SLIDING WINDOW BASED MICRO-EXPRESSION SPOTTING: A BENCHMARK

Thuong-Khanh Tran, Xiaopeng Hong, and Guoying Zhao

Center for Machine Vision and Signal Analysis,
University of Oulu, Finland,
{khanh.tran, xiaopeng.hong, guoying.zhao}@oulu.fi

Abstract. Micro-expressions are very rapid and involuntary facial expressions, which indicate the suppressed or concealed emotions and can lead to many potential applications. Recently, research in micro-expression spotting obtains increasing attention. By investigating existing methods, we realize that evaluation standards of micro-expression spotting methods are highly desired. To address this issue, we construct a benchmark for fairer and better performance evaluation of micro-expression spotting approaches. Firstly, we propose a sliding window based multi-scale evaluation standard with a series of protocols. Secondly, baseline results of popular features are provided. Finally, we also raise the concerns of taking advantages of machine learning techniques.

Keywords: Affective computing, Micro-expression spotting, Evaluation protocols, Multi-scale analysis, Sliding window based.

1 Introduction

Micro-expressions are brief involuntary facial expressions which occur when people are trying to hide true feelings or conceal emotions. Micro-expressions play an important role in psychology which helps understand spontaneous emotions. Indeed, studying micro-expression facilitates to build an useful human behavior understanding system which can be applied in various fields such as medicine, criminal investigation, and business [1, 2]. Therefore, this topic has been attracting attentions from diverse areas, such as psychology and computer science.

The research of micro-expression can be briefly divided into two tasks: recognition and spotting, which determine the category of emotional states and locate positions of micro-expression in videos, respectively. To our best knowledge, most of the existing micro-expression studies focus on the recognition task [3–6], while the spotting problem is rarely addressed. For real world applications, micro-expression position must be determined as the first step before emotion recognition or interpretation. Spotting remains an extremely difficult task because of the subtle and short-duration facial movements.

This paper focuses on the micro-expression spotting task. There are previous studies involved in this topic [1, 2],[6, 7]. However, there are still several issues remaining.

First, it's difficult to make fair comparisons between existing methods under different test setups. In addressing this issue, experiments must be conducted under the same evaluation settings. It requires an establishment of protocol to standardize the performance evaluation.

Second, subtle movement such as small head movement and illumination effects usually cause previous spotting methods to have failures. They are small changes occurring between continuous frames, just as micro-expressions. Thus, these extrinsic changes may lead to false alarms on spotting. To deal with this issue, we need a method which can distinguish micro-expressions from other changes in consecutive frames. Thus, we tackle the spotting problem by detecting consecutive frames (which is termed a scanning window hereinafter) rather than one single frame. By leveraging machine learning techniques, we build models to classify scanning windows into two types: micro-expressions and non micro-expressions. Recent advance in machine learning provides powerful tools for robust classification to avoid the effect of small head movement and illumination.

Third, existing methods still spot micro-expression in single-scale. These methods can face many challenges caused by the heterogeneity of micro-expression length. To overcome this issue, we propose detecting micro-expression on various lengths of video. Thus, we apply multi-scale analysis in micro-expression spotting.

Overall, in this paper, we make the following contributions:

Evaluation protocols: We standardize the comparisons of spotting methods by building a series of protocols to make the evaluation more consistently and effective.

Multi-scale sliding window based approach: We propose a multi-scale sliding window framework. To our best knowledge, there are no previous studies combining sliding window based method and multi-scale analysis to spot micro-expression.

With the proposed framework and the designed protocols, we evaluate several widely-used methods. We target two goals at 1) selecting potential detectors and 2) offering the baseline results for micro-expression spotting. As a result, it provides a benchmark of micro-expression spotting for future studies.

This paper is organized as follows: Section 2 surveys methods of micro-expression spotting in previous studies. In Section 3, proposed benchmark is detailed. In Section 4, we report the results of performance evaluation under various of parameters and test setups. Finally, Section 5 concludes our research and discusses the future works.

