sciendo

CAIM
ISSN 2038-0909

Commun. Appl. Ind. Math. 15 (2), 2024, 102–123 DOI: 10.2478/caim-2024-0016

Research Article

A visual perception-guided data augmentation method for efficient machine learning-based detection of facial micro-expressions

Vittoria Bruni[1], Salvatore Cuomo[2], Domenico Vitulano[1*]

[1]Department of Basic and Applied Sciences for Engineering, Sapienza Rome University, Italy

[2]Department of Mathematics and Applications "R. Caccioppoli", University of Naples Federico II, Italy

[*]Email address for correspondence: domenico.vitulano@uniroma1.it

Communicated by Clemente Cesarano

Received on 06 25, 2024. Accepted on 10 3, 2024.

Abstract

Automatic spotting and classification of facial Micro-Expressions (MEs) in 'in-the-wild' videos is a topic of great interest in different fields involving sentiment analysis. Unfortunately, automatic spotting also represents a great challenge due to MEs quick temporal evolution and the lack of correctly annotated videos captured in the wild. In fact, the former makes ME difficult to grasp, while the latter results in the scarcity of real examples of spontaneous expressions in uncontrolled contexts. This paper proposes a novel but very simple spotting method that mainly exploits MEs perceptual characteristics. Specifically, the contribution is twofold: i) a distinguishing feature is defined for MEs in a domain that can capture and represent the peceptual stimuli of MEs, thus representing a suitable input for a standard binary classifier; ii) a proper numerical strategy is developed to augment the training set used to define the classification model. The rationale is that since MEs are visible by a human observer almost regardless of the specific context, it stands to reason that they have some sort of perceptual signature that activates pre-attentive vision. In this work this fingerprint is called Perceptual Emotional Signature (PES) and is modelled using the well-known Structural SIMilarity index (SSIM), which is a measure based on visual perception. A machine learning based classifier is then appropriately trained to recognize PESs. For this purpose, a suitable numerical strategy is applied to augment the training set; it mainly exploits error propagation rules in accordance with perceptual sensitivity to noise. The whole procedure is called PESMESS - Perceptual Emotional Signature of Micro-Expressions via SSIM and SVM. Preliminary studies show that SSIM can effectively guide the detection of MEs by identifying frames that contain PESs. Localization of PESs is accomplished using a properly trained Support Vector Machine (SVM) classifier that benefits from very short input feature vectors. Various tests on different benchmarking databases, containing both 'simulated' and 'in-the-wild' videos, confirm the potential and promising effectiveness of PESMESS when trained on appropriately perception-based augmented feature vectors.

Keywords: `Perceptual-based data augmentation, Facial Microexpression spotting, Perceptual fingerprint, Structural SIMilarity Index, Support Vector Machine.`

AMS subject classification: `68T10 - 68U10`

1. Introduction

Sentiment analysis is becoming a topic of increasing interest and an active research field. It finds applications in different fields, ranging from psychology to marketing and investigation, thanks to both new communication paradigms and the significant proliferation and diffusion of multimedia contents. One of the most stimulating and challenging task in this context is certainly represented by the automatic understanding of people emotional state [42]. Based on the pioneering study of Albert Mehrabian in 1972 [38], this task can be accomplished by focusing on paraverbal and nonverbal communication, that represent the main components in human conversation (38% and 55% respectively) compared to the verbal communication (whose contribution is only 7%). Nonverbal communication includes postures, facial expressions, eye gaze, gestures, etc., and research on this topic is mainly based on the results achieved by Haggard and Isaacs [22] first, and Ekman and Friesen later [17]. In particular, facial Micro Expressions (MEs) are *very brief, subtle, and involuntary facial expressions which normally occur when*

a person either deliberately or unconsciously conceals his or her genuine emotions [19,56]. Therefore, they convey the real emotional state of a person, including the one he tries to hide — Figure 1 shows two examples. Due to their key role in the non verbal communication, they are currently under study in very different fields like marketing, medicine, forensics and security [18,54], in which it is essential to have an effective computational framework to aid the analysis process. Although desirable, this task is difficult especially in the detection stage [44], which is challenging even for a human being, mainly for two reasons:*i)* very short duration of MEs, ranging from 1/25 to 1/5 of a second (recently relaxed to a maximum duration of 1/2 second) [56]; *ii)* very low intensity, as MEs can be masked by partial facial movements.

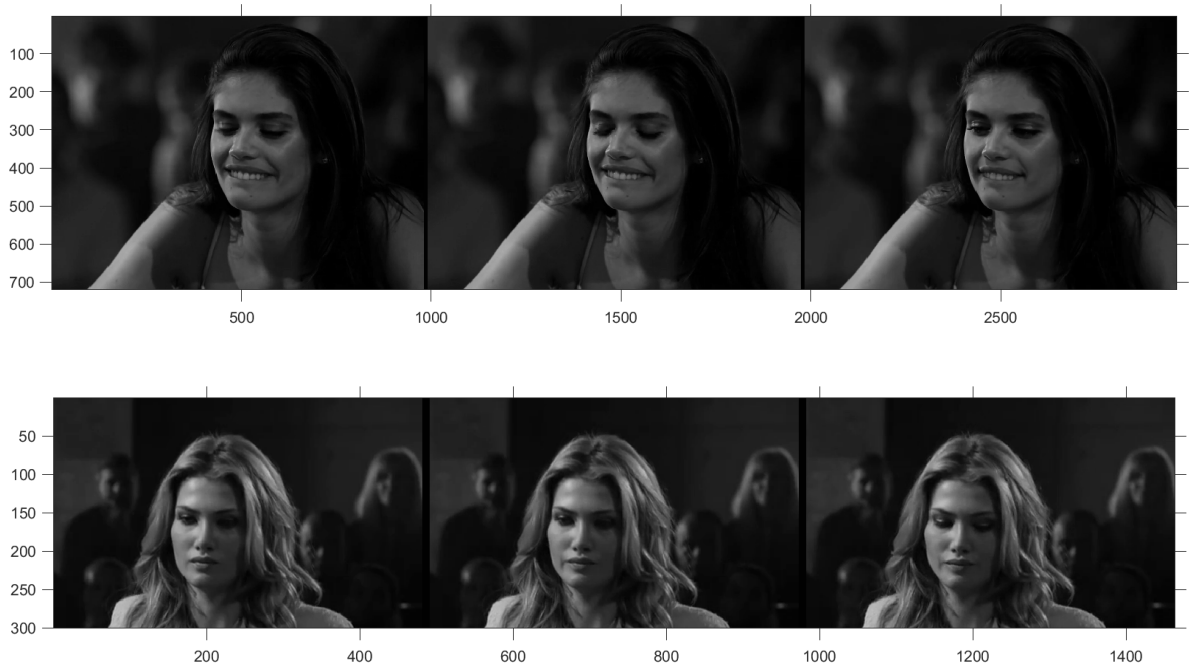


Figure 1. MEVIEW database [39]. Some frames relative to two MEs: *Top*) *happiness* (Video 11.3); *Bottom*) *contempt* (Video 14.1).

Therefore, the literature on MEs analysis can be broadly divided into two classes: *i)* MEs spotting and *ii)* MEs recognition. The former consists of identifying those (few) frames capturing MEs. The latter assigns a specific emotional state, such as fear, disgust, sadness, etc., to the detected ME [17,22]. This paper focuses on ME spotting. As Figure 2 shows, it usually consists of cascaded steps that include facial points registration and tracking, salient face regions detection under noisy conditions, ME feature description and recognition. The cascaded approach makes the procedure sensitive to error spreading over successive blocks [42]. In addition, some of these blocks require computing time that is not suited for real-time applications, especially those involving long (25-60 frames per second) and complicated in the wild videos (e.g., corrupted by adverse light conditions, additional unavoidable face movements etc.).

Based on the first empirical and pioneering studies in [12,13], this paper proposes a very simple approach oriented to detect video subsequences containing MEs with high probability. It mainly aims at simulating the sixth sense that an observer employs for detecting MEs during a conversation. Modeling this sense is a hard task since the observer catches something strange without recognizing specific details, appearing that no specific computable features are used. To reach this goal, the proposed approach uses the Structural SIMilarity index (SSIM) [51] over pairs of subsequent frames. SSIM is a well-known image quality assessment metric defined through three terms that relate to the Human Visual System. Specifically, luminance and contrast (first two terms) account for the two visual channels of vision sys-

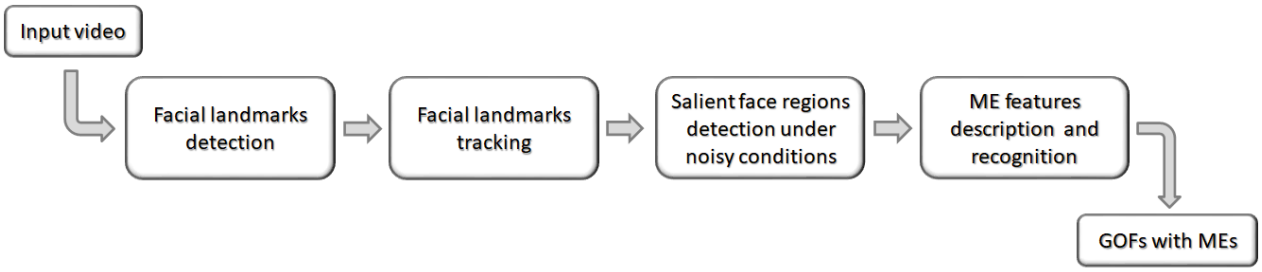


Figure 2. Block scheme of the main steps of MEs spotting framework. The output consists of groups of frames (GOFs) where MEs occur.

tem [5,36,51,55]; the third term (the structural one) captures frames correlation and locally 'replaces' a temporal motion estimator. As already shown in [12,13], the probability distribution of SSIM at the frames capturing MEs shows a peculiar asymmetrical behaviour that is caused by the lack (or reduction) of motion before ME's apex (maximum expression of ME); this is mainly due to the interlocutor's need to conceal his emotional state. Therefore, the movement caused by ME can be seen as an outlier in SSIM distribution.

