IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 16, 2019, accepted June 25, 2019, date of publication July 1, 2019, date of current version July 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925916

Efficient Early Event Detector for Streaming Sequence

LIPING XIE[1,2], JUNSHENG ZHAO[3], HAIKUN WEI[1], ZHUN FAN[4], (Senior Member, IEEE), AND GUOCHEN PANG[2]

[1]Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China
[2]School of Automation and Electrical Engineering, Linyi University, Linyi 276005, China
[3]School of Mathematical Science, Liaocheng University, Liaocheng 252059, China
[4]Key Laboratory of Digital Signal and Image Processing of Guangdong Province, Shantou University (STU), Shantou 515063, China

Corresponding author: Haikun Wei (hkwei@seu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1500802, in part by the National Natural Science Foundation of China under Grant 61802059 and Grant 61773118, in part by the Natural Science Foundation of Jiangsu under Grant BK20180365, in part by the Open Project Program of the State Key Lab of CAD&CG, Zhejiang University, under Grant A1911, in part by the Innovation Fund of Key Laboratory of Measurement and Control of Complex Systems of Engineering through Southeast University under Grant MCCSE2018B01, in part by the Innovation Fund of Key Lab of Digital Signal and Image Processing of Guangdong Province under Grant 2018GDDSIPL-04, and in part by the Innovation Fund of Jiangsu Key Laboratory of Image and Video Understanding for Social Safety through Nanjing University of Science and Technology, Nanjing, China, under Grant 30918014107.

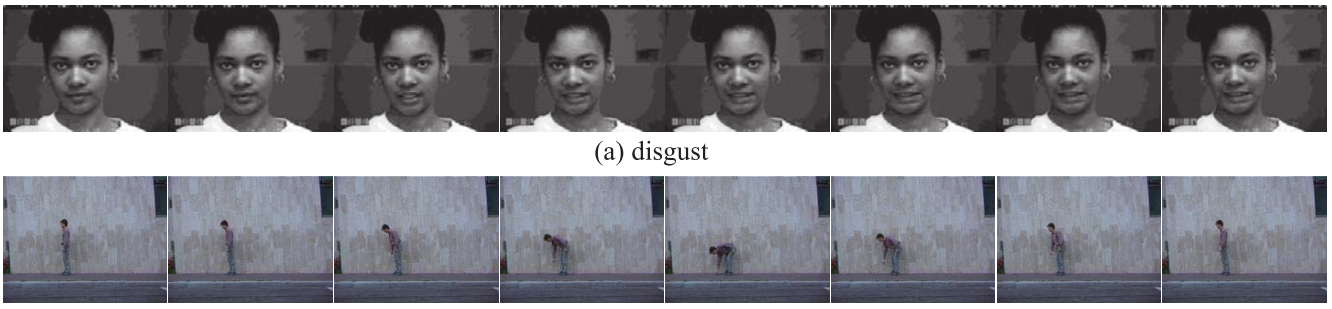
ABSTRACT Extensive research has been paid for event detection in the past decades. However, the timeliness requirement of real-world applications cannot be satisfied by these approaches. Early event detector is thus proposed recently to deal with this issue. Early detection aims to recognize the target as early as possible, i.e., it can detect partial events and create a monotonous function to rank them. Although important and practical, few studies have been given for early detection due to its complexity. Max-margin Early Event Detector (MMED) is a well-known approach, which achieves satisfied performance in identifying partial events. However, the MMED works in an offline manner and may fail in this era of streaming sequence. In addition, the large memory consumption and high retraining time cost of the MMED are hard to be satisfied in general platform conditions. In this paper, we introduce an online learning technique with max-margin to early event detection. The proposed model could be adapted to the changing data distribution of the streaming sequences. No historical data need to be stored. Therefore, both the memory requirement and retraining time cost are decreased significantly. We evaluate the proposed approach on three benchmark datasets with various complexities. The extensive results demonstrate both the effectiveness and efficiency of the proposed framework.

INDEX TERMS Online learning, early event detection, streaming sequence.

I. INTRODUCTION

Extensive attention has been paid for sequence-based tasks at various fields, such as action recognition [1], [2], gesture classification [3], [4], facial expression recognition [5], and so on [6]. Numerous approaches have been proposed [7], [8] and large numbers of outstanding achievements have been obtained [9]. However, early event detection is still a relatively new problem, which receives few attentions. In practice, each temporal sequence has a duration as illustrated in Figure 1. For example, a complete facial expression video contains the onset, the peak, and the offset; a dynamic hand gesture usually has three overlapping phases: preparation, nucleus, and retraction. Conventional sequence-based approaches make detection or classification after the sequence ends. This is not desirable in real-world applications since the timeliness is ignored [10]–[13]. In contrast, early event detection aims to identify the objective event after it starts and before it ends. The timeliness is thus guaranteed to make decision as early as possible. Early event detection has extensive potential applications. For example, human-computer interaction is becoming more prevalent in recent years, the timeliness is especially critical to improve the comfort and communication efficiency. Therefore, early event detection is an important technique which has proved useful in many applications.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.



(a) disgust

(b) bend

FIGURE 1. Examples of events which include the duration. The goal of early event detection is to detect the event as soon as possible after it starts and before it ends. (a) Expression "Disgust" of CK+ dataset from the onset to the peak frame; (b) Event "bend" of Weizmann dataset from the onset, peak and offset frame.

Different with conventional sequence-based approaches [14], [15], which only give a recognition result for the whole event, early event detection has three major challenges: 1) several partial events are contained in each sequence, and the specific number of partial events is determined by the length of the training sample. Therefore, each sequence corresponds to a different number of partial events, and the detector needs to detect all the partial segments. Meanwhile, the size of training set is thus augmented. 2) the information available in segments are different from each other. For the partial events from one sequence, the longer the segment, the more information it contains. An effective detector must describe the monotonicity of this relationship to give a reliable detection. 3) the location of the detected event, i.e., the start and the end, is required to report.

There are several works trying to solve this challenging problem. A probabilistic reliable-inference framework is proposed in [16] to give the probability ration test. Although the early recognition result is given, HMM adopted is still trained on the whole sequence, which leads to an unconvincing result. Similarly, the situation and the strategy utilized in [17] suffer the same problems. In 2013, a RankBoost-based approach [18] is proposed to make an early expression detection. The major limitation of this method is that the length of sequence needs to be known before testing, which is unpractical. In 2014, hidden Markov [19] is adopted to deal with this issue. Similarly, the unpractical part of this method is that it assumes each sequence starts with a neutral frame. Max-margin Early Event Detector (MMED) is then proposed and can be applied in real-world applications.

Subsequently, various deep learning methods have been put forward. [20] develops an autoregressive convolutional neural network to predict the semantic information for unseen frames. [21] combines VAEs and GANs to exploit human pose with the utilization of unlabeled data in unconstrained settings. Encoder-decoder neural network is adopted by [22] to predict and learn from dog behavior. Reference [23] employs a self-consistency model to localize the manipulations by detecting the anomalous cues. Reference [24] proposes to predict multiple human trajectories with no supervision. Shou *et al.* [25] make attempts for online detection of action start by addressing several specific challenges in realistic videos. Shyamal *et al.* [26] design a deep architecture (Single-Stream Temporal Action Proposals, SST) to generate temporal action proposals for long and untrimmed videos. Achal *et al.* [27] propose a recurrent predictive-corrective network to update per-frame predictions and intermediate activations by exploiting motion cues within videos. Xu *et al.* [28] present an end-to-end temporal proposal classification network, i.e., region convolutional 3D network for activity detection. Gao *et al.* [29] deal with the problem of temporal action proposal generation by designing a novel temporal unit regression network model.

Although various methods have been designed for different tasks, MMED [30] is still a representative and general model for early event detection. The structured output SVM is utilized to deal with the augmented training set, and the ranking relationship of the constraints is used to build the monotonicity of the function. The incoming problem of MMED is that the memory consumption is large, and the retraining cost is expensive when a new training sequence comes. Therefore, MMED may fail in situation with streaming sequences and large-scale applications. The general computation platform cannot meet these high requirements. To deal with these problems, we introduce online learning to MMED to make further improvements. An online framework with max-margin for early event detector is proposed in this paper, and we term it as OMED. OMED is updated by the sequential sample one by one. No historic data need to be stored, which greatly decrease the memory consumption. In addition, efficient algorithm is designed to optimize the proposed model. The effectiveness and efficiency are validated on three benchmark datasets with various complexities. In brief, we summarize the main contributions of our work in the following:

- An online framework with max-margin for early event detectors, which is termed OMED, is proposed. Not only the retraining time cost and memory consumption are reduced greatly, the changing data distribution is also modeled.

