Spotting Macro- and Micro-expression Intervals in Long Video Sequences

Authors: Ying He, Su-Jing Wang, Jingting Li, Moi Hoon Yap, Su-Jing Wang

Date: 2019-12-20T00:00:00+00:00

Abstract

This paper presents baseline results for the Third Facial Micro-Expression Grand Challenge (MEGC 2020). Both macro- and micro-expression intervals in CAS(ME)^2 and SAMM Long Videos are spotted by employing the method of Main Directional Maximal Difference Analysis (MDMD). The MDMD method uses the magnitude maximal difference in the main direction of optical flow features to spot facial movements. The single frame prediction results of the original MDMD method are post processed into reasonable video intervals. The metric F1-scores of baseline results are evaluated: for CAS(ME)^2, the F1-scores are 0.1196 and 0.0082 for macro- and micro-expressions respectively, and the overall F1-score is 0.0376; for SAMM Long Videos, the F1-scores are 0.0629 and 0.0364 for macro- and micro-expressions respectively, and the overall F1-score is 0.0445. The baseline project codes is publicly available at https://github.com/HeyingGithub/Baseline-project-for-MEGC2020_spotting.

Full Text

Preamble

Spotting Macro- and Micro-expression Intervals in Long Video Sequences

Ying He[1], Su-Jing Wang[1], Jingting Li[1], and Moi Hoon Yap[2]
[1] Key Laboratory of Behavior Sciences, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China
[2] Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, M1 5GD, UK

Abstract—This paper presents baseline results for the Third Facial Micro-Expression Grand Challenge (MEGC 2020). We employ the Main Directional Maximal Difference Analysis (MDMD) method to spot both macro- and micro-expression intervals in CAS(ME)[2] and SAMM Long Videos. The MDMD

method uses the magnitude maximal difference in the main direction of optical flow features to detect facial movements. The single-frame predictions from the original MDMD method are post-processed into reasonable video intervals. The baseline results achieve F1-scores of 0.1196 for macro-expressions and 0.0082 for micro-expressions on CAS(ME)², with an overall F1-score of 0.0376. On SAMM Long Videos, the F1-scores are 0.0629 for macro-expressions and 0.0364 for micro-expressions, with an overall F1-score of 0.0445. The baseline project code is publicly available at https://github.com/HeyingGithub/Baseline-project-for-MEGC2020_spotting.

I. Introduction

Facial expressions are important non-verbal cues that convey emotions. Macro-expressions are the common facial expressions we encounter in daily life and are the types most people are familiar with. A special category of expressions, known as "micro-expressions," was first discovered by Haggard and Isaacs [**?**]. Micro-expressions (MEs) are involuntary facial movements that occur spontaneously when a person attempts to conceal an experienced emotion in a high-stakes environment. These expressions are extremely brief, typically lasting less than 500 milliseconds [**?**, **?**]. The close connection between micro-expressions and deception makes research in this area highly significant for applications such as medical care [**?**] and law enforcement [**?**].

Expression spotting aims to identify the moments when expressions occur within complete video sequences. The Second Micro-Expression Spotting Challenge (MEGC 2019) [**?**] explored methods for spotting micro-expression intervals in long videos, building upon several decades of research in this area [**?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**]. However, micro-expressions often accompany macro-expressions, and both types of expressions are valuable for affective analysis. Consequently, developing methods capable of spotting both macro- and micro-expressions represents the main theme of MEGC 2020.

In this paper, we provide the baseline method and results for the Third Facial Micro-Expression Grand Challenge (MEGC 2020), focusing on spotting macro- and micro-expression intervals in long video sequences from the CAS(ME)² and SAMM Long Videos datasets. Our primary method is Main Directional Maximal Difference Analysis (MDMD) [**?**]. The original MDMD method only predicts whether individual frames belong to facial movements. To obtain target intervals, we form intervals from adjacent frames that are consistently predicted as macro- or micro-expressions, then remove intervals that are excessively long or short. Parameters are adjusted for specific expression types and datasets. We evaluate performance using F1-scores on both long video datasets.

The remainder of this paper is organized as follows: Section II presents the methodology and performance metrics, Section III introduces detailed experimental results, and Section IV concludes the paper.