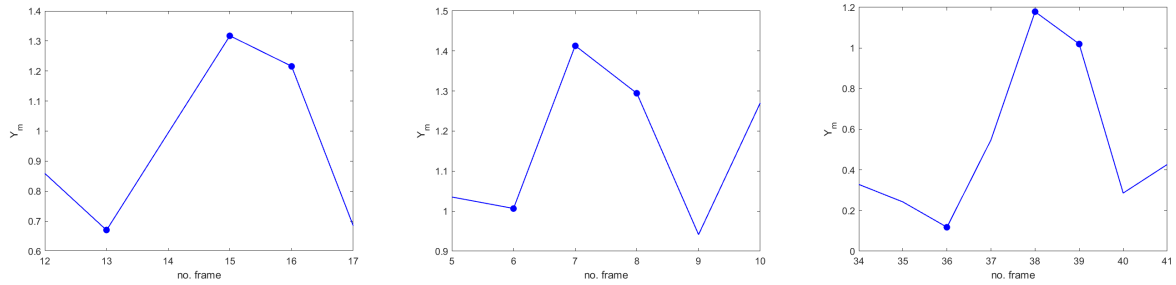


Figure 3. Three examples of PES (whose samples are marked by dots) corresponding to different kinds of ME.

Based on this observation, in [12] the Perceptual Emotional Signature (PES) of a ME was defined as the temporal waveform of the asymmetry index of SSIM. On the other hand, a detailed analysis allows us to observe that different MEs have very similar PESs — some examples are shown in Figure 3. This enables us to design a simple Support Vector Machine (SVM) classifier that discriminates true MEs from false alarms when properly trained on a well-populated dataset. To compensate the lack of databases having genuine (i.e., not simulated) MEs, a data augmentation strategy is also adopted ([2], pg. 337). It mainly consists in adding some noise over the original PES profile; noise intensity is subject to satisfying certain conditions arising from perceptual constraints on the original video sequences. Specifically, the modification of PES is studied in dependence on some distortion applied to the original sequence, where the distortion is constrained to visual masking effects.

The proposed PESMESS acts as a preprocessing step in MEs analysis by selecting GOFs (Group Of Frames) where MEs occur with high probability. Existing and more sophisticated tools [42] can then run only on the selected GOFs. The main advantage of PESMESS then consists of a drastic reduction of the computational complexity of the whole MEs spotting procedure, thus contributing to build a real time framework. Moreover, to the best of the authors' knowledge, this is the first attempt in which visual perception tools, specifically SSIM, are used to manage non verbal communication.

The remainder of the paper is organized as follows. The next section contains a very short review on the state of the art of ME spotting. Section 3 illustrates the proposed method along with the formal description of the visual perception-based augmentation procedure, while Section 4 shows some experimental results. The last section draws the conclusions.

Table 1. Broad classification of some of the most performing approaches for MEs spotting with relative strategies and possible limits in critical acquisition conditions.

Reference	Strategy	Possible Drawbacks
[33]	manual selection of facial landmarks	sensitivity to face position, occlusions etc
[3,15,26,37,40]	automatic selection of facial landmarks	sensitivity to face position, occlusions etc
[27,48]	face masking and region retrieval (mouth and eye)	possible generation of false alarms due to sensitivity to eye blinking
[28,31,32]	face masking and region retrieval (eye, eyebrow, mouth)	increased risk of excluding too much meaningful information
[43]	spotting as frame-by frame classification	not too robust to spontaneous MEs
[47]	temporal method: analysis of the temporal strain	more robust to spontaneous MEs but limited precision (dependence on threshold values)
[41,52]	features difference for facial motion variations	dependence on predefined temporal window
[16]	separation between eye blinking and MES through the analysis of phase variations between frames .	limited precision in the wild
[35,59]	apex frame detection by geometric and/or appearance features	limited precision in the wild
[30,31,33,50,53]	temporal spotting using intervals from onset to offset frame	dependence on variability of contexts and operational scenarios
[23,29,61]	apex spotting (machine and deep learning)	dependence on variability of contexts and operational scenarios

2. A short review on MEs spotting

Research on facial MEs has increased so much in recent years that it is difficult to present a comprehensive and exhaustive review of all the approaches available in the literature. There are several tutorials on the topic, such as [20,21,42,62], which can be read to get a broad overview of ME detection and recognition. Since we are interested in an innovative approach to ME detection, this section is limited to this topic. To help reading, Table 1 summarizes some of the referenced works by indicating their main strategy and limitations.

MEs spotting mainly aims at partitioning a video sequence into binary subsequences: those containing MEs and the ones that didn't. As already mentioned in the introduction, although ME spotting may seem to be a simple operation, it is in practice a very complicated process that involves multiple steps, such as preprocessing, feature description and ME identification [42]. Since one of the goals of the proposed approach is to deal with long, possibly "in-the-wild", videos which usually contain both MEs and other visual components like macro-expressions (longer than 10 seconds [45,60]), the spotting process becomes even more difficult in practice [34]. Therefore some specific procedures are needed to reduce the computational complexity and improve the performance of the whole spotting process, which includes landmark detection, landmark tracking, registration, masking and region retrieval [42]. Facial landmark detection involves locating landmark points from the face under examination. Several approaches have been proposed, starting from a manual selection of landmarks (originally there were only 12 [33]) to more general techniques (i.e. involving landmark tracking and registration) and automatic methods based on Constrained Local Model (CLM) [15], Dlib [26], Active Shape Model (ASM) [40], Discriminative Response Maps Fitting (DRMF) [3] and Face++ [37]. Face masking and region retrieval deal with the huge amount of false alarms caused by undesired facial movements. This phase accounts for specific face regions like eye [27] or mouth and eye [48]. Eyes are often considered because of their frequent movement in addition to blinking. Other approaches have shown that suitable region selection (for instance, eye, eyebrow, mouth) can contribute to increase the accuracy of the final spotting result. Some examples can be found in [28,31,32]. All the mentioned steps play a crucial and delicate role, as very high precision is required to prevent error amplification each time they are combined to get the final spotting output.

ME spotting approaches can be split into two groups: classifier-based methods and rule-based methods. The former includes frame-by-frame methods [43], which are not too robust to spontaneous MEs, and temporal methods [47], which take into account the relationships between frames by defining temporal strain maps, i.e. the amount of deformation incurred during motion. The latter methods are usually more effective for spontaneous MEs detection. Features difference analysis is used in [41,52], where facial motion variations are taken into account. However, they suffer from the dependence on the predefined temporal window that limits their adaptivity to videos having different rates. There are some approaches oriented to distinguish between eye blinking, usually seen as a false alarm and true MEs. An example is provided in [16], where an effective approach to distinguish between MEs and eye movements is presented, and it is based on the analysis of phase variations between frames through the Riesz Pyramid. A different class of methods focuses on apex frame detection [35,59] by using geometric and/or appearance features of specific face components.

It is worth noting that the enormous effort on this topic allows for several classifications to be defined. Again, they can be divided into two main classes [34]. The first includes methods that involve temporal spotting to locate the ME intervals from onset to offset frame; the other consists of methods that take into account apex spotting to detect only the apex frame. To the first class also belong approaches using machine learning tools, like [30] where SVM is used to classify ME and non-ME frames by means of Local Temporal Patterns, or deep learning approaches, such as Histogram of Oriented Optical Flow (HOOF) in [50], Micro-Expression Spotting Network (MESNet) in [53], Shallow Optical Flow Three-stream CNN (SOFTNet) in [33]. Among apex frame based approaches, it is worth mentioning the method in [31], which is based on a divide-and-conquer strategy, the method in [29], in which the detection of amplitude variations of the maximum frequency is exploited, and SMEConvNet [61], which adopts a deep learning strategy. Finally, it is worth mentioning the approach in [23], in which long videos are analysed using Main Directional Maximal Difference (MDMD).

Despite the proliferation of different and increasingly high-performance methodologies and strategies, MEs identification remains an open problem because of not only the inherent MEs characteristics but also the variability of contexts and operational scenarios that are not adequately represented by the currently existing benchmark datasets. In this work we are then interested in investigating the possibility of identifying a kind of robust and characteristic fingerprint of MEs by exploiting the sensitivity, albeit unconscious, of the human visual system to changes in the observed scenario during the pre-attentive phase. In addition, we aim to measure how this fingerprint is affected and masked by external factors and facial movements; this formal result is used to construct a sufficiently populated dataset that is used to train a machine learning classifier to recognize this fingerprint, thus helping to compensate for the lack of a dataset of annotated but authentic expressions.

Figure 4. Sequence 15.1 in MEVIEW dataset. Three key frames of a ME respectively corresponding to: Onset, Apex and Offset — see text for details.