- Theoretical analysis of the proposed algorithm is offered. The computational complexity of OMED is much lower than that of MMED.
- Extensive experiments on various datasets are conducted to demonstrate both the effectiveness and efficiency of the proposed method.

We organize this paper as follows. The related work of online learning is reviewed in Section II. Some preliminary knowledge and the well-known MMED model are presented in Section III. In addition, the details of the proposed online framework are also given in this section. Section IV provides theoretical analysis of the proposed framework. The experimental results are shown in Section V. Finally, we conclude this paper in Section VI.

II. RELATED WORK

Most of the previous approaches [18], [19], [30] for early event detection operate in an offline manner. The models need to retrain on the entire training set whenever a new sample comes. Subsequently, the computation cost is expensive. Moreover, the memory demand is large since the training set is augmented for building the ranking relationships. These high requirements are hard to be satisfied for conventional computers. Therefore, we introduce online learning technique to early event detection.

Different with batch learning techniques [31]–[34], the model with online learning is updated by the data one-by-one, i.e., the data come sequentially. It is a common solution where the computation is infeasible on the entire training data set. Except for the significantly reduced time cost, the dynamic change of new data distribution is well modeled by online learning. Therefore, it is particularly well suited to early event detection, where both the training data set and memory consumption are intolerable.

Over the past decades, numerous works for online learning have been created, a large set of which perform with linear functions. Some popular linear online algorithms include: Perceptron algorithm [35], the family of passive-Aggressive (PA) learning algorithms [36], the Online Gradient Descent (OGD) algorithms [37], the Soft Confidence Weighted algorithms (SCW) algorithms [38], etc. There are also some "online kernel learning" approaches [39], including Randomized Budget Perceptron (RBP) [40], Budgeted Passive Aggressive Nearest Neighbor (BPA-NN) [41], and Budget Stochastic Gradient Descent (BSGD-M) algorithm in [42].

There are also numerous methods proposed for online SVM, such as HULLER [43], Relaxed Online SVM (ROSVM) [44], Markov sampling based online SVM [45], online structured output SVM [46], *et al.* However, these models cannot be utilized directly to the application of early event detection due to its complex characteristics as we described in Introduction.

III. PRELIMINARY

A. PROBLEM SETUP

Some Notations for Early Event Detection: Suppose we are given a set of training sequences and their associated ground truth annotations for the events of interest $(X^1, y^1), (X^2, y^2), \ldots, (X^n, y^n)$. Here, two elements are contained in $y^i = [s^i, e^i]$ to denote starting and ending of the event in training sequence X^i. n is the total number of training sequences. We adopt l^i to indicate the length of training sequence X^i. For each time $t = 1, \ldots, l^i$, let y_t^i be the partial event of y^i which has occurred already, i.e., $y_t^i = y^i \cap [1, t]$, which is possibly empty. We denote $\mathcal{Y}(t)$ be the set of all possible time intervals from the 1^{st} to the t^{th} frame: $\mathcal{Y}(t) = \{y \in \mathbb{N}^2 | y \subset [1, t]\} \cup \{\emptyset\}$. The empty segment $y = \emptyset$, indicates no event occurrence. For a sequence X of length l, $\mathcal{Y}(l)$ is the set of all possible locations of an event. For an arbitrary time interval $y = [s, e] \in \mathcal{Y}(l)$, let X_y indicate the segment of X from the s-th to e-th frames.

For a sequence segment X_y, its detection score is denoted as:

$$f(X_y, w, b) = \begin{cases} w^T \varphi(X_y) + b & \text{if } \mathcal{Y} \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases}$$

where f is a linear function, and $\varphi(X_y)$ denotes the feature descriptor of segment X_y. In the following, we use $f(X_y)$ to represent $f(X_y, w, b)$ for the sake of brevity.

B. MAX-MARGIN EARLY EVENT DETECTOR

Max-margin Early Event Detectors: MMED [30] is a representative approach in the field of early event detection and satisfactory performance has been achieved. The comparative information of two different segments is learnt by extending SOSVM in MMED. The basic idea of MMED is to guarantee that: the output score of the partial event, which has been seen at time t, is larger than that of any other occurred segment by a margin, i.e., $f(X^i_{y^i_t}) \geq f(X^i_y) + margin$. The detailed learning formulation of MMED is as follows:

$$\begin{aligned} &\min_{\{w,b,\xi^i \geq 0\}} \frac{1}{2}\|w\|_F^2 + \frac{C}{n}\sum_{i=1}^{n}\xi^i, \\ &\text{s.t.} \quad f(X^i_{y^i_t}) \geq f(X^i_y) + \Delta(y^i_t, y) - \frac{\xi^i}{\mu(\frac{|y^i_t|}{|y^i|})}, \\ &\forall i, \quad \forall t = 1, \ldots, l^i, \ \forall y \in \mathcal{Y}(t). \end{aligned} \tag{1}$$

Here, $|\cdot|$ represents length function, and $\mu(\cdot)$ is a rescaling factor of slack variable, which denotes the importance of a correct detection. In this paper, we adopt the piece-wise linear function followed [30]:

$$\mu(x) = \begin{cases} 2x & 0 < x \leq 0.5, \\ 1 & 0.5 < x \leq 1 \ \ and \ \ x = 0. \end{cases}$$

$\mu(0) = \mu(1) = 1$ emphasizes the importance of true rejection and true detection of a complete event. $\bigtriangleup(\cdot)$ is the margin of the pairwise segments. Here, $\bigtriangleup(y^i_t, y) = 1 - overlap(y^i_t, y)$.

Some important notations utilized in this article are summarized in Table 1.

TABLE 1. Important notations used in this article.

Notation	Description
X^i	the i-th sequence sample
l^i	the length of the i-th video sequence
$y^i = [s^i, e^i]$	the label information of the i-th sample, which indicate the start and the end
$y^i_t = y^i \cap [1, t]$	the partial event that has occurred at time t
$\mathcal{Y}(t)$	the set of all possible segments at time t
$\mathcal{Y}(l)$	the set of all possible segments for the whole sequence sample
X_y	an arbitrary segment indicated by y from sequence X
n	the total number of training sequences
w	the weight vector contained in the score function
b	the bias term of the score function
$\varphi(\cdot)$	the feature vector

IV. OMED: ONLINE FRAMEWORK WITH MAX-MARGIN FOR EARLY EVENT DETECTION

A. FORMULATION

Although MMED [30] achieves early detection effectively sometimes, the expensive computation and large memory requirement are usually hard to satisfy. Moreover, the high retraining cost for a batch learning method is intolerable in real-world applications. To deal with these problems, we propose an online learning framework with max-margin for early event detection, which is termed OMED. Each time, OMED deals with only one training sample. We use (X^i, y^i) to denote the sequence received at time i, the updating model can be written as follows:

$$\begin{aligned} &\min_{\{w,b,\xi^i\geq 0\}} \frac{1}{2}\|w - w_{i-1}\|_F^2 + C\xi^i, \\ &\text{s.t.} \quad f(X^i_{y^i_t}) \geq f(X^i_y) + \Delta(y^i_t, y) - \frac{\xi^i}{\mu(\frac{|y^i_t|}{|y^i|})}, \\ &\forall i, \quad \forall t = 1, \ldots, l^i, \ \forall y \in \mathcal{Y}(t). \end{aligned} \tag{2}$$

Whenever receiving a new training sequence, the weight vector w in function f is updated with only this new sample without any historic data. Hence, only the constraints contained in this new data are needed for updating. The retraining process and memory consumption are thus decreased greatly with online learning.