II. Methodology

This section describes the benchmark datasets, baseline method, and performance metrics.

A. Datasets

CAS(ME)² [**?**]: Part A of the CAS(ME)² database contains 22 subjects and 98 long videos. Facial movements are classified as macro- and micro-expressions, with video samples potentially containing multiple macro- or micro-expressions. The onset, apex, and offset indices for these expressions are provided in an Excel file, and eye blinks are labeled with onset and offset times.

SAMM Long Videos [**?**]: The original SAMM dataset [**?**] contained 159 micro-expressions and was used in the past two micro-expression recognition challenges [**?**, **?**]. Recently, the authors released the SAMM Long Videos dataset, which consists of 147 long videos containing 343 macro-movements and 159 micro-movements. The onset, apex, and offset frame indices for both micro- and macro-movements are outlined in the ground truth Excel file. More detailed comparative information about these two datasets is presented in [TABLE:I].

B. Baseline Method

1) Preprocessing Expression spotting focuses on facial regions. We preprocess every video sample by cropping and resizing facial regions across all frames. For each video, we locate a rectangular bounding box that exactly encloses the facial region in the first frame, then crop and resize all video frames according to this box. We determine the bounding box using facial landmarks detected by the corresponding function in the "Dlib" toolkit [**?**], as we found that applying a face detection algorithm directly does not perform well. The preprocessing details are as follows.

First, we use the landmark detection function in the "Dlib" toolkit to obtain 68 facial landmarks on the face in the first video frame, as illustrated in Figure 1: see original paper showing the first frame of s23_0102 from CAS(ME)². The landmarks are labeled as L , L , ···, L in the sequence returned by the Dlib landmark detection function, with corresponding coordinates (x , y), (x , y), ··· , (x , y). The coordinate system is consistent with that in the OpenCV toolkit [**?**], where the x-axis represents the horizontal direction from left to right and the y-axis represents the vertical direction from top to bottom. The green dots in Figure 1: see original paper show the landmarks, with some serial numbers marked in red text.

Second, to form a rectangular box that bounds the facial region precisely, we identify the leftmost, rightmost, topmost, and bottommost landmarks as L , L , L , L with coordinates (x , y), (x , y), (x , y), (x , y), respectively. Rather

than forming the box directly from these extreme landmarks, we create two points: A(x , y - (y - y)) and B(x , y) to obtain box B with A as the upper-left corner and B as the lower-right corner. The coordinate y - (y - y) means the upper edge of the box is shifted upward by a relative distance to retain more region around the eyebrows. In Figure 1: see original paper, box B is shown as the blue rectangle.

Third, as shown in Figure 1: see original paper (the region within B), we found that for several subjects in both datasets, the bottom region is too large due to inaccuracies in landmark detection. Therefore, we detect landmarks again on the region of the first frame within B to crop faces more precisely, as shown in Figure 1: see original paper. We then obtain a new bottommost landmark L' (x' , y'). Box B is updated to B' (x , y), where y is the smaller of y and y' . A new rectangular box B' is formed with A as the upper-left corner and B' as the lower-right corner. In Figure 1: see original paper, box B' is shown as the blue rectangle, and the region of the first frame within B' is illustrated in Figure 1: see original paper, where we can see the facial region is better localized.

Finally, after obtaining box B' , we crop all video frames within rectangular box B' to extract the facial regions, which are then resized to 227 × 227 pixels.

2) MDMD The Main Directional Maximal Difference Analysis (MDMD) method was proposed in [**?**]. The main idea is that when an expression occurs, the face experiences a process of forming an expression and returning to a neutral state. The main movement directions will be opposite in this process, and by analyzing this phenomenon, expressions can be spotted. Here we review the MDMD method.

Given a video with n frames, the current frame is denoted as F . F is the k-th frame before F , and F is the k-th frame after F . We compute the robust local optical flow (RLOF) [**?**] between frame F (Head Frame) and frame F (Current Frame), denoted as (u , v). For convenience, (u , v) represents the displacement of any point. Similarly, the optical flow between frame F (Head Frame) and frame F (Tail Frame) is denoted as (u , v). Then, (u , v) and (u , v) are converted from Euclidean coordinates to polar coordinates (,) and (,), where and represent magnitude and direction, respectively.

