Semi-supervised Early Event Detection

Liping Xie[1,2(✉)], Chen Gong[2], Jinxia Zhang[1], Shuo Shan[1], and Haikun Wei[1]

[1] Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China
{lpxie,jinxiazhang,230189536,hkwei}@seu.edu.cn

[2] Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
chen.gong@njust.edu.cn

Abstract. Early event detection is among the keys in the field of event detection due to its timeliness in widespread applications. The objective of early detection (ED) is to identify the specified event of the video sequence as early as possible before its ending. This paper introduces semi-supervised learning to ED, which is the first attempt to utilize the domain knowledge in the field of ED. In this setting, some domain knowledge in the form of pairwise constraints is available. Particularly, we treat the segments of complete events as must-link constraints. Furthermore, some segments do not overlap with the event are put together with the complete events as cannot-link constraints. Thus, a new algorithm termed semi-supervised ED (SemiED) is proposed, which could make better early detection for videos. The SemiED algorithm is a convex quadratic programming problem, which could be resolved efficiently. We also discuss the computational complexity of SemiED to evaluate its effectiveness. The superiority of the proposed method is validated on two video-based datasets.

Keywords: Semi-supervised learning · Domain knowledge · Early event detection

1 Introduction

While event detection has received rapid accumulation over the past decades [15, 19], little attention has been given to early detection (ED) [10,23,25]. ED aims to make a reliable early identification for video sequences, where the temporal information needs to be well exploited. Different with event detection where the result is given after the video ends, ED gives sequential results, i.e., each frame contained in the video corresponds to an independent detection score. An efficient algorithm can obtain an early identification as soon as possible based on these frame-level outputs, and the accuracy is also guaranteed. The potential applications of ED can range form healthcare [4], environmental science [21], artificial intelligence [14,24], etc. For example, the timeliness in human-computer

S. Palaiahnakote et al. (Eds.): ACPR 2019, LNCS 12047, pp. 678–690, 2020.
https://doi.org/10.1007/978-3-030-41299-9_53

interaction is the key issue to provide an efficient and comfortable communication especially in this age of artificial intelligence. Therefore, ED will become more and more important in the future.

Over the past few years, several works have been proposed for ED. The max-margin early event detector (MMED) [10] is the first and also the most representative model for ED. Subsequently, [23] applies multi-instance learning to early detection based on MMED and obtain better performance in both of accuracy and timeliness. In addition, [23] further extends the model in an online manner to deal with streaming sequences and make it available in large-scale applications. Some frameworks with several assumptions are also presented in recent years. [7] tries to give a probability ration test based on probabilistic reliable-inference, where the model is still trained on the whole video sequences. [16] designs a RankBoost-based approach where the length of the testing video needs to be provided before testing. [20] suffers the same problem with [16], which is also unpractical to be utilized in real-world applications.

The application of domain knowledge to various data mining fields attracts increasing attention due to its effectiveness [26]. Normally, the forms of domain knowledge could be expressed in several ways: class labels, pairwise constraints and other prior information [6]. In most cases, the domain knowledge in the form of pairwise constraints is utilized since the cost of labelling is expensive in real-world applications. In contrast, whether the pairs of samples belong to the same type of class or not is relatively practical to obtain. In general, the pairs from the same class are treated as the must-link constraints, and the pairs form different classes are for cannot-link constraints. Moreover, the pairwise constraints can be derived from labeled data but not vice versa. Pairwise constraints have been applied in many applications such as facial expression recognition [11], image retrieval [2,5], dimension reduction [6,26] and image segmentation [12,13].

Although studied extensively of domain knowledge in various applications, few attentions have been given to ED. Introducing domain knowledge [22] to ED raises a natural and new problem: how to combine the detection function with the provided domain knowledge to obtain a more accurate score for each frame. There are two key issues existed in semi-supervised early event detection. The most critical one is to make full use of the domain knowledge to help ED model learning. The objective of ED is to improve the performance in terms of both timeliness and accuracy. In addition, how to define the pairwise constraints is another problem since the label is given for video sequence and not for segments. Therefore, the segments utilized for model training have no label information.

