Technical Report

MICRO-EXPRESSION SPOTTING: A BENCHMARK

Xiaopeng Hong,* Thuong-Khanh Tran, Guoying Zhao
Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland

Micro-expressions are rapid and involuntary facial expressions, which indicate the suppressed or concealed emotions. Recently, the research on automatic micro-expression (ME) spotting obtains increasing attention. ME spotting is a crucial step prior to further ME analysis tasks. The spotting results can be used as important cues to assist many other human oriented tasks and thus have many potential applications. In this paper, by investigating existing ME spotting methods, we recognize the immediacy of standardizing the performance evaluation of micro-expression spotting methods. To this end, we construct a micro-expression spotting benchmark (MESB). Firstly, we set up a sliding window based multi-scale evaluation framework. Secondly, we introduce a series of protocols. Thirdly, we also provide baseline results of popular methods. The MESB facilitates the research on ME spotting with fairer and more comprehensive evaluation and also enables to leverage the cutting-edge machine learning tools widely.

This technical report is extended from an ACIVS17 paper. We are now expanding this work and updating this report when there are substantial achievements. The following citing information may be used for reference:

   'Thuong-Khanh Tran, Xiaopeng Hong, Guoying Zhao. Sliding-window based micro-expression spotting: A benchmark. In Proc. Advanced Concepts for Intelligent Vision Systems (ACIVS), 2017'.

1 Test

2 Introduction

Micro-expressions are brief involuntary facial expressions which occur when people are trying to hide true feelings or conceal emotions. Micro-expressions assist to understand spontaneous emotions and thus play an important role in psychology. Indeed, studying micro-expression facilitates to build a human behavior understanding system which can be useful in various fields such as medicine, criminal investigation, and business [10, 11]. Therefore, this topic has been attracting increasing attention from diverse areas, such as psychology and computer science.

   The research on micro-expression analysis can be briefly divided into two tasks: spotting and recognition, which locate positions of micro-expression in videos and determine the category of emotional states, respectively. To our best knowledge, most of the existing micro-expression studies focus on the recognition task [16, 17, 22, 24]. It may be because spotting is probably even more challenging than recognition. An oblique reference can be found in

---

   Emails: xiaopeng.hong@oulu.fi

the human test experiments in [11], where the automatic ME recognition algorithm outperformed general human beings while its accuracy became slight lower than human beings when the spotting task is added to the ME analysis system.

However, for real-world applications, micro-expression positions must be determined first before further emotion recognition or interpretation. To fill in this gap, in this paper, we focus on the micro-expression spotting task.

There are only several studies in ME spotting [11, 14, 15, 23]. Moilanen *et al.* utilized the Chi-Square distance of Local Binary Pattern (LBP) to spot micro-expressions in fixed-length scanning windows [14]. Patel *et al.* calculated optical flows for small local spatial regions, then used heuristics algorithm to filter out non-micro-expression [15]. Wang *et al.* suggested a method named Main Directional Maximal Differences to utilize the magnitude of maximal difference in the main direction of optical flow [21]. Basically, almost all methods focus on finding differences between non-micro and micro frames. They usually rely on empirically-set thresholds to eliminate false alarms caused by, for example, head movement or illumination effect. Xia *et al.* made the first attempt in utilizing machine learning for micro-expression spotting [23]. 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.

Obviously, there are issues remaining for these ME spotting methods. Firstly, it's difficult to make fair comparisons between existing methods as they were usually evaluated with different test setups. Secondly, subtle movements such as slight head movement and illumination effects are small changes occurring in continuous frames, just as micro-expressions. Thus, these extrinsic movements may lead to false alarms. Thirdly, existing methods usually spot micro-expression in single-scale. It probably causes problems as the length of micro-expression varies greatly.

In addressing the first issue, it is important that experiments can be conducted under the same evaluation settings. It is thus required to standardize the performance evaluation. For the second issue, we tackle the spotting problem by examining a sequence of consecutive frames (which is termed a scanning window hereinafter) rather than one single frame. Recent state-of-the-art machine learning methods thus can be leveraged to avoid the effect of extrinsic variations. To deal with the third issue, we introduce multi-scale analysis in micro-expression spotting. As a result, the efforts in address these issues give rise to a benchmark of micro-expression spotting for future studies.

The contributions of this paper are three-fold:

Multi-scale sliding window based approach: We propose a multi-scale sliding window framework. To our best knowledge, this is the first micro-expression spotting study combining the sliding window mechanism and multi-scale analysis.

Evaluation protocols: We standardize the comparisons of spotting methods by building a series of protocols.