3. The Proposed Model: PESMESS

As already pointed out, the duration of MEs is very short: it ranges from 100/166 ms to 500 ms. Moreover, facial MEs are mainly local: they can affect only a part of the speaker's face. Taking into account that the early visual system plays a major role in the first few hundreds ms [36], it is obvious

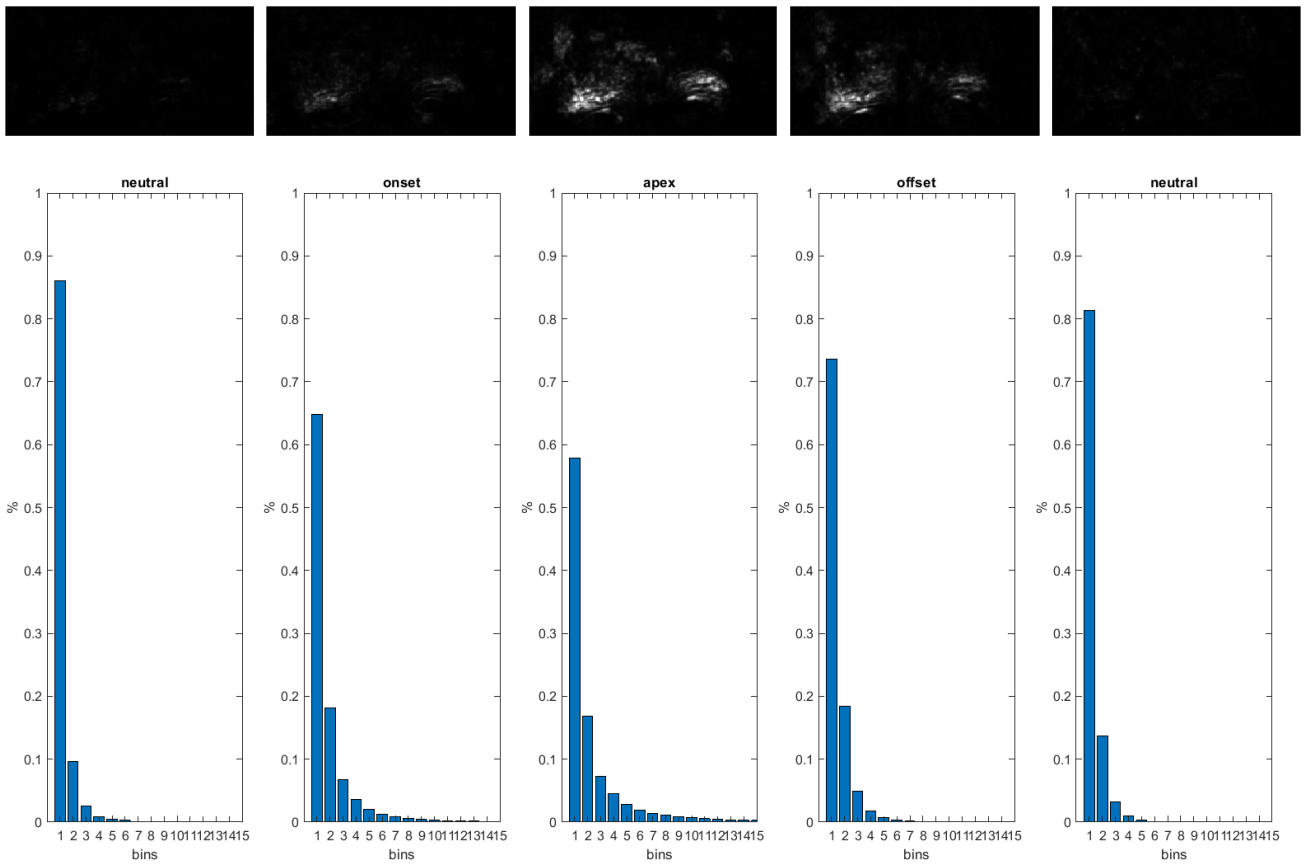


Figure 5. Sequence 15.1 from MEVIEW database [39]. *Top*) S map of key frame pairs of the most prominent ME time steps (lighter pixels indicate where the ME occurs). *Bottom*) Distributions of the corresponding S maps where the bin-size has been set equal to 0.01 — the Figure depicts a zoom of the distributions close to zero.

that MEs escape anyone attention. The human brain, as well as the human visual system, is calibrated to pick up slow and fast movements, in agreement with the theoretical formulation and formalization provided by Adelson and Bergen in their early papers — see for instance [1]. In the case of MEs, however, face movement is too fast to get a clear sense of what is happening — at least with the naked eye. Fortunately, modern cameras and computational devices can help us to face with this problem. Looking at test video frames, it can be seen that each ME is composed of three main phases, namely *onset, apex,* and *offset*, which are generally preceded and followed by a neutral phase (see Figure 4) [49]. A deeper analysis shows that while the onset and apex behave similarly, regardless of the type of ME, the offset may have a different duration: very short for spontaneous MEs, or longer when the subject is trying to control his or her expression. In the latter case, there are usually more frames in which the offset expression remains almost 'frozen' to mask the emotional state [13]. Complementing the preliminary study in [12], this paper will show that the information above is sufficient to give a good ME characterization. The underlying idea is to exploit the same tools adopted by the Human Visual System to detect the presence of MEs. Specifically, it can be shown that any ME has a sort of fingerprint that can be suitably recognized exploiting a global visual perception-based measure. If the latter is implemented through the well known *Structural SIMilarity* index (SSIM) [51], the resulting fingerprint has a shape very similar for any ME (see Figure 3). This fingerprint is denoted by PES (Perceptual Emotional Signature) and is achieved by considering the temporal symmetry of SSIM distribution, as detailed in the next section. Furthermore, by modeling external factors and other facial movements as noise, formal bounds for variability of PES are derived to define the constraints under which it can be recognized. Finally, it is shown how these

constraints can be used to augment the training set and properly train an SVM-based classifier using PES as input features.

3.1. *SSIM-based PES definition*

SSIM (*Structural SIMilarity* index) [4,51] is a full-reference image quality assessment measure that computes the visual difference between corresponding blocks (I and J) of two images as it follows

$$SSIM(I,J) = \underbrace{\frac{2\mu_I\mu_J + C_1}{\mu_I^2 + \mu_J^2 + C_1}}_{\substack{luminance\ adaptation}} \cdot \underbrace{\frac{2\sigma_{IJ} + C_2}{\sigma_I^2 + \sigma_J^2 + C_2}}_{\substack{contrast\ masking \\ + \\ spatial\ correlation}}, \tag{1}$$

where μ_I, μ_J, σ_I and σ_J respectively are the sample means and standard deviations of I and J, σ_{IJ} their covariance, while C_1 and C_2 are numerical stabilizing constants that depend on the dynamic range L of pixel vaules. Specifically, $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$ where $k_1 = 0.01$ and $k_2 = 0.03$ by default. Though its simplicity, this measure is able to properly combine three basic features to which human eye is mostly sensitive, i.e. luminance, contrast and structure, and the corresponding mechanisms, i.e. luminance adaptation, contrast masking, contrast and structures sensitivity. In particular, luminance and contrast gain control allows to maintain perceptual sensitivity under different lighting conditions, while contrast masking mainly measures to what extent an object having specific frequencies is visible whenever embedded in a background having similar frequencies [9]. SSIM values belong to the range $[-1, 1]$: the greater this value, the better the quality. To have values consistent with the concept of distance, the quantity

$$S(I,J) = 1 - SSIM(I,J) \tag{2}$$

can be considered — $S(I,J) \geq 0$, $S(I,J) = 0$ if and only if $I = J$, while the more $S(I,J)$ far from zero, the larger the difference between I and J. Since S can be computed locally, an S map is created when SSIM is computed at the neighborhoods of each pixel pairs in two subsequent frames. As a result, it plays a key role in revealing facial movements between subsequent frames and thus also MEs. Again, the rationale is that MEs can be seen as a specific, local and structural distortion between two consecutive frames. It turns out that MEs contribute to change the temporal distribution of S.

To fix this point better, Figure 5 shows the S maps computed between consecutive frames containing an ME for the sequence in Figure 4. As can be observed, the only contribution in the S map is due to ME at the face region, as the man mainly tries to remain impassive and expressionless. In particular, as Figure 5.bottom shows, the distribution of S values is unimodal, while its mode is close to the minimum admissible value (i.e., 0). This is because the skin appears almost homogeneous and makes no contribution to the S map. In the presence of ME, the distribution of S values shows considerable changes toward its tails: the more perceptually significant ME is (for example, *apex*), the more significant the change in the distribution of S. As empirically studied in [12], robust statistical estimators of distribution asymmetry are able to consistently capture this property. In particular, absolute indices based on order statistics have been shown to be more robust than Fisher's most famous definition of skewness [25], since they are more sensitive to the presence of extreme values, in which we are interested here. Thus, they have been shown to be sensitive to the observed change in the distribution of S when ME occurs. Some representative examples are:

- the second *Pearson's coefficient* of skewness

$$A_p = (E - Q_2),$$

  where E is the mean and Q_2 is the median value of S values;

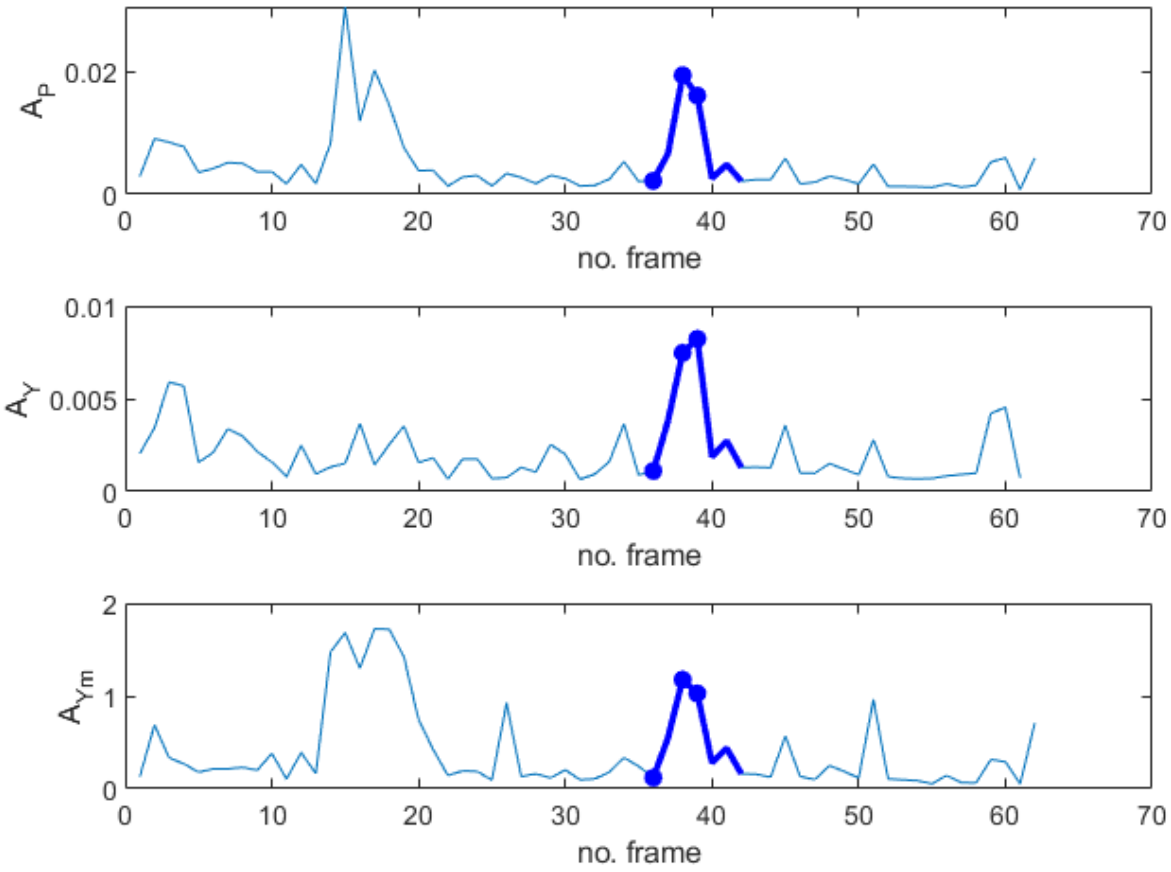


Figure 6. Behavior of Pearson (A_P), Yule (A_Y) and modified Yule (A_{Ym}) absolute asimmetry indices for sequence 15.1 of MEVIEW dataset. Bold lines indicate the ME, while dots are in correspondence to its significant temporal phases and identify ME perceptual fingerprint (PES).