In this paper, we study early detection where the domain knowledge is available in the form of pairwise constraints. A novel algorithm, termed semi-supervised early detection (SemiED), is proposed. Inspired by the success of MMED, we follow the framework of structured output SVM (SOSVM), and the monotonicity of the detection function is learnt by the pairwise constraints. Based on MMED, we introduce semi-supervised learning to SOSVM. Particularly, the pairwise constraints for domain knowledge in SemiED consists of two parts: (1) each video sequence from the datasets contains a complete event.

We thus treat these complete events as the must-link constraints since they all correspond to the greatest value of the detection function; (2) we extract some segments before the event fires or after it ends, i.e., these segments have no overlap with the complete events. We put these segments together with the complete events as the cannot-link constraints. We develop an efficient algorithm for problem optimization. The comparison of computational complexity of MMED and the proposed SemiED is also provided. The superiority of SemiED is validated on two popular video-based datasets with various complexities.

2 SemiED: Semi-supervised Early Event Detection

Notations: In this paper, the video sequences for training and their associated ground truth annotations of interest are denoted as (X^1, y^1), $(X^2, y^2), ..., (X^n, y^n)$. Here, two elements are contained in $y^i = [s^i, e^i]$ to indicate the beginning and ending of the event for the i-th training sample X^i. n denotes the total number of training videos. The length of the i-th training video X^i is represented as l^i. For every frame time $t = 1, ..., l^i$, we adopt y^i_t to denote the partial event of y^i that has occurred, i.e., $y^i_t = y^i \cap [1, t]$. y^i_t may be empty if no event is contained. In addition, $\mathcal{Y}(t)$ is utilized to denote all the possible segments from the 1-st to the t-th frames: $\mathcal{Y}(t) = \{y \in \mathbb{N}^2 | y \subset [1, t]\} \cup \{\emptyset\}$. The segment $y = \emptyset$, indicates y is empty and there is no event occurs. If a video sequence X has the number of frames, which is denoted as l, then $\mathcal{Y}(l)$ indicates the set that all possible segments are contained. Note that, for an arbitrary segment $y = [s, e] \in \mathcal{Y}(l)$, X_y is the subset of X from the s-th to the e-th frames.

2.1 Formulation

Following [10], we employ structured output SVM for early event detection. The monotonicity of the detection function is achieved by extracting various pairwise constraints and ranking them based on the information contained. Before the formulation, we first give the detection function as follows:

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

where $\varphi(X_y)$ is the feature representation of segment X_y. In this paper, we utilize $f(X_y)$ to denote $f(X_y, w, b)$ for brevity. The fundamental principle of ED is that the score of a positive training sample $X^i_{y^i}$ is greater than the that of any other segment from the same video sequence, i.e., $f(X^i_{y^i}) > f(X^i_y), \forall y \neq y^i$. The early event detection based on SOSVM can be written as follows:

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

where $|\cdot|$ represents the length of the segments, and $\mu(\cdot)$ is a rescaling factor of slack variable. Following [10], the piece-wise function with linearity is employed as follows:

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

$\triangle(\cdot)$ is an adaptive margin of the pairwise segments and denoted as the loss of the detector for outputting y when the desired output is y^i, i.e, $\triangle(y_t^i, y) = 1 - overlap(y^i, y)$.

In this paper, we use M and C to denote the number of must-link and cannot-link constraints respectively. We first denote $\Phi = [\varphi(X_y^1), \varphi(X_{y^1}^1), \varphi(X_y^2), \varphi(X_{y^2}^2)...\varphi(X_y^n), \varphi(X_{y^n}^n)] \in \mathbb{R}^{d\times 2n}$, $\varphi(X_y^i)$ is the segment from the 1-st frame to the event beginning. $\varphi(X_{y^i}^i)$ represents the feature vector of the complete event in video sample X^i. For brevity, we use $\Phi = [\varphi(y_1), \varphi(y_2)...\varphi(y_{2n})] \in \mathbb{R}^{d\times 2n}$. Then, the objective function of the domain knowledge can be written as minimizing $J(w)$:

$$\begin{aligned} & J(w) \\ &= \frac{\beta}{2n_M} \sum_{(y_i,y_j)\in M} (f(y_i) - f(y_j))^2 - \frac{\alpha}{2n_C} \sum_{(y_i,y_j)\in C} (f(y_i) - f(y_j))^2 \\ &= \frac{\beta}{2n_M} \sum_{(y_i,y_j)\in M} (w^T\varphi(y_i) - w^T\varphi(y_j))^2 - \frac{\alpha}{2n_C} \sum_{(y_i,y_j)\in C} (w^T\varphi(y_i) - w^T\varphi(y_j))^2 \end{aligned} \tag{2}$$