Based on the directions { }, all optical flow vectors {(,)} are divided into a directions. [Figure 2: see original paper] illustrates the case when a = 4. The Main Direction Θ is the direction containing the largest number of optical flow vectors among the a directions. The main directional optical vector M is the optical flow vector (,) that falls within the Main Direction Θ.

$$\{(\rho_{HC}^{M}, \theta_{HC}^{M})\} = \{(\rho_{HC}, \theta_{HC}) | \theta_{HC} \in \Theta\}$$

The optical flow vector corresponding to (,) between frames F and F is denoted as (,):

$\{(\rho_{HT}^M, \theta_{HT}^M)\} = \{(\rho_{HT}, \theta_{HT}) | (\rho_{HT}, \theta_{HT})$ and $(\rho_{HC}^M, \theta_{HC}^M)$ are two different vectors of the same point in $F_{i-k}\}$

After sorting the differences - in descending order, the maximal difference d is defined as the mean difference value of the first 1/3 of the differences - to characterize frame F :

$$d_i = \frac{\sum\{\rho_{HC}^M - \rho_{HT}^M\}}{|\{(\rho_{HC}, \theta_{HC})\}|}$$

where g = |{(,)}| is the number of elements in the subset {(,)}, and max S denotes a set comprising the first m maximal elements in subset S.

Since our method is block-based analysis, the cropped facial region of each frame is divided into b × b blocks, as shown in [Figure 3: see original paper]. We calculate the maximal difference d (j = 1, 2, ···, b^2) for each block in frame F . For frame F , there are b^2 maximal differences d due to the b × b block structure. We then arrange the b^2 maximal differences in descending order, where đ is the first s maximal difference and characterizes the frame F feature:

$$\bar{d_i} = \sum_{j=1}^{s} \max_s \{d_i^j\}, \quad j = 1, 2, \cdots, b^2$$

If a person maintains a neutral expression at F , their emotional expression (such as disgust) starts at the onset frame between F and F , is repressed at the offset frame between F and F , and then the facial expression returns to neutral at F , as shown in Figure 4: see original paper. In this case, the movement between F and F is more intense than the movement between F and F because the expression is neutral at both F and F . Therefore, the đ value will be large. Another situation occurs when a person maintains a neutral expression from F to F . Here, the movement between F and F is similar to the movement between F and F , resulting in a small đ value. In long videos, sometimes an emotional expression starts at the onset frame before F and is repressed at the offset frame after F , as shown in Figure 4: see original paper. In this case, the đ value will also be small if k is set to a small value. However, k cannot be set to a large value because this would influence the accuracy of optical flow computation.

We employ a relative difference vector to eliminate background noise, computed by:

$$r_i = \frac{d_i}{\sum_{t=i-k+1}^{i+k-1} d_t}, \quad i = k+1, k+2, \cdots, n-k$$

Therefore, frame F is characterized by r . A threshold is used to obtain frames with peaks representing facial movements in a video:

$$\text{threshold} = r_{\text{mean}} + p \times (r_{\text{max}} - r_{\text{mean}})$$

where

$$r_{\text{mean}} = \frac{\sum_{i=k+1}^{n-k} r_i}{n-2k} \quad \text{and} \quad r_{\text{max}} = \max_{i=k+1}^{n-k} \{r_i\}$$

p is a variable parameter in the range [0, 1]. Frames with r larger than the threshold are identified as frames where expressions appear.

3) Parameter Settings and Post-Processing In [**?**], several parameter combinations were explored to spot micro-expressions on the CAS(ME)² dataset. For spotting both macro- and micro-expressions on the two datasets for MEGC 2020 (CAS(ME)² and SAMM Long Videos), we select the best combination of blocks and directions explored in [**?**] and set other parameters according to the FPS of each dataset. Since the original MDMD only predicts whether a frame belongs to facial movements, we add post-processing to output the target intervals required by MEGC 2020.

The number of blocks is set to 6 × 6 and the number of directions a is set to 4. In the CAS(ME)² dataset, k is set to 12 for micro-expressions and 39 for macro-expressions. In the SAMM Long Videos dataset, k is set to 80 for micro-expressions and 260 for macro-expressions.