When testing, the score of the sequence is computed for each frame and stored in memory. The beginning and ending of an event are determined by the pre-defined thresholds. We can adjust these thresholds to trade off the high TPR for low FPR and vice versa. The detailed definitions of TPR and FPR can be found in the following. In this paper, the experimental results are obtained with FPR as 0.1. When the ending of the event is detected, the previous memory of this sequence is cleared. For the current frame, all segments of the occurred part of the sequence are used for computation, and the maximum score is returned as the output. Therefore, the iterative computation can be adopted to avoid double computing. The score of the frame at time t can be written as follows:

$$\max_{y\in\mathcal{Y}(t_0,t)} f(X_y) = \max\{\max_{y\in\mathcal{Y}(t_0,t-1)} f(X_y), f(X_{\widehat{y}_t})\}, \tag{3}$$

where t_0 is the starting frame of the video sequence in consideration, $\mathcal{Y}(t_0, t-1)$ denotes the set of available segments from time t_0 to time t, $\widehat{y}_t$ is the incremental segments which terminate at t.

- **TPR:** the true positive rate, it measures the proportion of the actual positive samples that recognized as positive ones.
- **FPR:** the false positive rate, it measures the proportion of the negative samples that recognized as positive ones.

B. OPTIMIZATION

We adopt Lagrange multiplier method to optimize the proposed model (2). During the optimization, the bias b is absorbed into the weight vector w by adding one dimension "1" of the feature vector. The function can thus be written as $f(x) = w^T\varphi + b = [w^T\ b][\varphi; 1]$. We use M to define the number of constraints contained in each video sequence. Therefore, the model needs to be updated M times for each coming training data. The problem (2) with one constraint can thus be rewritten as follows:

$$\begin{aligned} &\min_{\{w,\xi^i\geq 0\}} \frac{1}{2}\|w - w_{i-1}\|_F^2 + C\xi^i, \\ &\text{s.t.} \ f(X^i_{m_t}) \geq f(X^i_m) + \Delta(m_t, m) - \frac{\xi^i}{\mu(\frac{|m_t|}{|m|})}. \end{aligned} \tag{4}$$

To simplify the formulation, we use $(X^i_{m_t}, X^i_m)$ to denote $(X^i_{y^i_t}, X^i_y)$, i.e., the m-th constraint of sample (X^i, y^i), and (m_t, m) denotes (y^i_t, y) respectively. Then the Lagrangian is defined as:

$$\begin{aligned} &L(w, \xi^i, \lambda^i_m, \gamma^i_m) \\ &= \frac{1}{2}\|w - w_{i-1}\|_F^2 + C\xi^i - \gamma^i_m\xi^i \\ &\quad + \lambda^i_m\{\mu(\frac{|m_t|}{|m|})[f(X^i_m) - f(X^i_{m_t}) + \Delta(m_t, m)] - \xi^i\}, \end{aligned} \tag{5}$$

where $\lambda_m^i \geq 0$, $\gamma_m^i \geq 0$, $\forall m$ are Lagrangian multipliers. Then we set the partial derivative of (5) with respect to w to zero, and obtain the following formulation:

$$w - w_{i-1} + \lambda_m^i \mu(\frac{|m_t|}{|m|})[\varphi(X_m^i) - \varphi(X_{m_t}^i)] = 0. \quad (6)$$

After denoting $P_m^i = \mu(\frac{|m_t|}{|m|})[\varphi(X_{m_t}^i) - \varphi(X_m^i)]$, we obtain:

$$w = w_{i-1} + \lambda_m^i P_m^i. \quad (7)$$

Then we set the partial derivative of (5) with respect to ξ^i to zero, the following formulation is achieved:

$$\frac{\partial L(w, \xi^i, \lambda_m^i, \gamma_m^i)}{\partial \xi^i} = C - \gamma_m^i - \lambda_m^i = 0. \quad (8)$$

Since $\gamma_m^i \geq 0$, we obtain $0 \leq \lambda_m^i \leq C$. The Lagrangian of (5) with respect to λ_m^i can be written as:

$$L(\lambda_m^i) = \frac{1}{2}\|\lambda_m^i P_m^i\|_F^2 + \lambda_m^i \mu(\frac{|m_t|}{|m|})[(w_{i-1} + \lambda_m^i P_m^i)^T \varphi(X_m^i) - (w_{i-1} + \lambda_m^i P_m^i)^T \varphi(X_{m_t}^i) + \Delta(m_t, m)]. \quad (9)$$

Further, by setting $\frac{\partial L(\lambda_m^i)}{\partial \lambda_m^i} = 0$,we obtain:

$$\frac{\partial L(\lambda_m^i)}{\partial \lambda_m^i} = \mu(\frac{|m_t|}{|m|})[f_{i-1}(X_m^i) - f_{i-1}(X_{m_t}^i) + \Delta(m_t, m)] - \lambda_m^i \|P_m^i\|_F^2 = 0. \quad (10)$$

We leave the detailed formula derivation of (9) and (10) to Appendix. Thus, we have:

$$\lambda_m^i = \frac{\mu(\frac{|m_t|}{|m|})[f_{i-1}(X_m^i) - f_{i-1}(X_{m_t}^i) + \Delta(m_t, m)]}{\|P_m^i\|_F^2}. \quad (11)$$

Combining the fact that $0 \leq \lambda_m^i \leq C$, we obtain:

$$\lambda_m^i = min\{C, \frac{l_{i-1}(f, X_{m_t}^i, X_m^i)}{\|P_m^i\|_F^2}\}. \quad (12)$$

The loss $l(f, X_{m_t}^i, X_m^i) = max(0, \mu(\frac{|m_t|}{|m|})[f_{i-1}(X_m^i) - f_{i-1}(X_{m_t}^i) + \Delta(m_t, m)])$. To sum up, when the new training sample with the number of constraints M comes, the model OMED is updated as follows:

$$w_i = w_{i-1} + \sum_{m=1}^{M} \lambda_m^i P_m^i, \quad (13)$$

where λ_m^i is given by Eq.(12), and $P_m^i = \mu(\frac{|m_t|}{|m|})[\varphi(X_{m_t}^i) - \varphi(X_m^i)]$. Therefore, the parameter λ_m^i can be computed directly, which leads to an efficient optimization. The detailed procedure of the proposed OMED can been seen in Algorithm 1.

Algorithm 1 Online Framework With Max-Margin for Early Event Detection

Input: Training set (X^i, y^i), $i = 1, \ldots, n$ and its subsets for constraints; the corresponding parameters: $M, C, \mu(\frac{|m_t|}{|m|})$, $\Delta(m_t, m)$, $\varphi(X_{m_t}^i)$, $\varphi(X_m^i)$;
M: the number of constraints for each sequence;
$C > 0$: the regularization parameter;
$\mu(\frac{|m_t|}{|m|})$: the function to compute the proportion of the occurred event;
$\Delta(m_t, m)$: the margin denoted by $X_{m_t}^i$ and X_m^i;
$\varphi(X_{m_t}^i)$,$\varphi(X_m^i)$: the feature representation of $X_{m_t}^i, X_m^i$ with additional dimension;

Output: a set of weight vectors: $w = (w_1, w_2, \ldots, w_n)$.

1: Initialize $w_0 = 0$.
2: **For** $i = 1, \ldots n$
3: **For** $m = 1, \ldots M$
4: Compute $P_m^i = \mu(\frac{|m_t|}{|m|})[\varphi(X_{m_t}^i) - \varphi(X_m^i)]$;
5: Compute $f_{i-1}(X_{m_t}^i) = w_{i-1}^T \varphi(X_{m_t}^i)$;
6: Compute $f_{i-1}(X_m^i) = w_{i-1}^T \varphi(X_m^i)$;
7: Compute $l_{i-1}(f, X_{m_t}^i, X_m^i) = \mu(\frac{|m_t|}{|m|})[f_{i-1}(X_m^i) - f_{i-1}(X_{m_t}^i) + \Delta(m_t, m)]$;
8: Compute $\lambda_m^i = min\{C, \frac{l_{i-1}(f, X_{m_t}^i, X_m^i)}{\|P_m^i\|_F^2}\}$;
9: **End for**
10: Compute $w_i = w_{i-1} + \sum_{m=1}^{M} \lambda_m^i P_m^i$;
11: $i = i + 1$.
12: **End for**

C. COMPUTATIONAL COMPLEXITY ANALYSIS

1) MMED

The standard QP (quadratic programming) is adopted for the optimization of MMED [30]. According to the conclusion described in [47], the computational complexity of MMED with QP algorithm requires $O(Mnd^2)$, where n denotes the number of training sequences, M is the number of augmented constraints extracted from each sample, and d represents the feature dimensionality. Therefore, the complexity of MMED is linear with respect to the number of training sample and constraints, but quadratic with respect to the feature dimension.

