22^{nd} *Computer Vision Winter Workshop*
Nicole M. Artner, Ines Janusch, Walter G. Kropatsch (eds.)
Retz, Austria, February 6–8, 2017

Spotting Facial Micro-Expressions "In the Wild"

Petr Husák, Jan Čech, Jiří Matas
Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University in Prague
{husakpe1,cechj,matas}@cmp.felk.cvut.cz

Abstract. *Micro-expressions are quick facial motions, appearing in high stake and stressful situations typically when a subject tries to hide his or her emotions. Two attributes are present - fast duration and low intensity. A simple detection method is proposed, which determines instants of micro-expressions in a video. The method is based on analyzing image intensity differences over a registered face sequence. The specific pattern is detected by an SVM classifier. The results are evaluated on standard micro-expression datasets SMIC-E and CASMEII. The proposed method outperformed competing methods in detection accuracy. Further, we collected a new real micro-expression dataset of mostly poker game videos downloaded from YouTube. We achieved average cross-validation AUC 0.88 for the SMIC, and 0.81 on the new challenging "in the Wild" database.*

1. Introduction

Micro-expressions (MEs) are defined as very brief, involuntary facial expressions. MEs tend to occur in high stake and stressful situations, when something valuable can be gained or lost and the emotions are either deliberately or unconsciously concealed [2, 5, 6]. Thus the MEs are a promising cue in catching a liar and their detection could be important for police inquiring or psychological examinations. The principle is that, even if a subject attempts to conceal unwanted emotions, a very subtle facial expression often leaks [2].

Various reasons exist to hide emotions. The context is always fundamental. Causes might be cultural conventions, social behavior or even deception. People learn to control their emotions from childhood, e.g. smile when it is appropriate, do not express anger in public, or fake an emotion to trick an opponent [8]. Ekman studied MEs in the context of lies and showed MEs are promising in detecting deception [5]. Compared to a polygraph, revealing a liar with a camera is not invasive and could be applied even without awareness of the subject. However MEs do not provide proof of lying. MEs only reflect the current emotional state.

Recognizing MEs can be helpful in many disciplines. Teachers could teach their students more efficiently. Business courses are promoted to improve the communication skills by reading facial expressions and thus increasing the emotional intelligence.

An experimental study [24] discussing the duration of MEs considered the maximum time internal of MEa less than 500ms and showed that the mean value was 314ms. According to psychologists, the ME duration lasts between 1/25s and 1/3s [5, 14]. Due to the short duration and often insignificant changes on the face surface, it is challenging to notice MEs for humans. Haggard and Isaacs, the first ME discoverers, claimed that for people it is not possible to spot MEs in real time [11]. Later, Ekman published a study [2] showing the ability can be trained and improved by practicing. The technique is developed and reported in[3, 14].

MEs, similarly as common facial expressions, can be divided into six basics groups [4] - anger, disgust, fear, joy, sadness and surprise. Each of them has a specific appearance in the face which can be described by Facial Action Coding System (FACS) [7]. The FACS describes 64 facial muscle motions by Action Units (AUs). According to evolutionary psychology, emotions and in particular facial expressions have a physiological meaning to prepare the body for upcoming events [8]. For instance, the eyebrows are lowered and the sight is fixed to the source of anger to prepare the body to launch an attack, or, the eyes are wide open in fear to increase the amount of light, which should prepare the body for a possible self defense.

In computer vision, ME research is mainly focusing on the isolated recognition problem [10, 16, 20, 21]. The recognition refers to a classification of the emotion type from a pre-segmented video sequence of the ME. While in this paper, we are interested in detection of MEs in longer real videos.

Currently, several spontaneous databases with MEs are available - CASME [23, 25], SMIC [13] and recently new SAMM [1]. They all were recorded in laboratory conditions. The expressions are real, but induced artificially. The subjects were recorded in the frontal head pose while watching emotional videos being asked to keep as much neutral expression as possible. Moreover, to simulate the stress factor, a reward or a punishment followed, filling a boring form in case of apparent failure of suppressing the emotions. However, the databases are mainly designed for ME recognition and the ME intervals are clipped tightly. Two databases can be used for ME detection. CASMEII contains slightly larger margin around the ME and SMIC was later extended (SMIC - E) and contains longer video content before and after ME occurrences. SAMM database offers high quality videos of ME, however containing the ME events for isolated recognition only.