2 Related Work

Micro-expression spotting is to automatically detect the frames where micro-expression takes place from a continuous sequence. There are previous studies involved to this topic. However, these methods still have some limitations. In order to get knowledge about remaining issues, we briefly survey existing methods of micro-expression spotting.

Moilanen et al. utilized Chi-Square distance of Local Binary Pattern (LBP) to spot spontaneous micro-expression in fixed-length scanning windows [7]. Their results were evaluated on both CASME and SMIC-VIS-E dataset. Patel et al. proposed calculating optical flow vector for small local spatial regions, then using heuristics algorithm to filter out non-micro-expression [8]. Wang et al. suggested a method named Main Directional Maximal Differences which utilizes the magnitude of maximal difference in the main direction of optical flow [9]. Their experiments were carried out on CASME2 dataset. Generally, almost all the methods focus on finding differences between non-micro and micro frames. They often calculate threshold to eliminate false alarms: for example, head movement or illumination effect. There are few studies using sliding window based method or machine learning techniques to detect micro-expression. Xia et al. made the first attempt utilizing machine learning for micro-expression spotting [10]. In this research, Adaboost was utilized to predict whether the probability of a duration of frames belonging to a micro expression or not. Random walk functions were used to integrate and refine the output of Adaboost and obtain the final result. The experiments were conducted on the SMIC and CASME dataset. However, study of Xia [10] still has limitation: using only one test setup. Author uses splitting dataset into training and testing subsets by a ratio of 30%/70%. Thus, it's difficult to make fair comparisons with other test setups, for example, leave-one-subject-out test setup.

Comparing to previous works, our study improves the drawbacks by suggesting a benchmark including three contributions: 1) standardizing evaluation protocols, 2) combining sliding window based method and multi-scale to spot micro-expression, 3) providing baseline results based on the proposed method and designed protocols.

3 Proposed Benchmark

As introduced, we present the proposed benchmark including multi-scale sliding window based framework and evaluation protocols in this Section. We begin with the overview of framework in first Sub-section. In second Sub-section, sampling data is detailed. Next, classification method is introduced. Then, we describe test setups and performance measurement in Sub-section 3.4 and 3.5, respectively.

3.1 Overview of framework

The framework of micro-expression spotting has four main components as illustrated in Fig. 1 : 1) multi-scale analysis by Temporal Interpolation Model (TIM),

2) sliding window based sampling, 3) binary classification, and 4) the integration and refinement of the output of the classifiers by for example non-maximum suppression.

In the first component, multi-scale analysis is employed to scale video sequence in temporal space. Temporal Interpolation Model (TIM) [11] is utilized in this component. TIM inputs original video sequence and interpolates the frames to produce a particular number of output frames. As illustration in Fig. 1, we input a video sequence then we obtain video sequences with interpolated frames by scaling values. In utilizing multi-scale analysis, micro-expression can be detected on various lengths of video sequence.

In second component, we carry out sampling data based on sliding window based method. In this technique, a scanning window is slid across positions of video sequences on all scales to obtain samples. These samples are categorized as either micro or non-micro. We label samples by calculating intersection of scanning window sample and ground truth. For more detail, we describe sampling data in next Section.

After sampling data, binary classification is employed to predict micro expression positions. The classification model is trained by a modern machine learning techniques such as SVM, Adaboost. This component has sub-tasks which are explained clearly in Sub-section 3.3. After this step, micro-expression positions are predicted by the responses of classifier. Thus, the evaluation can be conducted to measure performance of classifiers in this stage.

However, there are two drawbacks which can affect the evaluation results in whole video. First, binary classification step is returning multiple micro-expression samples around the ground truth, instead of only one. Second, results are existing false positive samples. For addressing these issue, we suggest post-processing task including two targets: 1) merging or integrating the nearby detected samples, 2) removing the redundant and false positive samples. Thus, we utilize Non Maximal Suppression (NMS) in last component. The output of last component is the final result of whole video evaluation.