Baseline results: With the proposed framework and the designed protocols, we evaluate several widely-used methods and offer baseline results for micro-expression spotting.

This paper is organized as follows: In Section 2, proposed benchmark is detailed. In Section 3, we report the results of performance evaluation under various of parameters and test setups. Finally, Section 4 concludes our research and discusses the future works.

3 Multi-scale Sliding-window ME Spotting

We detail the proposed benchmark including multi-scale sliding window based framework and the evaluation protocols in this Section.

3.1 Overview of framework

The framework of micro-expression spotting has four main mechanisms as illustrated in Fig. 1 : 1) multi-scale analysis, 2) sliding window based sampling, 3) binary classification, and 4) the integration and refinement of the outputs of binary classifiers.

Firstly, we employ multi-scale analysis to rescale video sequences in temporal space, so that micro-expression of various lengths can be detected using a fix scanning window. Temporal Interpolation Model (TIM) [26] has been
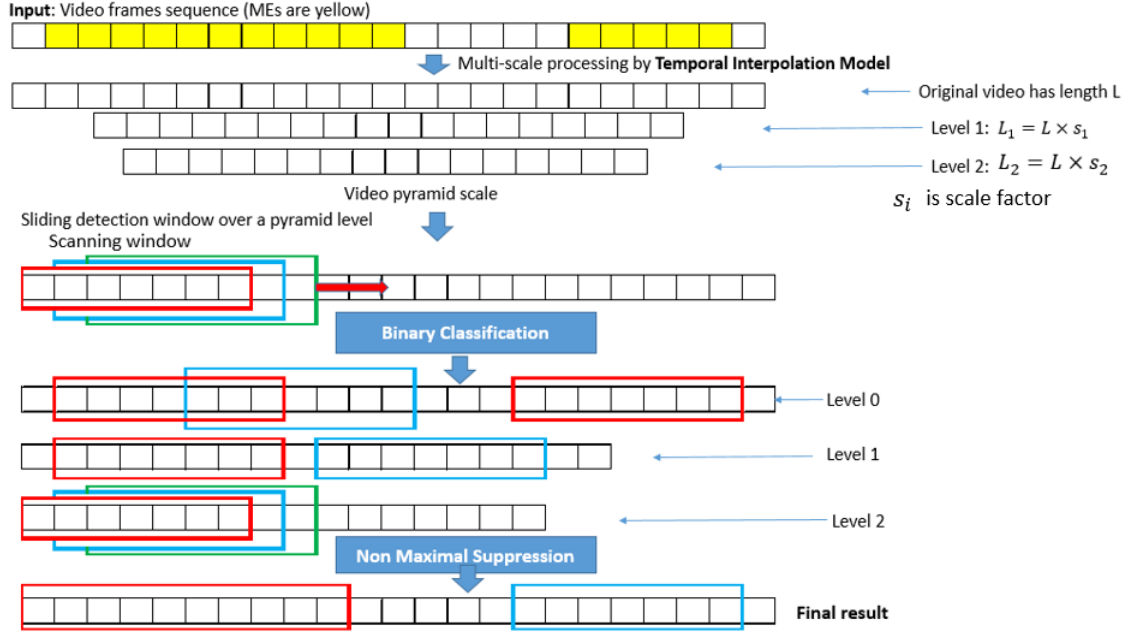
Figure 1: *Illustration of sliding window based method for micro-expression spotting.*

successfully applied to ME analysis [11] and thus is utilized. It takes a video sequence as input and interpolates the frames to produce a sequence with a particular number of frames.

Secondly, we sample data by using a fix-length scanning window, which is slid along all (allowed) positions of video sequences across all scales of interests to obtain samples. These samples are labeled as micro or non-micro. More detail will be described in the next sub-section.

Thirdly, provided a set of samples with *known* binary labels (*a.k.a* the training set), it is feasible to use modern machine learning techniques such as SVM [1], Boosting [6], random forest [13], or the cutting-edge deep learning methods [9] to optimize a binary classifier, so that the labels of *unknow* samples can be predicted. This sub-tasks will be explained further in sub-section 3.3.

Finally, in order to get the unique final result, it is required to integrate the binary decisions made on the densely sampled and probably overlapping scanning windows. We thus suggest a post-processing mechanism for 1) merging or integrating the nearby detected samples, and 2) removing the redundant and false positive samples. It can be accomplished by a temporal Non Maximal Suppression (NMS), which can be easily implemented by slightly modifying from the version for spatial domain [3, 18].