In this paper, we propose a simple method for ME detection with promising results. The algorithm is based on detecting characteristic image intensity changes from registered face over time that occur due to MEs. The facial regions are registered by using facial landmarks around facial features automatically detected in every frame. An SVM classifier is trained to detect MEs from the intensity signal. To the best of our knowledge, we achieved the highest reported accuracy in the detection task on the SMIC and CASMEII databases.

Unlike existing methods, tested on laboratory videos, we collected a dataset of more realistic uncontrolled videos. To demonstrate the potential of the proposed method, we present a dataset of challenging “in-the-wild” videos downloaded from the Internet. We call the dataset MEVIEW (Micro-Expression VIdEos in the Wild). The content is captured mostly from poker games and TV interviews.

2. Related work

Most of the related work focuses on the recognition problem; not many papers deal with the detection of MEs, although a fast detection is necessary in real situations due to rare occurrences of MEs. Especially in real videos, it is challenging to distinguish brief facial motions from neutral faces and to avoid false alarms caused by global facial movements, speaking, occlusions, etc.

Polikovski et al. [17, 18] first described the problem in the Computer Vision area investigating both the detection and recognition problem using 3D histograms of gradients. However, the results were shown only on a small database which has never been published and the MEs were not real but posed by non-professionals volunteers.

Shreve et al. [19] performed another research on the ME detection problem using optical flow and estimating the spatio-temporal strain. However the results are not comparable since were evaluated on a private database without publishing it.

Moilanen et al. [15] proposed a method based on feature difference analysis. Features were obtained dividing the face into a 6×6 grid and each box described by LBP. They reported results on the SMIC and the CASME database.

Further, Li et al. [12] proposed a system for automatic ME analysis consisting of both spotting and recognition (MESR).

Compared to previous works, the proposed detector is designed to be operable on uncontrolled “in-the-wild” videos.

3. Proposed method

We propose a method which spots facial MEs from videos captured by a standard camera, i.e. at a frame rate 25fps. A realistic, non-laboratory, setup is expected. A subject is not cooperating, but he or she moves freely, the viewpoint of the camera is not always frontal and stable and the illumination of the scene may change.

Our algorithm is based on the assumption that (usually a tiny) muscle contraction during the ME results in a sudden measurable change of intensity in a particular region inside the face image. Two effects are possible: (1) texture changes, e.g. wrinkles may appear, or (2) surface normal changes, when a larger textureless region is moved. Either of the effects or both of them are present. Since the magnitude is small, and the phenomenon can easily be confused with a global head/camera motion, speaking, eye-blinking, or presenting a normal (controlled) expression, we proposed the algorithm that first reg-

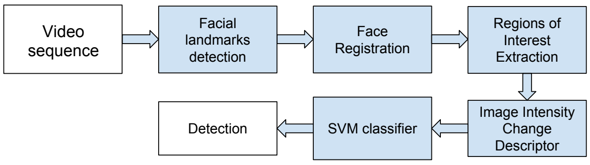
Video sequence
Facial landmarks detection
Face Registration
Regions of Interest Extraction
Image Intensity Change Descriptor
SVM classifier
Detection

Figure 1. A flowchart of the proposed automatic ME detection algorithm.

ister the face to undo the global motion, and then use a classifier to detect typical patterns of the ME from registered intensity difference signal over time.

The pipeline is depicted in Fig. 1. The face is first found and landmarks are detected in the video sequence. Then the face image is warped into a canonical coordinate system and split into regions of interest describing the facial parts, where a motion caused by MEs is expected. Each ROI measures the image intensity change within a temporal sliding window. To distinguish the ME from false intensity changes caused by other motions and illumination changes, the SVM classifier was trained.

3.1. Face rectification

To remove the global motion, a transformation into a canonical coordinate system (of the generic frontal landmark configuration) is found and the face image is warped accordingly. Facial regions are thus registered, the pixels are as much as possible corresponding over time.

3.1.1 Landmarks

Intraface [22], the robust facial landmark detector, was used to detect the face and 49 facial points in every video sequence frame t

$$\mathbf{x}^t = [(x_1^t, y_1^t), (x_2^t, y_2^t), \dots (x_{49}^t, y_{49}^t)]. \qquad (1)$$

This gives a standard set of landmarks defined on a contour of the eyes, eyebrows, nose and mouth.