- the *Yule's coefficient*

$$A_Y = Q_3 - 2Q_2 + Q_1,$$

  where Q_i is the $i-$th quartile;
- the modified *Yule's coefficient*

$$A_{Ym} = M - 2Q_2 + m, \tag{3}$$

  where $(M - m)$ is the range of S values.

With reference to Figure 6, it is worth observing that, regardless of their specific definition, all of the absolute indices mentioned above produce a similar profile marking the time phases of a ME: neutral, onset, apex, offset, neutral. In particular, a quick and significant change occurs between onset and apex, while a lighter one appears between apex and (the starting point of) the offset phase. In contrast, the temporal behavior of the perceptual index from offset to neutral state is less predictable. This is mainly due to the unconscious intention of a speaker to mask his own emotional state. The temporal profile of the asymmetry index can then be seen as an ideal fingerprint for a ME, and it will be defined as the *Perceptual Expression Signature* (PES).

3.2. *Definition of PES-based features for ME identification*

Because of the inherent nature of ME, PES shows almost the same, though not equal, profile for different types of ME. On the other hand, it has to be recognized among different profile types related to different clips of the video under consideration. This involves solving two different subproblems when defining a binary classifier:

- definition of the input feature vector;
- building a proper training set.

It is worth observing that the definition of the best classifier is out of the scope of this paper and the widely used Support Vector Machine (SVM) will be employed in the experimental results [2].

The considerations made at the end of the previous section may guide the definition of the feature vector. Denoting by $A(t)$ the generic asymmetry index at time instant t, PES feature vector is expected to be composed by the value of A at the *onset*, the *apex* and the *offset*. To make the recognition step more robust, an additional component v_4 should take into account the temporal perceptual index behavior from the offset to the neutral state. Therefore, PES feature vector is defined as

$$V = [A(t_1)\ A(t_2)\ A(t_3)\ v_4], \tag{4}$$

where t_1 refers to the *onset* and the remaining ones to *apex* and *offset* respectively, while $v_4 \in \{-1, +1\}$ and indicates the sign of the derivative of A from the offset to the neutral state. In fact, although the temporal perceptual index behavior from the offset to the neutral state is less predictable and measurable (see signal after PES), it usually shows a decreasing trend for MEs. Therefore, v_4 helps to discard false alarms by distinguishing profiles similar to PES but not related to MEs.

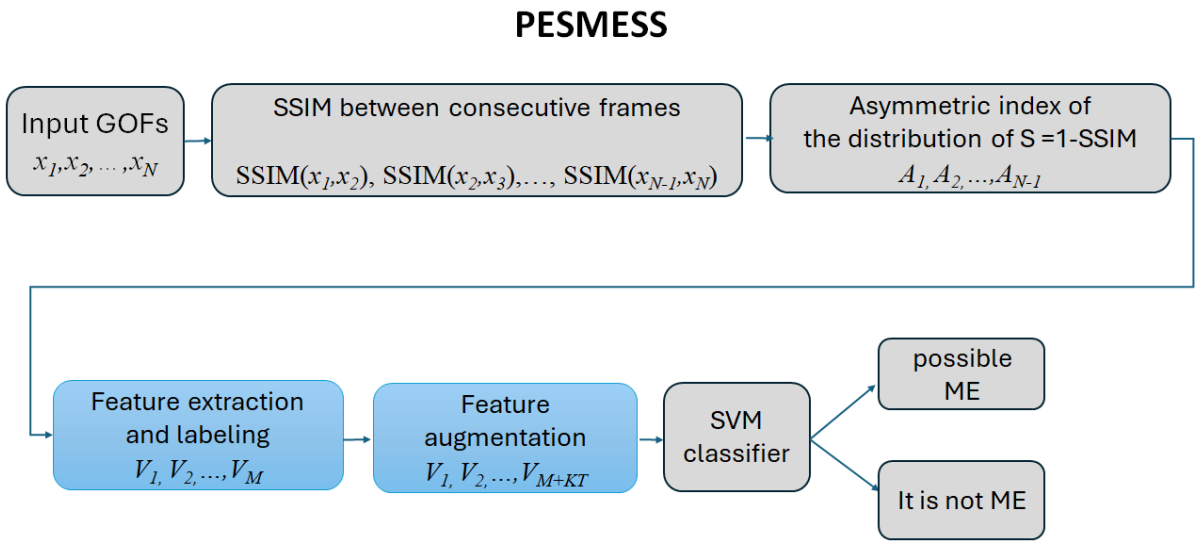
The second sub-problem represents one of the main criticisms in MEs analysis since it is mainly related to the lack of videos showing real MEs. The reason for this deficiency is basically due to the fact that MEs are produced when a speaker tries to lie or to hide his/her emotions; therefore, it is not easy to find many real examples where it happens. On the other hand, simulated MEs do not have the same authenticity of in-the-wild MEs. Based on this consideration and taking into account the relatively limited variability of PESs, the second sub-problem is reformulated as *building an appropriate training set of feature vectors*. For this purpose, three basic PESs have been selected from a video of the MEVIEW database [39], which contains in-the-wild MEs. Then a very simple data augmentation technique was employed to populate the training set. Specifically, to replicate the variability of the relationships between onset, apex and offset, noisy versions of the three basic shapes were generated based on some visual perception rules [6–8,10,55]. Perceptual noise was added to these basic shapes to produce many other plausible shapes, by setting the fourth feature equal to -1 for each new MEs.

Figure 7 shows a block scheme of the proposed PESMESS method, including details concerning the adopted procedures for feature extraction and feature augmentation. Next section provides formal details concerning the generation of perceptual noise.

3.3. *Visual MEs Augmentation*

Data Augmentation (DA) is a powerful tool often adopted in machine learning and artificial intelligence [2]. DA consists of a set of techniques oriented to increase the amount of information i.e., new data useful to reduce/eliminate overfitting when training a machine learning framework [46]. It increases examples by adding either slightly modified copies of already existing ones or new synthetic ones starting from existing data. The well-known problem is how to increase information avoiding annoying drawbacks that may invalidate the final result. The solution found in this paper is to slightly modify PES by adding a subtle visual noise (see Figure 8), thus replicating what usually happens in in-the-wild videos. This is due to the global nature of frame information where PES is influenced by causes that can be:

- **objective:** quantization noise, change of light etc.



Feature extraction (PES) and labeling

- Detect the local maxima in the sequence A and let $\{t_1, \ldots, t_M\}$
- For each local maximum at t_k detemine:
  - The value of the closest and preceding local minimum of A, i.e. $A(t_{mk})$
  - The successive value in $A(t_k+1)$
  - If t_k is the apex of a ME then set v=-1, otherwise set v=1
  - Define the feature vector composed of four components
    $V_k = \{A(t_{mk}), A(t_k), A(t_k+1), v\}$

Feature augmentation

For each V_k labeled as ME (let K the number of MEs):

- Generate a new vector as it follows:
  - Modify its second and third component by randomly generating values of the parameter γ constrained to eq. (10) and using the rule in eq. (7)
  - Add the new vector to the training set of features

Repeat previous steps T times

Figure 7. Block scheme of PESMESS (*top*). Pseudocode of the procedures for feature extraction (*middle*), and feature augmentation (*bottom*).

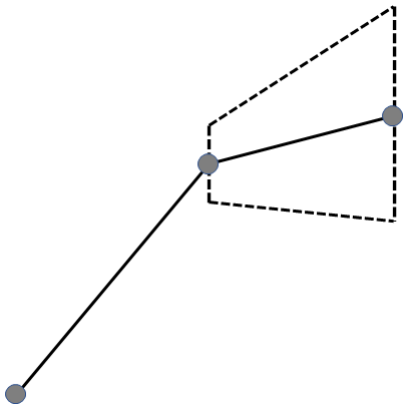


Figure 8. Sketch of the augmentation process. Initial PES (solid line with gray circles); the 2nd and 3rd sample of the new PESs fall into the dashed region.

- **subjective:** head movements, involuntary or voluntary contractions of speaker's face muscles oriented to mask his/her own emotional state etc.

It is worth outlining that, from a perceptual point of view, an observer is able to perceive MEs among all the aforementioned disturbing components, that can be all classified as disturbing noise to be accounted for in order to produce new information. In the sequel, a formalization of this scenario will be given: the proposed formalism can rigorously explain how and how much to increase information for the training set.

Based on the considerations above, if an ME occurs in the k-th frame x_k, then

$$x_k = f(x_{k-1}) + \varepsilon \tag{5}$$

where f denotes the transformation representing the pure ME, while ε represents the noise sources, that are modelled as additive and independent contributions to the analysed frame. As ME is visible, ε is such that the visual masking constraint is met [6,8,55], i.e.

$$0 \leq \frac{\sigma_\epsilon^2}{\sigma_{f(x_{k-1})}^2} \leq \rho, \tag{6}$$

where ρ is the just noticeable visibility threshold. As a result, equation (5) represents a DA rule whenever different noise realizations constrained to equation (6) are considered.