3.2 Data Sampling

In this paper, we tackle micro-expression spotting by considering it as a binary classification task, which is carried out on a set of samples, namely fixed-length sequences of frames. A set of samples, especially a training set of samples with *known* labels should be built first to optimize the parameters of a chosen model. More specifically, given a video of T frames, we slide a scanning window of L frames to have a sample set V as:

$$V = [A_1, B_1] \cup ... \cup [A_i, B_i] \cup ... \cup [A_M, B_M] \tag{1}$$

where $[A_i, B_i]$ is a sample starting from the frame A_i to B_i. $A_{i+1} = A_i + s$ and $B_i = A_i + L - 1$ for $(i = 1..M)$ are indexes of the first frames and last frames of sample, respectively, and s is the sliding stride. Obviously, the number of samples $M = floor(\frac{T-L}{s}) + 1$.

Then, by calculating the Intersection over Union of the X_W and ground truth X_G (as widely suggested by the object detection challenges such as the popular PASCAL VOC challenge [5]), we determine the labels of samples as positive (for micro-expressions) and negative (for non micro-expressions):

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

We set $\varepsilon = 0.5$ following the suggestion of [5].

3.3 Binary Classification

Binary classification plays a central role in micro-expression spotting as it enables to leverage the advance in machine learning techniques for extracting the features and making binary decisions. We follow [11] and employ three descriptors of the scanning windows and linear support vector machine to classify the samples.

- Local Binary Pattern for Three Orthogonal Planes (LBP-TOP) [25]: LBP-TOP is an extension of LBP in spatial-temporal do-main. A recent efficient implementation of LBP-TOP can be found in [8].

- Histogram of Oriented Gradient for Three Orthogonal Planes (HOG-TOP): HOG-TOP is extended from HOG to 3D to calculate oriented gradients on three orthogonal planes XY, XT, and YT for modeling the dynamic texture in video sequence [11].

- Histogram of Image Gradient Orientation for Three Orthogonal Planes (HIGO-TOP) [11]: HIGO-TOP is a 3D extension of Histogram of Image Gradient Orientation (HIGO). HIGO, as a degraded variant of HOG, ignores the magnitude and counts the responses of histogram bins.

3.4 Test Setup

In order to standardize the evaluation, cross-validation should be performed for splitting the set of samples into a training set and a testing set. Two test setups, namely "normal random sampling test" and "subject independent test" are suggested.

Normal random sampling test: Samples are randomly selected for the training set and the testing set. Division process is carried out by three suggested various splitting rates: 30%/70%, 50%/50% and 70%/30%. It is recommended to carry out 10-fold cross-validation for reliable performance evaluation.

Subject independent test: In the each fold of the leave-one-subject-out setup which is previously established for ME recognition [11], samples of one subject is used for testing and the others are for training.

3.5 Performance Measurement

In this sub-section, we discuss on the measurement for comparing micro-expression detectors. According to different semantic levels, there are two measurement.

The first one is used to directly evaluate performance of the binary classification. It measures the accuracy on detected positive samples and negative samples. As the objects it counts are sample which are obtained by sliding windows, this measurement is named per-window measure.

It is pointed out in the field of pedestrian detection that the per-window measure has its limitations [4]. Recall that the target of spotting is to locate when a ME starts and ends, it is the final outputs that the performance of spotting can only be accurately evaluated by examining. However, as the per-window measure is obtained before NMS is carried out, it can only briefly reflect the performance of spotting. Fig. 2 illustrates the cases which are out of the scope of per-window evaluation: multiple detected samples around ground truth, false positive samples and false negative samples caused by incorrect scales and positions.

For these reasons, we introduce the per-video measure as the second measurement. Evaluation is carried out on the final outputs of detected windows, after handling the post-processing tasks and mentioned requirements. False positive, true positive and missing samples are determined and counted by Eq. 2.
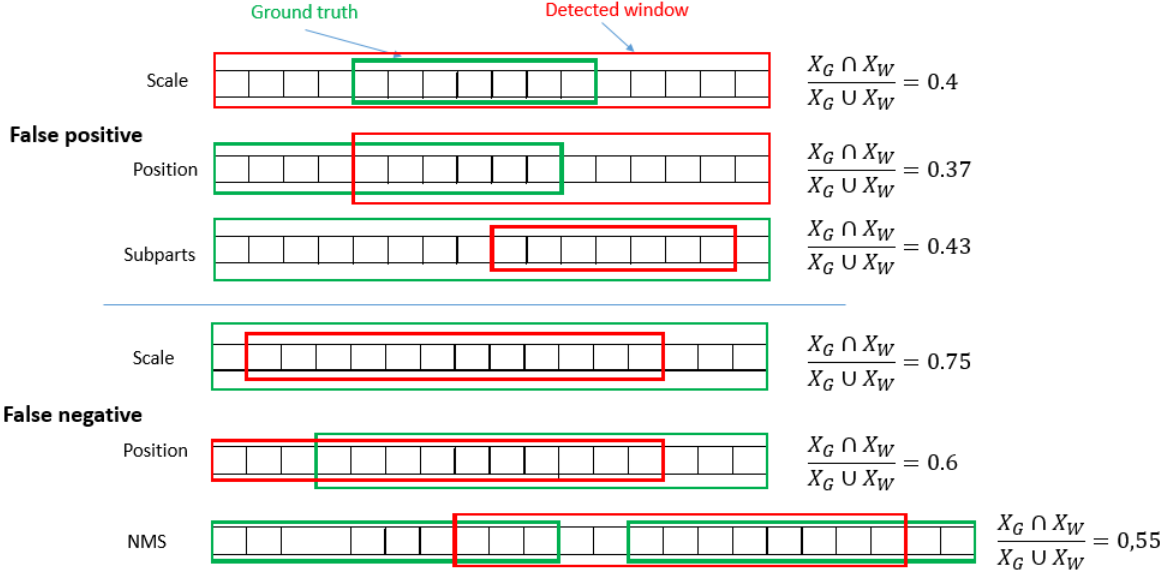