3.1.2 Transformation

Every face image in time t is transformed by similarity into the generic canonical shape model

$$\bar{\mathbf{x}} = [(\bar{x}_1, \bar{y}_1), (\bar{x}_2, \bar{y}_2), \dots (\bar{x}_{49}, \bar{y}_{49})], \qquad (2)$$

which is an average shape of the landmarks. The generic shape is distributed with the Intraface detector.

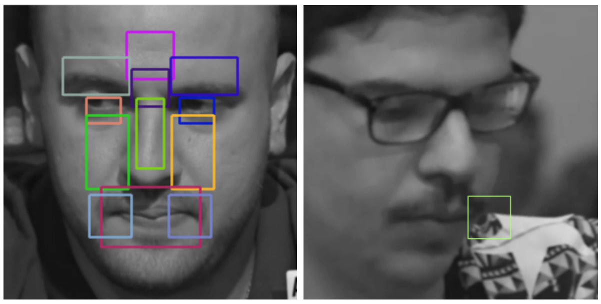
(a) (b)

Figure 2. Rectified face with twelve regions of interests depicted. Regions of Interests are designed to envelope important facial muscles to determine the location of the ME (a). The occlusion caused by a head yaw. The ROI does not envelope the facial region and is excluded from the classification (b).

The similarity transformation is given by scale s, rotation angle φ and translation vector $\mathbf{x}_0$

$$\mathcal{S}(\mathbf{x}; s, \varphi, \mathbf{x}_0) = \begin{pmatrix} s\cos\varphi & -\sin\varphi \\ \sin\varphi & s\cos\varphi \end{pmatrix} \mathbf{x} + \mathbf{x}_0. \qquad (3)$$

The problem is formulated in the least squares sense over all landmark points

$$\mathcal{S}^* = \arg\min_{s,\varphi,\mathbf{x}_0} \sum_{i=1}^{49} ||\mathcal{S}(\mathbf{x}_i; s, \varphi, \mathbf{x}_0) - \bar{\mathbf{x}}_i||_2^2. \qquad (4)$$

Minimizing the objective is solved by Procrustes analysis [9]. Finally, the image is converged into grayscale, smoothed by Gaussian filter with $\sigma = 1$ (to remove high-frequency noise of the camera) and warped by the estimated transformation $\mathcal{S}^*$ and cropped into 300×300 pixels image.

Note that the generic mean shape model may differ from the person specific landmark configuration. The shapes cannot be fully registered by the similarity transformation, that captures only in-the-plane head/camera rotation, translation motion and camera zooming. Nevertheless, this is not a problem, since due to a fast nature of the ME, we compare intensities between nearby frames. Their landmark configurations are not very different. Despite the rectified image is not perfectly frontal for off-the-plane rotations, nearby frames are transformed similarly. More complex transformation, e.g. using a piece-wise planar 3D model, or elastic registration might result in artifacts caused by landmark fluctuations. However, small landmark estimation errors are filtered out by the simple transformation of four parameters.

3.2. Regions of Interest

The structure of facial muscles produces various facial expressions. It is observed that certain emotions always trigger a subset of muscles and these motions are described by facial action units [7]. Polikovski et al. [18] had the idea dividing the face into regions. As the occurrence of MEs is local, i.e. region-dependent, and manifests only in a small part of the face.

Following [18], twelve Regions of Interests (ROIs) were defined from the landmark positions on the face, see Fig. 2a. Each ROI envelopes a group of muscles, their contraction causes a specific facial motion and thus a change of expression.

3.3. Face description

Facial ROI $k \in \{1, \ldots, 12\}$ with an image subwindow of size $m \times n$ is vectorized

$$I_t^k \in \mathbb{R}^{mn}. \tag{5}$$

As the regions are processed independently, the index k is neglected in the subsequent text.

For each frame, every region I_t is photometrically normalized to suppress possible global illumination changes. The normalization is performed such that mean μ is mapped to 0 and standard deviation σ to 1. Moreover, the result is divided by the total number of pixels in the ROI to balance the different sizes of ROIs

$$\hat{I}_t = \frac{1}{mn} \frac{I_t - \mu \mathbf{1}}{\sigma}. \tag{6}$$

The squared Euclidean difference between image regions at frame t_1 and frame t_2 is given by

$$d_{t_1,t_2} = ||\hat{I}_{t_1} - \hat{I}_{t_2}||_2^2. \tag{7}$$

We define the intensity descriptor for face region in frame t by collecting differences d_{t,t_1} over the sliding window $t_1 \in \{1, \ldots, T\}$

$$\Phi_t = [d_{t,t+1}, d_{t,t+2}, \ldots, d_{t,t+T}]. \tag{8}$$

The ME is observed for a short period of time. According to the maximum considered ME duration, a sliding window of length $T = 0.5$s is used. The temporal descriptor measures the changes of image intensities within the sliding window for every ROI, see Fig. 3.

