Real-Time Micro-Expression Detection From High Speed Cameras

Diana Borza, Razvan Itu, Radu Danescu
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Diana.Borza@cs.utcluj.ro, Razvan.Itu@cs.utcluj.ro, Radu.Danescu@cs.utcluj.ro

***Abstract*—This work presents an original real time, robust micro-expression detection algorithm. The algorithm analyses the movement modifications that occur around the most prominent facial regions using two absolute frame differences. Next, a machine learning algorithm is used to predict if a micro-expression occurred at a given frame *t*. Two classifiers were evaluated: decision tree and random forest classifier. The robustness of the proposed solution is increased by further processing the preliminary predictions of the classifier: the appropriate predicted micro-expression intervals are merged together and the interval that are too short are filtered out. The proposed solution achieved an 86.95% true positive rate on CASME2 dataset. The mean execution time of the proposed solution on 640x480 images is 9 milliseconds. (*Abstract*)**

Keywords—micro-expression; detection; frame difference; random forest classifier; decision tree;

I. INTRODUCTION

The analysis of facial expression dates back to the Vth century B.C., when physiognomy was used to assess the character or personality of a person from its face traits, and it has been intensively studied ever since. In the early 1970`s a major breakthrough has been achieved, when Paul Eckman and his colleagues used facial expressions to recognize hidden emotions [1]. His studies had a great impact in the development of the current facial expression recognition systems. Eckman defined several facial cues to detect deceit: micro-expressions, squelched expressions, reliable facial muscles, expression asymmetries and various parameters related to the dynamics of the expression. Nowadays, automatic expression and micro-expression analysis has a strong impact on a variety of applications, ranging from human computer interaction to surveillance systems, biometry etc.

Micro-expressions (ME) are considered the most reliable sources of deceit detection. In the United States, within the SPOT program ([2]), airport employees are trained in ME recognition in order to detect the passengers with suspicious behavior. ME are short facial expressions (with a duration between 1/5 and 1/25 of a second), that usually occur when people try to hide their feelings (in cases of both deliberate and non-conscious) concealment. A micro-expression can be described by its time evolution – onset (the moment when the ME starts), apex (the moment of maximum amplitude) and offset (the moment when it fades out) – its amplitude and its symmetry.

Recently, the automatic analysis of ME has attracted the attention of researchers in the computer vision field. There are several challenges that need to be addressed in this relatively new field: first, as ME are involuntary, training and test datasets are hard to gather. However, several ME databases are available [3, 4, 5], but they only contain video sequences captured in controlled scenarios. Another difficulty is related to data labelling, as this is a time consuming and subjective process. As a result, some ME databases [5] classify the expressions only into three categories: positive, negative and surprise. Finally, ME are very fast movements and are visible only for a limited number of frames. Therefore, high speed camera and accurate motion and tracking algorithms are required in the analysis of ME.

In this paper, we propose a fast and robust micro-expression detection framework based solely on the movement magnitudes that appear on certain regions of the face. The detection process determines if a ME has occurred at a certain time moment, while the recognition process establishes the type of the micro-expression. For the detection part, we use a sliding window to iterate over the movement variations of the video sequence and we compute the minimum and maximum response for each window position. The resulting feature vector is fed to a classifier in order to determine if a ME occurred at the center of the window. The raw result from the classifier is further processed in order to filter out false positives and to merge responses corresponding to the same ME.

This work has the following structure: in Section 2 the recent advances in the field of ME detection and recognition are presented. The outline of the proposed solution is illustrated in Section 3 and detailed in Section 4. The experimental results are presented and discussed in Section 4. Finally, this work is concluded in Section 5.

II. STATE OF THE ART

Although automatic ME detection and recognition is not as widely studied as macro-expression analysis, with the recent advances in computer vision, several works addressed this problem. A ME analysis framework usually consists of three main tasks: (1) the selection of the relevant face regions, (2) the

978-1-5386-3368-7/17/$31.00

extraction of spatiotemporal features and (3) the detection and recognition of ME using machine learning algorithms.

The first module is related to the selection of the facial areas where the MEs are more likely to occur. The Facial Action Coding System (FACS) is a methodology used to classify facial expressions based on the muscles that produce them and it is used by trained human practitioners. For the automatic ME analysis, the face is usually segmented according to the most prominent facial elements (eyes, mouth corners and nose) ([7, 8, 9]), or a complex deformable model is used to divide the face into more precise regions ([10, 6]). Another approach is to split the face into *n* equal cells ([11, 12]).

