Deep Spiking Neural Network for Video-based Disguise Face Recognition Based on Dynamic Facial Movements

Daqi Liu, Nicola Bellotto, *Member, IEEE*, and Shigang Yue, *Senior Member, IEEE*

***Abstract*—With the increasing popularity of social media and smart devices, the face as one of the key biometrics becomes vital for person identification. Amongst those face recognition algorithms, video-based face recognition methods could make use of both temporal and spatial information just as humans do to achieve better classification performance. However, they cannot identify individuals when certain key facial areas like eyes or nose are disguised by heavy makeup or rubber/digital masks. To this end, we propose a novel deep spiking neural network architecture in this study. It takes dynamic facial movements, the facial muscle changes induced by speaking or other activities, as the sole input. An event-driven continuous spike-timing dependent plasticity learning rule with adaptive thresholding is applied to train the synaptic weights. The experiments on our proposed video-based disguise face database (MakeFace DB) demonstrate that the proposed learning method performs very well - it achieves from 95% to 100% correct classification rates under various realistic experimental scenarios.**

***Index Terms*—Deep learning, spiking neural network, event-driven STDP, continuous learning, video-based disguise face recognition.**

I. Introduction

FACE is a vital biometric for person identification. Due to the satisfying recognition performances, the face biometrics-based identification systems are becoming ubiquitous, i.e. the Face ID unlocking mechanism of the iPhone X or the electronic passports allowing the automatic verification of traveller identity. However, such systems can only work well under relative constraint circumstances and are vulnerable for suspicious anti-surveillance behaviours. For instance, the face unlocking schemes used in current smart devices, under some common applying circumstances, cannot recognize the owners if they wear certain necessary disguises (such as respirator masks or specific sun-glasses). For public safety, many law enforcement agencies need to scan faces from the surveillance footages. However, the suspects can easily fool the current face biometrics-based identification systems by simply wear different disguise accessories, such as mask, hat or sun-glasses. The failure of identification could potentially jeopardize the public safety.

To address those issues, various disguise face recognition methods are proposed [1], [2], [3], [4]. However, they are solely applied for still images and only incorporate spatial facial features, which are not enough for the current information age, especially when social media and smart devices are widely used. In contrast, several video-based face recognition (VFR) methods [5], [6], [7], [8] are proposed by using spatiotemporal facial features. Unfortunately, all methods mentioned above require the authentic context information around certain facial key-points, which are often hard to obtain in the disguised scenarios. They would not work or perform poorly if some of the key contextual information are missing. For instance, to perform disguise face identification, [4] needs to first detect 14 facial key-points. However, the 10 key-points required around eyes can be easily covered by wearing a black sun-glass. Even though it is possible to apply additional compensation strategy to approximate the key-points, the recognition performance would be greatly devastated. Moreover, the challenging disguise variations are not considered by current video face datasets, such as CMU MoBo [9], Honda/UCSD [10] or YouTube Faces DB [11].

Instead of relying on the local facial context information, in this paper, a novel identification paradigm is proposed by combining the deep learning models with the global dynamic facial movements. Here, the global dynamic facial movements represent the facial muscle changes induced by talking or other activities, and the deep learning model could apply a spiking neural network (SNN) architecture or a conventional convolutional neural network (CNN) framework. It is known that one is hard to mimic other's subconscious actions such as facial muscle changes [12]. The proposed identification paradigm could still recognize the disguised individuals even some key facial context information are missing.

To balance the identification performance and the processing speed, instead of using the existing CNN models, we propose a novel deep SNN framework in this paper. It takes the dynamic facial movements as the sole input and uses the differential video frames [13], [14] to characterize them. However, such simple linear operation is not enough to substantially reduce the huge data redundancy of the input visual stimuli, which implies that there are large volumes of intra-class variations within the dynamic facial movements. Since ventral stream could extract complex abstractions from its lower layers and reduce the related intra-class variances [15], a deep feature

Daqi Liu is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. This work was done during the author was affiliated with University of Lincoln. e-mail: daqi.liu@surrey.ac.uk