The concise form of $J(w)$ can be further denoted as follows:

$$J(w) = \frac{1}{2}\sum_{i,j}(w^T\varphi(y_i) - w^T\varphi(y_j))^2 S_{ij} \tag{3}$$

where

$$S_{ij} = \begin{cases} -\dfrac{\alpha}{n_C}, & if(y_i, y_j) \in C, \\ \dfrac{\beta}{n_M}, & if(y_i, y_j) \in M, \\ 0, & otherwise. \end{cases} \tag{4}$$

Then we reformulate (3) as:

$$\begin{aligned} J(w) &= \frac{1}{2}\sum_{i,j}(w^T\varphi(y_i) - w^T\varphi(y_j))^2 S_{ij} \\ &= \frac{1}{2}\sum_{i,j}(w^T\varphi(y_i)\varphi(y_i)^T w + w^T\varphi(y_j)\varphi(y_j)^T w - 2w^T\varphi(y_i)\varphi(y_j)^T w)S_{ij} \\ &= \sum_{i,j}(w^T\varphi(y_i)\varphi(y_i)^T w - \sum_{i,j} w^T\varphi(y_i)\varphi(y_j)^T w)S_{ij} \\ &= \sum_{i,j} w^T\varphi(y_i)S_{ij}\varphi(y_i)^T w - \sum_{i,j} w^T\varphi(y_i)S_{ij}\varphi(y_j)^T w \\ &= \sum_{i} w^T\varphi(y_i)D_{ii}\varphi(y_i)^T w - w^T\Phi S\Phi^T w \\ &= w^T\Phi(D-S)\Phi^T w \\ &= w^T\Phi L\Phi^T w \end{aligned} \tag{5}$$

where D denotes a diagonal matrix where the entries are rows sums of S, i.e., $D_{ij} = \sum_j S_{ij}$. $L = D - S$ is called the Laplacian matrix. Therefore, the regularizer (5) can further be simplified as minimizing $J(w)$, where

$$J(w) = w^T\Phi L\Phi^T w \tag{6}$$

where L could be computed by the domain knowledge. Therefore, the semi-supervised early event detection (SemiED) can be formulated as follows:

$$\begin{aligned} &\min_{\{w,b,\xi^i\geq 0\}} \frac{1}{2}\|w\|_F^2 + \gamma\sum_{i=1}^{n}\xi^i + w^T\Phi L\Phi^T w, \\ &\text{s.t.}\quad f(X^i_{y^i_t}) \geq f(X^i_y) + \Delta(y^i_t, y) - \frac{\xi^i}{\mu(\frac{|y^i_t|}{|y^i|})}, \\ &\qquad \forall i, \forall t = 1, ..., l^i, \forall y \in \mathcal{Y}(t). \end{aligned} \tag{7}$$

During the experiments, we set $\alpha = 1$ and $\beta = 20$ in SemiED, which are the empirical values. The detailed information of optimization and analysis of computational complexity can be seen in the following.

2.2 Optimization

Before the detailed optimization procedure, we first reformulate the problem (7) as follows:

$$\begin{aligned} &\min_{\{w,b,\xi^i\geq 0\}} \frac{1}{2}w^T(I_d + \Phi L\Phi^T)w + \gamma\sum_{i=1}^{n}\xi^i, \\ &\text{s.t.}\quad f(X^i_{y^i_t}) \geq f(X^i_y) + \Delta(y^i_t, y) - \frac{\xi^i}{\mu(\frac{|y^i_t|}{|y^i|})}, \\ &\qquad \forall i, \forall t = 1, ..., l^i, \forall y \in \mathcal{Y}(t). \end{aligned} \tag{8}$$