Regarding the threshold, p varies from 0.01 to 0.99 with a step-size of 0.01, and final results are reported under the setting of p = 0.01. The original MDMD only predicts whether a frame belongs to facial movements. To output target intervals, adjacent frames consistently predicted as macro- or micro-expressions form an interval, and intervals that are too long or too short are removed. The number of micro-expression frames is limited between 7 and 16 for CAS(ME)² and between 47 and 105 for SAMM Long Videos. The number of macro-expression frames is defined as larger than 16 for CAS(ME)² and larger than 105 for SAMM Long Videos.

C. Performance Metrics

To avoid inaccuracy caused by annotation, we propose to evaluate spotting results per interval in MEGC 2020.

1. True Positive Definition per Interval in One Video The true positive (TP) per interval in one video is first defined based on the intersection between the spotted interval and the ground-truth interval. The spotted interval W is considered a TP if it satisfies the following condition:

$$\frac{W_{\text{spotted}} \cap W_{\text{groundTruth}}}{W_{\text{spotted}} \cup W_{\text{groundTruth}}} > k$$

where k is set to 0.5, and W T represents the ground truth of the macro- or micro-expression interval (onset-offset). If the condition is not fulfilled, the spotted interval is regarded as false positive (FP).

2. Result Evaluation in One Video Suppose there are m ground-truth intervals in the video, and n intervals are spotted. Based on overlap evaluation, the TP count in one video is a (where a m and a n). Therefore, FP = n - a and FN = m - a. The spotting performance in one video can be evaluated using:

$$\text{Recall} = \frac{a}{m}, \quad \text{Precision} = \frac{a}{n}, \quad F\text{-score} = \frac{2 \times a}{m + n}$$

However, real-life videos present complicated situations that influence per-video evaluation: (1) There might be no macro- or micro-expressions in the test video (m = 0), making the recall denominator zero; (2) If no intervals are spotted (n = 0), the precision denominator becomes zero; (3) It is impossible to compare two spotting methods when both TP amounts are zero (all metric values equal zero), even if one method spots fewer intervals than the other. To avoid these situations, for single-video spotting result evaluation, we only record the TP, FP, and FN counts; other metrics are not considered for individual videos.

3. Evaluation for Entire Database Suppose in the entire dataset: - There are V videos containing M macro-expression (MaE) sequences and M micro-expression (ME) sequences, where M = Σ m and M = Σ m ; - The method spots N MaE intervals and N ME intervals in total, where N = Σ n and N = Σ n ; - There are A TPs for MaE and A TPs for ME in total, where A = Σ a and A = Σ a .

The dataset can be considered as one long video. Results are first evaluated separately for MaE spotting and ME spotting, then overall results for macro- and micro-expression spotting are evaluated. Recall and precision for the entire dataset are calculated as follows:

- For macro-expression:

$$\text{Recall}_{MaE}^{D} = \frac{A_1}{M_1}, \quad \text{Precision}_{MaE}^{D} = \frac{A_1}{N_1}$$

- For micro-expression:

$$\text{Recall}_{ME}^{D} = \frac{A_2}{M_2}, \quad \text{Precision}_{ME}^{D} = \frac{A_2}{N_2}$$

- For overall evaluation:

$$\text{Recall}_D = \frac{A_1 + A_2}{M_1 + M_2}, \quad \text{Precision}_D = \frac{A_1 + A_2}{N_1 + N_2}$$

Then, F1-scores for all three evaluations are obtained based on:

$$F1\text{-score} = \frac{2 \times (\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}}$$

The challenge champion will be determined by the best overall score for spotting both micro- and macro-expressions.

III. Results and Discussion

For parameter p, we studied evaluation results by varying p from 0.01 to 0.99 with a step-size of 0.01. The results from 0.01 to 0.20 are shown in [TABLE:II]. In [TABLE:II], we list TP counts and F1-scores for macro- and micro-expression spotting separately. We observe that for both expression types in both datasets, the number of TPs decreases as p increases. Regarding F1-score, it also shows a decreasing trend in SAMM Long Videos. However, in CAS(ME)^2, the F1-score initially increases before beginning to decrease. This initial increase in CAS(ME)^2 occurs because the total number of predicted intervals (n) becomes smaller as p increases, causing precision (a/n) to increase.