Nicola Bellotto is with the School of Computer Science, University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, UK. e-mail: nbellotto@lincoln.ac.uk

Shigang Yue is with Machine Life and Intelligence Research Centre, and School of Mechanical and Electronic Engineering, University of Guangzhou, Guangzhou, China, and the School of Computer Science, University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, UK. e-mail: syue@lincoln.ac.uk (*Corresponding author: Shigang Yue)

extracting architecture simulating ventral stream is applied to obtain the desired feature vectors in this paper. Via the proposed spiking encoding scheme, the above feature vectors are further converted into spiking pattern sequences so that they can be trained by an event-driven continuous STDP learning rule. Besides, an adaptive thresholding strategy is also applied to achieve a stable learning procedure. The proposed video-based disguise face recognition (VDFR) method could still identify the individual even certain facial areas are disguised. Experimental results on the newly created video face database MakeFace DB demonstrate that it achieves very satisfying performances - from 95% to 100% correct classification rates under various realistic experimental scenarios.

The outline of this paper is summarized as follows: Section II presents related work and the framework of the proposed method is introduced in Section III. Section IV elaborates the neuron model and the learning method. The experimental results and discussions are illustrated in Section V, Section VI, respectively. Finally, Section VII concludes this paper.

II. Related Work

As far as we know, the current disguise face recognition methods [1], [2], [3] are designed solely for still images. Typically, given the related facial patches, intensity and texture encoder (ITE) or principal component analysis (PCA) are often applied to produce the associated feature vectors. Furthermore, local binary pattern (LBP) or support vector machine (SVM) are generally employed to classify different individuals.

Instead of only using spatial facial features like the above algorithms, the video-based face recognition methods [5], [6], [7], [8] incorporate associated temporal features. Unfortunately, the ideal face frames, in real-world scenarios, are often hard to obtain since one may wear heavy makeup or even use rubber/digital mask to disguise oneself. Such scenarios present huge challenges for the current VFR methods as it becomes extremely hard to capture the required feature vectors.

Deep learning models, especially CNN architectures, are widely applied to various applications in recent years. In [16], the authors introduce a novel CNN model to efficiently implement a multi-scale and sliding window approach. To obtain a deeper network with better generalization capability, [17] incorporates a micro neural network with more complex structures and [18] extends the knowledge distillation approach to allow the training of a student that is deeper and thinner than the teacher.

Unlike the conventional deep learning models, various SNN models [19], [20], [21], [22], [23], [24] are proposed during the last decade to investigate the spatiotemporal information. Within these SNN models, different learning methods are used to learn the synaptic weights, such as back propagation (BP) algorithm [24], Tempotron rule [23] or STDP learning method [20]. Among all current learning methods, STDP is widely believed that it underlies learning and information storage in the brain, as well as the development and refinement of neuronal circuits during brain development [25], [26]. It is shown to have several interesting computational capabilities, such as enabling the neurons to detect hidden causes of their inputs [27] or performing dimensionality reduction on their inputs [28]. The feasibility and efficiency of using STDP as learning methods in SNNs are demonstrated in our previous works [29], [30], [31], [32], in which [32] provides a prototype for the proposed VDFR method.

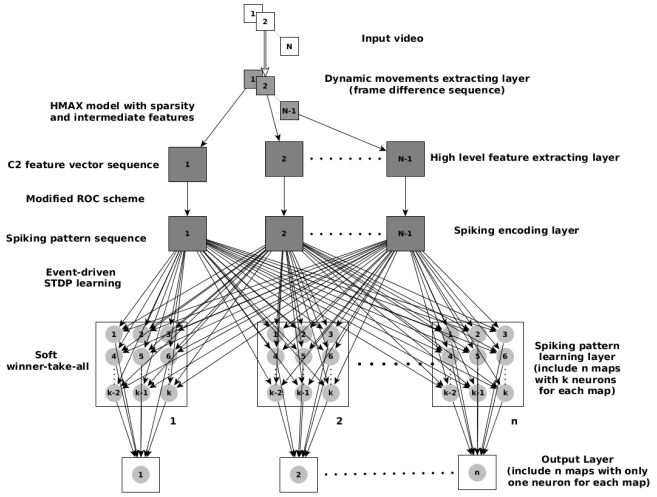


Fig. 1: Schematic diagram of the proposed feed-forward SNN. The input are the video clips from the MakeFace DB and the output are the spikes fired by the neurons within the output layer. The input video clip belongs to a certain class if its related map in the output layer fires a spike first. For better viewing, the lateral inhibition connections in the spiking pattern learning layer are not included.