Figure 2: *Untested cases of per-window evaluation. These are several cases of not testing during per-window evaluation that can lead to false positive or false negative in whole video evaluation. False positive can arise from detection at incorrect scales or positions. False negative can arise from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

To provide a comprehensive performance evaluation of ME spotters, we utilize Detection Error Tradeoff (DET) curve. As a result, there are two types of DET curves accordingly: 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 provide 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.

4.1 Implementation note

We use the SMIC-VIS-E dataset [12] to build the benchmark and conduct experiments. This dataset has 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. These videos are annotated with onset and offset frames.

All the frames on this dataset are pre-processed as follows [11]: face detector and KTL tracking algorithm are utilized to locate the whole faces through video [20]. 68 land-marks points are located by Active Shape Model method [2] on the first frame of each video and they are aligned to the model face by using Local Weighted Mean (LWM) [7]. After that, face area is cropped by using the defined rectangles from the eye landmarks.

To determine the length of the scanning windows, we count the frequency of the length of ground truths. Based on the plotted histogram (in Fig. 3), the length of scanning windows is set to 9 as it has the most number of micro-expression lengths. For those MEs having length other than the fixed length, we slightly modify and then fix their lengths to 9 for simplification. In our experiments, ε in matching between ground truth and final detected window is selected by 0.5.
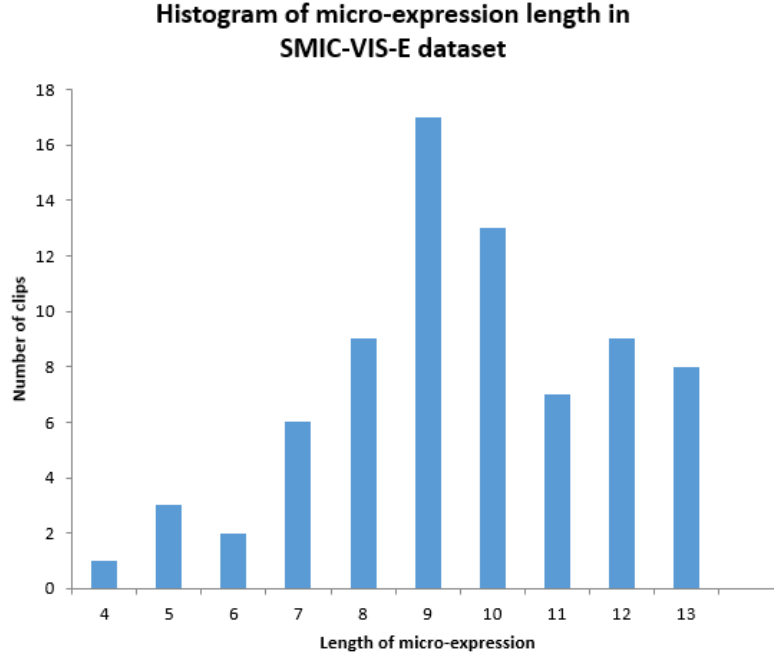
Figure 3: *Histogram of micro-expression sizes in dataset.*

Note that a feature usually has parameters which highly 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 $6 \times 6 \times 4$. There are three optional overlap rates: 0.2, 0.3 and 0.5. The number of bins for HIGO-TOP and HOG-TOP are selected from: 8, 12 and 16. In Fig. 4 and the following figures, 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.*

The evaluation experiments are conducted under both normals random sampling and subject test setups.

