



Stochastic-Gradient-Descent-Based Max-Margin Early Event Detector

Hong-Wen Wang, Dai-Yi Chiu, Wei-Chung Chan, and Zhi-Fang Yang^(✉)

Department of Computer Science and Information Engineering, National Taipei University, New Taipei City 23741, Taiwan, R.O.C.
zfyang@mail.ntpu.edu.tw

Abstract. Max-margin-based early event detection is first solved by max-margin early event detector (MMED) proposed by Hoai and Torre [10]. In this study, the stochastic gradient descent mechanism is used to replace the quadratic programming solver in [10] to achieve early event detection. Three datasets are tested, including synthetic data, the extended Cohn-Kanade dataset (CK+), and Australian sign language data (ASL). The experimental results show that the proposed approach is feasible, and that the performance is comparable to that obtained in MMED.

Keywords: Early event detection · Quadratic programming · Stochastic gradient descent

1 Introduction

Event detection is one widely researched field in computer engineering [1, 2, 4, 6–10, 13–17]. Generally, event detection can be divided into complete event detection and partial event detection [15, 16]. In partial event detection, early event detection is focused on finding the target event as soon as possible. Early event detection can be applied to many applications, such as facial expression detection [8–10, 15, 16] and sign language detection [10, 15, 16].

Hoai and Torre [10] propose the max-margin early event detector (MMED) based on the structured output support vector machine (SOSVM) which is the generalized support vector machine (SVM) [3, 5]. Later, Huang et al. [4] design the sequential MMED (SMMED) to extend MMED to deal with multi-classes. For MMED, solving the quadratic programming problem is the major work [3]. However, if the number of constraints is very huge, it is too time consuming to be feasible. Thus, alternative solvers may be considered, for instance, Shai et al. [12] propose the primal estimated sub-gradient solver (Pegasos) in which the stochastic gradient descent method is utilized to solve the quadratic programming problem in the support vector machine instead. Furthermore, it is proven to be able to converge efficiently [12].

Thus, due to the huge number of constraints in MMED, about three thousand constraints for one simple synthetic training instance [10, 12], in this study, the stochastic gradient descent method is used to replace the quadratic programming solver in MMED.

Three data sets are used in the experiments as those used in [10], and comparable experimental results are obtained to verify the feasibility of the proposed approach.

2 Proposed Approach

In order to incorporate the stochastic gradient descent mechanism, Pegasos [12], into MMED [10], the objective function of MMED [10] needs to be rewritten in the gradient-descent style [12]. In this study, for a given training sample $S = \{(x^i, y^i) | x^i \in X, y^i \in Y, i = 1, \dots, n\}$, where X is the input training instance set, and Y the set of early events, the proposed gradient-descent-style objective function is derived to be:

$$\min \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{n} \sum_{i=1}^n \max \left\{ 0, \mu \left(\frac{|y_t^i|}{|y^i|} \right) \left[f(x_{y-y_t^i}^i) + \Delta(y_t^i, y) \right] \right\}, \quad (1)$$

where \mathbf{w} is the weight vector of the linear early event detector in [10], λ the trade-off coefficient in [12], y_t^i the early event observed at time t of the complete event y^i in the training instance x^i [10], y a time segment in x^i [10], $\mu \left(\frac{|y_t^i|}{|y^i|} \right)$ the scaling function for soft margin [10], $\Delta(y_t^i, y)$ the loss function of mistaking y as y_t^i [10], and $f(x_{y-y_t^i}^i) = f(X_y^i) - f(X_{y_t^i}^i)$ the difference of the detection scores of $f(X_y^i)$ and $f(X_{y_t^i}^i)$ [10, 12]. The proposed algorithm to achieve the optimization of Eq. (1), denoted as Pegasos-based MMED (PMMED), is as follows.