The aim of this section is then to infer this DA rule over the asymmetry index of the temporal S distribution to directly generate reliable PES profiles[a]. To this aim, let

- P_k and $\bar{P}_k$ respectively be the distribution of $S_k = 1 - SSIM_k$ and $\overline{S}_k = 1 - \overline{SSIM}_k$, where $SSIM_k = SSIM(x_{k-1,b}, x_{k,b})$ is the SSIM evaluated between two corresponding blocks in two consecutive frames, while $\overline{SSIM}_k = SSIM(x_{k-1,b}, f(x_{k-1,b}))$ is the SSIM value computed in the case of pure frames (without noise sources); and let
- A_k and $\bar{A}_k$ be the corresponding asymmetry indices, computed as in equation (3);

then, a DA rule for PES can rely on the following model

$$A_{k+1} = \bar{A}_{k+1} \pm \gamma \bar{A}_k, \tag{7}$$

where k and $k+1$ refer to two specific temporal events: ME apex (t_k) and ME offset (t_{k+1}), i.e. $\bar{\beta}_k = \frac{\bar{A}_k}{\bar{A}_{k+1}} > 1$. γ is the augmentation parameter that is constrained to noise sources meeting the masking condition in eq. (6).

[a] It is worth observing that, by definition, the asymmetry index of S has the same value of the asymmetry index of $SSIM$ but opposite sign.

This approach is twofold advantageous. First, it allows to easily generate copies of PES, that is a time dependent signal, only in the time interval where ME occurs; secondly, it is considerably faster than the augmentation model in eq. (5) that involves ME video frames.

We are now interested in determining admissible ranges or bounds for the augmentation parameter γ, in agreement with the masking condition. To this aim we take advantage of the following Proposition, whose proof is in the Appendix:

Proposition 3.1. *Let set* $SSIM_k = SSIM(x_{k-1,b}, x_{k,b})$ *and* $\overline{SSIM}_k = SSIM(x_{k-1,b}, f(x_{k-1,b}))$*, where* x_k*,* x_{k-1} *and* $f(x_{k-1})$ *are defined as in eq. (5), and let the masking condition in eq. (6) hold true, then*

$$SSIM_k = \alpha_{k,b}\, \overline{SSIM}_k,$$

where

$$\alpha_{k,b} = \frac{\sigma^2_{x_{k-1}} + \sigma^2_{f(x_{k-1})} + C_2}{\sigma^2_{x_k} + \sigma^2_{x_{k-1}} + C_2} \tag{8}$$

with

$$\alpha_{k,b} \in \Omega_\alpha = \left[\frac{1}{1+\rho}, 1\right], \tag{9}$$

and $\alpha_{k,b} = 1 \Leftrightarrow \sigma_\epsilon = 0$.

Proposition 3.1 states that a noise source in the original data decreases SSIM absolute value by introducing a multiplicative factor $\alpha_{k,b} \leq 1$ in the measure. As a result, S value increases; in addition, for each image block, the masking condition in eq. (6) provides the interval of admissible values for $\alpha_{k,b}$ — i.e., the ones corresponding to masked changes.

Based on these observations, the following Proposition can be proved.

Proposition 3.2. *Let* $\varepsilon_{A_{k+1}} = \frac{A_{k+1} - \bar{A}_{k+1}}{\bar{A}_{k+1}}$ *be the percentage error committed when using* A_{k+1} *instead of the noiseless* $\bar{A}_{k+1}$*, and let define* ε_{A_k} *accordingly. If* $\bar{\beta}_k = \frac{\bar{A}_k}{\bar{A}_{k+1}}$ *is the ratio of the asymmetry measure of noiseless* $\bar{S}$ *distribution in correspondence to ME (k) apex and offset (k + 1), with* $\bar{\beta}_k > 1$*, then*

$$|\gamma| = \frac{|\varepsilon_{A_{k+1}}|}{|\bar{\beta}_k|} \sim \frac{1}{|\bar{\beta}_k|} \frac{\rho}{2(1+\rho)} < \frac{\rho}{2(1+\rho)} \tag{10}$$

where ρ *is the masking threshold.*

The Proof is in the Appendix. Eq. (10) provides admissible values for augmentation parameter to use for directly generating feasibile realizations of PES related to visible MEs in noisy environment, subjected to the visual masking condition in eq. (6). This makes it possible to produce new PES visually equivalent to the original one, as shown in Figure 9, thus compensating for the lack of availability of authentic ones.

4. Experimental Results

The proposed PESMESS has been tested on various sequences from different databases. In this section, we have selected some examples from two publicly available datasets: CASME II [58] and MEVIEW (Micro Expression VIdEo in the Wild)) [39].

CASME II [58] is an improved version of CASME dataset [57]. It offers 247 videos, recorded using high frame-rate cameras (200 fps) and containing many spontaneous and brief MEs (usually less than .2 secs). MEVIEW [39] meets the need of more unconstrained MEs embedded in "in-the-wild" situations with more realistic high-stake scenarios. It has been built by collecting mostly poker game videos downloaded from YouTube in order to catch players when trying to conceal or fake their true emotions. In fact, poker made players very competitive and then prone to mask/hide their true emotions. With a more real

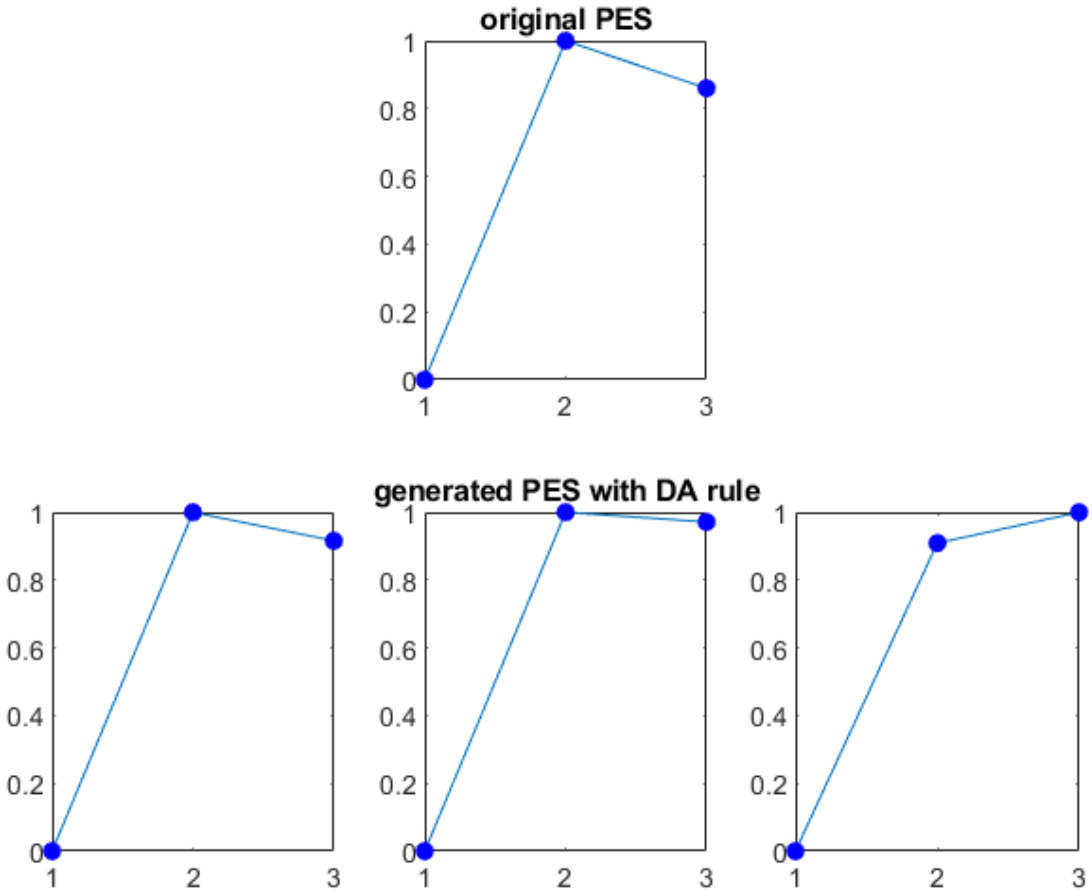


Figure 9. Original PES (*top*) and some new PESs (*bottom*) generated using the DA rule in eq. (7) using the admissible range for the augmentation parameter in eq. (10).

camera view (switching from the entire shot to a single face), 31 videos have been captured, having a 3s duration on average and referring to 16 individuals. A METT-trained annotator has also been used for labeling onset and offset of involved MEs — with successive FACS coding.

The choice of these two databases is motivated by the need to consider a broader scenario for identifying ME. Since the fundamental difference between posed and spontaneous MEs lies in the relevance between expressed facial movement and the underlying emotional state [24,42], the inclusion of databases of both posed and spontaneous MEs, or rather of in-the-wild MEs, enables a better understanding of the potential of the proposed approach. Sequences taken by different devices have not been considered as PESMESS is almost independent of them: the only requirement is a sufficient resolution in terms of pixels number and frame/rate — as is usual for any modern device.

All results have been achieved by implementing PESMESS in Matlab environment. With reference to the flowchart in Figure 7, an automatic face detector (vision.CascadeObjectDetector) has been applied to each video frame and S distribution has been computed only considering the sub-image containing the main subject face and a small part of background pixels. As shown in Figure 10, the shape of PESs for any of the considered asymmetry absolute indices (Pearson, Yule and modified Yule) remains almost unchanged and somewhat robust to external visual components, including the case of inaccurate cutting in face selection. For the sake of simplicity and without loss of generality, the following results will only refer to Yule index.