3.4. Micro-expression detection

Two methods are proposed. The first one, described in Sec. 3.4.1, does not involve any training and is based only on the fact that the highest image intensity changes are in the apex frame of the ME. The other method is supervised and is trained to capture the pattern of MEs. The latter is more robust to filter other events in the face that cause image intensity changes, see Sec. 3.4.2.

3.4.1 Baseline method

Consider an example of the ME in Fig. 3 and the corresponding shape of Φ_t. In the most emotional ME frame (the apex frame), there is supposed to be the largest difference between the apex frame and the preceding frame. Therefore, a score is obtained by aggregating the differences

$$b(t) = \sum_{j=t-T}^{t-1} ||\hat{I}_t - \hat{I}_j||_2^2. \tag{9}$$

The signal $b(t)$ is thresholded. The maxima usually coincide with the ME apex frames. The idea behind is that in case of the ME, Φ_t signals over reference frames t tend to be coherent and produce a strong response $b(t)$ as demonstrated in Fig. 4.

However, the drawback is that other local maxima are generated by other events causing rapid changes in the image, e.g. a global motion or blinking. This weakness is partially mitigated by training a classifier as described in the following section.

3.4.2 SVM classifier

The classifier is trained to distinguish MEs from other false intensity changes. The SVM classifier with RBF kernel was trained on the SMIC_E_HS database. Videos in SMIC have a 100fps frame rate. The feature vectors Φ_t were sub-sampled by taking every 4-th element, since our test videos are only 25fps.

The positive samples were collected from the training sequences. Let t_i be all onset frames of the annotated MEs for every region, $i = 1, \ldots, N$. Then, the positive sample set is defined as

$$P = \bigcup_{i=1}^{N} \{\Phi_{t_i-15}, \Phi_{t_i-14}, \ldots, \Phi_{t_i+5}\}.$$

The data augmentation, where 15 frames before and 5 frames after the onset frame, ensures that the apex