As ME are brief facial movements their analysis requires robust spatiotemporal image descriptors. Various descriptors have been used in the literature: Local Binary Patterns in Three Orthogonal Planes (LBP-TOP) [6], 3D histogram of oriented gradients (HOG) [7, 8], dense optical flow ([10]), optical strain [11]. Finally, using the appropriate features, ME can be classified using supervised ([6]) or non-supervised ([7, 8]) machine learning algorithms.

Several works perform both ME detection and recognition. In [12], the authors propose a general micro expression analysis framework that performs both micro expression detection and recognition. The detection phase does not require any training and exploits frame difference contrast to determine the frames where movement occurred. For the recognition phase, several descriptors (LBP-TOP, HOG and Histogram of Image Gradient Orientation (HIGO)) are fed to a support vector machine classifier. In [11] optical strain is weighted with LBP-TOP features in order to detect and recognize ME.

III. Solution Outline

Figure 1 shows the outline of the proposed solution.

The method analyzes the motion variation that occurs across the high-speed video sequence. Two absolute image differences are computed: the difference between the current frame *t* and the frame *t-3* (that describes the noise variation) and the difference between the current frame and the previous frame at distance $\Delta t/2$ (that describes the motion information).

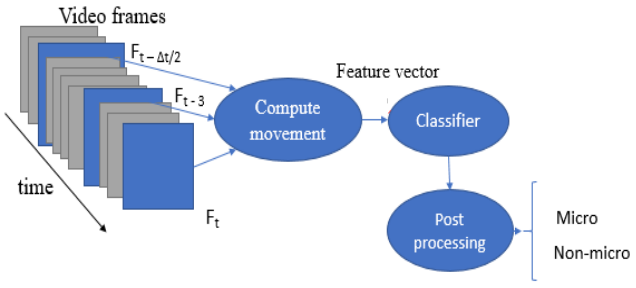

Fig. 1. Solution outline

The movement magnitude is computed by dividing the second difference image to the first difference image. Next, the mean magnitude variation around the most prominent parts of the face (eyebrows, eye corners, mouth corners, chin) is computed and a classifier is used to determine if a ME occurred at the current frame *t*. Finally, the response of the classifier is further processed in order to increase the robustness of the solution.

IV. Solution Description

In this section, each module of the proposed solution is described in detail.

A. Selection of relevant face regions

The proposed solution analyses the movement magnitude variation in 10 equally-sized regions of interest on the face. These cells were selected based on the Facial Action Coding System methodology, such that all the muscles that are involved in the occurrence of MEs are comprised. First, 68 facial landmarks are localized on the face using constrained local models [13] and the cells location are computed based on these interest points. Three cells are positioned in the eyebrow area, four cells are positioned around the outer eye corners and mouth corners respectively. Two cells are set around the nostrils, and finally, one cell is used on the chin area. The width and height of a cell are equal to half the mouth width. Figure 2 shows the 10 cells that are analyzed by the ME detection and recognition algorithm.

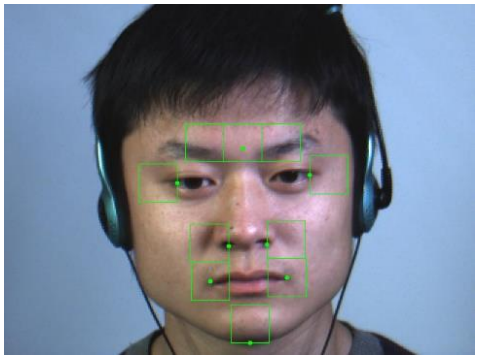

Fig. 2. Facial regions of interest. 10 regions of interest are selected around the the most prominent facial areas, where the ME are likely to cause muscle movements.

B. Feature extraction

We propose a simple method for estimating the motion variation that occurs during a ME. Let Δt denote the average ME duration (expressed in number of frames) for a given dataset. As the ME video sequences are captured with a high-speed camera, practically, there should be no facial movement variation between consecutive frames (0.005 s).

The movement magnitude for a ME is very low, so we need to consider the noise as a normalization factor. Therefore, for each frames *t* we compute two absolute difference images: ΔME (the difference between the frame *t* and the frame $t - \Delta t/2$) and $\Delta\varepsilon$ (the difference between the frame *t* and the frame $t - 3$); these differences are illustrated in Figure 3 (a) and (b). The $\Delta\varepsilon$ image describes the noise that occurs at frame *t*.