Inspired by the biological retina, a novel frame-free event-driven dynamic vision sensor (DVS) is devised in recent years, which generates asynchronous address events (spikes) that signal scene reflectance changes at the times they occur. Various methods [33], [34], [23] are proposed to address different applications in real-time by using such asynchronous address events. For instance, [34] proposes an event-driven CNN model to quickly recognize rotating human silhouettes or high speed poker card symbols. [23] introduces an event-driven feed-forward categorization system based on tempotron classifier, which consumes much less simulation time while still maintaining comparable performance on several datasets.

III. Framework of the proposed VDFR method

Motivated by the motion sensitive neuron models [13], [14], the dynamic facial movements (the facial muscle changes induced by talking or other behaviors) become the sole input in this paper. They are generally characterized by the absolute differential video frames. However, the abundant intra-class variances within these differential video frames prevent themselves from accomplishing the VDFR applications. To reduce these intra-class variances, a deep SNN architecture simulating ventral stream is proposed in this paper. Ventral stream is capable of extracting different level of features [15], which are invariant for various transformations [35]. As demonstrated in Fig.1, the proposed hierarchical feed-forward SNN framework

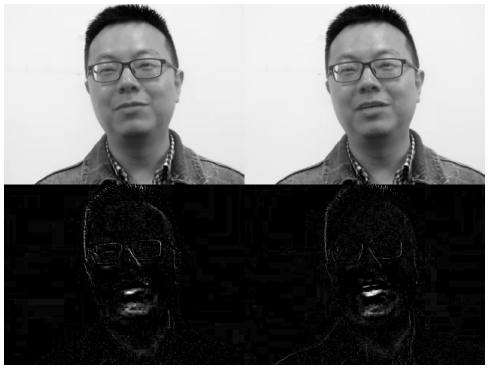


Fig. 2: Two adjacent video frames and their corresponding dynamic movements. For better viewing, the dynamic movements are processed by Min-Max normalization.

consists of five layers, in which each layer will be elaborated in the following subsections, respectively.

A. Dynamic Movements Extracting Layer

For the current VFR methods to work properly, the first and most important requirement is that they can extract different feature vectors from the input video frames. To fulfill this requirement, certain areas within the face need to be visible. However, within the VDFR applications, such areas are more likely to be covered by other objects, such as sunglasses or masks. It is known that one is hard to mimic other's subconscious actions such as facial muscle changes [12]. In this paper, the dynamic facial movements are used to differentiate different individuals.

To capture the dynamic facial movements, the subjects would be asked to speak during recording and the absolute differential video frames $\mathrm{d}I(x,y,t)^n/\mathrm{d}t^n$ are applied to characterize such dynamic facial movements:

$$\frac{\mathrm{d}I^n(x,y,t)}{\mathrm{d}t^n} = \left| \frac{\mathrm{d}I^{n-1}(x,y,t)}{\mathrm{d}t^{n-1}} - \frac{\mathrm{d}I^{n-1}(x,y,t-1)}{\mathrm{d}t^{n-1}}) \right| \quad (1)$$

where $I(x,y,t-1)$ and $I(x,y,t)$ depict two consecutive input video frames in time domain. For efficiency, in this paper, n is set to 1 and the corresponding absolute differential video frame $\mathrm{d}I(x,y,t)/\mathrm{d}t$ or $I^{'}(x,y,t)$ is computed as follows:

$$I^{'}(x,y,t) = \frac{\mathrm{d}I(x,y,t)}{\mathrm{d}t} = |(I(x,y,t) - I(x,y,t-1))| \quad (2)$$

Fig.2 demonstrates two adjacent video frames and their corresponding dynamic movements (applied Min-Max normalization for better viewing).