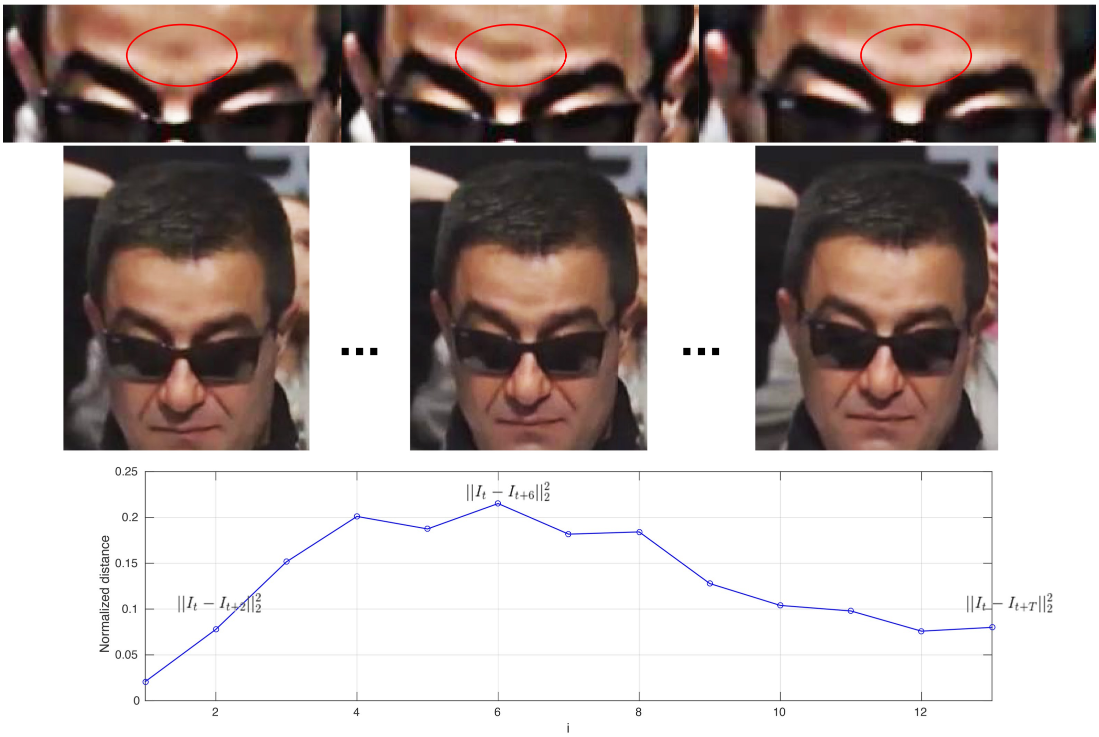

Figure 3. The micro-expression (a surprise emotion) after the poker gamer uncovered his cards. The ME starts with the neutral expression and returns after (12 frames) back to the neutral expression. In the middle frame, a raised eyebrow and wrinkles on the forehead can be noticed. The plot shows elements of the descriptor Φ_t for the forehead ROI.

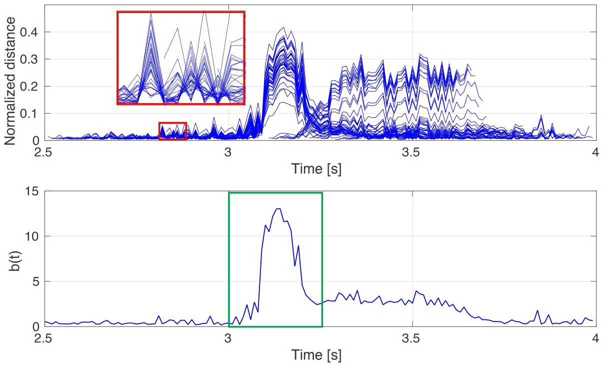

Figure 4. The upper plot shows the expanded intensity descriptors Φ_t over reference frames t. The coherence of the signals and the largest difference in the apex frame results in a peak of the aggregated response $b(t)$.

frame of the ME is always present within the feature vectors. The equal number of negative samples were collected from the rest of the video sequences.

Note that a specific SVM classifier was trained for every region of interest $k = 1, \ldots, 12$. The reason for independent processing is that the intensity pattern may be different due to e.g. amount of texture within the region. Finally, SVMs were trained from 800–2000 samples depending on the occurrence of the ME in a particular ROI in the training data.

3.5. Implementation details

A few tricks to improve the detection were implemented. The regions of the eyes were finally removed from the set of detectors as the eye blinking causes rapid movement of approximately same duration as micro-expressions. However, filtering eye blinks is not possible since the spontaneous blinking is often partially overlapping with the MEs.

In the “In the Wild” videos, the participants are not always shot from the frontal pose. Then a subset of ROIs becomes fully or partially occluded due to a head yaw, see Fig. 2b. Therefore, we detect a non-frontal head pose and the features from the occluded parts of the face are not considered in further processing. The head pose is estimated simply by computing the natural logarithm of the ratio between the inner eye landmarks and the top nose landmark. In the frontal pose, the log-ratio is approximately 0, while the non-frontal pose is detected when the magnitude of the log-ratio exceeds empirically set threshold 0.5.

Finally, the detector response was modified by non-maximal suppression with a 20 frames window to avoid multiple responses for the same event in the data.

4. Experiments

4.1. Datasets

In literature, there are databases of spontaneous ME collected by asking participants to watch an emotive video in a laboratory environment in front of a high speed camera. Then participants' task was to conceal any emotion. As a motivation, they can either get a money reward CASMEII [23], or be punished by filling a long form in SMIC [13] in case of apparent failure.

The original SMIC [13] consists of 164 MEs with 16 participants. The resolution is 640×480 px and frame rate 100 Hz. The videos are clipped only to the duration of ME with no margin before and after the ME. Therefore, it is suitable for isolated recognition. The labels contain three categories - positive, negative and surprise. The expressions are not FACS coded. Later, an extended version was published with longer videos, having an average length of 5.9s, with annotated onset and offset frames of the ME. This version was released especially for the detection problem.

The CASME database [25] consists of 195 elicited MEs from 19 participants captured in 60 Hz frame rate. MEs are FACS-coded and contain the emotion labels. The CASME data were recorded by two cameras. The first with resolution 1280×720 (CASME-A) and the other one with 640×480 (CASME-B).

