to the best of our knowledge, this is the first study where physiological signals are estimated from this dataset. The two modalities explored are also the only ones we can use on that dataset.

The paper is organized as follows: we describe M/M-FEs and PRV-based feature extraction methods used for this study in Sect. 2. Experiments are presented and discussed in Sect. 3 and a conclusion is given in Sect. 4.

## 2 Feature Extraction Methods

In order to recognize emotional states, we have to go through two steps: feature extraction (M/M-FEs and PRV) and classification. In this section, we present the estimation of M/M-FEs and PRV-based features.

### 2.1 M/M-FEs-based Feature Extraction

We use the *Local Binary Pattern Three Orthogonal Planes* (LBP_TOP) operator, which is the baseline descriptor used as reference in most papers studying micro-expressions [16,17] to describe M/M-FE videos. The LBP operator is a type of visual descriptor that was originally designed for texture description [18]. The general idea is to threshold a small area around each pixel in order to build a binary code. This code is obtained by comparing neighbour pixel values with the center pixel: values superior or equal to the center pixel's value get assigned a 1 while smaller values get assigned a 0. The choice of the surrounding area directly affects the kind of edges it is possible to detect in an image. For pixel neighborhoods referring to $N$ sampling points on a circle of radius $R$, we generally use the notation $LBP_{N,R}$, whose value for a pixel $c$ can be given by:

$$LBP_{N,R} = \sum_{p=0}^{N-1} t(g_p - g_c)2^p. \tag{1}$$

Here $g_c$ represents the gray value of the center pixel $c$ while $g_p$ represents the gray values of equally spaced pixels on a circle of radius $R$, $t$ defines a thresholding function $t(x) = 1$ if $x \geq 0$ and $t(x) = 0$ otherwise. The feature vector representing an input image is calculated by extracting the histogram distribution of the LBP. We can consider LBP as texture primitives that include different types of curved edges, spots, flat areas, and so on. For an efficient facial representation, images usually get divided into local blocks from which we extract the LBP histograms and concatenate them into an enhanced feature histogram [19]. Local texture can then be described using said histograms of the binary values for a block in the image. The conventional LBP only serves for spatial data in 2D images. To describe data in the 3D spatio-temporal domain, the basic LBP is extracted from the three planes XY, XT and YT for each pixel as shown in Fig. 1. The resulting three histograms are then concatenated into a feature vector describing the video.

LBP_TOP can only describe a single M/M-FE. The method to describe a whole video containing several M/M-FEs is explained in Sect. 3.2.

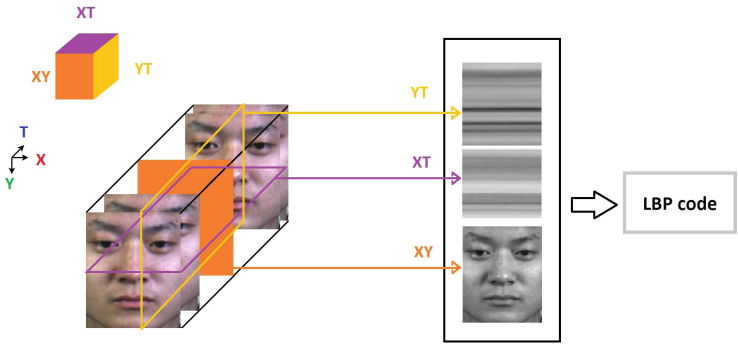


**Fig. 1.** Illustration of a spatiotemporal volume of a video [16]

### 2.2 PRV-Based Feature Extraction

Three main steps are followed to obtain the PRV signal as summarized in Fig. 2: face detection and tracking, RPPG pulse extraction and PRV estimation. First, the Viola-Jones face detection algorithm is used to detect the region of interest (ROI) for each video frame. The location of the ROI is then tracked and predicted with a linear Kalman filter. Next, the skin is detected by using the method Conaire *et al.* proposed in [20], allowing to select pixels that are spatially averaged. This yields to a unique RGB triplet for each frame. The triplets are then concatenated to form the RGB temporal traces.