B. High Level Feature Extracting Layer

Due to their intrinsic intra-class variations, the dynamic facial movements cannot be directly used to classify the individuals. A high level feature extracting layer is required to obtain the balanced feature vectors in terms of invariance and distinguishability [36]. Various computational models [37], [38], [39], [40] are available to fulfil the above requirement, and most of them are inspired by visual cortex. For instance, [37] proposes a feed-forward HMAX model and obtains satisfying results on several datasets; Motivated by the simple/complex cells [41] in the primary visual cortex V1, [38], [39] build a feature extracting model by alternating between a template matching and a maximum pooling operation; [40] further incorporates sparsity constraints and local intermediate abstractions into the above model, and obtains quite satisfying performances accordingly.

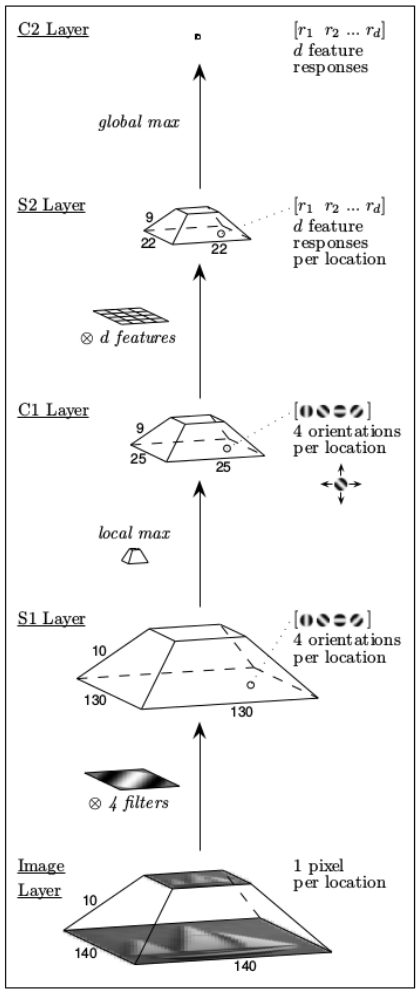


Fig. 3: The base computational model in [40] includes 5 layers with each constructs from the preceding layer via alternating between template matching and maximum pooling operations, except the first layer. Here, d is the number of elements in a $C2$ feature vector while $\otimes$ represents the template matching.

Inspired by the base computational model (as demonstrated in Fig.3) in [40], a GPU-based cortical network simulator (CNS) [42] is applied to construct the high level feature extracting layer used in this paper. It includes five feed-forward sub-layers, which are briefly introduced in the following subsections:

1) Input image layer: We first convert the input image into a grayscale one, and then resize it so that the shorter edge equals to 140 pixels, while retaining the same aspect ratio as the input image. The ultimate aim is to build an image pyramid $I^{'}(x,y,t)^{\sigma}$ containing 10 scales with each a factor of $2^{1/4}$ smaller than the last, in which σ is the scale.

2) Gabor filter (S1) layer: In this layer, Gabor response $F^{\sigma,\theta}_{(x,y)}$ is employed to mimic simple cell in primary visual cortex V1, where (x, y) represents the location, σ the scale and θ the orientation. Here, four orientations $(0^o, 45^o, 90^o, 135^o)$ are selected. For an input image $I^{'}(x, y, t)^{\sigma}$, the Gabor response is computed as follows:

$$F^{\sigma,\theta}_{(x,y)} = \exp\left(-\frac{\left(x_0^2+\gamma^2 y_0^2\right)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda}x_0\right), \quad \text{s.t.} \\ x_0 = xcos\theta + ysin\theta; \quad y_0 = -xsin\theta + ycos\theta \tag{3}$$

where γ is the aspect ratio, λ the wavelength, x_0 and y_0 the coordinates after rotating θ. We select the identical Gabor filter size (11×11) for all scales and normalize these Gabor responses to prevent the impacts of varying illumination.