Algorithm 1: PMMED

```

input:  $S, \lambda, T$ 
1. Set  $w_1 = 0$  ;
2. for  $t \leftarrow 1$  to  $T$  do
3.   Choose  $i \in \{1, \dots, |S|\}$  uniformly at random;
4.   Set  $y_t^i = \max_{y \in Y} f(x_y^i)$  ;
5.   Set  $\eta_t = \frac{1}{\lambda t}$  ;
6.   if  $f(X_{y-y_t^i}^i) + \Delta(y_t^i, y) > 0$  then
7.     Set  $w_{t+1} \leftarrow (1 - \eta_t \lambda) w_t - \eta_t \mu \left( \frac{|y_t^i|}{|y^i|} \right) f'(X_{y-y_t^i}^i)$  ;
8.   else
9.     Set  $w_{t+1} \leftarrow (1 - \eta_t \lambda) w_t$  ;
10.  end
11. end
output:  $w_{T+1}$ 

```

For each iteration in the above algorithm, a training instance (x^i, y_t^i) is chosen randomly, the corresponding gradient is computed, and the weight vector w_i is updated. Note that η_t is the learning rate at iteration t .

As shown in Line 7 of Algorithm 1, the gradient ∇_t is

$$\nabla_t = \lambda w_t + 1 \left[f(\mathbf{X}_{y-y_t}^i) + \Delta(y_t^i, \mathbf{y}) > 0 \right] \mu \left(\frac{|y_t^i|}{|y^i|} \right) f'(\mathbf{X}_{y-y_t}^i). \quad (2)$$

The derivation of Eq. (2) is as follows. First, recall the objective function of MMED in [10]:

$$\begin{aligned} & \min \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^n \xi^i \\ & \text{subject to } f(\mathbf{X}_{y_t^i}^i) \geq f(\mathbf{X}_y^i) + \Delta(y_t^i, \mathbf{y}) - \frac{\xi^i}{\mu \left(\frac{|y_t^i|}{|y^i|} \right)} \\ & \forall i = 1, \dots, n, \forall t = 1, \dots, l^i, \forall \mathbf{y} \in Y(t). \end{aligned} \quad (3)$$

Since the slack variables $\xi^i \geq 0$, the violated constraint for the objective function shown in Eq. (3) can be rewritten to be the following.

$$0 \leq \xi^i < \mu \left(\frac{|y_t^i|}{|y^i|} \right) \left[f(\mathbf{X}_y^i) - f(\mathbf{X}_{y_t^i}^i) + \Delta(y_t^i, \mathbf{y}) \right] \quad (4)$$

Based on the hinge loss idea [3], Eq. (4) can be expressed as:

$$\xi^i = \max \left\{ 0, \mu \left(\frac{|y_t^i|}{|y^i|} \right) \left[f(\mathbf{X}_y^i) - f(\mathbf{X}_{y_t^i}^i) + \Delta(y_t^i, \mathbf{y}) \right] \right\} \quad (5)$$

By using Eq. (5), Eq. (3) can be rewritten to derive the following equation.

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{n} \sum_{i=1}^n \max \left\{ 0, \mu \left(\frac{|y_t^i|}{|y^i|} \right) \left[f(\mathbf{X}_{y-y_t^i}^i) + \Delta(y_t^i, \mathbf{y}) \right] \right\}. \quad (6)$$

Finally, change the trade-off coefficient from C in the second term to λ in the first term in Eq. (6), and Eq. (1) is derived. Then, the gradient ∇_t can be derived as shown in Eq. (2).

3 Experimental Results

In this study, three datasets are utilized, including synthetic data, the extended Cohn-Kanade dataset (CK+), and Australian sign language data (ASL) [10]. The ROC curve, the AMOC curve, and the F1-score curve are used to demonstrate detection precision, timeliness, and location ability of the proposed approach, respectively [10].

3.1 Synthetic Data

Following MMED [10], the synthetic dataset is used to check correctness of the proposed approach. One time series contains a target event, two non-target events, and four connecting segments [10]; the total length of one signal sequence is between 160 and 180 numerical values. Some examples of the synthetic data sequence can be found in Fig. 1, in which the green ones are the target events.

In the experiments, one training/testing dataset consists of 100 time series. A total of 100 pairs of training and testing datasets are generated. For each pair, the proposed PMMED and the MMED method [10] are applied, and the testing results are expressed via ROC, AMOC, and F1-score curves.

Some samples of the testing results are given in Fig. 1, in which the target event is shown in green, and the fire frames are marked as red and blue vertical lines for PMMED and MMED, respectively. It can be seen that PMMED detects the target event earlier than MMED.