3.2 Data Sampling

Based on proposed approach, we tackle micro-expression detection as a binary classification problem based on a window sliding across positions in video sequence. Therefore, sampling data needs to be focused first. One sample is considered as one position of scanning window.

Given a video having T frames, our method determines the sample set V as:

$$\begin{aligned} V &= [A_1, B_1] \cup ... \cup [A_i, B_i] \cup ... \cup [A_M, B_M] \\ M &= floor(\frac{T-L}{s}) + 1 \end{aligned} \tag{1}$$

where $[A_i, B_i]$ is sample starting from frame A_i to frame B_i, $A_1 = 1$, $A_{i+1} = A_i + s$, $A_i (i = 1..M)$ is index of first frame in one sample, $B_i = A_i + L - 1$ is

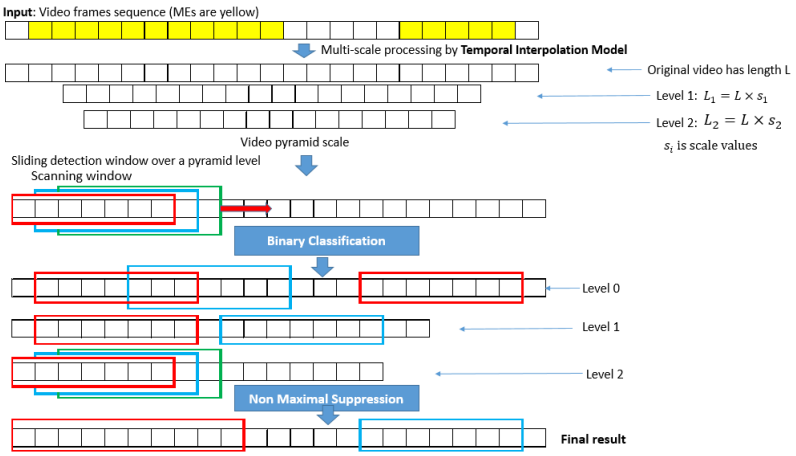


Fig. 1. Illustration of sliding window based method for micro-expression spotting.

the index of last frame in one sample, M is the number of samples, L is length of scanning-window and s is stride value.

Then, by using supervised learning approach, samples need to be labeled positive or negative (for micro-expression spotting, it is micro or non-micro) in dataset. We propose using overlap rate of scanning-window sample X_W and ground truth X_G in video to decide the positive and negative of samples:

$$\begin{aligned} positive &: \frac{X_G \cap X_W}{X_G \cup X_W} \geq \varepsilon \\ negative &: \frac{X_G \cap X_W}{X_G \cup X_W} < \varepsilon \end{aligned} \tag{2}$$

In our research, we select $\varepsilon = 0.5$ according to the study in [10]. This selection has a drawback when the overlap rate of ground truth and scanning window is less than 0.5 in any cases on one video. For example, if the ground truth has length 4 and scanning window has length 9, they totally can't match positive samples by formula (2). Thus, it can causes missing of positive samples on one video. In dealing with this issue, we propose extending or normalizing ground truths to same size of scanning windows.

3.3 Binary Classification

Now, we present the classification method utilized to provide baseline results. As mentioned, our method for dealing with micro-expression spotting is sliding window based method. A scanning window is slid across positions of video sequence. In each position, we run a micro or non-micro classifier to determine if the current clip is micro-expression or not.

There are four main steps as followings: (1) Face detection, (2) Face alignment, (3) Feature extraction, (4) Micro expression classification.