Here, I_d represents the identity matrix with size of $d \times d$. Then, we set $u = [w^T\ b\ \xi^1\ \xi^2 ... \xi^n]^T \in \mathbb{R}^N$, $N = d + 1 + n$. d represents the length of the feature vector, n is the total number of sequences. Problem (8) can be rewritten as follows:

$$\begin{aligned} &\min_u \frac{1}{2} u^T H u + hu, \\ &\text{s.t.} \quad f(X^i_{y^i_t}) \geq f(X^i_y) + \Delta(y^i_t, y) - \frac{\xi^i}{\mu(\frac{|y^i_t|}{|y^i|})}, \\ &\forall i, \forall t = 1, ..., l^i, \forall y \in \mathcal{Y}(t), \\ &u_i \geq 0, i \in (d+2, d+1+n). \end{aligned} \tag{9}$$

Here, H is a block diagonal matrix with the 1-st block as $I_d + \Phi L \Phi^T$ and 0 for others. $h = [0...0\ 0\ \gamma ... \gamma]$ with n nonzero elements. We can see that this is a convex quadratic programming problem which has massive pairwise constraints. In this paper, the constraint generation strategy utilized in [18] is employed, which reduces the memory consumption by several iterations. It has been validated that the convergence to the global minimum could be guaranteed.

2.3 Computational Complexity

This section presents the discussion of the computational cost for the proposed algorithm SemiED with the comparison to MMED, which is the most representative framework for ED and the proposed SemiED is designed based on MMED.

(1) MMED: The optimization of MMED is a standard quadratic programming problem. According to the computational complexity theory stated in [17], the complexity of MMED is $O((n + d + 1)^3 S)$ and roughly estimated by $O((n + d)^3 S)$. Here, n is the number of training samples, d is the dimension of the feature and S denotes the size of the problem encoding in binary.

(2) SemiED: Similar to MMED, the solution of SemiED is a quadratic programming problem, and the complexity of optimization procedure is also $O((n + d)^3 S)$. In addition, SemiED needs to compute the regularizer and the most expensive operation contained is the computation of $(I_d + \Phi L \Phi^T)$, which needs the complexity of $O(n^2 d + n d^2)$. Therefore, the computational complexity of SemiED is $O((n+d)^3 S)$. We can see that SemiED shares the same complexity with MMED although domain knowledge is adopted in SemiED.

3 Experiments

This section validates the performance in terms of accuracy, timeliness and training time cost on two video-based datasets: Weizmann dataset and UvA-NEMO dataset. There are 10 different actions and two smiles in Weizmann and UvA-NEMO respectively. During our experiments, the specified event "Bend" on Weizmann dataset and the event "spontaneous smile" on UvA-NEMO dataset

are chosen for detection. In the following, we first introduce the datasets, experiment setup, as well as evaluation criteria before the analysis of experimental results.

3.1 Datasets

Weizmann Dataset. It [3] is created from 9 subjects and 90 video sequences are contained, i.e, every subject performs 10 actions. The specified actions are: Bend, Jack, Jump, Pjump, Run, Side, Skip, Walk, Wave1 and Wave2. Following [10], we concatenate all actions of the same subject to construct a longer video sequence. Particularly, we put the targeted event "Bend" at the end of each longer sequence. In this paper, the AlexNet architecture [9] is utilized for frame-level feature extraction, in which the features have the dimensionality of 4096. Then, we apply PCA for feature dimension reduction, and 1000-dimensional features are obtained. Note that we utilize the leave-one-out cross validation for Weizmann dataset due to the limitation of video samples.