For normal random sampling test, we carry out 10 times random tests and report the means and standard deviations of the miss rates which are calculated corresponding to a series of fixed false positive rates.

For subject independent test, we carry out leave one subject out cross validation. As the tests for random sampling, We also report results with respect to both FPPW and FPPV measure in this test. Nevertheless, there is one issue worth of mentioning for reporting results in FPPV. As some subject only has two videos available, there is not enough points to draw a smooth DET curve for the experiment of this fold.

To address this issue, we use Eqs. 3 & 4 for computing the (overall) FPPV and miss rate over all videos on the dataset, rather than the mean FPPV and miss rate of those on each fold of tests.

$$FPPV_{OverAll} = \frac{\sum_{j=1}^{8} FP_j}{N} \tag{3}$$

$$MissRate_{OverAll} = 1 - \frac{\sum_{j=1}^{8} TP_j}{N^+} \tag{4}$$

where N is number of videos in dataset ($N = 76$ in SMIC-VIS-E), FP_j is number of false-positive in testing subject j^{th}, FP_j is number of true-positive in testing subject j^{th}, and N^+ is the total number of ground truth MEs.
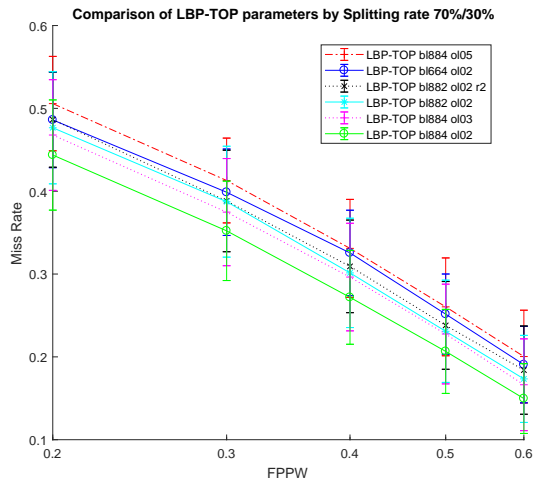
4.2 Per-window results

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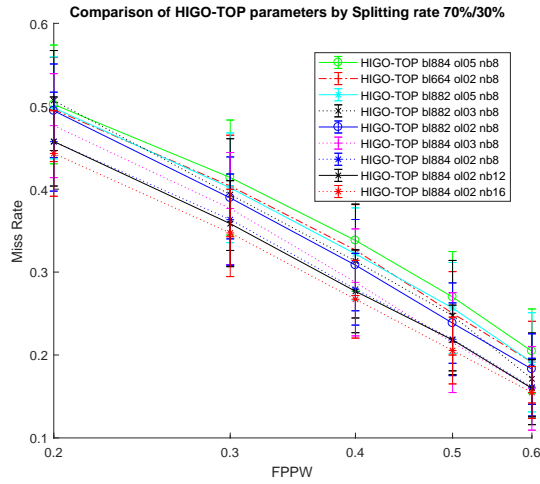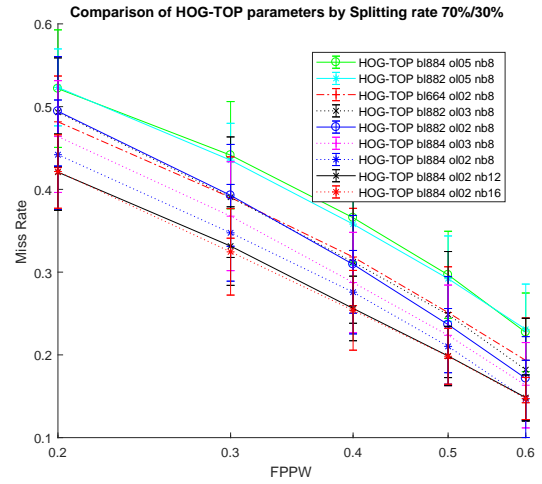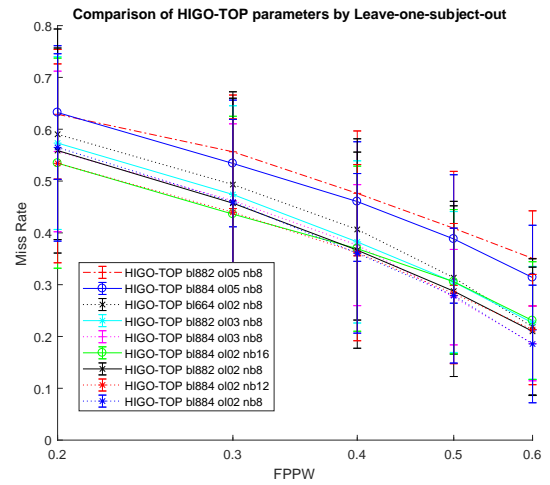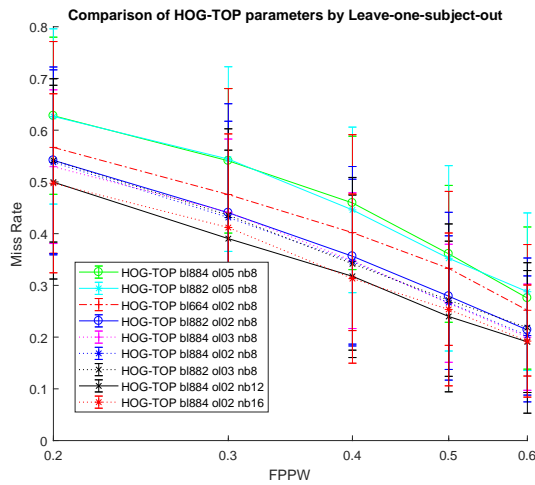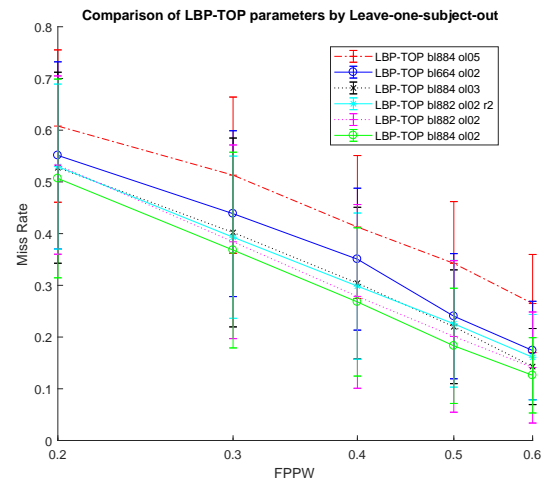
Figure 4: *Performance of binary classifiers on per-window measurement. (a), (b) and (c) present results evaluated by normal test setup of features: HIGO-TOP, HOG-TOP and LBP-TOP, respectively. (d), (e) and (f) present results evaluated by leave-one-subject-out setup of features: HIGO-TOP, HOG-TOP and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