Later, the database CASMEII was released as an extension of the original dataset. The database contains 255 videos with 26 subjects, having onset, apex and offset frame labels of the MEs; FACS coding and emotion type is annotated. The resolution is 640x480 and frame rate is 200 Hz. The average length of the video clips is 1.25s [23].

We collected data for our new MEVIEW dataset. The dataset consists of videos from poker games and TV interviews downloaded from the Internet. The advantage of poker games is the stress factor and the need to hide emotions. Players try to conceal or fake their true emotions, which is a scenario where MEs are likely to appear. The MEs are still rare, since the TV show post production often cut the most valuable moments such as the detail of a player's face while the cards are being uncovered or someone raises, calls or folds his/her cards. The average length of a video in the dataset is 3s. The camera shot is switched often, and we took the entire shot with a single face in the video clip.

MEs were manually searched in videos. A special attention was paid around key moments of the game in case of poker videos, or while a person is listening to a hard question in case of the TV interviews. These suspicious events were checked carefully. The annotator participated in the Ekman's online course [1] and was able to detect several MEs in the videos with a reasonable confidence. The onset and offset frames of the ME were labeled in long videos, FACS coding and the emotion types were also annotated. In total, 31 videos with 16 individuals were collected.

4.2. Accuracy Evaluation

The proposed algorithm was tested on both SMIC an CASME II datasets and also on our MEVIEW dataset. Both the cross-validation and cross-dataset experiments were performed. The algorithm was tested against the competing method of Li et al. [12] with favourable results. We measured impact of face registration on the detection accuracy on all three datasets.

For evaluation, we strictly follow the evaluation protocol of [12]. Having annotated onset and offset frames of the ME, then a frame is considered to be correctly detected if it is in range [onset-($N/4$), offset+($N/4$)], where N is the maximal considered length of a ME, where $N = 64$ for CASMEII (200 Hz), $N = 32$ for SMIC (100 Hz), and $N = 8$ for MEVIEW dataset (25 Hz). It means, the interval is expanded by a small margin to tolerate certain uncertainty of the annotation in the precise ME interval. All such correctly detected frames are counted as True Positives (TP). If a detection is out of the range, then all N frames are counted as False Positives (FP). The True Positive Rate (TPR) is TP divided by number of annotated positive frames, and the False Positive Rate (FPR) is FP divided by the total number of frames in a sequence without the number of annotated positive frames. The performance is evaluated by receiver operating characteristic (ROC) curve and the area under the curve (AUC).

For our method, at each frame, we have twelve detectors that are spotting MEs. The detection of the ME is considered if at least one detector fires. cw

4.3. Cross-validation on SMIC dataset

In the SMIC_E_HS database, there are 20 participants. Therefore, a 20-fold cross-validation, always

[1] http://www.paulekman.com/micro-expressions-training-tools/

Table 1. Summary results and face rectification influence.

	SMIC-E-HS		CASMEII		in the Wild	
	Baseline	SVM	Baseline	SVM	Baseline	SVM
Similarity	0.85	0.88	0.90	0.97	0.83	0.81
W/o transformation	0.80	0.87	0.90	0.94	0.82	0.73

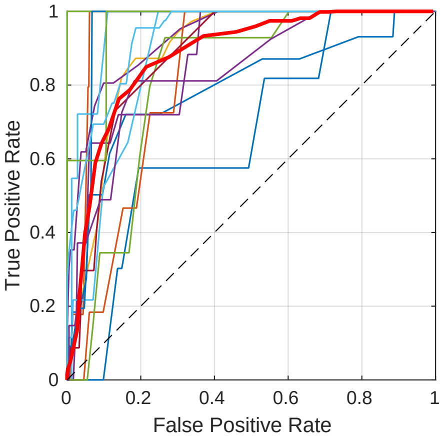

Figure 5. The 20-fold cross-validation on SMIC_E_HS dataset, each trial in a different color. The mean ROC curve, highlighted in red, has AUC = 0.88.

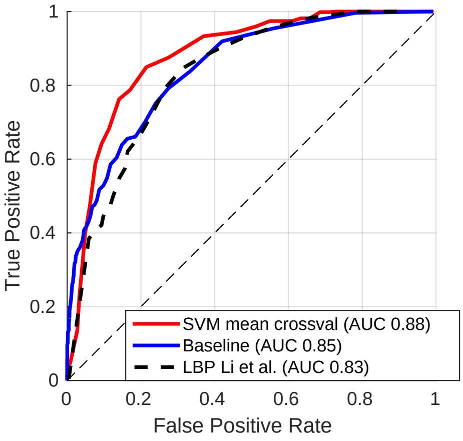

Figure 6. Comparison among the mean SVM cross-validation ROC, the baseline method the LBP-based method by Li et al. [12] on SMIC_E_HS.

leaving all videos of one participant, was performed to train the SVM classifier and to estimate the detection accuracy.