Firstly, face detector and KTL tracking algorithm are utilized to locate the whole faces through video [12]. To carry out face alignment, 68 land-marks points are located by Active Shape Model method [13] on the first frame of each video and they are aligned to the model face by using Local Weighted Mean (LWM) [14]. After that, face area is cropped by using the defined rectangles from the eye landmarks. Next, the feature descriptors based on the scanning windows are extracted. Our method utilizes three descriptors mentioned in the research of Li [1, 2]:

- Local Binary Pattern for Three Orthogonal Planes (LBP-TOP): this feature was proposed by Zhao and Pietikäinen [15] and considered as the extension of LBP for dynamic texture analysis in spatial-temporal do-main. It was widely utilized for facial-expression recognition and also for micro-expression recognition [1, 2, 15].
- Histogram of Oriented Gradient (HOG) for Three Orthogonal Planes (TOP). This feature is extended from HOG to 3D to calculate oriented gradients on three orthogonal planes XY, XT, and YT (TOP) for modeling the dynamic texture in video sequence.
- 3D extension of Histogram of Image Gradient Orientation (HIGO) which is a degraded variant of HOG. It ignores the magnitude and counts the responses of histogram bins. HIGO is also extended to three orthogonal planes (TOP) by the same way of HOG-TOP to utilize in video scenes.

In last step, Linear Support Vector Machine is used to classify the label of video clips: micro or non-micro.

3.4 Test Setup

We suggest using different test setups to obtain more informative and consistent results. Most of micro-expression datasets are organized by subjects. Each subject has a certain number of videos. So data can be splitted into training and testing subsets in terms of subject or video. Different setups can cause inconsistency in comparing methods. In order to standardize the evaluation, we design two setups for evaluation: "normal test" and "subject independent test" for splitting by videos and subjects, respectively.

Normal test: Videos from a dataset are selected randomly for the training set and the testing set. Division process is carried out by three various splitting rates: 30%/70%, 50%/50% and 70%/30%. Diverse splitting rates can help us analyze the effects of *train/test* ratio on the performance of detectors. Each detector is evaluated 10 times with different sets to estimate the accuracy.

Subject independent test: Also the leave-one-subject-out setup, one subject is used as testing set and other subjects are used as training set. A N-fold cross-validation is performed to train classifiers and estimate the accuracy of detectors.

3.5 Performance Measurement

In this Sub-section, we discuss methods which are used to compare micro-expression detectors. We describe performance measurements as below.

As introduced, our results are provided from binary classification step. The binary classification plays a central role in micro-expression spotting as it is the important step that leverages the advance in machine learning techniques for extracting the features and making decisions. Common methodology for evaluating binary classifiers is to measure their per-window performance on detected positive samples and negative samples. Per-window measurement is widely used to compare classifiers or to evaluate systems that perform sliding window based method [16].

However, we realize that per-window is not effective and accurate method for our target. Reconsideration, we require a spotting system having following steps: input a video, performing the multi-scale processing in video, returning the location and score of each detection and utilizing NMS to merge nearby detected-windows. Per-window does not take involvement in NMS and it does not measure errors caused by incorrect scales or positions. Fig. 2 illustrates the untested cases of per-window evaluation: multiple detected samples around ground truth, false positive samples and false negative samples caused by incorrect scales and positions. Additionally, not all proposed methods are based on sliding-window or binary classifier. Therefore, per-window measurement is hard to be applied for comparing these methods. Above reasons require us to select other measurement: per-video. Evaluation is carried out on the final list of detected windows, after handling the post-processing tasks and mentioned requirements. Counting false positive, true positive and missing samples are done by overlap matching between detected-window and ground truth. We re-use the formula (2) to determine the cases of matching (changing positive to matched and negative to unmatched): unmatched detected is false positive, unmatched ground truth is false negative and matched ground truth is true positive.

Overall, we suggest using two mentioned test setups and two performance measurements. To demonstrate the performance of detectors, we utilize Detection Error Tradeoff (DET) curve to compare evaluated detectors. We plot two types of measurement: miss rate against false-positive per-window (FPPW) and miss rate versus false-positive per-video (FPPV).

4 Implementation and Results

In this section, we perform evaluation results of selected features (HIGO-TOP, HOG-TOP, LBP-TOP) for spotting problem. Experimental results are reported by two test setups and two measurements, that is expected to provide the baseline results for future studies.