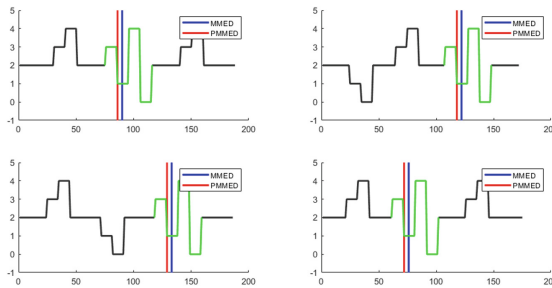


Fig. 1. Samples of testing results of synthetic data. The horizontal axis shows the timing of the time series, and the vertical axis indicates the signal value. (Color figure online)

In Fig. 2, the performance of PMMED and MMED are shown and compared. As shown in Fig. 2(a), all the 200 testing datasets for PMMED and MMED detect the target event with 100% accuracy with the area under ROC curve of 1. In Fig. 2(b), the AMOC curves of the set of the 100 testing datasets for PMMED and MMED are shown by thin red curves and by thin blue ones, respectively; the average AMOC curves are shown by thick ones. It can be found that, for timeliness performance, MMED is a little better than PMMED. The reason may be that the stochastic nature of PMMED selects not-so-good answers sometimes. By using the same illustration style for the F1-score curves in Fig. 2(c), it can be found that PMMED is a little better than MMED on the location ability. The reason may be that required by MMED is the optimal solution, which delays the detection of the event location. However, as shown in Fig. 2, the performance of PMMED and that of MMED are comparable for the synthetic datasets.

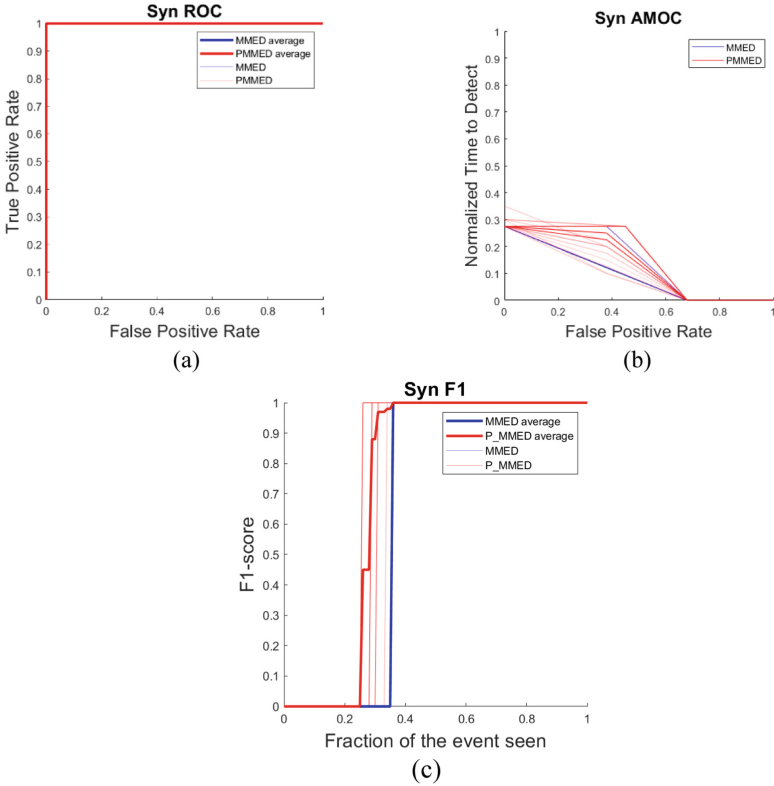


Fig. 2. Performance of synthetic data: (a) ROC curves; (b) AMOC curves; (c) F1-score curves. (Color figure online)

3.2 Extend Cohn-Kanade Dataset

In CK+, there is a total of 327 CAPP facial image sequences from 123 subjects, including 4 kinds of negative emotions and 3 kinds of positive emotions [10]. In the experiments, 100 positive and 100 negative emotion sequences are randomly selected as a training dataset, and the remaining 127 sequences are the testing dataset. The procedure is repeated 100 times to generate 100 pairs of training and testing datasets. Then both PMMED and MMED are applied to all the pairs to get the testing results. Note that the negative emotion is the target.