miss rates at a fixed FPPW of 0.4 as a reference points when comparing various methods.
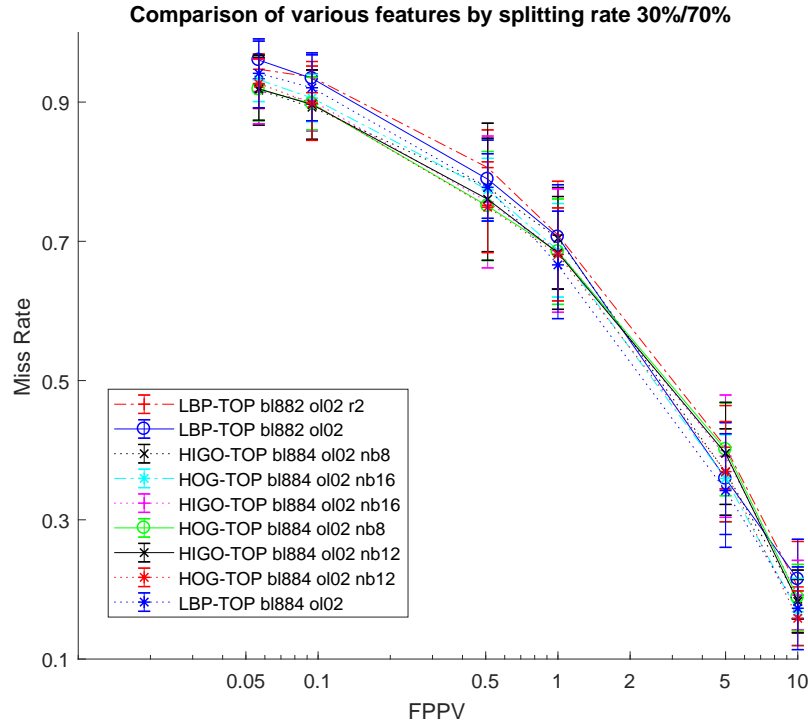

Figure 5: *Evaluation results on per-video measurement, using a splitting rate of* 30%/70%.

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 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. As the case with respect to the per-window measure, we also use the mean miss rates at a fixed FPPV of 1.0 as a reference points when comparing various methods.