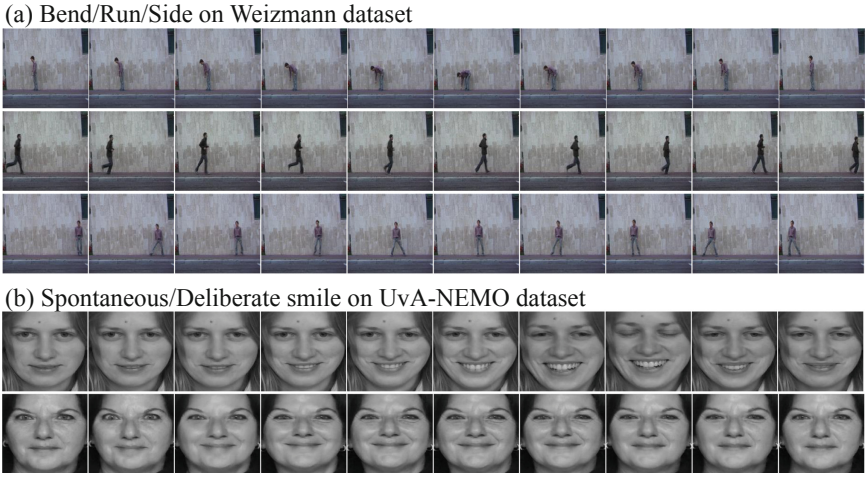

Fig. 1. Examples from two video-based datasets. Note that the length of each sequence is set to the same as "10" for better illustration, which does not represent the actual length. (a) Bend/Run/Side on Weizmann dataset; (b) Spontaneous/Deliberate smile on UvA-NEMO dataset.

Table 1. Statistics of benchmark datasets.

	#Sample					#Feature
	Training				Test	
	Video	Pairwise constrnts	M constrnts	C constrnts		
Weizmann	8	120	16	8	1	1000
UvA-NEMO	800	1200	1600	800	440	944

UvA-NEMO Dataset. It is built for the analysis of the dynamic difference from spontaneous/deliberate smiles. It consists of 1240 smile videos, where 597 are spontaneous. All the videos contained begin with a neutral or near-neutral frame, and the same situation with the ending frame. During the experiments, the Local Binary Patterns (LBP) [1] is adopted for frame-level feature extraction. Particularly, we crop each frame into 4×4 blocks and the number of neighboring points is set to 2^3. Thus, the dimensionality of the extracted feature is $4 \times 4 \times 59 = 944$.

We present some examples for illustration from each dataset in Fig. 1, and summarize the statistics in Table 1.

3.2 Experiment Setup

Compared Approaches. The results of SemiED are compared with two baseline methods (FrmPeak, FrmAll) [23] and MMED [10]. FrmPeak and FrmAll are frame-based SVMs, where the detection result is obtained by classifying each frame contained in the video. The difference between FrmPeak and FrmAll is the samples used for training. FrmPeak utilizes only the peak frames for training while in FrmAll, all the frames are employed. MMED and the proposed SemiED are segment-based methods.

Experiment Setting. In this paper, we tune the parameter γ contained in the proposed SemiED model and the parameter in SVM on set $\{10^i | i = -5, -4, ..., 3, 4, 5\}$, and the best performance is reported. Note that the SVMs utilized for FrmPeak and FrmAll are linear. During the experiments, the overlap of the extracted pairwise constraints is set to be lower than 0.7. This is to guarantee the discrimination for the training set. We conduct all the compared experiments on the computer with the following configurations: Intel(R) Xeon(R) Core-20, CPU E5-2650 v3-2.3 GHz, Memory 48 GB, LINUX operating system with Matlab 2015a.

3.3 Evaluation Criteria

This section presents the introduction of the evaluation criteria employed during our experiments: F-score [9], AUC [27], AMOC [8] and the training time curve. F-score is a measurement for detection accuracy, which is usually adopted for binary classification in statistical analysis. The two main variables in F-score are the precision p and the recall r on the testing set, which is computed as follows: $p = \frac{|y \cap y^*|}{|y|}$, $r = \frac{|y \cap y^*|}{|y^*|}$, where y^* is the segment contains the event and y is the output event that detected. AUC is short for the area under the Receiver Operating Characteristic (ROC) curve, which is also used for the evaluation of accuracy. In AUC, TPR denotes that the model fires during the event of interest, and FPR is that the model fires before the beginning or after the ending. AMOC represents the Activity Monitoring Operating Curve, which is used to evaluate the timeliness of detection. In AMOC, Normalized Time to Detection

(NTtoD) is computed, and the value is the lower the better. NTtoD is denoted as NTtoD= $\frac{t-s+1}{e-s+1}$, where t, s, and e denote the current-, the starting- and the ending-frame, respectively.