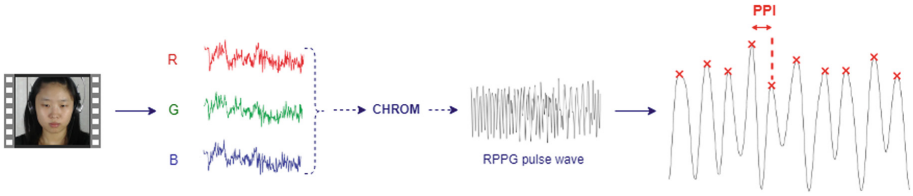


**Fig. 2.** Framework to obtain a PRV signal using RPPG

The second step consists in computing the pulse signal from the RGB traces. Many advanced and complex techniques have been proposed recently [11,21]. In this work, we use the chrominance algorithm proposed by De Haan *et al.* [22], and denoted as CHROM in Fig. 2. The principal advantage of this method lies in its computational simplicity, owing to its analytic formulation. After normalizing the RGB traces (let $R_n$, $G_n$ and $B_n$ be the relative normalized traces), two orthogonal chrominance signals $X_s$ and $Y_s$ are built as: $X_s = 3R_n - 2G_n$ and $Y_s = 1.5R_n + G_n - 1.5B_n$ . $X_s$ and $Y_s$ are then band-pass filtered with a Butterworth filter (cut-off frequencies of 0.7 and 3.5 Hz) to give two signals $X_f$ and $Y_f$. The pulse signal $S$ is then obtained as: $S = X_f - \alpha Y_f$, where $\alpha = \frac{\sigma(X_f)}{\sigma(Y_f)}$, and $\sigma(.)$ is

the standard deviation operator. Including the ratio $\alpha$ minimizes disturbances due to motion, since they alter the amplitudes of the chrominance signals $X_s$ and $Y_s$ in the same way while the cardiac pulse signal does not. The third step is to interpolate and resample (we used a sampling rate of 125 Hz) the RPPG pulse wave in order to increase the time domain resolution and to facilitate peak detection. *Pulse-to-Pulse Interval* (PPI) time series is then measured to constitute the PRV signal.

From the PRV signal we extract time-domain, frequency-domain and statistical features, which are defined as follows:

- **time-domain:** features that are computed in this study are the standard deviation of the pulse-to-pulse intervals (SDPP) and the square root of the mean squared of successive differences (RMSSD) of the PPI series. SDPP is the square root of PPI variance, and reflects the effect of all the components that induce the pulse variability during the video recording. RMSSD describes the evolution of consecutive pulse-to-pulse time intervals and reflects the high-frequency component of the PRV. SDPP and RMSSD are obtained as:

$$SDPP = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(PP_i - \overline{PP})^2} \tag{2}$$

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{i=2}^{N}(PP_i - PP_{i-1})^2} \tag{3}$$

  where $N$ is the total number of PP intervals, $PP_i$ is the $i^{th}$ PPI and $\overline{PP}$ is the mean value of the PPI series.
- **frequency-domain:** the density spectrum analysis of the PRV gives the low (LF) and high (HF) frequency components of the pulse variability. Since the interpretation of the role of the LF component in describing the autonomic function is complex, as it reflects both the sympathetic and the parasympathetic activities; we focused on the HF component. The HF range includes frequencies between 0.15 Hz and 0.4 Hz, and represents the PNS activity.
- **statistical analysis:** the TIPP describes a tiangular iterpolation of the PPI histogram. It is the width of the triangular function that best fits the PPI histogram.

## 3 Experiments and Discussion

### 3.1 Dataset Presentation

The number of scientific papers dealing with the automatic analysis of micro-expressions is rather limited. One of the reasons for it can be attributed to the lack of datasets containing real micro-expressions. Fortunately, this is beginning to change and recently a new dataset for facilitating the study on spontaneous

micro-expressions spotting and recognition has been made public. This dataset, named $CAS(ME)^2$ [16], is the first dataset that can be used for the study of facial micro-expressions, and offers at the same time the possibility to estimate, by video analysis, physiological parameters such as PRV. Indeed, the videos are sufficiently long and with no inter-frame compression, allowing to use RPPG techniques.

22 participants in total were filmed while watching 3 different kinds of excitation videos (disgust, anger and happiness). Each candidate was asked not to show his emotions. This was done to minimize the number of easily noticeable macro-expressions and to maximize the number of micro-expressions they would show. The dataset was originally proposed for automatic M/M-FE spotting and recognition, and while micro-expressions spotting has been getting good results [23], their recognition still presents many challenges.

$CAS(ME)^2$ offers 2 kinds of annotations for M/M-FEs. The first one is done according to the facial muscle movements based on the *Action Units* (AU) following the Facial Action Coding System (FACS) proposed by Ekman. The second annotations are based on the self-reported emotions from the candidates. These two annotations are not in agreement for all the videos. Furthermore, in some cases, the emotional type of the elicitation video and the annotations based on AU are sometimes contradictory (some subject would show a negative facial expression in front of a happiness-inducing video). 24.05% of the facial expressions are classified as *others* (*i.e.* where related AU are not discriminative). Some subjects would also show contradictory facial expressions on the same video. These observations encouraged us to propose for the first time the use of the emotional type of elicitation videos, also called excitation videos, of $CAS(ME)^2$ as our ground truth. We work under the assumption that the emotional state of the person watching a video would be equivalent to the emotion that video was made to induce. Our motivation behind this decision is to use a labeling that would be more straightforward and less inclined to cause confusion.