(a) Δε (b) ΔME (c) MM

Fig. 3. Frame movement computation. (a) Difference between the current frame and the preious frame at three frames distance. (b) Difference between the current frame and the *t* the frame *t* - *Δt*/2. (c) Movement magnitude

The first difference image ΔME describes the movement variation that occurred within the *Δt*/2 interval, while the *Δε* is considered a neutral reference image (as there is practically no facial movement that is captured within the interval of 3 frames) and it is used as a normalization factor. Therefore, the movement magnitude MM (Figure 3(c)) at each frame *t* is computed as:

$$MM = \frac{\left|frame_t - frame_{t-\frac{\Delta t}{2}}\right| + 1}{[frame_t - frame_{t-3}] + 1}$$

For each one of the 10 face cells, we compute the average value of the MM image within that region of interest. For example, Figure 4 illustrates the average value of the MM image for the middle eyebrow region.

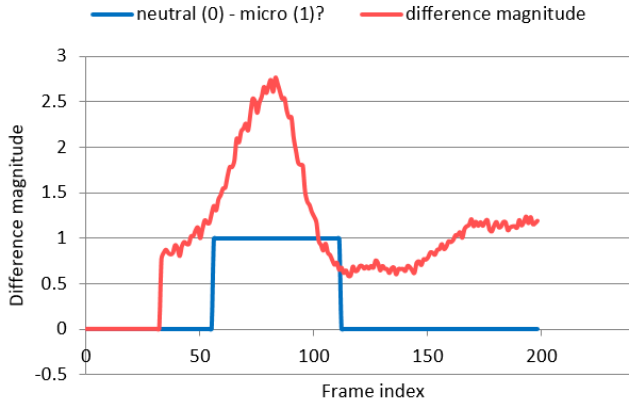

Fig. 4. Difference variation of the middle eyebrow face cell. The ground truth labeling of the ME sequence is marked with a blue step, and the difference variation is depicted in grey.

A sliding time window is used to iterate through the responses for all the cells and the minimum and maximum value within the time frame are saved to a feature vector that will be further analyzed by a classifier in order to detect if a ME occurred. For each cell, we compute the average minimum and maximum value within the sliding window and we concatenate them to the feature vector. The dimensionality of the feature vector is 20 (10 cells x 2 vales per cell).

$$feature_t = \underset{c_i \in cell}{||} \left(\max_{t \in sz} \langle MM_t[c_i]\rangle, \quad \min_{t \in sz} \langle MM_t[c_i]\rangle\right.$$

, where $\langle MM_t[c_i]\rangle$ represents the average value of the MM image within the region of interest c_i, at frame *t* and || represents the concatenation operator.

In order to make the algorithm more robust to illumination changes and to eliminate the lighting bias, we also convolved the input frame image with the Laplace kernel:

$$L = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Figure 5 shows the image obtained after the Laplacian filtering. The results obtained using differences the raw images and the Laplacian filtered images will be discussed in Section 5.

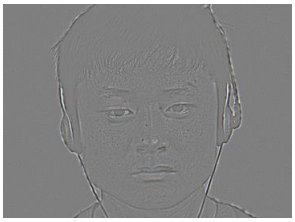

Fig. 5. Laplace filtering

C. Classification

The extracted feature vectors are used as input for a classification algorithm that will determine the state (ME or non-ME) at each frame *t*. We performed the classification using two classifiers: *decision tree* and *random forest classifier*.

Decision trees [14] are non-parametric supervised learning algorithms that use a graph-like structure to determine classification rules. Each internal node contains a condition on an attribute and each edge represents the outcome of the "test" from the node. The class labels are encoded as leaves in the tree, while the paths from the root of the tree to each leaf represent the classification rules. Decision trees are computationally efficient (the prediction step is logarithmic in the number of data instances used to train the tree), easy to visualize and understand and require little or no data preprocessing. Their main disadvantage is that the learning algorithm can generate an over-complex tree that, in turn, leads to overfitting.

Random forest classifiers [15] are ensemble learning methods that were designed to cope with the problem of overfitting that occurs in decision trees. These classifiers generate multiple decision trees at training time and the final class label is the mode (the label that appears more often) of the classes of the individual trees.

D. Post processing

The preliminary result (R_t) obtained from the classifier is further analyzed in order to filter out false positive and to determine the time frame of the ME (onset, apex and offset moments). R_t contains the predicted classes (0 – non-ME class and 1 – ME class) for each frame from the input video sequence. We make the assumption that the preliminary result vector should contain agglomerations of ME class predictions around the apex frame a ME, and the singular predictions of ME class correspond to false positives. Therefore, we first determine all the contiguous intervals that contain only ME class predictions.