2) OMED

As it can be seen in Algorithm 1, the computational complexity of OMED mainly includes three parts: the computation of P_m^i,$f_{i-1}(X_{m_t}^i)$,and $f_{i-1}(X_m^i)$, i.e., the steps 4, 5, and 6, which share the same complexity $O(d)$. Since the iterations of internal and external loops are M and n respectively, Algorithm 1 requires $O(Mnd)$. Obviously, it is linear with respect to all parameters (n, M, s), which achieves much lower complexity compared with that of MMED.

V. EXPERIMENTS

In this section, we validate the effectiveness and efficiency of the proposed OMED on three benchmark

video datasets with various complexities: Weizmann dataset [48], [49], CK+ dataset [50], and UvA-NEMO dataset [51]. Weizmann dataset contains 10 actions, CK+ dataset has six prototypic expressions, and UvA-NEMO dataset contains two smiles: spontaneous and deliberate smiles. In our experiments, we aim to detect action ''Bend'' on Weizmann dataset; ''negative facial expressions'' (Anger, Disgust, Fear and Sadness) on CK+ dataset; and ''spontaneous smile'' on UvA-NEMO dataset, respectively. Prior to the analysis of experimental results, we present the datasets, experiment setup, as well as evaluation criteria first.

A. DATASETS

In our experiments, three widely utilized video datasets with various complexities are adopted for evaluation.

1) WEIZMANN DATASET

90 video sequences from 9 persons are contained in this dataset. Each person performs 10 actions: Bend, Jack, Jump, Pjump, Run, Side, Skip, Walk, Wave1 and Wave2 [48], [49]. One video sequence consists of one single action. The length of the sequences varies from 28 to 146. Similar with [30], we concatenate 10 actions of each person to create a longer video. The event of interest video ''Bend'' is put at the end of the longer sequence. In this paper, we adopt AlexNet architecture [52] to extract the frame-level features with the dimensionality as 4096. Then PCA is employed to obtain the low-dimensional features, and we set it as 1000 in the experiments. Due to the limitation of samples, leave-one-out cross validation is adopted during the experiments. 8 videos are used for training, and the remaining as a testing sample.

2) CK+ DATASET

CK+ is the extended Cohn-Kanade dataset proposed in [50]. There are 210 adults, aging from 18 to 50 years old, contained in this dataset. Some statistics are summarized to illustrate the variety as follows: 69% of the sequences are female, 81% are Euro-American, 13% are Afro-American, and 6% are other groups. Each adult performs 23 different expressions, in which six prototypic emotions are involved. In this paper, the prototypic emotions from 327 sequences are utilized. Each sequence varies from the onset to the peak of some expression. We randomly choose 200 sequences for training, in which the number of optimistic and negative samples are equal. The canonical normalized appearance feature (CAPP) provided in [50] is adopted to describe the frame-level features, and the dimensionality of CAPP is 1656.

3) UVA-NEMO DATASET

This dataset [51] is created by the Science Center NEMO. It is built for analyzing the dynamic difference of spontaneous/deliberate smiles. There are 1240 smile video sequences contained, where 597 are spontaneous smiles and remaining are deliberate ones. These sequences are from 400 adults, where 185 are female. All the video sequences are normalized to the same resolution of 1920×1080 pixels under controlled illumination conditions. The deliberate smiles are obtained by asking the subjects to pose as realistically as possible. In contrast, the spontaneous smiles are gained by showing funny video segments. There is a rule that all the video sequences start and end with neutral or near-neutral expressions. In this paper, we extract the Local Binary Patterns (LBP) [53] for each frame. 4×4 blocks are adopted with the neighboring points as 2^3. The feature dimension of LBP is $4 \times 4 \times 59 = 944$.

Some examples of the sequences from three datasets can be seen in Figure 2, and the statistics of these datasets are summarized in Table 2. During the experiments, five-fold cross-validation strategy is adopted for CK+ and UvA-NEMO datasets to tune the parameters. All the reported results are repeated five times, and the average values are reported.

TABLE 2. Statistics of three benchmark datasets.

	#Sample			#Feature
	Training		Test	
	video	constraints		
Weizmann	8	120	1	1000
CK+	200	3000	127	1656
UvA-NEMO	800	1200	440	944

B. EXPERIMENT SETUP

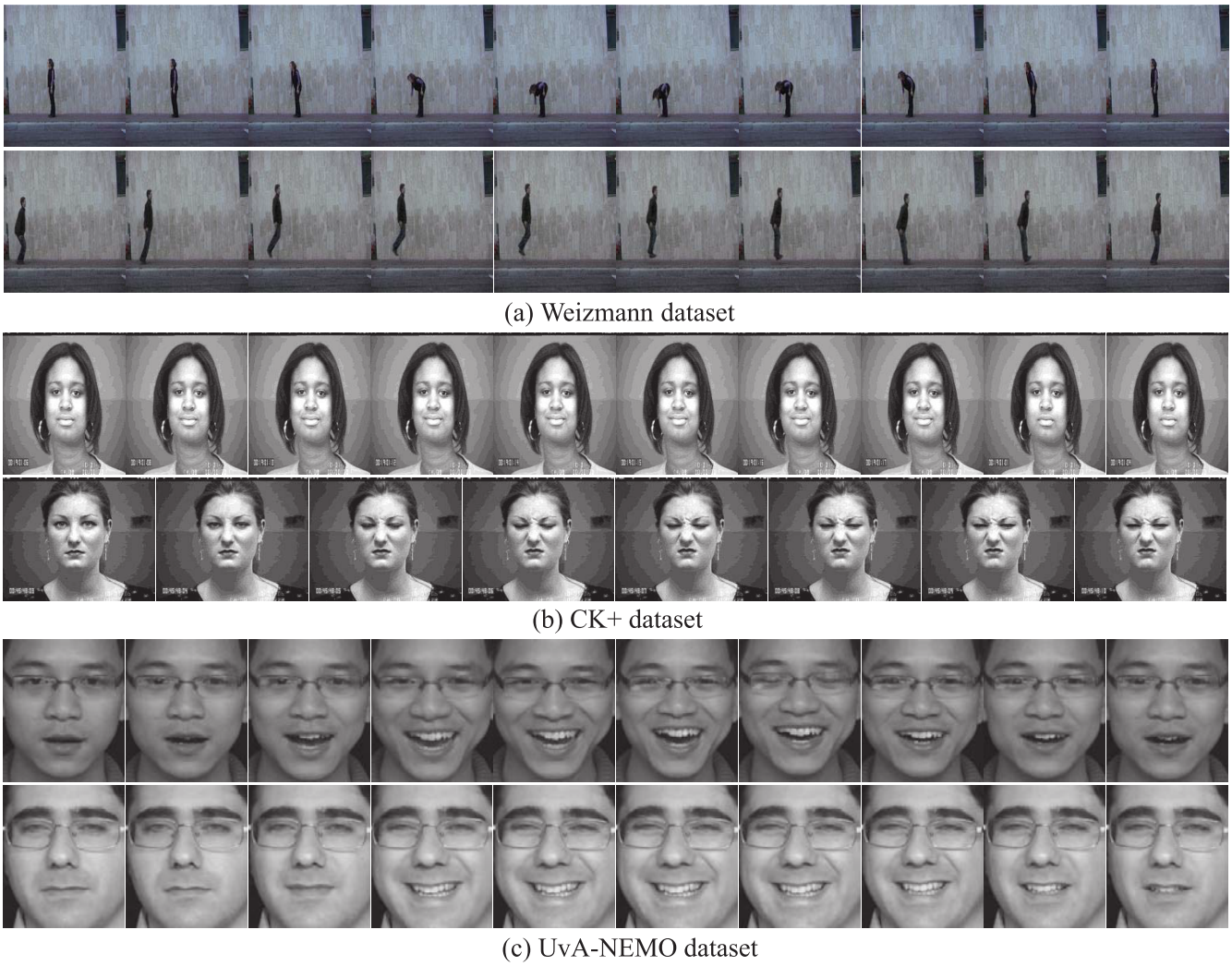
1) COMPARISON METHODS

The experimental results are compared with two baseline methods (FrmPeak, FrmAll) and the state-of-art approach: (MMED [30]). The baseline methods are frame-based SVMs, which make a detection by classifying each frame. The duration information is ignored in this way. The detailed setup of each approach can be seen in the following:

- **FrmPeak:** peak-frame-based SVM. Only the peak frame in each sequence is used to train SVM;
- **FrmAll:** all-frames-based SVM. All the frames contained in each sequence are adopted equally to train SVM;
- **MMED:** the max-margin early event detector proposed by [30];
- **OMED:** the presented online framework with max-margin for early event detection;

2) EXPERIMENT SETTING

During the experiments, all the trade-off parameters are tuned from the set $\{10^i | i = -5, -4, \ldots, 3, 4, 5\}$. The SVMs used in FrmPeak and FrmAll approaches are linear. For comparison, we use the same strategies of the constraints generation in MMED and OMED by considering the overlap among different segments. We set the number of constraints ''M'' for each video as 15, and the overlap is required lower than 0.7. All the experiments are conducted on a computer with the following specifications: Intel(R) Xeon(R) Core-20, CPU E5-2650



(a) Weizmann dataset

(b) CK+ dataset

(c) UvA-NEMO dataset

FIGURE 2. Examples from three benchmark datasets. Note that the number of frames in each video sequence is set as the same for presentation except for the expression on CK+ dataset since the length of this original sequence is "8". (a) Bend/Jump on Weizmann dataset; (b) Happiness/Disgust on CK+ dataset; (c) Spontaneous/Deliberate smile on UvA-NEMO dataset.

v3-2.3 GHz, Memory 48 GB, LINUX operating system with Matlab 2015a.

C. EVALUATION CRITERIA

In this section, we introduce several evaluation criteria used in our experiments: F-score [52], the area under the Receiver Operating Characteristic (ROC) curve (AUC) [54], the Activity Monitoring Operating Curve (AMOC) [55]) and the training time curve.

1) F-SCORE CURVE:

F-score [52] is utilized for evaluating the detection quality. It is a measurement considering both the precision p and the recall r of the testing samples. The harmonic average of the precision and recall is the F-score value: $F = \frac{p*r}{p+r}$. In our experiments, we define p and r as follows: the event of interest that the detector output at time t is y, and the ground truth (truncated event) is y^*. Then $p = \frac{|y \cap y^*|}{|y|}$. We output the F-score sequentially as the event of interest starting from 0% to 100%. F-score reaches best at 1 and worst at 0.

2) AUC CURVE:

AUC curve [54] demonstrates the results of accuracy comparison. AUC is the area under ROC curve, and ROC curve is created by plotting True Positive Rate (TPR) against False Positive Rate (FPR) by changing the threshold settings. In early event detection, we define TPR as the situation that the model makes a detection during the event of interest, and FPR is that the model fires before the event starts, or after it ends. The value in AUC curve is the higher the better.

3) AMOC CURVE:

AMOC [55] is applied to describe the timeliness of event detection. It is a function of timeliness and FPR by adjusting the threshold settings. In early event detection described of

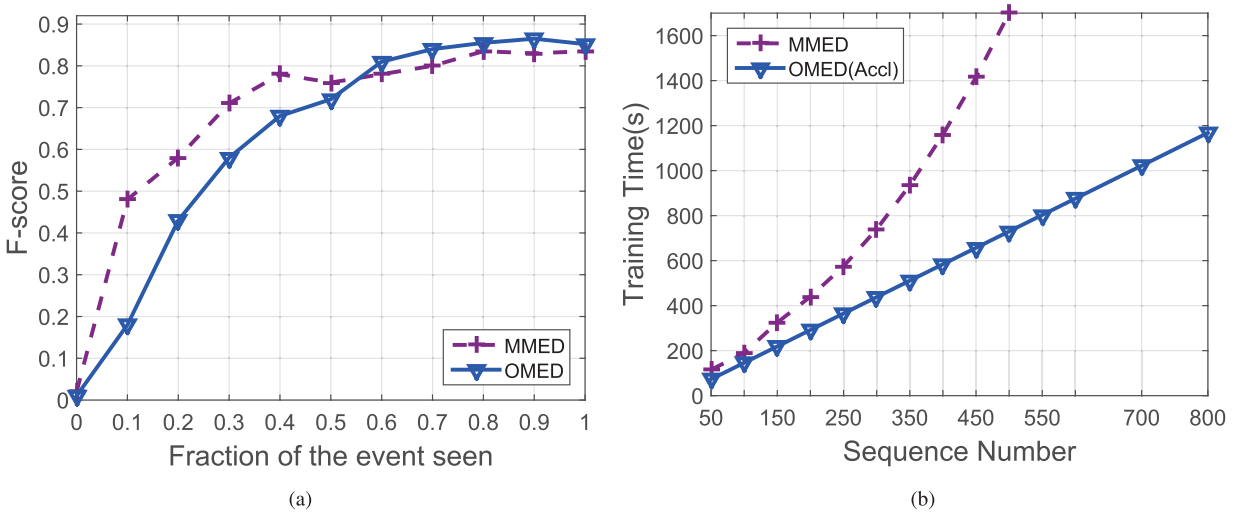


FIGURE 3. Experimental results on Weizmann dataset. (a) F-score curve; (b) The training time curve. Note that "Accl" refers to accumulative time cost.

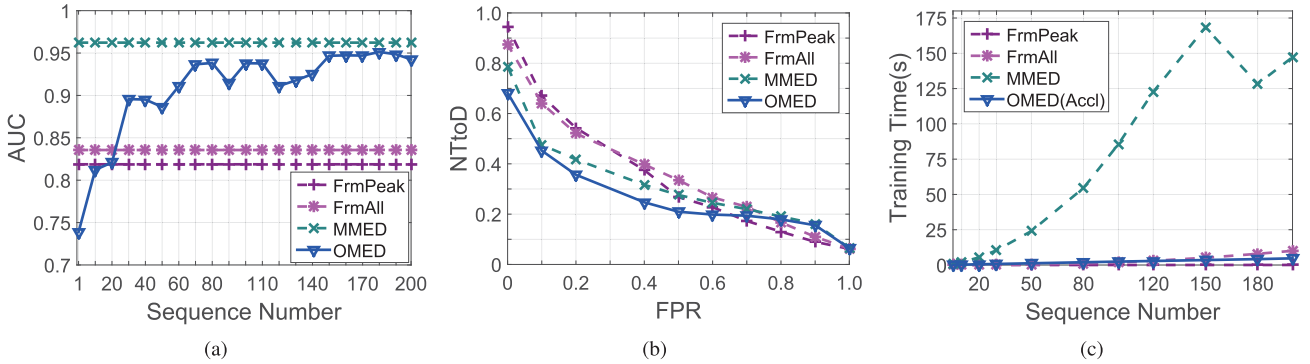


FIGURE 4. Experimental results on CK+ dataset. (a) AUC curve; Note that the "Sequence Number" only holds for OMED, and the results of offline approaches (FrmPeak, FrmAll, MMED) are achieved with 200 training sequences. (b) AMOC curve; (c) The training time curve. Note that "Accl" refers to accumulative time cost.