4.1 Implementation note

First, we briefly introduce our implementation. We select the public dataset SMIC-VIS-E extracted from SMIC [17] to conduct experiment. This dataset has

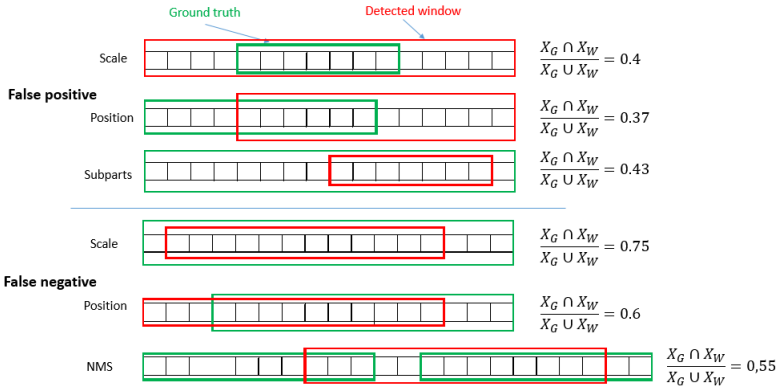
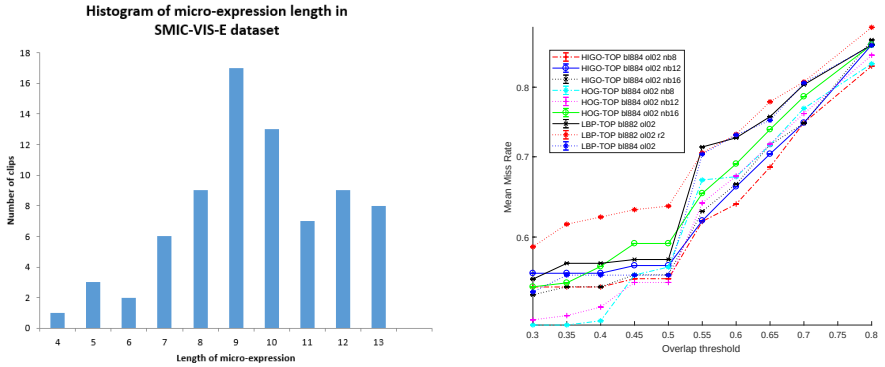


Fig. 2. Untested cases of per-window evaluation. These are several cases of not testing during per-window evaluation that can lead to false positive or false negative in whole video evaluation. False positive can arise from detection at incorrect scales or positions. False negative can arise from slight mis-alignments between detected window and ground truth positions or NMS.

76 video sequences with the frame size 640×480 pixels recorded at 25fps. It consists of 71 micro-expression videos and 5 non-micro videos. People in video are shown emotional clips and are instructed to hide their feelings. These videos are annotated with onset and offset frames. To select scanning window size, histogram of ground truth length is plotted. Based on the plotted histogram (see Fig. 3a), fixed-length of scanning window is set as 9 because it has the most number of micro-expression lengths. In SMIC-VIS-E dataset, there are some ground truths having size less than fixed-length. We extend these ground truths to 9 for adapting formula (2). To select value of ε in matching between ground truth and final detected window, diverse thresholds from 0.3 to 0.7 were tested. Fig. 3b shows the performance of nine detectors with mean miss rates corresponding to thresholds of ε. Based on plotted results on Fig. 3b, ε is determined below 0.55 for reducing miss rates. In our experiments, ε is selected by 0.5.