In total, the number of available videos is 62, with 14 videos provoking happiness, 24 for disgust and 24 for anger. The video length ranged from 1 min (1800 frames) to approximately 2 min and 30 s (4500 frames).

### 3.2 Model Validation Protocol

The modalities we concentrate on were tested on the same videos following the *Leave One Subject Out* (LOSO) cross-validation protocol: one subject's data is used as a test set in each fold of the cross-validation. This is done to better reproduce actual use conditions where the encountered subjects are alien to the model when it was trained. Older studies would use k-fold cross-validation; however, this would result in a severe case of overfitting as the accuracies on the test sets would be much higher than with LOSO. This can be attributed to the fact that samples from the same subject would be present in both the training and testing sets. Considering the fact that the same subject can show the exact same expression many times (which may cause occurrences of the same expression from that subject to belong to the training and the test sets at the

same time), and that some subjects can be more inclined to show a specific type of emotion more often, using the LOSO protocol seems to be the most rigorous option.

### 3.3 Recognition with M/M-FEs and PRV

After extracting the M/M-FEs' features with LBP_TOP, a simple majority vote is applied on the different results of each M/M-FE in a video to get the person's general feeling over an extended period of time. The winning class would be used to represent a person's emotional state for the whole video.

The overall classification then needs four SVMs with *Radial Basis Function* (RBF) kernels: one that was trained on the three classes, and three that were trained on each possible pairing of two classes from the three initial ones. The 2-class SVMs were used when we had a perfect equality between 2 classes using the 3-class emotional state classifier. The parameters for the spatio-temporal radii of LBP were equivalent to the ones used in [16]. Concerning the PRV, the features were obtained from the extracted PRV waveforms of each video, and concatenated to form (SDPP, RMSSD, HF, TIPP) vectors. Emotional state classification was then realised based on these values using a non-linear SVM, with an RBF kernel.

PRV-based features encapsulate slow changes in the heart rate. As a consequence, the entire videos are needed to obtain meaningful PRV information. This is not the case of M/M-FEs, which appear for very small durations (between 1/2 and 4 s for macro and less than 1/4 s for micro-expressions). Besides, each video contains different M/M-FEs, which impelled us to apply an aggregation process in order to describe the videos.

### 3.4 Results and Discussion

The final results of emotion classification using macro and micro-expressions and the pulse rate variability are given by Table 1. Confusion matrices give the rate of successful and unsuccessful predictions of the emotional states in order to estimate to what point a classifier confuses two classes. Accuracy rates describe how reliable the classifiers are at predicting emotional states correctly. The voting process on M/M-FEs had an overall accuracy of 42.74% while PRV's accuracy was 59.79 %. If we compare the results of LBP_TOP on excitation videos (for the emotional state) and its original use for labels based on AU, we can see that with 40.95% [16] on the 4 AU-based classes and 42.74% on the excitation video-based labeling, the scores are comparable.

The observed results are interesting since PRV performances surpass those of M/M-FEs, which is quite surprising as $CAS(ME)^2$ was originally made for M/M-FEs recognition but not for PRV estimation. Besides, the dataset does not propose videos with a reference neutral emotional state, preventing us from normalizing the PRV-based features as is usually done to mitigate the impact of the user-dependence effect [24].

**Table 1.** Emotional state classification confusion matrix and accuracies for M/M-FEs and PRV. Present excitation emotions in $CAS(ME)^2$ are Disgust, Anger and Happiness. Results are expressed in percentage (%)

| | M/M-FEs | | | PRV | | |
|---|---|---|---|---|---|---|
| True | Predicted | | | | | |
| | Disgust | Anger | Happiness | Disgust | Anger | Happiness |
| Disgust | 57.4 | 40.4 | 2.1 | 75.0 | 4.0 | 21.0 |
| Anger | 44.7 | 55.3 | 0.0 | 0.0 | 75.0 | 25.0 |
| Happiness | 56.7 | 43.3 | 0.0 | 35.0 | 20.0 | 45.0 |
| **Accuracy** | **42.74** | | | **59.79** | | |

A possible explanation for the discrepancy of the results obtained with the two modalities could be that M/M-FEs adequately describe sudden changes in emotions, contrary to PRV features that cannot describe these rapid variations in emotions. The difference in temporality between these two modalities is quite obvious; however, it suggests that the use of M/M-FEs for the description of emotions felt over long periods, *e.g.* as in [25], might not be the most appropriate modality. Moreover, the relationship between the emotional states and the occurrences of M/M-FEs is actually rather complex. An interesting illustration of this hypothesis is that we observed cases of contradictory facial expressions on the same video.