The videos were manually checked and only ROIs with an observable muscle motion in the annotated ME interval was considered as a valid sample in the training set. Cross-validation ROC curves are shown in Fig. 5, together with a mean ROC curve. It is seen, that ROC curves have relatively high variance for cross-validation runs. The reason might be that MEs of certain participants are much easier/more difficult detectable than for others. For certain subjects, spotting their MEs is difficult in a manual inspection of the video played slowly and repeatedly. On the hand, we noticed movements that might be micro-expressions that were not annotated in the dataset.

The mean ROC curve is compared to the baseline method, Sec. 3.4.1, and the result of [12] in Fig. 6. The mean ROC curve outperforms both the baseline thresholding and the method [12]. Nevertheless, the average result of the cross-validation is shown.

4.4. Cross-database experiment

The SVM detectors were trained on the entire SMIC_E_HS database and the resulting classifiers were evaluated on CASMEII, see Fig. 7 and on the MEVIEW database, see Fig. 8.

The CASMEII contains videos in 200 Hz frame rate. Therefore, the videos were sub-sampled taking every 8th frame. The proposed SVM detector outperformed the other methods. For CASMEII, we can see that the SVM significantly outperformed the baseline thresholding, and is slightly more accurate than [12].

Results on our MEVIEW dataset are generally inferior to the results of CASMEII and SMIC. The reason is that our MEVIEW dataset is much more challenging and includes many “in-the-wild” phenomena that cause false positives. Nevertheless, the strongest detections in both the SVM and baseline methods, belongs to true MEs. An example of detection scores of all twelve ROIs for several frames of the poker-game video, around an event shown in Fig. 3, is presented in Fig. 9. We can see the true MEs have the highest score, while several high scoring events, false events appear. I might be due to improperly compensated motion.

The proposed algorithm produces some false positives. They are often caused by eye blinking. The eye blink duration is similar to MEs. Excluding the eye regions helped to reduce the false alarms. However, we observed that the eye blinks somehow causes a shift of the entire set of landmarks and thus influence the transformation, which may result in false detection in other regions than around the eyes. Completely filtering the eye blink instances would on the

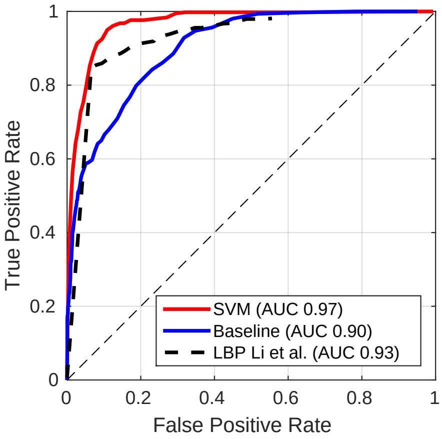

Figure 7. The ROC curves of the SVM, baseline method and LBP by Li et al. [12] on CASMEII.

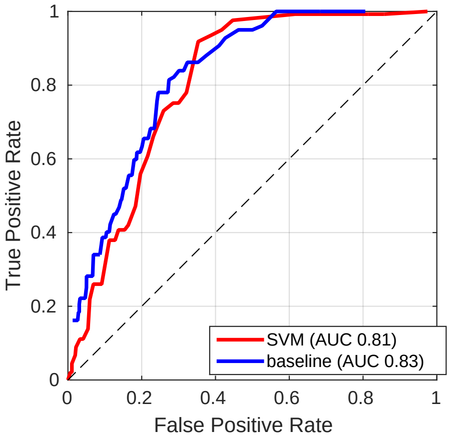

Figure 8. Comparison of the ROC between the SVM and baseline method on MEVIEW database.

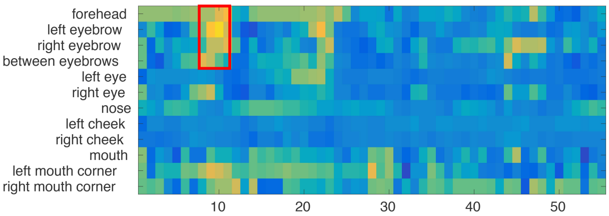

Figure 9. The output of the SVM detectors corresponding to the poker gamer in Fig. 3. The ME is marked by a red rectangle.

other hand decrease the true positive rate, since the eye blinks often co-occur with MEs or MEs often follow shortly after the eye-blinks.