3) Local invariance (C1) layer: Similar to the complex cells in V1, the units in C1 layer tend to retinotopically pool over previous S1 units with the same orientation/scale band. Given m afferents $\left(F^{\sigma,\theta}_{(x_1,y_1)}, \cdots F^{\sigma,\theta}_{(x_m,y_m)}\right)$, the maximum response $r^{\sigma,\theta}_{(x,y)}$ is used to characterize the related C1 unit:

$$r^{\sigma,\theta}_{(x,y)} = \underset{j=1\cdots m}{max}\, F^{\sigma,\theta}_{(x_j,y_j)} \tag{4}$$

The C1 unit, given the same position/scale, would only retain the dominant orientation via applying a lateral inhibition mechanism among S1/C1 units with different orientations.

4) Intermediate feature (S2) layer: Like [38], we randomly sample prototype patches (patch means a set of processing units) from the previous C1 layer as the templates, and then apply template matching between every possible patch of C1 units and each of the prototype patches. Essentially, a prototype patch is a patch of C1 units sampled from a random position/scale within the previous C1 layer. It has varying sizes and includes four orientations at each position.

Given a patch of C1 units X and a potential S2 prototype P (both have the same size of $n \times n \times 4$ with $n \in \{4, 8, 12, 16\}$), we use a Gaussian radial basis function to compute the corresponding response:

$$R(X, P) = exp\left(-\frac{\|X-P\|^2}{2\sigma^2\alpha}\right) \tag{5}$$

where σ represents the standard deviation and α is a normalizing factor for varying patch size. Inspired by the fact that the neurons tend to prefer certain potential inputs, we can further reduce the number of inputs for a S2 feature to one per C1 position by retaining the magnitude of the *dominant* orientation and its identity at every position in the patch. This would improves the generalization capability and ensure the S2 units more robust to local clutters.

5) Global invariance (C2) layer: Unlike classical HMAX model [38], we incorporate certain limits into the de facto global invariance mechanism so that the final C2 features are not globally invariant to all positions and scales. This is inspired by the fact that the neurons within visual area V4 and inferior temporal cortex (IT) are known to have limited receptive fields and do not exhibit global invariance.

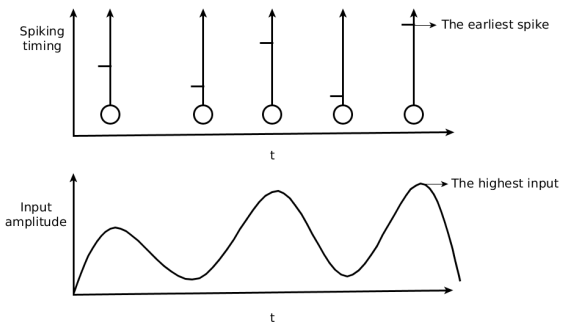


Fig. 4: A schematic diagram of ROC coding scheme. The short horizontal line within the spike part represents the latency of firing a spike. It can be seen that the spiking time has a inverse relationship with the intensity of the input visual stimulus.

C. Spiking Encoding Layer

In ventral stream, neurons would fire continuous spiking pattern sequences after being activated by the successive inputs. This motivates us to propose a spiking encoding layer to transform the above analog feature vectors into the continuous spiking pattern sequences. Essentially, spiking pattern sequences are series of spiking timings used to represent the spatiotemporal information of the input analog feature vectors. From an information-theoretic point of view, the spiking coding scheme achieves the information format transformation task.

Traditionally, both spiking rate and specific spiking timings are considered to convey information about the input stimuli. However, the former suffers two main drawbacks: 1) the spiking rate would only become meaningful when the processing time window is long enough, yet [43] shows the mammalian brain only use a quite short time window (millisecond scale) to process the input visual stimuli; 2) it is impossible to differentiate different input stimuli if incorporating background noise with the same spiking rate as the input stimulus into the rate-based SNN [44]. Thus, in this paper, we choose a simple yet efficient spiking time-based encoding scheme, rank order coding (ROC) [45], [46], [47], to produce the spiking pattern sequences. As demonstrated in Fig.4, within ROC, the spiking time has a inverse relationship with the intensity of the input visual stimulus. Even though ROC can only generate the first spike wave, it is enough for further processing in most cases.