The intervals that are too close to each other (their distance is less than $\Delta t/4$) are merged together, and next, all the intervals that are too short (their width is lower than $\Delta t/10$) are considered false positives as filtered out. The remaining intervals are considered ME intervals and their centroid is selected as the apex frame of the ME.

Figures 6 and 7 show the raw response of a classifier on an input video sequence and the filtered response of using the proposed algorithm. The ground truth onset, apex and offset frames of the video sequence are also marked on the plot.

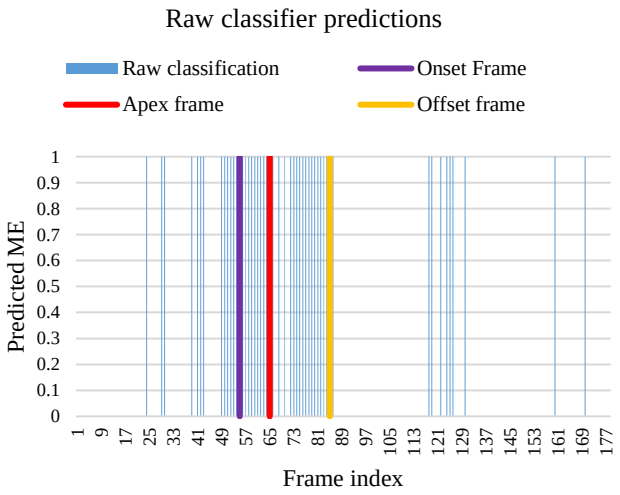

Fig. 6. Raw classifier prediction. The predictions are depicted in blue vertical lines; the ground truth onset and the apex and offset frames are depicted in violet, red and yellow respectively.

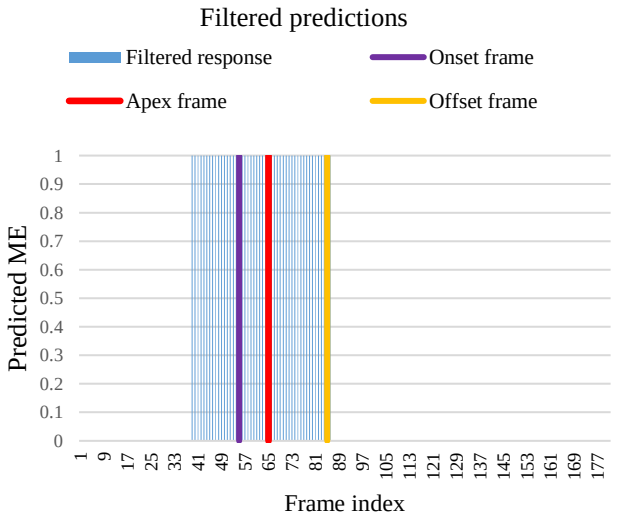

Fig. 7. Post processing of the classifier result. The retained classifier predictions are depicted in blue vertical lines and the ground truth onset, apex and offset frames are depicted in violet, red and yellow respectively

V. EXPERIMENTAL RESULTS

The proposed solution was trained and evaluated on the CASME II [5] database. This dataset contains 247 video sequences of spontaneous micro-expressions, captured from 26 participants. The mean age of the participants is 22.03 years, with 1.6 standard deviation. The video sequences were captured by a high-speed camera (200 fps), with a resolution of 640×480 pixels. The video sequences are labeled with the onset, apex and offset moments, and with one of following ME types: happiness, disgust, surprise, repression and tense.

For the evaluation part, we used "leave one subject out cross validation" (LOSOCV): we randomly selected two subjects and all the video sequences belonging to these two subjects were used only to evaluate the performance of the proposed solution.

To label the data for detection module, a sliding time window is iterated through the video sequence. If Δt is the average micro-expression duration (67 frames), and t_{apex} is the ME ground truth apex frame, the current frame t is labeled using the following rule:

- If $t \in [0, t_{apex} - \delta \cdot \Delta t]$ or $t \in [t_{apex} + \delta \cdot \Delta t]$, then the frame t is labelled as non-micro-expression frame (neutral frame or macro-expression);
- If $t \in (t_{apex} - \delta \cdot \Delta t, t_{apex} + \delta \cdot \Delta t)$, then frame t is considered a ME frame.

Table I shows the performance of the algorithm on the CASME2 dataset. TPR stands for True Positive Rate, FPR for False Positive Rate, FNR stands for False Negative Rate and TNR represents the True Negative Rate.