3.4 Results on Weizmann Dataset

The compared performance with different methods can be seen in Figs. 2, 3 and Table 2. F-score is adopted on this dataset instead of AUC and AMOC since there is no negative training sequences contained and the return type of FPR is void. Similarly, only the results of segment-based methods are reported since SVM could not be used for FrmAll and FrmPeak due to the lack of negative samples. From the results, we can see that: (1) The F-score of SemiED is obviously better than that of MMED on the whole event fraction. This demonstrates the effectiveness of SemiED and is consistent with the analysis that the utilization of domain knowledge could help model learning for better detection function. (2) The training time cost of SemiED is comparable to that of MMED. This demonstrates that the additional computation of regularizer has no increasing for the complexity. This is also consistent with the analysis of computational complexity in Sect. 2.3. Note that the training cost of SemiED does not strictly increase with the training size. This is because the number of iterations needed for each quadratic optimization is not fixed, and it is varied from the number of training sequences. (3) The comparison of timeliness is shown in Fig. 3. The results further demonstrate the effectiveness of the proposed SemiED. The NTtoD of SemiED is obviously lower than that of MMED, which is used to denote the normalized detection time.

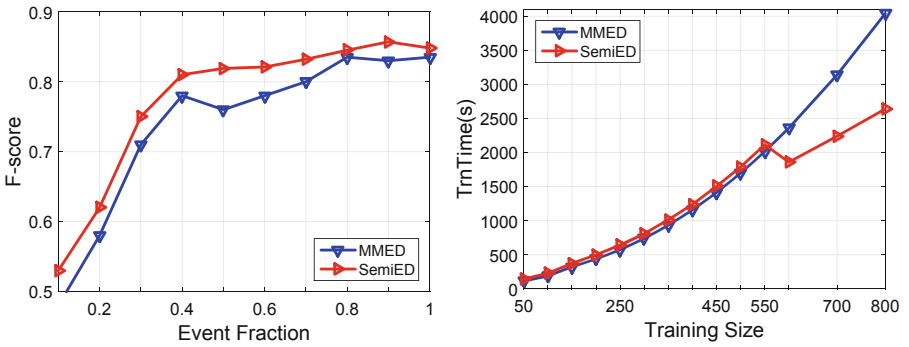

Fig. 2. Experimental results on Weizmann dataset. (a) F-score comparison for the performance of accuracy. The score of F-score is the higher the better. (b) Comparison of training time cost for the performance of efficiency. The value of time cost is the lower the better.

Table 2. Comparisons of F-score (mean and deviation) value and the training time cost (s) of segment-based methods on Weizmann dataset. The F-score results are provided with the best event fraction seen. The training time cost is reported by setting the number of training sequences as 450.

	MMED		SemiED	
	F-score	Time	F-score	Time
Weizmann	0.835 ± 0.020	1415	0.857 ± 0.121	1505

3.5 Results on UvA-NEMO Dataset

AUC, AMOC and training cost are utilized for comparison on UvA-NEMO dataset. From the results in Figs. 3, 4, and Table 3, some conclusions could be obtained: (1) Segment-based methods obviously outperform baselines in terms of both accuracy and timeliness. The time cost of FrmAll is obviously much greater than other methods. Although the cost of FrmPeak is much lower than the segment-based methods, the accuracy and timeliness of FrmPeak could not be accepted. (2) Compared with MMED, SemiED achieves much better performance of the timeliness, especially when the value of FPR is less than 0.3. Note that we usually give attention to the low FPR since the value is meaningless when greater. (3) The time cost of SemiED is approximately the same as that of MMED, in which both MMED and SemiED are much lower compared with FrmAll. This demonstrates that the utilization of domain knowledge brings no additional complexity. (4) Figure 3 illustrates the NTtoD comparison of MMED and the proposed SemiED on UvA-NEMO dataset. The lower value NTtoD is, the better timeliness the detection is. The results demonstrate that the

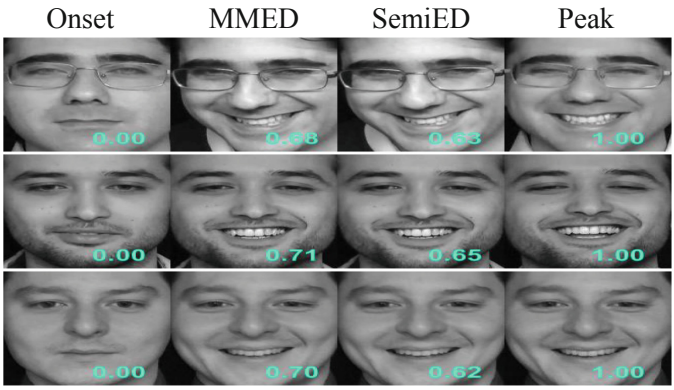

Fig. 3. NTtoD of some examples from UvA-NEMO dataset. The four frames in each row from left to right are: the onset, the ones MMED and SemiED make a detection and the peak one. The number on each frame denotes the value of NTtoD. For a testing sequence, the length from the onset to the peak frame is normalized as "1". Therefore, the value of NTtoD is the lower the better.