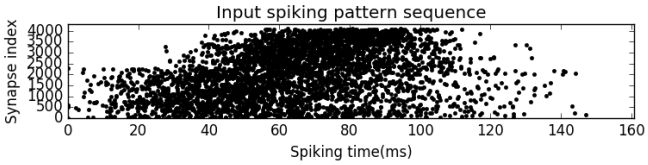
Unlike the classical ROC coding scheme, we use absolute spiking timings, instead of relative firing orders, to represent the input visual stimuli. This is because various spiking patterns may share a same relative firing order and we cannot distinguish these patterns if using classical ROC scheme. The proposed ROC coding scheme is demonstrated as follows:

$$t = p(max(r) - r) \tag{6}$$

where t (with the unit s) represents the corresponding spiking timing, r a possible C2 feature vector, $max(r)$ the maximum r value in the receptive field, $p \in [0, 1]$ a constant to specify the actual time span of a spiking pattern. Due to the normalization, for a given C2 feature vector, the maximum

(a) One input differential frame.

(b) Enhanced input differential frame.



(c) The corresponding spiking pattern.

Fig. 5: A simple example using the proposed ROC encoding scheme. Given the input differential frame (a), the corresponding spiking pattern (c) can be obtained through high level feature extracting layer and spiking encoding layer. For better viewing, (a) is enhanced to (b) using Min-Max normalization.

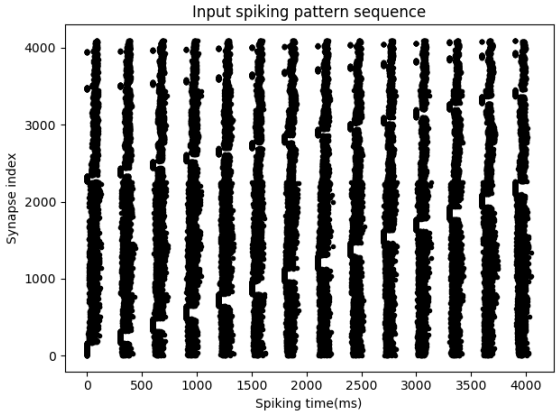


Fig. 6: One continuous input spiking pattern sequence. For better viewing, only 14 spiking patterns are included in this image. The actual continuous sequence contains much more spiking patterns.

t will be $1s$ given $p = 1$. Since a C2 feature vector has 4096 elements, we incorporate 4096 neurons in the spiking encoding layer so that each neuron can fire at most once given its corresponding element, as shown in Fig.5. Due to the limited space, the spiking pattern plot shown in Fig.5 may not be that noticeable. Moreover, [48] shows a versatile SNN requires continuous stimulus presentation. Therefore, an intuitive sequential continuous input mechanism is applied to construct the continuous spiking pattern sequence, as shown in Fig.6. Note, an interval with time span T_s is inserted into the two adjacent spiking patterns with identical time span $T_i = T_s$.

D. Spiking Pattern Learning Layer

Spiking pattern learning layer is vital since, after learning, the neurons within this layer would become selective to different input spiking patterns. Such selectivity is essentially obtained from the learned efficiency matrix. Specifically, it is initialized to random values within a predefined range and gradually converges over the learning procedure. The learned efficiency matrix (selectivity) is then applied to distinguish different individuals in the testing period.

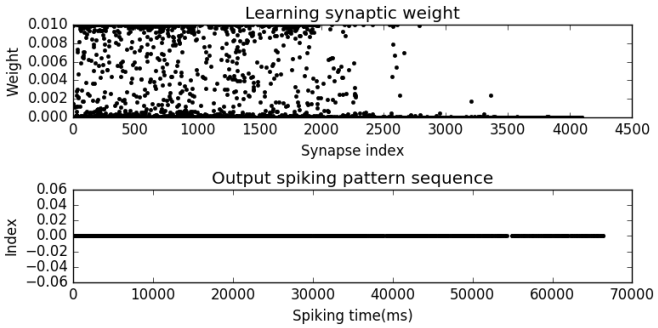


Fig. 7: The learning results of one continuous input spiking pattern sequence.