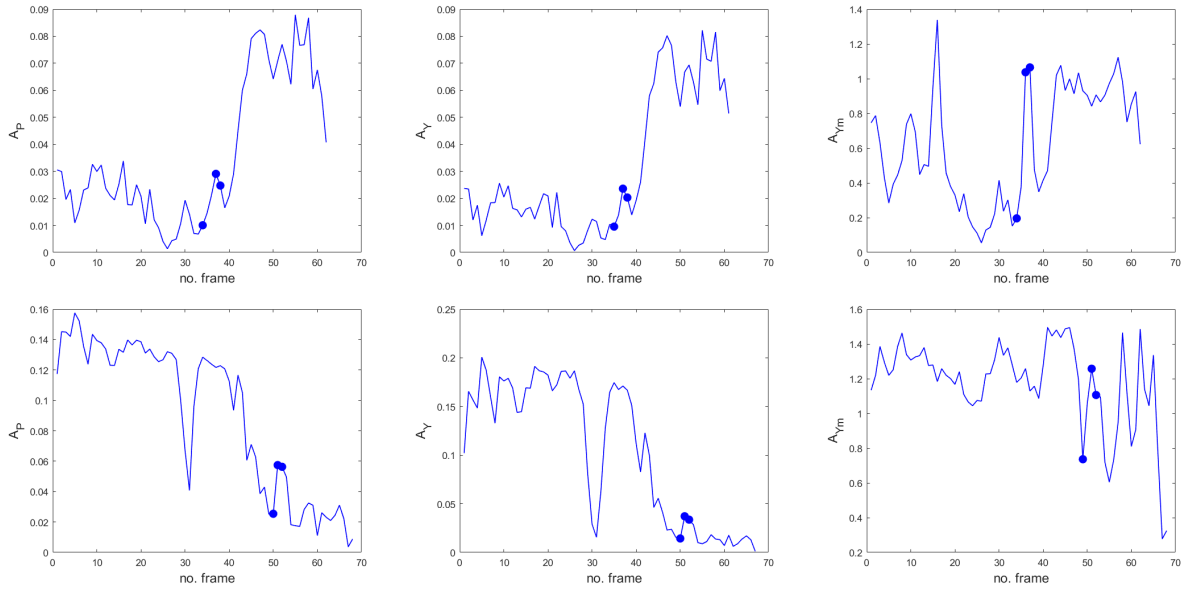


Figure 10. Examples of PES (marked samples) for the three asymmetry indexes on two videos from MEVIEW database: Sequence 14.1 (*top*) and Sequence 11.2 (*bottom*).

Regarding the population of the training set using the data augmentation procedure described in the previous section, it has been built starting from the sequence 15.1 from MEVIEW. This video contains a more evident ME located at frame no. 38 (the reference frame is the first frame of the considered ME) and two just perceivable MEs at frame no. 3 and 23 — see Figure 11. The feature vectors of these three subsequences concatenated with -1 (the sign of the decreasing derivative) represent the group of positive examples in the training set. The negative ones are all triplets from sequence 15.1 except for the ones in 3, 23 and 38. Even in the latter case the values $\{-1, 0, 1\}$ have been added to the feature vector of the negative examples. All subsequences have been rescaled after the subtraction of the minimum in the range [0 1]. Data augmentation has been then adopted to increase especially the positive examples. For each of the three positive cases, 33 additional examples have been produced by means of a 5% additive white noise (uniform distribution) and rescaling them in the interval [0 1]. It is worth observing that this tuning is in agreement with the theoretical constraint in eq. (10). In agreement with [6,8,14,55], by setting the masking threshold equal to the value $\rho = 0.33$[b], we have $|\gamma| < \frac{0.33}{2(1+0.33)} = 0.1241 = 12\%$ that approximately corresponds to percentage of noise in the range $[\frac{\gamma}{3}, \frac{\gamma}{2}] = [4, 6]$. It is worth outlining that the noise has been only added to the 2nd and 3rd sample of the feature vector. This choice is motivated by the fact that the first sample has a much lower amplitude than the other two. It turns out that the relation between samples 1-2 and 1-3 does not change. In addition, after rescaling the first sample is equal to 0. The 4th sample of the feature vector does not change as it represents the derivative sign. The final set of positive examples is then composed of 102 elements. With regard to the negative cases, the same process has been applied, leading to additional 102 elements. This training set has been then used for defining the binary SVM-based classification model.

The following tests demonstrate that, using the learnt model, spotting is possible even when ME is not the only relevant event in a group of frames. In this sense, videos from MEVIEW are the best suited to evaluate the robustness of PESMESS. In addition, despite its simplicity, PESMESS can recognize almost any ME in a video, which may be even more than those labeled in the available databases.

Table 2 contains some detection results for MEVIEW database, while Table 3 refers to CASME II. The tables report the position (apex) and the analysed frame area for each PES. Table 2 clearly shows the high sensitivity of PESMESS in detecting any strange reaction of the subject of the video under

[b] $\rho = 0.33$ represents a good trade off able to take into account for contrast masking effect, which is responsible for missing/hiding information in complex backgrounds.

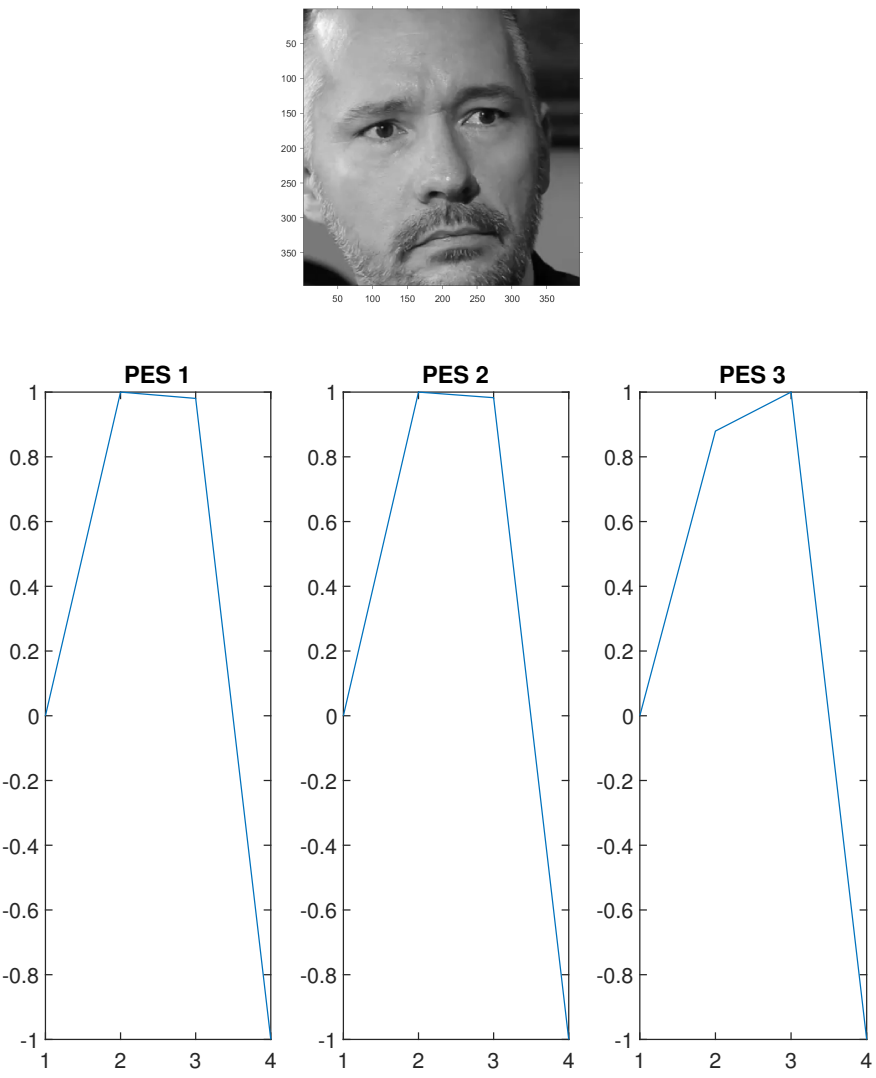


Figure 11. Sequence 15.1 from MEVIEW. *Top*) A frame of sequence 15.1. *Bottom*) The three selected PESs.

analysis. All MEs contained in the video sequence are detected: the most evident and annotated ME as well as other less evident MEs. This is not surprising because MEs definition is not completely coded. On the other hand, it is also possible to have some false alarms, as in sequence 1.1, mainly due to video zooming or stabilization, but eye blinking is never detected. Similar comments can be made for the results presented in Table 3. It is worth observing that in this case the videos have higher resolution (200 fps); hence, to be consistent with the input required by the trained model, the video sequences have been sampled using a sample step equal to 8. This subsampling allows us to have PES waveforms consistent with those estimated from MEVIEW dataset. In this case, it is interesting to note that a single ME can have multiple detection — usually no more than 2 PESs for each ME. This allows us to conclude that PESMESS is capable of capturing non-uniform velocities of the facial muscles during the ME generation process and that they can only be detected when using high-resolution (higher frame-rate) cameras.

Finally, PESMESS also benefits from the need for very simple operations, making it an optimal candidate for speeding up the ME detection and recognition processes. To reinforce this property, some speed-up methods [11] for SSIM computation can also be used in the future.

Table 2. MEVIEW dataset: for each sequence, the location and the relative ME description are reported.