A feature has parameters which can affect results, thus we conduct experiments under various parameter settings to find the optimal combination. In our implementation, the block division is evaluated under three settings: $8 \times 8 \times 4$, $8 \times 8 \times 2$ and $6 \times 6 \times 4$. The overlap rate has values: 0.2, 0.3 and 0.5. Number of bins in features HIGO-TOP and HOG-TOP select values: 8, 12 and 16. On Fig. 4 and Fig. 5, name of descriptors explains feature used and parameter combination, *e. g.*, *HIGO-TOP bl884-ol02-nb8 means feature HIGO-TOP, block division* $8 \times 8 \times 4$, *overlap rate* 0.2 *and number of bin is 8.* In performing results, validation of each descriptor is conducted many times (both on normal and subject test setup). Normal test carries out 10 times with different random sets and subject independent test carries out 8 times by the number of subjects. Therefore, we report results by the mean values and standard deviation. FPPW selects 0.4 and FPPV selects 1 as the reference points for comparison.



(a) (b)

Fig. 3. Histogram of micro-expression sizes in dataset (a), and evaluation results of overlap rate thresholds between ground truth and final detected window (b).

In experiment, there were unexpected results when we were calculating FPPV of leave-one-subject-out setting. A subject only has two videos in dataset, thus FPPV is calculated incorrectly when evaluating by subject independent test setup. In addressing this issue, we propose formula (3) for computing FPPV in leave-one-subject-out protocol.

$$FPPV = \frac{\sum_{i=1}^{V} FP_i}{V} \tag{3}$$

where V is number of videos in dataset, FP_i is number of false-positive in video i^{th}. Meaning that we don't use mean and standard deviation values in this protocol. False-positive values are cumulated from all subject tests.

4.2 Per-window results

At first, we report the results that are evaluated on per-window measurement. These results demonstrate the performance of binary classifiers and effects of parameter combinations.

Normal test setup: We discuss on the performance of classifiers evaluated by normal test setup. Fig. 4a, Fig. 4b and Fig. 4c plot DET curves from the performance of each feature. We present results of three features separately for comparing parameter combinations. In this setup, we only report comparison from splitting rate 70%/30% because they got the best ones. From HIGO-TOP feature, optimal block division and overlap rate are determined by $8 \times 8 \times 4$ and 0.2. HIGO-TOP bl884-ol02-nb16 is best with mean miss rate of 26.73%. In HOG-TOP, HOG-TOP bl884-ol02-nb12 and HOG-TOP bl884-ol02-nb16 are the bests with mean miss rates of 25.61% and 25.39.10%, respectively. In LBP-TOP, combination of block division $8 \times 8 \times 4$ and overlap rate 0.2 outperforms other

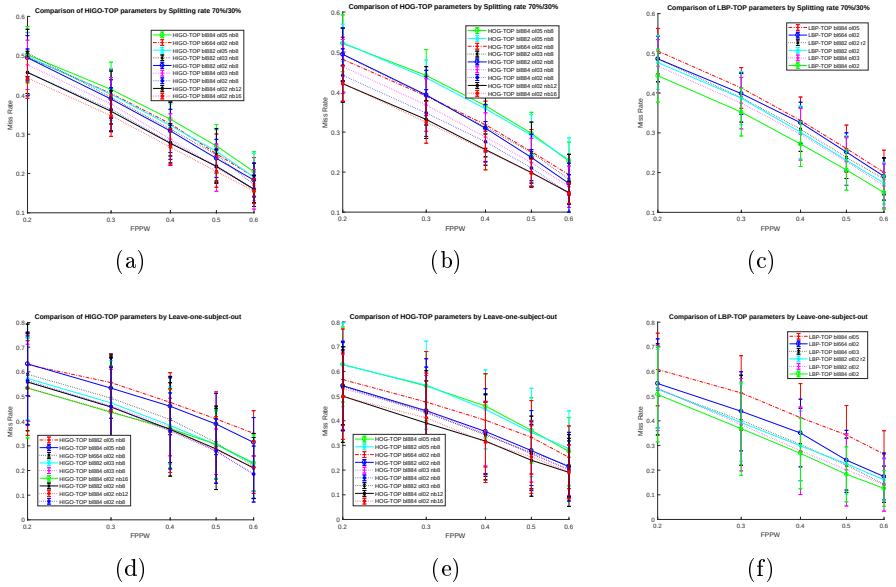


Fig. 4. Performance of binary classifiers on per-window measurement. (a), (b) and (c) present results evaluated by normal test setup of features: HIGO-TOP, HOG-TOP and LBP-TOP, respectively. (d), (e) and (f) present results evaluated by leave-one-subject-out setup of features: HIGO-TOP, HOG-TOP and LBP-TOP, respectively. Legends are ordered by performance (lower is better).