In Fig. 3, the illustration style is the same as that of Fig. 2, that is, the red curves are for PMMED, the blue ones for MMED, and each thick curve for the average of 100 thin curves. As shown in Fig. 3(a) and Fig. 3(b), the areas occupied by the red curves and the blue ones are almost the same. However, it can be said that the performance of MMED is a little better than that of PMMED if the two average curves in either subfigure are carefully compared. That is, for accuracy and timeliness, the performance of PMMED and MMED are quite comparable on the CK+ dataset. However, as for the F1-score curve shown in Fig. 3(c), the performance of MMED is better than that of PMMED. The reason may be that the strategy of optimal solution finding for MMED is more

suitable for event location detection when tackling more complicated events like facial expressions.

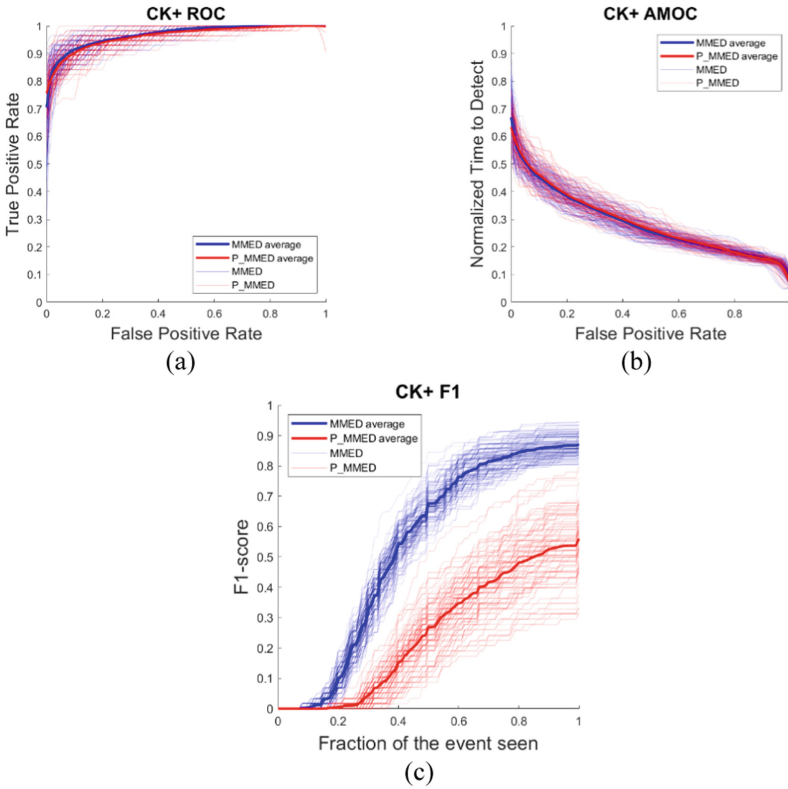


Fig. 3. Performance of CK+ dataset: (a) ROC curves; (b) AMOC curves; (c) F1-score curves.

3.3 Auslan Dataset

The Auslan dataset (ASL) consists of 95 words; for each word, 27 instances can be found, and there are 22 features derived from hands are used to represent an instance [10]. In this study, based on the experimental setting in MMED [10], the target was “I love you”, and two sets of 15 randomly chosen words were added before and after the target, respectively. For each word, the first 15 instances were used for training and the remaining 12 ones for testing. The pre-processing of the training/testing data followed that designed in MMED [10].

A total of 100 sentences were generated as the training set, and a total of 200 sentences as the test set. Then PMMED and MMED were applied, and the performance was analyzed. As shown in Fig. 4, all the ROC curve, the AMOC curve, and the F1-score curve indicates that MMED works better than PMMED. Since the stochastic nature of

PMED is to find a solution based on only one instance chosen randomly, it was natural for MMED to get better results compared with PMMED. However, it is still can be claimed that PMMED and MMED are comparable in terms of accuracy according to Fig. 4(a).