Since the TP count is an important metric for spotting result evaluation, we select results under p = 0.01 as the final baseline results. Detailed final baseline results for spotting macro- and micro-expressions are shown in [TABLE:III]. For CAS(ME)^2, the F1-scores are 0.1196 for macro-expressions and 0.0082 for micro-expressions, with an overall F1-score of 0.0376. For SAMM Long Videos, the F1-scores are 0.0629 for macro-expressions and 0.0364 for micro-expressions, with an overall F1-score of 0.0445. Additional details regarding the number of ground-truth labels, TPs, FPs, FNs, precision, recall, and F1-scores for various conditions are shown in [TABLE:III].

IV. Conclusions

This paper addresses the challenge of spotting macro- and micro-expressions in long video sequences and provides baseline methods and results for the Third Facial Micro-Expression Spotting Challenge (MEGC 2020). We employ Main Directional Maximal Difference Analysis (MDMD) [**?**] as the baseline method, adjusting parameter settings for CAS(ME)^2 and SAMM Long Videos for the MEGC 2020 spotting challenge. Slight modifications are made to predict more reasonable intervals during post-processing. Experiments were conducted and

predicted results were evaluated using MEGC 2020 metrics. The results demonstrate that the MDMD method can produce reasonable performance, but there remains a significant challenge in reducing the number of false positives.

References

[1] G. Bradski. The OpenCV Library. *Dr. Dobb' s Journal of Software Tools*, 2000.
[2] 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*, 9(1):116–129, 2018.
[3] J. Endres and A. Laidlaw. Micro-expression recognition training in medical students: a pilot study. *BMC Medical Education*, 9(1):47, 2009.
[4] M. Frank, D. Kim, S. Kang, A. Kurylo, and D. Matsumoto. Improving the ability to detect micro expressions in law enforcement officers.
[5] 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. 1966.
[6] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(3):1755–1758, 2009.
[7] J. Li, C. Soladie, R. Sguier, S. Wang, and M. H. Yap. Spotting micro-expressions on long videos sequences. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–5, 2019.
[8] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*, 9(4):563–577, 2018.
[9] 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. In *IAPR Asian Conference on Pattern Recognition*, pages 665–669, 2015.
[10] D. Matsumoto and H. S. Hwang. Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion*, 35(2):181–191, 2011.
[11] A. Moilanen, G. Zhao, and M. Pietikäinen. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *International Conference on Pattern Recognition*, pages 1722–1727, 2014.
[12] 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*, 96(1):81–92, 2013.
[13] 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*, 9(4):424–436, 2017.
[14] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang. MEGC 2019—the second facial micro-expressions grand challenge. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–5, 2019.

[15] T. Senst, V. Eiselein, and T. Sikora. Robust local optical flow for feature tracking. 22(9):1377–1387, 2012.
[16] M. Shreve, J. Brizzi, S. Fefilatyev, T. Luguev, D. Goldgof, and S. Sarkar. Automatic expression spotting in videos. *Image and Vision Computing*, 32(8):476–486, 2014.
[17] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar. Macro- and micro-expression spotting in long videos using spatio-temporal strain. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 51–56, 2011.
[18] M. Shreve, S. Godavarthy, V. Manohar, D. Goldgof, and S. Sarkar. Towards macro- and micro-expression spotting in video using strain patterns. In *Workshop on Applications of Computer Vision*, pages 1–6, 2009.
[19] S.-J. Wang, S. Wu, X. Qian, J. Li, and X. Fu. A main directional maximal difference analysis for spotting facial movements from long-term videos. *Neurocomputing*, 230:382–389, 2017.
[20] 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, 2011.
[21] 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.
[22] C. H. Yap, C. Kendrick, and M. H. Yap. A spontaneous facial micro- and macro-expressions dataset. preprint arXiv:1911.01519, 2019.
[23] M. H. Yap, J. See, X. Hong, and S.-J. Wang. Facial micro-expressions grand challenge 2018 summary. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 675–678. IEEE, 2018.
[24] 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*, 6:71143–71151, 2018.

Note: Figure translations are in progress. See original paper for figures.