Table 3. Comparisons of AUC (mean and standard deviation) value and the training time cost (s) of the different methods on UvA-NEMO dataset. The training time cost is reported by setting the number of training sequences as 200.

	FrmPeak		FrmAll		MMED		SemiED	
	AUC	Time	AUC	Time	AUC	Time	AUC	Time
UvA-NEMO	0.62 ± 0.02	0.09	0.65 ± 0.01	4705	0.78 ± 0.01	713	0.81 ± 0.01	759

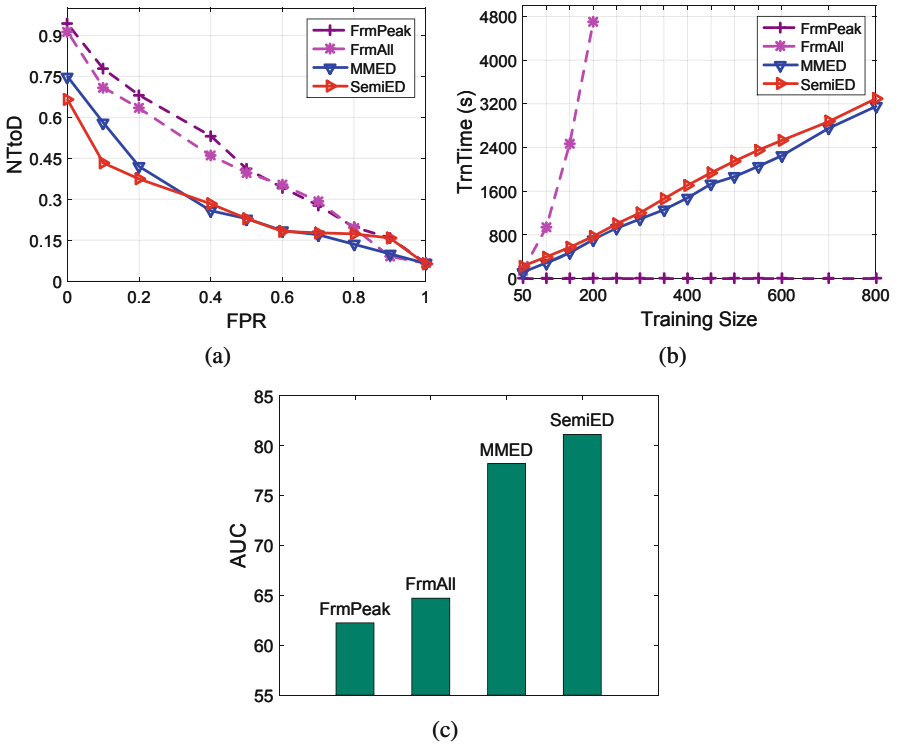

Fig. 4. Experimental results on UvA-NEMO dataset. (a) AUC comparison for the performance of accuracy. The value of AUC is the higher the better. (b) AMOC comparison for the performance of timeliness. The value of AMOC is the lower the better. (c) Comparison of training time cost for the comparison of efficiency. The value of time cost is the lower the better.

exploitation of domain knowledge makes the detector achieve better performance of not only the accuracy but also the timeliness, which is especially important for ED.

4 Conclusions

Early event detection is a relatively new and challenging problem, which receives few attentions in the past years. Motivated by the semi-supervised learning and the success of representative MMED, we propose an efficient semi-supervised early event detection algorithm called SemiED in this paper. SemiED exploits both the must-link and cannot-link constraints for the early detection function. The domain knowledge is thus well utilized to obtain better detection performance. We also provide the theoretical analysis to demonstrate that the domain knowledge utilized has no additional computation of the complexity compared with MMED. Extensive experiments and comparisons on two video-based datasets illustrate that the proposed SemiED enjoys much better performance in terms of both accuracy and timeliness than that of MMED while the training time cost is comparable with each other.