4.5. Impact of Facial rectification

In the following experiment, we measured the impact of the facial rectification to the final accuracy of the method. We skip the face rectification by similarity transformation and the ROIs were defined from landmarks independently. The ROIs were placed based on the location of their landmarks, the size of the regions was derived from the inter-ocular distance and their orientation was defined by the vector connecting the eye centers.

The results of all three datasets are shown in Tab. 1. It can be seen that using the facial rectification by similarity transformation improves results. The difference is significant for the SVM method on our MEVIEW dataset, where subjects move freely and mostly a hand-held camera is used. On laboratory SMIC and CASMEII datasets, with fixed camera and a static subject, the difference is negligible. On the other hand, the rectification might compensate landmark fluctuations, according to the results on the SMIC dataset.

5. Conclusion

In this paper we studied ME spotting and proposed a method to detect MEs in realistic long videos. The method is based on measuring the brief local intensity changes caused by facial muscle contractions. We presented our MEVIEW dataset of “in-the-Wild”, mostly poker tournament videos downloaded from the Internet. The proposed method was evaluated on two standard datasets (SMIC, CASMEII) and on our challenging MEVIEW dataset. The proposed algorithm was compared against our baseline method and against a method of [12] with favorable results.

As a limitation of the proposed method, we see a number of false alarms the algorithm produces. Nevertheless, since the true MEs tend to give a very high score, more examples will be collected by manual inspection of high scoring events in long videos by experienced human annotators. Much more data would surely allow to design more sophisticated classifier with higher detection accuracy.

Acknowledgements

The research was supported by CTU student grant SGS15/155/OHK3/2T/13 and by the Technology Agency of the Czech Republic research program TE01020415 (V3C).

References

[1] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2016. 2

[2] P. Ekman. Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, 1000(1):205–221, 2003. 1

[3] P. Ekman. *Micro Expressions Training Tool*. Emotionsrevealed. com, 2003. 1

[4] P. Ekman. *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. Macmillan, 2007. 1

[5] P. Ekman. Lie catching and microexpressions. *The philosophy of deception*, pages 118–133, 2009. 1

[6] P. Ekman and W. V. Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88–106, 1969. 1

[7] P. Ekman and W. V. Friesen. Facial action coding system. 1977. 1, 4

[8] P. Ekman and W. V. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003. 1

[9] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):285–339, 1991. 3

[10] Y. Guo, Y. Tian, X. Gao, and X. Zhang. Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 3473–3479. IEEE, 2014. 2

[11] 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. 1

[12] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen. Reading hidden emotions: spontaneous micro-expression spotting and recognition. *arXiv preprint arXiv:1511.00423*, 2015. 2, 6, 7, 8

[13] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6, April 2013. 2, 6

[14] D. Matsumoto and H. S. Hwang. Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion*, 35(2):181–191, 2011. 1

[15] A. Moilanen, G. Zhao, and M. Pietikäinen. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1722–1727. IEEE, 2014. 2

[16] T. Pfister, X. Li, G. Zhao, and M. Pietikäinen. Recognising spontaneous facial micro-expressions. In *2011 International Conference on Computer Vision*, pages 1449–1456. IEEE, 2011. 2

[17] 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*, 96(1):81–92, 2013. 2

[18] S. Polikovsky, Y. Kameda, and Y. Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In *Crime Detection and Prevention (ICDP 2009), 3rd International Conference on*, pages 1–6, Dec 2009. 2, 4

[19] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar. Macro-and micro-expression spotting in long videos using spatio-temporal strain. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 51–56. IEEE, 2011. 2

[20] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, and J. Tao. Micro-expression recognition using color spaces. *IEEE Transactions on Image Processing*, 24(12):6034–6047, 2015. 2

[21] Q. Wu, X. Shen, and X. Fu. The machine knows what you are hiding: an automatic micro-expression recognition system. In *International Conference on Affective Computing and Intelligent Interaction*, pages 152–162. Springer, 2011. 2

[22] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 3

[23] 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. 2, 6

[24] 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*, 37(4):217–230, 2013. 1

[25] 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. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7, April 2013. 2, 6