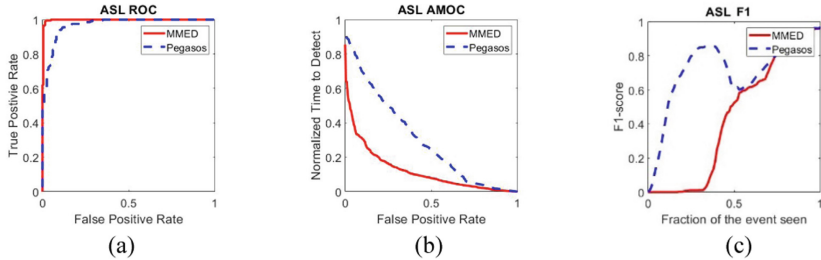


Fig. 4. Performance of ASL dataset: (a) ROC curves; (b) AMOC curves; (c) F1-score curves.

4 Conclusion

In this study, an approach called PMMED has been proposed to achieve early event detection. In order to tackle with huge constraints, the core work of PMMED has been designed to use the strategy of stochastic gradient descent to solve the original quadratic programming problem in MMED [10]. The performance of MMED and PMMED has been compared through the experiments based on three datasets, including synthetic data, CK+, and ASL, and based on the ROC curves, the AMOC curves, and the F1-score curves. According to the experimental results, the performance of PMMED and MMED is quite comparable in terms of accuracy and timeliness, and MMED is better than PMMED in terms of the location ability.

References

1. Shahroudy, A., Ng, T., Gong, Y., Wang, G.: Deep multimodal feature analysis for action recognition in RGB+D videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(5), 1045–1058 (2018)
2. Tran, D., Yuan, J., Forsyth, D.: Video event detection: from sub-volume localization to spatio-temporal path search. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(2), 404–416 (2014)
3. Alpaydin, E.: *Introduction to Machine Learning*, 4th edn. MIT Press, London (2020)
4. Huang, D., Yao, S., Wang, Y., De La Torre, F.: Sequential max-margin event detectors. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *ECCV 2014*. LNCS, vol. 8691, pp. 410–424. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_27
5. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.* **6**(2), 1453–1484 (2005)
6. Azorin-López, J., Saval-Calvo, M., Fuster-Guilló, A., Oliver-Albert, A.: A predictive model for recognizing human behaviour based on trajectory representation. In: *Proceedings of International Joint Conference on Neural Networks (IJCNN)* (2014)

7. Wang, J., Wu, Y.: Learning maximum margin temporal warping for action recognition. In: Proceedings of ICCV 2013, pp. 2688–2695 (2013)
8. Su, L., Sato, Y.: Early facial expression recognition using early RankBoost. In: Proceedings of FG 2013, pp. 1–7 (2013)
9. Xie, L., Zhao, J., Wei, H., Zhang, K., Pang, G.: Online kernel-based structured output SVM for early expression detection. *IEEE Signal Process. Lett.* **26**(9), 1305–1309 (2019)
10. Hoai, M., De la Torre, F.: Max-margin early event detectors. *Int. J. Comput. Vision* **107**(2), 191–202 (2013). <https://doi.org/10.1007/s11263-013-0683-3>
11. Lu, S., Jin, Z.: Improved stochastic gradient descent algorithm for SVM. *Int. J. Recent Eng. Sci. (IJRES)* **4**(4), 28–31 (2017)
12. Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A.: Pegasos: primal estimated sub-Gradient SOLver for SVM. *Math. Program* **127**, 3–30 (2011)
13. Guha, T., Ward, R.K.: Learning sparse representations for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(8), 1576–1588 (2012)
14. Cao, Y., et al.: Recognize human activities from partially observed videos. In: Proceedings of CVPR 2013, pp. 2658–2665 (2013)
15. Yang, Z.F., Hung, C.T.: Feedback max-margin early event detector. In: Proceedings of 8th International Conference on Applied System Innovation (ICASI), pp. 115–118 (2022)
16. Yang, Z.F., Lin, Y.C.: Reduction in number of constraints in max-margin early event detectors. *Pattern Recogn.* **91**, 111–122 (2019)
17. Xu, Z., Yang, Y., Hauptmann, A.G.: A discriminative CNN video representation for event detection. In: Proceedings of CVPR 2015 (2015)