TABLE I. PERFORMANCE ON THE CASME 2 DATASET

Feature	Classifier	TPR	FPR	FNR	TNR
Raw pixels	Decision tree	68.18%	0.25%	31.81%	99.74%
Raw pixels	Random forest	72.72%	0.15%	27.27%	99.84%
Laplacian	Decision tree	76.19%	0.06%	23.80%	99.93%
Laplacian	**Random forest**	**86.95%**	**0.012%**	**13.04%**	**99.87%**

The best results are obtained using the Laplace filtering of the input image and a random forest classifier. An example of how the proposed algorithm detects the ME occurrence can be visualized at: https://drive.google.com/drive/folders/0ByAKFSXshk1ARXMxNFBYMU9hM3M?usp=sharing .

Our method is better than recent state of the art methods. In table II we present the comparison of the proposed solution with other state of the art works. ACC stands for accuracy, FPR – false positive rate and TPR – true positive rate.

Methods marked with an asterisk * were evaluated on SMIC [3] database. To detect the micro expressions, most of the works were only evaluated on SMIC database. Therefore, the numerical comparison with these methods might not be relevant.

TABLE II. COMPARISON WITH STATE OF THE ART WORKS

Method	Features	Performance
[3]	LBP-TOP	ACC: 65.49 %*
[5]	LBP-TOP	N/A
[11]	Optical Strain, LBP-TOP	ACC: 74.16%*
[12]	Frame differences	TPR*: 70%
Our solution	**Frame differences**	**TPR: 86.95%**

The execution time of the proposed solution is approximately 9 milliseconds on a 4th generation Intel i7 processor.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a fast and robust method for the detection of subtle expressions from high speed cameras. The method analyses the movement variations that occur in a given time frame using image differences. Two classifiers were used and evaluated to determine if a ME occurred at a given frame *t*. In order to ensure the robustness of the algorithm, the raw response of the classifier is further post-processed in order to filter out false positives and to merge the predictions that belong to the same ME zone. The proposed method is fast, robust and it achieves a high positive rate, while maintaining the false positive rate low.

As a future work, we plan to gather more data for the training process so that more data variation is present. Also, the method will be extended such that the actual class of the micro-expression is also recognized.

ACKNOWLEDGMENT

This work was supported by the MULTIFACE grant(Multifocal System for Real Time Tracking of Dynamic Facial and Body Features) of the Romanian National Authority for Scientific Research, CNDI–UEFISCDI, Project code: PN-II-RU-TE-2014-4-1746.

REFERENCES

[1] P. Ekman, Telling Lies: “Clues to Deceit in the Marketplace, Politics, and Marriage”, W. W. Norton & Company, 2009.

[2] Wikipedia (28.04.2017), “SPOT (TSA program)”, [Online], Available: https://en.wikipedia.org/wiki/SPOT_(TSA_program)

[3] H. Rautio (12.04.2017), "SMIC - Spontaneous Micro-expression Database," University of Oulu, [Online], Available: http://www.cse.oulu.fi/SMICDatabase.

[4] 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 FG. IEEE, 2013, pp. 1–7.

[5] Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, “CASME II: An improved spontaneous micro-expression database and the baseline evaluation,” PloS one, vol. 9, no. 1, 2014.

[6] T. Pfister, X. Li, G. Zhao and M. Pietikainen, "Recognising Spontaneous Facial Micro-expressions," in 2011 IEEE International Conference on Computer Vision (ICCV), Barcelona, 2011.

[7] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," in 3rd International Conference of Crime Detection and Prevention, 2009.

[8] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial Micro-Expression Detection in Hi-Speed Video Based on Facial Action Coding System (FACS)," IEICE TRANSACTIONS on Information and Systems, vol. E96, no. 1, 2013.

[9] S. Godavarthy, D. Goldgof, S. Sarkar M. Shreve, "Macro- and micro-expression spotting in long videos using spatio-temporal strain," in Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011.

[10] Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, and Guoying Zhao, "A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition," IEEE Transactions On Affective Computing, vol. 99, 2015.

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

[12] X. Li, H.O.N.G. Xiaopeng, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikainen. “Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-expression Spotting and Recognition Methods”. IEEE Transactions on Affective Computing, 2017.

[13] M. Cox, J. Nuevo, J.Saragih and S. Lucey, “CSIRO Face Analysis SDK“, AFGR 2013

[14] J.R, Quinlan. "Induction of decision trees." Machine learning 1, no. 1 (1986): 81-106.

[15] T.K. Ho. "Random decision forests." In Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, vol. 1, pp. 278-282. IEEE, 1995.