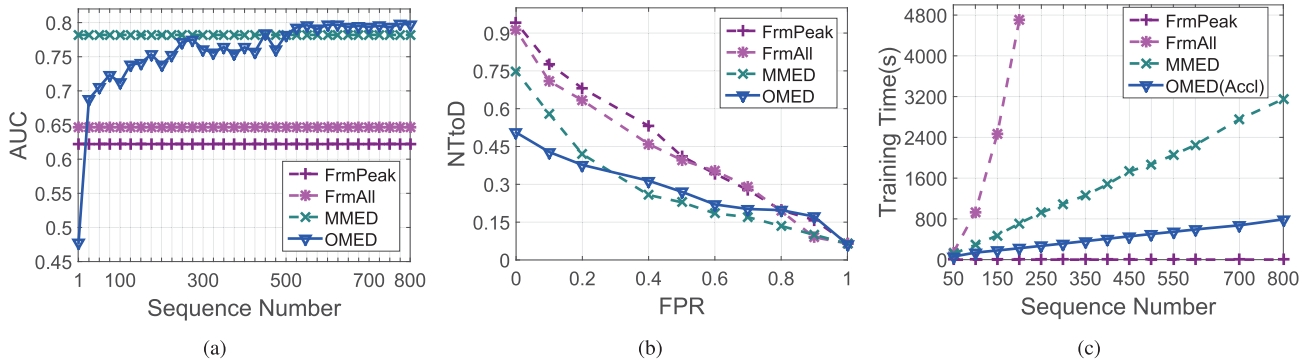
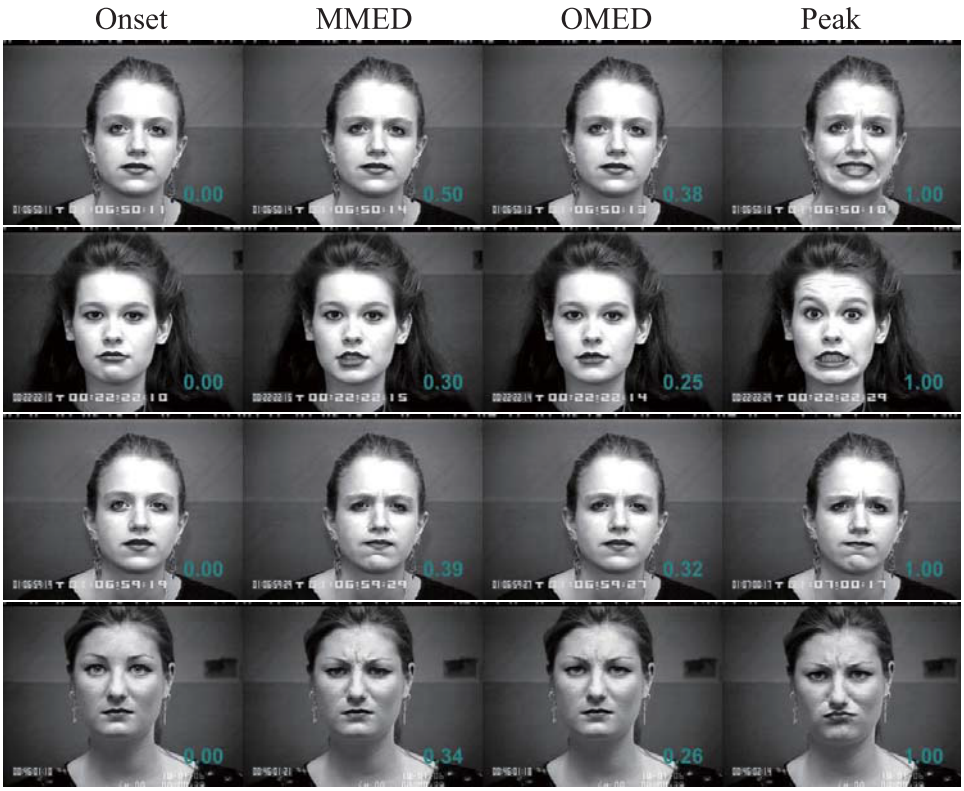
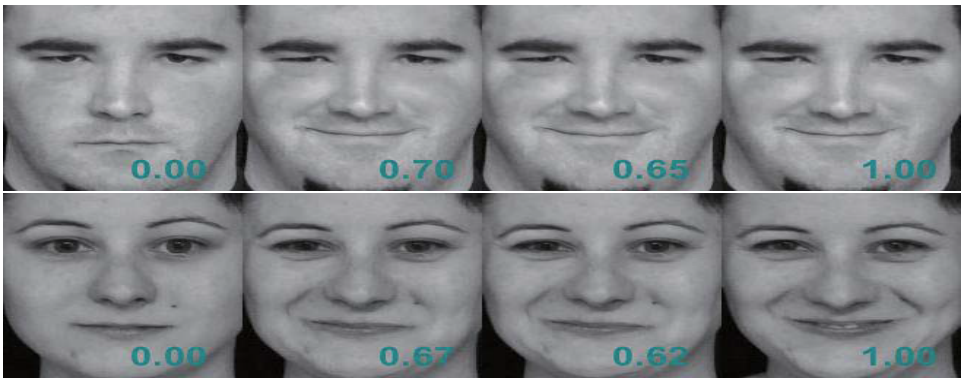


FIGURE 5. Experimental results on UvA-NEMO dataset. (a) AUC curve; Note that the "Sequence Number" only holds for OMED, and the results of offline approaches (FrmPeak, FrmAll, MMED) are achieved with 800 training sequences. (b) AMOC curve; (c) The training time curve. Note that "Accl" refers to accumulative time cost.

this paper, we adopt Normalized Time to Detection (NTtoD) to reflect the timeliness of detection. Suppose that the event of interest for a testing sequence is from the s-th to the e-th frame, and the model fires at time t (i.e., the t-th frame). If $s \leq t \leq e$, it is a successful detection, and NTtoD$= \frac{t-s+1}{e-s+1}$. If $t < s$ or $t > e$, it is a false detection, and NTtoD$= 0$ or NTtoD$= \infty$ respectively. The value in AMOC curve is the lower the better.



(a) CK+ dataset



(b) UvA-NEMO dataset

FIGURE 6. NTtoD of some examples from two benchmark datasets. (a) Fear/Fear/Anger/Anger on CK+ dataset; (b) Spontaneous/Spontaneous smile on UvA-NEMO dataset; The four frames in each row from left to right are: the onset, the ones MMED and OMED make a detection and the peak one. The number on each frame denotes the value of NTtoD. For a testing sequence, the length from the onset to the peak frame is normalized as "1".

4) THE TRAINING TIME CURVE:

The training time curve gives efficiency comparison. It demonstrates the tendency of training time cost for various number of training samples. In our experiments, the time of online method (OMED) refers the accumulative time cost, which is different with the instantaneous one. Since OMED is updated with only one sample when a new sequence arrives, the update time cost at various time is the same. Therefore, the value in the curve is the accumulative one. The value in training time curve is the lower the better.

D. RESULTS ON WEIZMANN DATASET

The performance of the compared approaches on Weizmann dataset are shown in Figure 3 and Table 3. Note that F-score is utilized instead of AUC and AMOC curves because there is no negative sample in Weizmann dataset. From the results, we observe that the F-score of OMED is comparable to that of MMED, especially when the fraction of the observed event is large as seen in Figure 3(a). This demonstrates the effectiveness of the online setting of MMED. Meanwhile, the retraining cost of OMED decreases significantly.

TABLE 3. F-score (mean and deviation) comparison on Weizmann dataset. The results are given at the best Fraction of the event seen. The training times (s) are given with the sample number as 450. Note that "Ins" refers to instantaneous time cost for online method, and "Accl" refers to accumulative time cost.

	MMED		OMED	
	F-score	Time	F-score	Time(Ins/Accl)
Weizmann	0.835±0.020	1415	0.865±0.164	**1.819/657**

TABLE 4. AUC (mean and standard deviation) and training time comparisons on two expression datasets. AUC of online method (OMED) is given at the best value corresponding to the number of training sequences. The training time(s) on CK+ and UvA-NEMO datasets are given with the sample number as 150 and 200 respectively. Note that "Ins" refers to instantaneous time cost for online method, and "Accl" refers to accumulative time cost.

	FrmPeak		FrmAll		MMED		OMED	
	AUC	Time	AUC	Time	AUC	Time	AUC	Time(Ins/Accl)
CK+	0.819±0.009	0.027	0.836±0.007	5.162	0.962±0.007	168	0.951±0.009	**0.025/3.528**
UvA-NEMO	0.622±0.015	0.092	0.647±0.008	4705	0.782±0.008	713	0.798±0.011	**1.136/226**

Note that the curve of MMED in Figure 3(b) stops early because the value is too large. In addition, the time cost of OMED is the accumulative value. The comparison of instantaneous cost when the 450-th training sample comes is illustrated in Table 3. The huge gap between the instantaneous value of MMED and OMED further validates the rapidity of training process of OMED.