Normal test setup: Fig. 5 provides 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%. In splitting rate 50%/50%, we get the new order on performance of detectors. Fig. 6 shows the evaluation results of this scenarios. HIGO-TOP-bl884-ol02-nb8 and HIGO-TOP-bl884-ol02-nb16 outperform other descriptors by
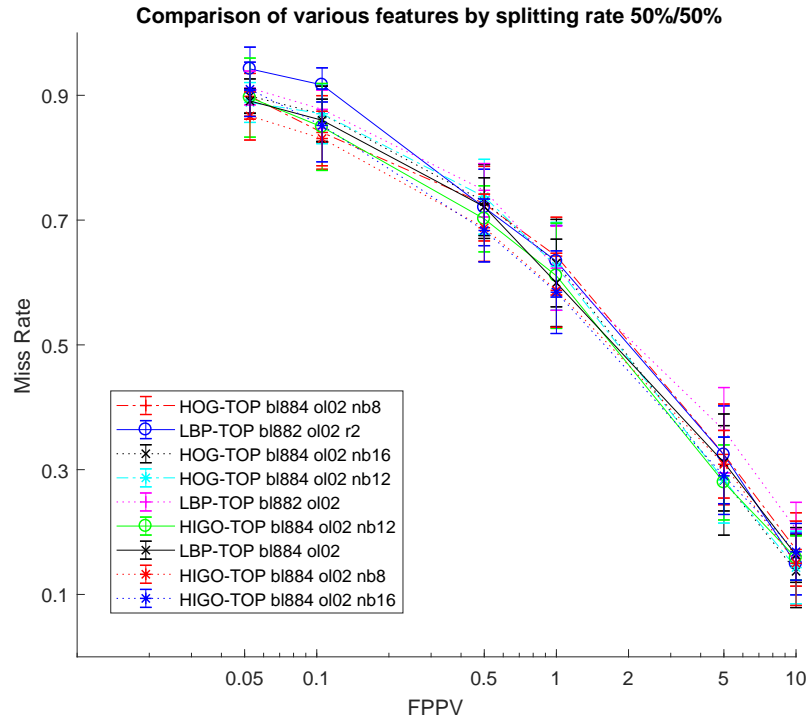
Figure 6: *Evaluation results on per-video measurement, using a splitting rate of* 50%/50%.
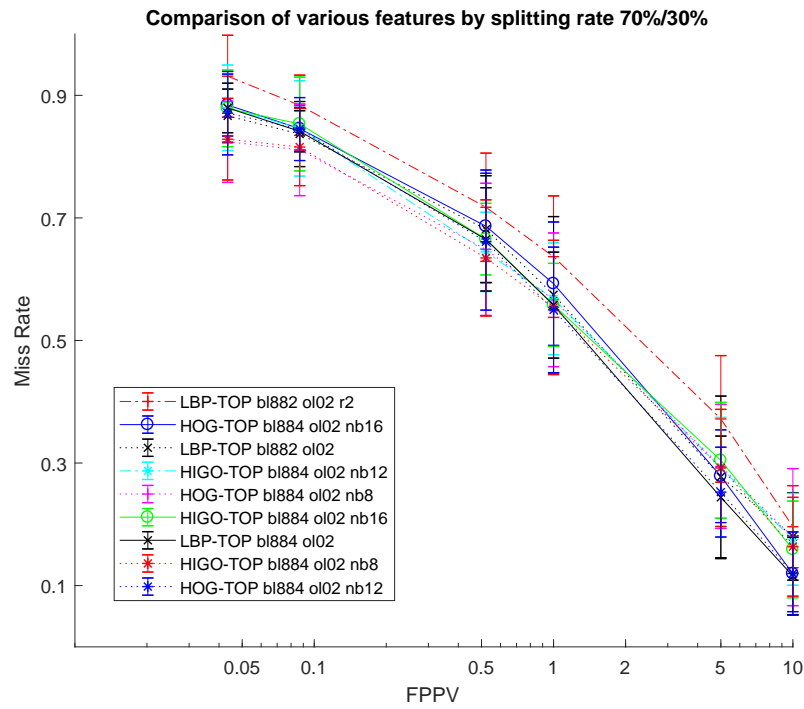