## 4 Conclusion

From simple videos it is possible to extract analytical and physiological features, including M/M-FEs and PRV. Our results show that PRV can be an interesting feature to estimate emotional states with a classification accuracy of about 60%. Although M/M-FEs yielded lower results, ways of improvement have to be delved into. We emphasize with this work that the use of M/M-FEs for the description of emotions must be considered carefully. Mainly because of the intricate relationship between the experienced emotion and its display medium. If both modalities are interesting for studies on emotion recognition, their complementarity will undoubtedly allow for a better apprehension of the complexity of the emotions felt and displayed.

Future works could include the use of other datasets that would take advantage of the complementarity of these two promising modalities with stimulus of various durations. Concerning PRV extraction, having neutral emotional state sequences would allow normalizing our features to mitigate the impact of the user-dependence. Eventually, we also plan a fusion of M/M-FEs and PRV features for emotional state prediction and to investigate the use of M/M-FEs features to alleviate the intra-user dependence of PRV features.

## References


1. Rothman, A.J., Salove, P., et al.: Emotional states and physical health. Am. Psychol. **55**(1), 110 (2000)
2. Tyng, C.M., Hafeez, U.A., et al.: The influences of emotion on learning and memory. Front. Psychol. **8**, 1454 (2017)
3. Chen, Y.-L., Chang, C.-L., Yeh, C.-S.: Emotion classification of youtube videos. Decis. Support Syst. **101**, 40–50 (2017)
4. Bargal, S., et al.: Emotion recognition in the wild from videos using images. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction (2016)
5. Vecchiato, G., Astolfi, L., Fallani, F.D.V.: On the use of EEG or MEG brain imaging tools in neuromarketing research. Comput. Intell. Neurosci. **2011**, 3 (2011)
6. Nass, C., Jonsson, M., Harris, H.: Improving automotive safety by pairing driver emotion and car voice emotion. In: Extended Abstracts on Human Factors in Computing Systems (2005)
7. Ekman, P., Friesen, W.V.: Nonverbal leakage and clues to deception. Psychiatry **32**(1), 88–106 (1969)
8. Ekman, P.: Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition). WW Norton & Company, New York (2009)
9. Haggard, E.A., Isaacs, K.S.: Methods of Research in Psychotherapy. Springer, New York (1966). https://doi.org/10.1007/978-1-4684-6045-2
10. McCraty, R., Atkinson, M., Tomasino, D.: Science of the Heart - Exploring the Role of the Heart in Human Performance, Institute of HeartMath (2001)
11. Sun, Y., Thakor, N.: Photoplethysmography revisited: from contact to non contact, from point to imaging. IEEE Trans. Biomed. Eng. **63**(3), 463–477 (2016)
12. McDuff, D.J., Estepp, J.R., Piassecki, A.M., Blackford, E.B.: A survey of remote optical photoplethysmographic imaging methods. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2015)
13. Gil, E., Orini, M., Bailon, R., et al.: Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions. Physiol. Measur. **31**(9), 1271 (2010)
14. Schafer, A., Vagedes, J.: How accurate is pulse rate variability as an estimate of heart rate variability? a review on studies comparing photoplethysmographic technology with an electrocardiogram. Int. J. Cardiol. **166**(1), 15–29 (2013)
15. Nitzan, M., Babchenko, A., Khanokh, B., Landau, D.: The variability of the photoplethysmographic signal - a potential method for the evaluation of the autonomic nervous system. Physiol. Measur. **19**(1), 93 (1998)
16. Qu, F., Wang, S.J., Yan, W.J., et al.: CAS(ME)$^2$: a database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Trans. Affect. Comput. **9**(4), 424–436 (2017)
17. Pfister, T., Li, X., Zhao, G., Pietikainen, M.: Recognising spontaneous facial micro-expressions. In: IEEE International Conference on Computer Vision (2011)
18. Ojala, T., Pietikäinenand, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **7**, 971–987 (2002)
19. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **12**, 2037–2041 (2006)

20. Conaire, C.O., O'Connor, N.E., Smeaton, A.F.: Detector adaptation by maximising agreement between independant data sources. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
21. Poh, M.Z., McDuff, D.J., Picard, R.W.: Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans. Biomed. Eng. **58**(1), 7–11 (2011)
22. De Haan, G., Jeanne, V.: Robust pulse-rate from chrominance-based rPPG. IEEE Trans. Biomed. Eng. **60**(10), 2878–2886 (2013)
23. Han, Y., Li, B., Lai, Y.K., Liu, Y.J.: CFD: a collaborative feature difference method for spontaneous micro-expression spotting. In: 25th IEEE ICIP (2018)
24. Benezeth, Y., Li, P., Macwan, R., et al.: Remote heart rate variability for emotional state monitoring. In: IEEE International Conference on Biomedical and Health Informatics (2018)
25. Li, X., Hong, X., Moilanen, A., et al.: Towards reading hidden emotions: a comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Trans. Affect. Comput. **9**(4), 563–577 (2018)