E. RESULTS ON CK+ DATASET

The experimental results on CK+ dataset are shown in Figure 4, Figure 6 and Table 4. It can be seen from the results that: 1) segment-based approaches (MMED, OMED) outperform frame-based approaches in both accuracy and timeliness. This demonstrates that the temporal information is well utilized in segment-based methods. 2) the AUC curve of OMED rises rapidly first, and then stays steady with small oscillations, which is comparable to that of MMED. This demonstrates the effectiveness of online setting. 3) the accumulative retaining cost of OMED is almost comparative to that of FrmAll and FrmPeak SVMs. In contrast, the retraining cost of MMED increases rapidly with respect to the number of samples. This demonstrates the efficiency of the proposed online method. Note that MMED curve has some oscillations because the number of iterations needed for convergence varies each time. The comparison of instantaneous time cost in Table 4 indicates that online method is even more efficient than frame-based methods.

F. RESULTS ON UVA-NEMO DATASET

The performance of the compared approaches on UvA-NEMO dataset are shown in Figure 5, Figure 6 and Table 4. From the results, we observe that: 1) the performance of FrmAll is not always better than that of FrmPeak because more information is contained. 2) all the segment-based methods outperform frame-based approaches. 3) the training cost of OMED is reduced significantly compared to MMED. Both the accuracy and timeliness are guaranteed. Note that the training time curve of FrmAll stops when the training number is 200. This is due to the intolerable value when the training number increases.

VI. CONCLUSIONS

Extensive attention has been paid on event detection in the past decades, but few works have been proposed to deal with early event detection. In this paper, an early event detector is designed to detect the partial events as early as possible. Compared with MMED, the proposed approach OMED has the following advantages: 1) OMED works in an online manner, i.e., it is updated by the sequential data one by one. Therefore, the changing data distribution contained in the streaming data is exploited. 2) the retraining time cost and memory consumption are thus decreased significantly to make the model available in large-scale applications. Extensive experiments on three benchmark datasets with various complexities have demonstrated both the effectiveness and efficiency of the proposed method.

APPENDIX

Formula Derivation of (9):

$$
\begin{aligned}
L(\lambda_m^i) &= \frac{1}{2}\|w - w_{i-1}\|_F^2 + C\xi^i - \gamma_m^i\xi^i \\
&\quad + \lambda_m^i\{\mu(\frac{|m_t|}{|m|})[f(X_m^i) - f(X_{m_t}^i) + \Delta(m_t, m)] - \xi^i\} \\
&= \frac{1}{2}\|(w_{i-1} + \lambda_m^i P_m^i) - w_{i-1}\|_F^2 + C\xi^i - \gamma_m^i\xi^i \\
&\quad + \lambda_m^i\{\mu(\frac{|m_t|}{|m|})[(w_{i-1} + \lambda_m^i P_m^i)^T\varphi(X_m^i) \\
&\quad - (w_{i-1} + \lambda_m^i P_m^i)^T\varphi(X_{m_t}^i) + \Delta(m_t, m)] - \xi^i\} \\
&= \frac{1}{2}\|\lambda_m^i P_m^i\|_F^2 + C\xi^i - \gamma_m^i\xi^i \\
&\quad + \lambda_m^i\{\mu(\frac{|m_t|}{|m|})[(w_{i-1} + \lambda_m^i P_m^i)^T\varphi(X_m^i) \\
&\quad - (w_{i-1} + \lambda_m^i P_m^i)^T\varphi(X_{m_t}^i) + \Delta(m_t, m)] - \xi^i\} \\
&= \frac{1}{2}\|\lambda_m^i P_m^i\|_F^2 + (C - \gamma_m^i - \lambda_m^i)\xi^i
\end{aligned}
$$

$$+\lambda_m^i\{\mu(\frac{|m_t|}{|m|})[(w_{i-1}+\lambda_m^i P_m^i)^T\varphi(X_m^i) - (w_{i-1}+\lambda_m^i P_m^i)^T\varphi(X_{m_t}^i)+\Delta(m_t,m)]\}$$
$$=\frac{1}{2}\|\lambda_m^i P_m^i\|_F^2+\lambda_m^i\mu(\frac{|m_t|}{|m|})[(w_{i-1}+\lambda_m^i P_m^i)^T\varphi(X_m^i) - (w_{i-1}+\lambda_m^i P_m^i)^T\varphi(X_{m_t}^i)+\Delta(m_t,m)]. \quad (14)$$

Formula Derivation of (10):

$$\frac{\partial L(\lambda_m^i)}{\partial\lambda_m^i}=\lambda_m^i\|P_m^i\|_F^2+\mu(\frac{|m_t|}{|m|})[(w_{i-1}+\lambda_m^i P_m^i)^T\varphi(X_m^i) - (w_{i-1}+\lambda_c^i P_m^i)^T\varphi(X_{m_t}^i)+\Delta(m_t,m)] + \lambda_m^i\mu(\frac{|m_t|}{|m|})[(P_m^i)^T\varphi(X_m^i)-(P_m^i)^T\varphi(X_{m_t}^i)]$$
$$=\lambda_m^i\|P_m^i\|_F^2+\mu(\frac{|m_t|}{|m|})[w_{i-1}^T\varphi(X_m^i)-w_{i-1}^T\varphi(X_{m_t}^i) +\Delta(m_t,m)]+2\lambda_m^i(P_m^i)^T\mu(\frac{|m_t|}{|m|})[\varphi(X_m^i)-\varphi(X_{m_t}^i)]$$
$$=\lambda_m^i\|P_m^i\|_F^2+\mu(\frac{|m_t|}{|m|})[f_{i-1}(X_m^i)-f_{i-1}(X_{m_t}^i) +\Delta(m_t,m)]-2\lambda_m^i(P_m^i)^T P_m^i$$
$$=\mu(\frac{|m_t|}{|m|})[f_{i-1}(X_m^i)-f_{i-1}(X_{m_t}^i)+\Delta(m_t,m)] -\lambda_m^i\|P_m^i\|_F^2=0. \quad (15)$$

REFERENCES

[1] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1933–1941.

[2] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[3] G. R. Naik, A. H. Al-Timemy, and H. T. Nguyen, "Transradial amputee gesture classification using an optimal number of sEMG Sensors: An approach using ICA clustering," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 8, pp. 837–846, Aug. 2016.

[4] G. R. Naik, D. K. Kumar, and Jayadeva, "Twin SVM for gesture classification using the surface electromyogram," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 301–308, Mar. 2010.

[5] K. Thar, N. H. Tran, T. Z. Oo, and C. S. Hong, "DeepMEC: Mobile edge caching using deep learning," *IEEE Access*, vol. 6, pp. 78260–78275, 2018.

[6] L. Xie, D. Tao, and H. Wei, "Joint structured sparsity regularized multiview dimension reduction for video-based facial expression recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, pp. 28:1–28:21, 2017.

[7] M. A. Mohandes and S. Rehman, "Wind speed extrapolation using machine learning methods and LiDAR measurements," *IEEE Access*, vol. 6, pp. 77634–77642, 2018.

[8] Y. M. Kassa, R. Cuevas, and Á. Cuevas, "A large-scale analysis of facebook's user-base and user engagement growth," *IEEE Access*, vol. 6, pp. 78881–78891, 2018.

[9] Y. Luo, Y. Wen, and D. Tao, "Heterogeneous multitask metric learning across multiple domains," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4051–4064, Sep. 2017.

[10] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence—Video to text," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4534–4542.

[11] J. Zhao, H. Wei, C. Zhang, W. Li, W. Guo, and K. Zhang, "Natural gradient learning algorithms for RBF networks," *Neural Comput.*, vol. 27, no. 2, pp. 481–505, Feb. 2015.

[12] V. P. Kour and S. Arora, "Particle swarm optimization based support vector machine (P-SVM) for the segmentation and classification of plants," *IEEE Access*, vol. 7, pp. 29374–29385, 2019.