To obtain the above selectivity, a specific competitive learning method - soft winner-take-all (WTA) - is applied in this paper. It incorporates lateral inhibition connections into the spiking pattern learning layer so that the neuron fires the first spike would inhibit other neurons within the same layer. In the spiking pattern learning layer, each neuron map contains k neurons, which correspond to k possible sub-classes. Each neuron is connected to all neurons within the spiking encoding layer. The specific learning method used in this paper will be introduced in the following section. Fig.7 shows the learning results of one continuous input spiking pattern sequence.

E. Output Layer

Since the above spiking pattern learning layer is only capable of providing decisions for separate video frames, we still need an output layer to fuse those decisions so that the input video can be classified ultimately. From this perspective, the output layer can be considered as a decision fusion hub.

Specifically, for each possible class, a specific neuron map is added in the output layer. Unlike the above spiking pattern learning layer, only one neuron is incorporated into the related map and it only connects with the associated neurons within the previous layer. Besides, the membrane potential would be increased without any reducing when activated by the presynaptic spikes. The input video would be recognized as the class if the neuron within the related map fires a spike first. A simple solution could be counting the number of times a neuron has been activated by the input video frames. If the number reaches a predefined threshold, the input video would be classified as that specific class.

IV. Neuron Model and Learning Method

For SNN-based methods, there are three fundamental building blocks: architecture, neuron model and learning rule. The architecture of the proposed VDFR method is already elaborated in the above section. To build a complete SNN-based method, we will introduce the neuron model and the

learning method in the following subsections, respectively. To implement the proposed SNN learning method, Brian simulator [49] is applied in this paper.

A. Neuron Model

For SNNs, neuron model is essential since it specifies the behaviours of the neurons when interacting with other neurons. In this paper, we use conductance-based leaky integrate-and-fire (LIF) [50], [51] as the neuron model due to its efficiency and biological plausibility. Such type of neuron model can be considered as a coincidence detector since only the most related pre-synaptic spikes can activate the post-synaptic neuron to fire a spike.

Like [50], [51], a neuron uses the excitatory/inhibitory synaptic conductance g_{ex}/g_{in} (dimensionless since they are measured in units of the leakage conductance of the neuron) to incorporate the excitatory/inhibitory membrane potential E_{ex}/E_{in} so that its membrane potential V can be obtained as follows:

$$dV/dt = (g_{ex}(E_{ex} - V) + g_{in}(E_{in} - V) + V_r - V)/\tau_m \tag{7}$$

where τ_m represents the neuron membrane time constant, V_r the resting membrane potential. Meanwhile, g_{ex}/g_{in} are computed as follows:

$$\begin{aligned} dg_{ex}/dt &= -g_{ex}/\tau_{ex} \\ dg_{in}/dt &= -g_{in}/\tau_{in} \end{aligned} \tag{8}$$

where τ_{ex}/τ_{in} is the the excitatory/inhibitory synaptic conductance time constant. When V surpasses the predefined threshold V_t:

$$V \geq V_t \tag{9}$$

a spike would be fired and V would be reset to V_r after entering the absolute refractory period (with time window T_{rf}). The description and value of each parameter mentioned above can be found in Table I.

TABLE I: Parameter settings of the proposed SNN.