Figure 7: *Evaluation results on per-video measurement, using a splitting rate of* 70%/30%.
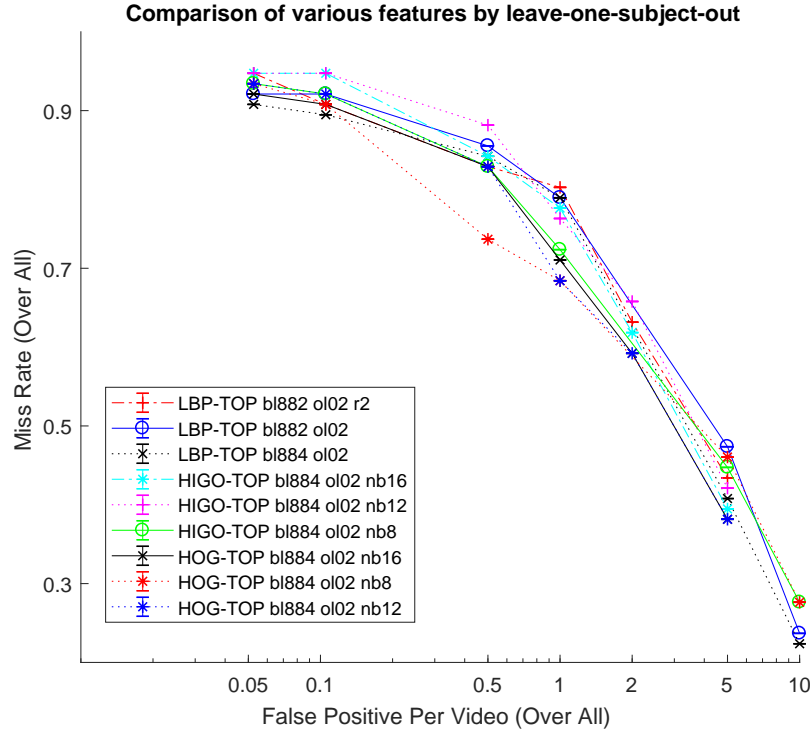
Figure 8: *Evaluation results on per-video measurement, under leave-one-subject-out tests.*

mean miss rates of 58%. In the last splitting rate (Fig. 7), 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.

Subject independent test setup: In last protocol, we show the performance of detectors evaluated by subject independent test. Fig. 8 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 construct a micro-expression spotting benchmark. A series of protocols and experimental settings such as the sliding window based mechanism and multi-scale analysis are introduced to standardize the evaluation. Baseline results of popular spotting methods are also provided. The experimental results strongly suggest that a saturation point of ME spotting methods is far from being reached. More cutting-edge spotting methods using, for example, Action Unit [19] and Deep learning techniques [16] are therefore worth of further exploration.

Acknowledgment

References

[1] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[2] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[4] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012.

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[6] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156, 1996.

[7] A. Goshtasby. Image registration by local approximation methods. *Image and Vision Computing*, 6(4):255–261, 1988.

[8] Xiaopeng Hong, Yingyue Xu, and Guoying Zhao. *LBP-TOP: A Tensor Unfolding Revisit*, pages 513–527. Springer International Publishing, Cham, 2017.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[10] 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:1511.00423*, 2015.

[11] 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 Trans. Affect. Comput.*, 2017.

[12] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Proc. FG*, 2013.

[13] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[14] A. Moilanen, G. Zhao, and M. Pietikäinen. Spotting rapid facial movements from videos using appearance-based feature difference analysis. In *Proc. ICPR*, 2014.

[15] D. Patel, G. Zhao, and M. Pietikäinen. Spatiotemporal integration of optical flow vectors for micro-expression detection. In *Proc. ACIVS*, 2015.

[16] Devangini Patel, X. Hong, and G. Zhao. Selective deep features for micro-expression recognition. In *Proc. ICPR*, 2016.

[17] S. Polikovsky, Y. Kameda, and Y. Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In *Proc. IET ICDP*, 2009.

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[19] Yan Tong, Wenhui Liao, and Qiang Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE transactions on pattern analysis and machine intelligence*, 29(10), 2007.

[20] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*. IEEE, 2001.

[21] S. Wang, S. Wu, and X. Fu. A main directional maximal difference analysis for spotting micro-expressions. In *ACCV 2016*, pages 449–461. Springer, 2016.

[22] Q. Wu, X. Shen, and X. Fu. The machine knows what you are hiding: an automatic micro-expression recognition system. In *Proc. ACII*, pages 152–162, 2011.

[23] Z. Xia, X. Feng, J. Peng, X. Peng, X. Fu, and G. Zhao. Spontaneous micro-expression spotting via geometric deformation modeling. *CVIU*, 147:87–94, 2016.

[24] F. Xu, J. Zhang, and J. Wang. Microexpression identification and categorization using a facial dynamics map. *IEEE Trans. Affect. Comput.*, 2016.

[25] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE TPAMI*, 29(6), 2007.

[26] Z. Zhou, G. Zhao, and M. Pietikäinen. Towards a practical lipreading system. In *CVPR 2011*, pages 137–144, June 2011.