		Sequence: 15.1		
PES no. 1	***Location***	3	***Description***	looks up
PES no. 2	***Location***	23	***Description***	open eyes
PES no. 3	***Location***	38	***Description***	ME (surprise)
		Sequence: 14.3		
PES no. 1	***Location***	5	***Description***	imperceptible smile
PES no. 2	***Location***	19	***Description***	ME (joy)
		Sequence: 1.1		
PES no. 1	***Location***	12	***Description***	zoom/stabilization
PES no. 2	***Location***	35	***Description***	zoom/stabilization
PES no. 3	***Location***	48	***Description***	ME begin
PES no. 4	***Location***	54	***Description***	ME reinforcement
PES no. 5	***Location***	64	***Description***	ME end
PES no. 6	***Location***	73	***Description***	mouth micro movement
		Sequence: 2.1		
PES no. 1	***Location***	29	***Description***	look down
PES no. 2	***Location***	81	***Description***	frozen eyes
PES no. 3	***Location***	99	***Description***	ME
		Sequence: 3.1		
PES no. 1	***Location***	27	***Description***	looks up
PES no. 2	***Location***	37	***Description***	frozen eyes
PES no. 3	***Location***	70	***Description***	mouth movement
PES no. 4	***Location***	78	***Description***	ME
		Sequence: 6.1		
PES no. 1	***Location***	3	***Description***	keep looking down
PES no. 2	***Location***	15	***Description***	ME
PES no. 3	***Location***	34	***Description***	eyes movement
PES no. 4	***Location***	47	***Description***	eyes and eyebrows
		Sequence: 7.9		
PES no. 1	***Location***	13	***Description***	mouth movement
PES no. 2	***Location***	27	***Description***	tongue movement
PES no. 3	***Location***	31	***Description***	tongue movement
PES no. 4	***Location***	46	***Description***	head movement
PES no. 5	***Location***	50	***Description***	head (FA)
PES no. 6	***Location***	60	***Description***	eyes movement
PES no. 7	***Location***	65	***Description***	tongue movement
PES no. 7	***Location***	65	***Description***	ME
		Sequence: 8.1		
PES no. 1	***Location***	13	***Description***	ME
PES no. 2	***Location***	22	***Description***	mouth (ME end)
PES no. 3	***Location***	32	***Description***	head movement
PES no. 4	***Location***	42	***Description***	eyes movement
PES no. 4	***Location***	46	***Description***	head movement
PES no. 5	***Location***	49	***Description***	head movement
PES no. 6	***Location***	56	***Description***	head movement
		Sequence: 8.2		
PES no. 1	***Location***	19	***Description***	ME
PES no. 2	***Location***	37	***Description***	head movement
PES no. 3	***Location***	43	***Description***	head and eyes movement

5. Concluding Remarks

In this paper a novel approach, namely PESMESS, to detect sub-sequences from videos containing MEs has been proposed. Based on simple operations that try to simulate the sixth sense of an observer when detecting MEs during a conversation, a perceptual fingerprint for MEs has been defined exploiting some peculiar features of the distribution of the Structural SIMilarity index computed over pairs of consecutive frames. To compensate the lack of annotated datasets for authentic MEs captured in in-the-

Table 3. CASME II dataset: for each sequence, the location and the relative ME description are reported.

Sequence: 3.2				
PES no. 1	***Location***	10	***Description***	head motion
PES no. 2	***Location***	16	***Description***	ME
Sequence: 4.2				
PES no. 1	***Location***	13	***Description***	ME
Sequence: 13.2				
PES no. 1	***Location***	9	***Description***	ME
PES no. 2	***Location***	12	***Description***	ME (continues)
Sequence: 13.4				
PES no. 1	***Location***	4	***Description***	ME
PES no. 2	***Location***	12	***Description***	ME (end)

wild videos, an augmentation procedure of the MEs perceptual fingerprint has been proposed. It relies on a guided perturbation of the perceptual fingerprint where bounds for the perturbation are rigorously derived using proper estimation error procedures and in accordance with visual masking mechanisms. Experimental results have demonstrated that the perception-guided augmentation procedure allows for the construction of a representative datatset which can be used to properly and effectively train a standard SVM classifier. PESMESS is indeed able to identify MEs in both real and simulated videos. The ability of PESMESS in detecting very short video clips containing MEs with high probability makes it a desirable pre-processing procedure to more sophisticated approaches which can validate the detected MEs [42]. This is advantageous especially in long videos analysis [34]. Finally, it is worth highlighting that eye blinking is never confused with MEs, while all 'emotional micro-expressions' are correctly detected by PESMESS. This can be a good starting point for further theoretical studies in which the definition of ME, which is actually only descriptive and not completely well defined, can be replaced by a formal one in terms of a mathematical formalism. This aspect, as well more practical issues, will be the subject of our future research.

Acknowledgments

This work has been partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU. The Authors are members of SIMAI (Societá Italiana di Matematica Applicata e Industriale), RITA (Research ITalian network on Approximation), the Italian national research group GNCS (INdAM) and UMI research groups TAA (Approximation Theory and Applications) and MAT AI&ML (Mathematics for Artificial Intelligence and Machine Learning).

Appendix

Proof of Proposition 3.1

From eq. (5) we have

$$\mu_{x_k} = \mu_{f(x_{k-1})} + \mu_\epsilon, \quad \sigma^2_{x_k} = \sigma^2_{f(x_{k-1})} + \sigma^2_\epsilon, \quad \text{and} \quad \sigma_{x_{k-1}x_k} = \sigma_{x_{k-1}f(x_{k-1})}.$$

Hence, by multiplying $SSIM(x_{k-1}, x_k)$ by $\frac{\sigma^2_{x_{k-1}}+\sigma^2_{f(x_{k-1})}+C_2}{\sigma^2_{x_{k-1}}+\sigma^2_{f(x_{k-1})}+C_2}$ it holds

$$SSIM(x_{k-1}, x_k) = SSIM(x_{k-1}, f(x_{k-1})) \underbrace{\frac{\sigma^2_{x_{k-1}} + \sigma^2_{f(x_{k-1})} + C_2}{\sigma^2_{x_{k-1}} + \sigma^2_{x_k} + C_2}}_{\alpha_{k,b}}.$$

Since $\sigma^2_{x_k} = \sigma^2_{f(x_{k-1})} + \sigma^2_\epsilon$ and $\sigma^2_* \geq 0$, we have $\alpha_{k,b} \leq 1$.

In addition, using the masking condition in eq. (6), we have

$$\alpha_{k,b} = \frac{1}{1+\frac{\sigma_\varepsilon^2}{\sigma_{x_{k-1}}^2+\sigma_{f(x_{k-1})}^2+C2}} \geq \frac{1}{1+\frac{\rho}{1+C_2+\frac{\sigma_{x_{k-1}}^2}{\sigma_{f(x_{k-1})}^2}}} \geq \frac{1}{1+\rho},$$

and then $\alpha_{k,b} \in \left[\frac{1}{1+\rho}, 1\right]$.

Proof of Proposition 3.2

By isolating γ in eq. (7) we have

$$\pm\gamma = \frac{A_{k+1}-\bar{A}_{k+1}}{\bar{A}_k} = \frac{A_{k+1}-\bar{A}_{k+1}}{\bar{A}_{k+1}}\frac{\bar{A}_{k+1}}{\bar{A}_k}.$$

Since $\varepsilon_{A_{k+1}} = \frac{A_{k+1}-\bar{A}_{k+1}}{\bar{A}_{k+1}}$ and $\bar{\beta}_k = \frac{\bar{A}_k}{\bar{A}_{k+1}}$, previous equation becomes $\pm\gamma = \frac{\varepsilon_{A_{k+1}}}{\bar{\beta}_k}$, and the leftmost equality in eq. (10) holds true.

For the estimation of $\varepsilon_{A_{k+1}}$, we observe that, as $S_k = 1 - SSIM_k$, by definition

$$A_{k+1} = -(M_{k+1} - 2\,Med_{k+1} + m_{k+1})$$

with

$$M_{k+1} = \max_b(SSIM_{k+1}), \quad Med_{k+1} = Q_2(SSIM_{k+1}), \quad m_{k+1} = \min_b(SSIM_{k+1}),$$

then

$$\varepsilon_{A_{k+1}} = -\left(\frac{M_{k+1}}{A_{k+1}}\varepsilon_{M_{k+1}} - 2\frac{Med_{k+1}}{A_{k+1}}\varepsilon_{Med_{k+1}} + \frac{m_{k+1}}{A_{k+1}}\varepsilon_{m_{k+1}}\right). \tag{11}$$

In addition, since

$$\varepsilon_{SSIM_k} = \frac{SSIM_k - \overline{SSIM}_k}{\overline{SSIM}_k} = \alpha_{k,b} - 1.$$

then

$$1 - \max_b \alpha_{k+1,b} \leq -\varepsilon_{M_{k+1}} \leq 1 - \min_b \alpha_{k+1,b} \qquad \Rightarrow \qquad -\varepsilon_{M_{k+1}} \sim 1 - \tilde{\alpha}_{k+1},$$

with $\tilde{\alpha}_{k+1} = \left(\frac{\max_b \alpha_{k+1,b} + \min_b \alpha_{k+1,b}}{2}\right)$.

The same arguments are valid for Med_{k+1} and m_{k+1}. As a result, eq. (11) can be rewritten as it follows

$$\varepsilon_{A_{k+1}} \sim 1 - \tilde{\alpha}_{k+1},$$

and eq. (7) becomes

$$|\gamma| \sim \frac{|1-\tilde{\alpha}_{k+1}|}{|\bar{\beta}_k|}.$$

Using the last result of Proposition 3.1, we have $\tilde{\alpha}_{k+1} = \left(1 + \frac{1}{1+\rho}\right)\frac{1}{2} = \frac{2+\rho}{2(1+\rho)}$. Since $\bar{\beta}_k > 1$, we get the last inequality in eq. (10), i.e. $|\gamma| \leq \frac{\rho}{2(1+\rho)}$.

References

1. H. E. Adelson, and J. R. Bergen, Spatiotemporal energy models for the perception of motion, *Journal of the Optical Society of America A* vol. 2, no. 2, 1985.
2. C.C. Aggarwal, *Neural Networks and Deep Learning*, A Textbook, Springer 2018.

3. A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, Robust discriminative response map fitting with constrained local models, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3444-3451, 2013.

4. D. Brunet, E.R. Vrscay, and Z. Wang, On the mathematical properties of the structural similarity index, *IEEE Transactions on Image Processing*, vol. 21, no. 4, 2012.

5. V. Bruni, G. Ramponi, A. Restrepo, and D. Vitulano, Context-Based Defading of Archive Photographs, *Journal of Image and Video Processing*, 2009.