Acknowledgments. This work was supported in part by the National NSF of China under Grant 61802059, in part by the NSF of Jiangsu under Grant BK20180365, in part by the Innovation Fund of Key Laboratory of Measurement and Control of CSE through Southeast University under Grant MCCSE2018B01 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 under Grant 30918014107.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 2037–2041 (2006)
2. Alemu, L.T., Pelillo, M.: Multi-feature fusion for image retrieval using constrained dominant sets. arXiv preprint arXiv:1808.05075 (2018)
3. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision (ICCV 2005) Volume 1, vol. 2, pp. 1395–1402. IEEE (2005)
4. Bonifait, L., et al.: Detection and quantification of airborne norovirus during outbreaks in healthcare facilities. Clin. Infect. Dis. **61**(3), 299–304 (2015)
5. Chung, Y.A., Weng, W.H.: Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. arXiv preprint arXiv:1711.08490 (2017)
6. Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: Advances in Neural Information Processing Systems, pp. 3079–3087 (2015)
7. Davis, J.W., Tyagi, A.: Minimal-latency human action recognition using reliable-inference. Image Vis. Comput. **24**(5), 455–472 (2006)
8. Fawcett, T., Provost, F.: Activity monitoring: Noticing interesting changes in behavior. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 53–62 (1999)
9. Hammerla, N.Y., Halloran, S., Plötz, T.: Deep, convolutional, and recurrent models for human activity recognition using wearables. arXiv preprint arXiv:1604.08880 (2016)

10. Hoai, M., De la Torre, F.: Max-margin early event detectors. Int. J. Comput. Vis. **107**(2), 191–202 (2014)
11. Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y.: Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 558–565. IEEE (2017)
12. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4500–4509 (2018)
13. Pape, C., Matskevych, A., Hennies, J., Kreshuk, A.: Leveraging domain knowledge to improve em image segmentation with lifted multicuts. arXiv preprint arXiv:1905.10535 (2019)
14. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artif. Intell. Rev. **43**(1), 1–54 (2015)
15. Satkin, S., Hebert, M.: Modeling the temporal extent of actions. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 536–548. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15549-9_39
16. Su, L., Sato, Y.: Early facial expression recognition using early rankboost. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–7 (2013)
17. Tseng, P., et al.: A simple polynomial-time algorithm for convex quadratic programming (1988)
18. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res. **6**(Sep), 1453–1484 (2005)
19. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542 (2015)
20. Wang, J., Wang, S., Ji, Q.: Early facial expression recognition using hidden markov models. In: International Conference on Pattern Recognition (ICPR), pp. 4594–4599 (2014)
21. Wilcox, T.M., et al.: Understanding environmental dna detection probabilities: a case study using a stream-dwelling char salvelinus fontinalis. Biol. Conserv. **194**, 209–216 (2016)
22. Xie, L., Tao, D., Wei, H.: Joint structured sparsity regularized multiview dimension reduction for video-based facial expression recognition. ACM Trans. Intell. Syst. Technol. (TIST) **8**(2), 28 (2017)
23. Xie, L., Tao, D., Wei, H.: Early expression detection via online multi-instance learning with nonlinear extension. IEEE Trans. Neural Netw. Learn. Syst. **30**(5), 1486–1496 (2019)
24. Xie, L., Wei, H., Zhao, J., Zhang, K.: Automatic feature extraction based structure decomposition method for multi-classification. Neurocomputing **173**, 744–750 (2016)
25. Xie, L., Zhao, J., Wei, H., Zhang, K., Pang, G.: Online kernel-based structured output svm for early expression detection. IEEE Sign. Process. Lett. **26**(9), 1305–1309 (2019)
26. Zhang, D., Zhou, Z.H., Chen, S.: Semi-supervised dimensionality reduction. In: Proceedings of the 2007 SIAM International Conference on Data Mining, pp. 629–634. SIAM (2007)
27. Zweig, M.H., Campbell, G.: Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. **39**(4), 561–577 (1993)