[13] L. Xie, H. Wei, and K. Zhang, "Behavioral modeling of nonlinear RF power amplifiers using ensemble SDBCC network," *Neurocomputing*, vol. 154, pp. 24–32, Apr. 2015.

[14] W. Guo, H. Wei, Y.-S. Ong, J. R. Hervas, J. Zhao, H. Wang, and K. Zhang, "Numerical analysis near singularities in RBF networks," *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 1–39, 2018.

[15] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for video sequence background subtraction," *Pattern Recognit.*, vol. 76, pp. 635–649, Apr. 2018.

[16] J. W. Davis and A. Tyagi, "Minimal-latency human action recognition using reliable-inference," *Image Vis. Comput.*, vol. 24, no. 5, pp. 455–472, 2006.

[17] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 1036–1043.

[18] L. Su and Y. Sato, "Early facial expression recognition using early RankBoost," in *Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–7.

[19] J. Wang, S. Wang, and Q. Ji, "Early facial expression recognition using hidden Markov models," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2014, pp. 4594–4599.

[20] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. LeCun, "Predicting deeper into the future of semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 648–657.

[21] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3332–3341.

[22] K. Ehsani, H. Bagherinezhad, J. Redmon, R. Mottaghi, and A. Farhadi, "Who let the dogs out? Modeling dog behavior from visual data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4051–4060.

[23] M. Huh, A. Liu, A. Owens, and A. A. Efros, "Fighting fake news: Image splice detection via learned self-consistency," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 101–117.

[24] D. Xie, T. Shu, S. Todorovic, and S.-C. Zhu, "Learning and inferring 'dark matter' and predicting human intents and trajectories in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1639–1652, Jul. 2018.

[25] Z. Shou, J. Pan, J. Chan, K. Miyazawa, H. Mansour, A. Vetro, X. Giro-i Nieto, and S.-F. Chang, "Online detection of action start in untrimmed, streaming videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 534–551.

[26] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, "SST: Single-stream temporal action proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2911–2920.

[27] A. Dave, O. Russakovsky, and D. Ramanan, "Predictive-corrective networks for action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 981–990.

[28] H. Xu, A. Das, and K. Saenko, " R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5783–5792.

[29] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia, "TURN TAP: Temporal unit regression network for temporal action proposals," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3628–3636.

[30] M. Hoai and F. De la Torre, "Max-margin early event detectors," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 191–202, 2014.

[31] W. Guo, Y.-S. Ong, Y. Zhou, J. R. Hervas, A. Song, and H. Wei, "Fisher information matrix of unipolar activation function-based multilayer perceptrons," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 3088–3098, Aug. 2019. doi: 10.1109/TCYB.2018.2838680.

[32] L. Xie, H. Wei, J. Zhao, and K. Zhang, "Automatic feature extraction based structure decomposition method for multi-classification," *Neurocomputing*, vol. 173, pp. 744–750, Jan. 2016.

[33] L. Xiang, G. Zhao, Q. Li, W. Hao, and F. Li, "TUMK-ELM: A fast unsupervised heterogeneous data learning approach," *IEEE Access*, vol. 6, pp. 35305–35315, 2018.

[34] Y. Luo, Y. Wen, T. Liu, and D. Tao, "Transferring knowledge fragments for learning distance metric from a heterogeneous domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 1013–1026, Apr. 2019.

[35] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6, p. 386, 1958.

[36] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 551–585, Dec. 2006.

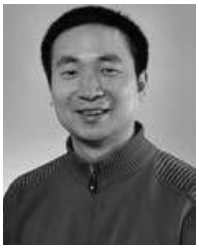
[37] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. Int. Conf. Mach. Learn.(ICML)*, 2003, pp. 928–936.
[38] J. Wang, P. Zhao, and S. C. H. Hoi, "Exact soft confidence-weighted learning," 2012, *arXiv:1206.4612*. [Online]. Available: https://arxiv.org/abs/1206.4612
[39] L. Xie, D. Tao, and H. Wei, "Early expression detection via online multi-instance learning with nonlinear extension," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1486–1496, May 2019.
[40] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile, "Tracking the best hyperplane with a simple budget perceptron," *Mach. Learn.*, vol. 69, no. 2, pp. 143–167, 2007.
[41] Z. Wang and S. Vucetic, "Online passive-aggressive algorithms on a budget," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 908–915.
[42] Z. Wang, K. Crammer, and S. Vucetic, "Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale SVM training," *J. Mach. Learn. Res.*, vol. 13, pp. 3103–3131, Oct. 2012.
[43] A. Bordes and L. Bottou, "The huller: A simple and efficient online SVM," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 2005, pp. 505–512.
[44] D. Sculley and G. M. Wachman, "Relaxed online SVMs for spam filtering," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2007, pp. 415–422.
[45] J. Xu, Y. Y. Tang, B. Zou, Z. Xu, L. Li, and Y. Lu, "The generalization ability of online SVM classification based on Markov sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 628–639, Mar. 2015.
[46] S. Hare, A. Saffari, and P. H. S. Torr, "Efficient online structured output learning for keypoint-based object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1894–1901.
[47] C. Alan, "The intrinsic computational difficulty of functions," in *Proc. Int. Congr.*, 1965, pp. 24–30.
[48] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vols. 1–2, Oct. 2005, pp. 1395–1402.
[49] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.
[50] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2010, pp. 94–101.
[51] H. Dibeklioğlu, A. Salah, and T. Gevers, "Are you really smiling at me? Spontaneous versus posed enjoyment smiles," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 525–538.
[52] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," 2016, *arXiv:1604.08880*. [Online]. Available: https://arxiv.org/abs/1604.08880
[53] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[54] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine," *Clin. Chem.*, vol. 39, no. 4, pp. 561–577, 1993.
[55] T. Fawcett and F. Provost, "Activity monitoring: Noticing interesting changes in behavior," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 1999, pp. 53–62.

LIPING XIE received the B.S. and Ph.D. degrees from the School of Automation, Southeast University, China, in 2011 and 2017, respectively. She was a Visiting Student with the Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia, from 2015 to 2017. She is currently a Lecturer with the School of Automation, Southeast University. Her research interests include computer vision, machine learning, and their applications to video understanding and analysis.

JUNSHENG ZHAO received the M.S. degrees from the School of Mathematical Science, Qufu Normal University, in 2006, and the Ph.D. degree from the School of Automation, Southeast University, China, in 2015. Since 2018, he has been an Associate Professor with the School of Mathematical Science, Liaocheng University. His research interests include singular learning dynamics of neural networks, estimation, and control and its applications.

HAIKUN WEI received the B.S. degree from the Department of Automation, North China University of Technology, China, in 1994, and the M.S. and Ph.D. degrees from the Research Institute of Automation, Southeast University, China, in 1997 and 2000, respectively.

He was a Visiting Scholar with the RIKEN Brain Science Institute, Japan, from 2005 to 2007. He was a member of the China's 27th to the Antarctic Expedition Team. He is currently a Professor with the School of Automation, Southeast University. His research interests include neural networks based machine learning, and its application in industry automation.



ZHUN FAN received the B.S. and M.S. degrees in control engineering from the Huazhong University of Science and Technology, Wuhan, China, in 1995 and 2000, respectively, and the Ph.D. degree in electrical and computer engineering from Michigan State University, Lansing, MI, USA, in 2004. He is currently a Full Professor with Shantou University (STU), Shantou, China. Before joining STU, he was an Associate Professor with the Technical University of Denmark (DTU), from 2007 to 2011, first with the Department of Mechanical Engineering, then with the Department of Management Engineering, and as an Assistant Professor with the Department of Mechanical Engineering, from 2004 to 2007. His major research interests include intelligent control and robotic systems, robot vision and cognition, MEMS, computational intelligence, design automation, optimization of mechatronic systems, machine learning, and image processing.

GUOCHEN PANG was born in Linyi, Shandong, China, in 1987. He received the M.S. degrees in operational Research and cybernetics from Qufu Normal University, Qufu, Shangdong, in 2012, and the Ph.D. degree in control theory and control engineering from Southeast University, China, in 2016. Since 2016, he has been a Lecturer with the School of Automation and Electrical Engineering, Linyi University. His research interests include fault detection, robust control actuator saturation, anti-disturbance control, and its applications.