6. V. Bruni, E. Rossi, and D. Vitulano, On the Equivalence Between Jensen-Shannon Divergence and Michelson Contrast, *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4278-4288, 2012.

7. V. Bruni, D. Vitulano, and Z. Wang, Special issue on human vision and information theory, *Signal, Image and Video Processing*, vol. 7, no.3,pp. 389-390, 2013.

8. V. Bruni, D. De Canditiis, and D. Vitulano, Speed up of Video Enhancement based on Human Perception, *Signal Image and Video Processing*, vol. 8, pp. 1199-1209, 2014.

9. V. Bruni, D. Panella, and D. Vitulano, Non local means image denoising using noise-adaptive SSIM, *Proceedings of the 23rd European Signal Processing Conference, EUSIPCO*, 2015.

10. V. Bruni, and D. Vitulano, Jensen shannon divergence as reduced reference measure for image denoising, *Lecture Notes in Computer Science*, vol. 10016, 2016.

11. V. Bruni, and D. Vitulano, An entropy based approach for SSIM speed up, *Signal Processing*, vol. 135, pp. 198-209, 2017.

12. V. Bruni, and D. Vitulano, SSIM Based Signature of Facial Micro-Expressions, *Proceedings of International Conference in Image Analysis and Recognition (ICIAR 2020)*, Lecture Notes in Computer Science, vol. 12131, 2020.

13. V. Bruni, and D. Vitulano, A Fast Preprocessing Method for Micro-Expression Spotting via Perceptual Detection of Frozen Frames, *Journal of Imaging*, MDPI, vol. 7, no. 4, 2021.

14. F.W. Campbell, and J.G. Robson, Application of Fourier analysis to the visibility of gratings, *Journal of Physiology*, vol. 197, no. 3, pp. 551-566, 1968.

15. D. Cristinacce, T. F. Cootes, et al., Feature detection and tracking with constrained local models, *Proceedings of the British Machine Vision Conference 2006*, Edinburgh, UK, vol. 1, 2006.

16. C. Duque, O. Alata, R. Emonet, A.C. Legrand, and H. Konik, Micro-expression spotting using the Riesz pyramid, *Proceedings of WACV*, Lake Tahoe, 2018.

17. P. Ekman, and M.V. Friesen, Nonverbal leakage and clues to deception, *Psychiatry*, vol. 32, pp. 88-106, 1969.

18. P. Ekman, Lie catching and microexpressions. The Philosophy of Deception, ed C. Martin (Oxford University Press), pp. 118-133, 2009.

19. P. Ekman, Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage, WW Norton & Company, 2009.

20. V. Esmaeili, M. Mohassel Feghhi, and S.O. Shahdi, A comprehensive survey on facial micro-expression: approaches and databases, *Multimedia Tools and Applications*, 2022.

21. W. Gong, and N.M. Elfiky, Deep learning-based microexpression recognition: a survey, *Neural Computing and Applications*, 2022.

22. E.A. Haggard, and K.S. Isaacs, Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy, *Methods of Research in Psychotherapy*. L. A. Gottschalk and H. Auerbach (Boston, MA: Springer), pp. 154-165, 1966

23. Y. He, S.J. Wang, J. Li, and M. H. Yap, Spotting macro-and micro-expression intervals in long video sequences, *Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture*

Recognition (FG 2020), pp. 742-748, 2020.

24. U. Hess, and R.E. Kleck, Differentiating emotion elicited and deliberate emotional facial expressions, *European Journal of Social Psychology*, vol. 20, pp. 369-385, 1990.
25. M. Kendall, and A. Stuart, The Advanced Theory of Statistics, Chareles Griffinn & Company Limited, 1976.
26. D. E. King, Dlib-ml, A machine learning toolkit, *The Journal of Machine Learning Research*, vol. 10, pp. 1755-1758, 2009.
27. L. Jingting, S.J. Wang, M. H. Yap, J. See, X. Hong, and X. Li, Megc2020-the third facial micro-expression grand challenge, *Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 777-780, 2020.
28. Y. Li, X. Huang, and G. Zhao, Can micro-expression be recognized based on single apex frame?, *Proceedings of the 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3094-3098, 2018.
29. J. Li, C. Soladie, and R. Seguier, Ltp-ml, Micro-expression detection by recognition of local temporal pattern of facial movements, *Proceedings of the 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 634-641, 2018.
30. J. Li, C. Soladie, and R. Seguier, Local temporal pattern and data augmentation for micro-expression spotting, *IEEE Transactions on Affective Computing*, 2020.
31. S.T. Liong, J. See, K. Wong, A.C. Le Ngo, Y.H. Oh, and R. Phan, Automatic apex frame spotting in micro-expression database, *Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 665-669, 2015.
32. S.T. Liong, J. See, K. Wong, and R. C.W. Phan, Automatic microexpression recognition from long video using a single spotted apex, *Proceedings of the Asian Conference on Computer Vision*, Springer, pp. 345-360, 2016.
33. G.B. Liong, J. See, and L.K. Wong, Shallow optical flow three- stream cnn for macro-and micro-expression spotting from long videos, *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, USA, pp. 2643-2647, 2021.
34. G.B. Liong, J. See, and C.S. Chan, Spot-then-recognize: A Micro-Expression Analysis Network for seamless evaluation of long videos, *Signal Processing: Image Communication*, vol. 110, 2023.
35. H. Ma, G. An, S. Wu, and F. Yang, A region histogram of oriented optical flow (RHOOF) feature for apex frame spotting in micro-expression, *Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 281–286, 2017.
36. V. Mante, R. A. Frazor, V. Bonin, W. S. Geisler, and M. Carandini, Independence of luminance and contrast in natural scenes and in the early visual system, *Nature Neuroscience*, vol.8, pp. 1690-1697, 2005.
37. I. Megvii, Face++ research toolkit, 2013.
38. A. Mehrabian, Nonverbal Communication, Publisher, ALDINE-ATHERTON, 1972 (eBook Published, 31 October 2017).
39. MEVIEW Homepage, urlhttp://cmp.felk.cvut.cz/ cechj/ME/.
40. S. Milborrow, and F. Nicolls, Active shape models with sift descriptors and mars, *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP)*, vol.2, pp. 380-387, 2014.
41. A. Moilanen, G. Zhao, and M. Pietikainen, Spotting rapid facial movements from videos using appearance-based feature difference analysis, *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR)*, pp. 1722–1727, 2014.

42. Y.H. Oh, J. See, A. C. Le Ngo, R.C. Phan, and V.M. Baskaran, A Survey of Automatic Facial Micro-Expression Analysis: Databases, Methods, and Challenges, *Frontiers in Psychology*, 2018.

43. S. Polikovsky, and Y. Kameda, Facial micro-expression detection in hi-speed video based on facial action coding system (facs), *IEICE Transactions on Information and Systems*, vol. 9, pp. 81-92, 2013.

44. S. Porter, and L. Ten Brinke, Reading between the lies identifying concealed and falsified emotions in universal facial expressions, *Psychological Science*, vol. 19, pp. 508-514, 2008.

45. F. Qu, S.J. Wang, W.J. Yan, H. Li, S. Wu, and X. Fu, Cas(me)2: a database for spontaneous macro-expression and micro-expression spotting and recognition, *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 424-436, 2017.

46. C. Shorten, and T.M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, *Journal of Big Data*, vol. 6, no. 60, 2019.

47. M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, Macro-and micro-expression spotting in long videos using spatio-temporal strain, *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pp. 51-56, 2011.

48. M. Shreve, J. Brizzi, S. Felatyev, T. Luguev, D. Goldgof, and S. Sarkar, Automatic expression spotting in videos, *Image and Vision Computing*, vol. 32, no. 8, pp. 476-486, 2014.

49. M.F. Valstar, and M. Pantic, Fully automatic recognition of the temporal phases of facial actions, *IEEE Transactions on Systems Man and Cybernetics Part B*, vol.42, pp. 28-43, 2012.

50. M. Verburg, and V. Menkovski, Micro-expression detection in long videos using optical flow and recurrent neural networks, *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1-6, 2019.

51. Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, Image Quality Assessment: From Error Visibility to Structural Similarity, *IEEE Transactions on Image Processing*, vol. 13, pp. 600-612, 2004.

52. 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*, vol. 230, pp. 382-389, 2016.

53. S.J. Wang, Y. He, J. Li, and X. Fu, Mesnet: A convolutional neural network for spotting multi-scale micro-expression intervals in long videos, *IEEE Transactions on Image Processing*, vol. 30, pp. 3956-3969, 2021.

54. S. Weinberger, Airport security: intent to deceive?, *Nature*, vol. 465, pp. 412-415, 2010.

55. S. Winkler, Digital Video Quality - Vision Models and Metrics, John Wiley & Sons, Ltd, 2005.

56. W.J. Yan, Q. Wu, J. Liang, Y.H. Chen, and X. Fu, How fast are the leaked facial expressions: the duration of micro-expressions, *Journal of Nonverbal Behavior*, vol. 37, pp. 217-230, 2013.

57. W.J. Yan, Q. Wu, Y.J. Liu, S.J. Wang, and X. Fu, Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces, *Proceedinds of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1-7, 2013.

58. W.J. Yan, X. Li, S.J. Wang, G. Zhao, Y.J. Liu, Y.H. Chen, et al., CASME II: an improved spontaneous micro-expression database and the baseline evaluation, *PLoS ONE*, vol. 9, 2014.

59. W.J.Yan, and Y.H Chen, Measuring dynamic micro-expressions via feature extraction methods, *Journal of Computational Science*, vol. 25, pp. 318-326, 2017.

60. C.H. Yap, C. Kendrick, and M.H. Yap, Samm long videos: A spontaneous facial micro-and macro-expressions dataset, *Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 771-776, 2020.

61. Z. Zhang, T. Chen, H. Meng, G. Liu, and X. Fu, Smeconvnet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos, *IEEE Access*, vol. 6, pp. 71143-71151, 2018.

62. H. Zhang, L. Yin, and H. Zhang, A review of micro-expression spotting: methods and challenges, *Multimedia Systems*, vol. 29, pp. 1897-1915, 2023.