descriptors by mean miss rate of 27.19% at 0.4 FPPW. Overall, HOG-TOP is better than other features in this test setup.

Subject independent test setup: Next, we discuss performance of classifiers evaluated by leave-one-subject-out setup. Similar to normal test, each feature is performed separately on Fig. 4d, Fig. 4e and Fig. 4f. In first feature, HIGO-TOP bl884-ol02-nb12 is best one with mean miss rate of 38.47% at 0.4 FPPW. In HOG-TOP, HOG-TOP bl884-ol02-nb12 performs best with mean miss rate of 33.67%. LBP-TOP descriptors are better than other features in this test setup. LBP-TOP bl884-ol02 and LBP-TOP bl882-ol02 obtain mean miss rates of 26.79% and 27.89% at 0.4 FPPW, respectively.

4.3 Per-video results

In this section, we report results measured by FPPV. In these protocols, comparisons between various descriptors and effects of training/testing ratios are presented. Effects of parameter combinations are not reported because we obtain descriptors extracted from similar parameters of per-window measurement.

Normal test setup: Fig. 5a performs results evaluated by splitting rate 30%/70%. LBP-TOP-bl884-ol02 is best with mean miss rate of 66.61% at 1 FPPV. HOG-TOP-bl884-ol02-nb12 is second with mean miss rates of 68.18%.

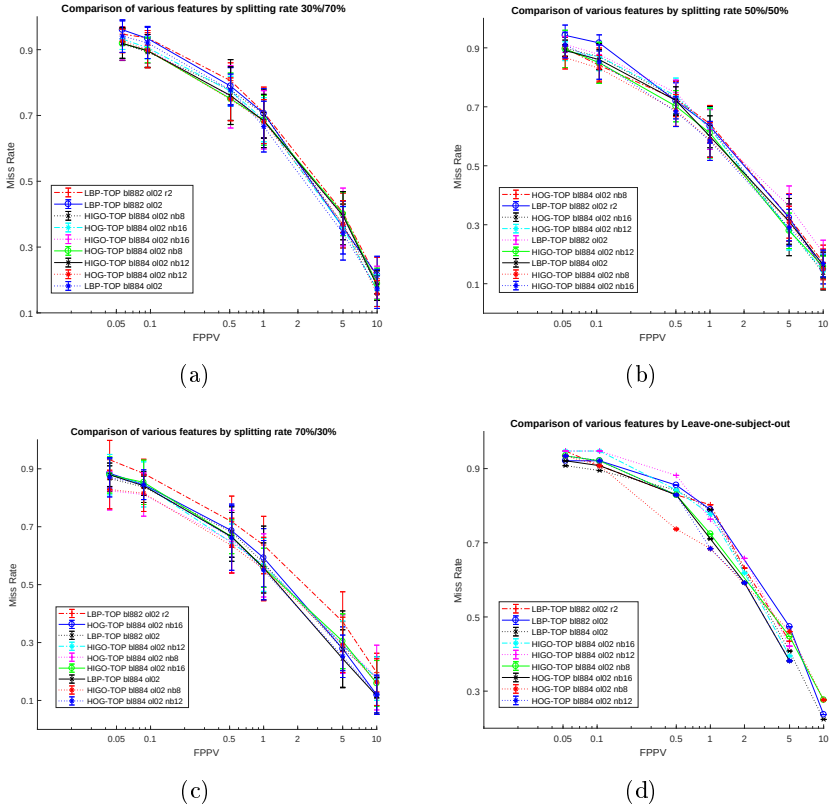


Fig. 5. Performance of detectors on per-video measurement. (a) , (b) , (c) show evaluation results using splitting rate 30%/70%, 50%/50% and 70%/30%, respectively. (d) Evaluation result by leave-one-subject-out test setup. Legends are ordered by performance (lower is better).

In splitting rate 50%/50%, we get the new order on performance of detectors. Fig. 5b shows the evaluation results of this scenarios. HIGO-TOP-bl884-ol02-nb8 and HIGO-TOP-bl884-ol02-nb16 outperform other descriptors by mean miss rates of 58%. In the last splitting rate, 70%/30%, HOG-TOP-bl884-ol02-nb12 is the best one with mean miss rate of 54.99% and HIGO-TOP-bl884-ol02-nb8 is second place with mean miss rate of 55.38% at 1 FPPV, see Fig. 5c.

Subject independent test setup: In last protocol, we show the performance of detectors evaluated by subject independent test. Fig. 5d demonstrates the results of last protocol. HOG-TOP outperforms other features in this test setup with mean miss rates of 68.42% at 1 FPPV (Both HOG-TOP-bl884-ol02-nb12 and HOG-TOP-bl884-ol02-nb12 have the same mean miss rate).

5 Conclusion and Discussion

In this paper, we propose a benchmark for micro-expression spotting to enable fair comparisons of different methods. Series of protocols and experimental settings such as the sliding window based scheme and multi-scale analysis are designed to standardize the evaluation. Baseline results of popular spotting methods are also provided.

In future, we plan to apply Action Unit detection to improve accuracy. Deep learning techniques [6] will also be explored for micro-expression detection.

References

1. Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikäinen, M.: Reading hidden emotions: spontaneous micro-expression spotting and recognition. arXiv preprint:1511.00423 (2015)
2. Li, X., Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikäinen, M.: Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Trans. Affect. Comput. (2017)
3. Polikovsky, S., Kameda, Y., Ohta., Y.: Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In: Proc. IET ICDP. (2009)
4. Wu, Q., Shen, X., Fu, X.: The machine knows what you are hiding: an automatic micro-expression recognition system. In: Proc. ACII. (2011) 152–162
5. Xu, F., Zhang, J., Wang, J.: Microexpression identification and categorization using a facial dynamics map. IEEE Trans. Affect. Comput. (2016)
6. Patel, D., Hong, X., Zhao, G.: Selective deep features for micro-expression recognition. In: Proc. ICPR. (2016)
7. Moilanen, A., Zhao, G., Pietikäinen, M.: Spotting rapid facial movements from videos using appearance-based feature difference analysis. In: Proc. ICPR. (2014)
8. Patel, D., Zhao, G., Pietikäinen, M.: Spatiotemporal integration of optical flow vectors for micro-expression detection. In: Proc. ACIVS. (2015)
9. Wang, S., Wu, S., Fu, X.: A main directional maximal difference analysis for spotting micro-expressions. In: ACCV 2016, Springer (2016) 449–461
10. Xia, Z., Feng, X., Peng, J., Peng, X., Fu, X., Zhao, G.: Spontaneous micro-expression spotting via geometric deformation modeling. CVIU **147** (2016) 87–94
11. Zhou, Z., Zhao, G., Pietikäinen, M.: Towards a practical lipreading system. In: CVPR 2011. (June 2011) 137–144
12. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. CVPR, IEEE (2001)
13. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models-their training and application. Computer vision and image understanding **61**(1) (1995) 38–59
14. Goshtasby, A.: Image registration by local approximation methods. Image and Vision Computing **6**(4) (1988) 255–261
15. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE TPAMI **29**(6) (2007)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. CVPR. (2005)
17. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikäinen, M.: A spontaneous micro-expression database: Inducement, collection and baseline. In: Proc. FG. (2013)