Parameter	Description	Value
d	the number of elements in a $C2$ vector [2]	4096
τ_m	membrane time constant [1]	10 ms
τ_v	potential threshold time constant [1]	20 ms
τ_{ex}	excitatory conductance time constant [1]	5 ms
τ_{in}	inhibitory conductance time constant [2]	10 ms
τ_+	pre-synaptic trace time constant [1]	20 ms
τ_-	post-synaptic trace time constant [1]	20 ms
p	a positive constant for ROC [2]	0.2
T_s	real processing time window [2]	150 ms
T_i	interval time window [2]	150 ms
T_{rf}	absolute refractory time [1]	1 ms
E_{ex}	excitatory membrane potential [1]	0 mV
E_{in}	inhibitory membrane potential [2]	-85 mV
V_r	resting membrane potential [1]	-74 mV
V_{thr}	membrane potential threshold [2]	-50 mV
V_i	increment for adaptive potential threshold [2]	5 mV
w^{min}	minimum synaptic weight [1]	0
w^{max}	maximum synaptic weight [2]	0.01
w^{in}	fixed inhibitory synaptic weight [2]	0.05
α_+	the pre-synaptic trace [1]	$0.01w^{max}$
α_-	the post-synaptic trace [1]	$-\alpha_+(\tau_+/\tau + -)1.05$

[1] Take the same value as [50].
[2] Optimized to achieve the best classification performance.

B. Learning Method

For SNNs, learning method is vital as it ensures the learned synaptic weights are selective to different visual stimuli. To achieve a biologically plausible, efficient and stable learning procedure, an event-driven STDP learning rule with adaptive thresholding strategy is applied in this paper. In the following subsections, we first introduce the event-driven STDP learning method and the adaptive thresholding rule, and then briefly elaborate the whole learning procedure.

1) Event-driven STDP Learning Method: To explain the learning mechanism of the neurons in brain, several postulates are proposed during the last several decades. Hebbian postulate [52] is perhaps the most well-known theory, which detects the causality between pre- and post-synaptic neurons [50] and can be summarized as "Cells that fire together, wire together" [52]. Unlike Hebbian postulate, spike-timing dependent plasticity (STDP) [53], [54], [55], [26], [56] is temporally asymmetric and it becomes the most popular learning method in recent decades. In STDP, according to the sign of the spiking time difference t between a pair of pre-/post-synaptic neurons, the associated synaptic efficiency $\mathrm{W}(\mathrm{t})$ would be potentiated/depressed through LTP (long-term potentiation)/LTD (long-term depression):

$$\begin{aligned} w(\mathrm{t}) &= A_+ \exp\left(-\frac{t}{\tau_+}\right) \quad for \;\; t > 0 \\ w(\mathrm{t}) &= -A_- \exp\left(\frac{t}{\tau_-}\right) \quad for \;\; t < 0 \end{aligned} \tag{10}$$

where the amplitude A_+ and the time constant τ_+ define the LTP learning window, while the amplitude A_- and the time constant τ_- specify the LTD learning window.

However, classical STDP method requires the neurons to sum over all pairs of spikes to update the associated synaptic weights, which is biologically implausible and not efficient. To this end, an event-driven STDP learning method is applied in this paper, which includes two on-line, local learning rules that can be immediately activated even when received/fired a single spike. It is achieved by incorporating the "traces" of pre- and post-synaptic activities α_+ and α_- (with time constants τ_+ and τ_-) into the learning procedure.

$$\begin{aligned} \tau_+ \frac{d}{dt}\alpha_+ &= -\alpha_+ \\ \tau_- \frac{d}{dt}\alpha_- &= -\alpha_- \end{aligned} \tag{11}$$

Due to biological reasons, in this paper, we define a $hardbound$ function to ensure the synaptic weight w remain-

ing between the minimum w^{min} and the maximum w^{max}:

$$hardbound(w, w^{min}, w^{max}) = \begin{cases} w^{max}, \ if \ w > w^{max} \\ w^{min}, \ if \ w < w^{min} \\ w, \ if \ w^{min} \leq w \leq w^{max} \end{cases} \tag{12}$$

Finally, the event-driven STDP learning method can be summarized as : when a neuron receives a pre-synaptic spike, the excitatory synaptic weight w^{ex} is updated as follows:

$$\begin{aligned} g_{ex} &= g_{ex} + w \\ A_+ &= A_+ + \alpha_+ \\ w^{ex} &= hardbound(w^{ex} + A_-, w^{min}, w^{max}) \end{aligned} \tag{13}$$

and when a neuron fires a post-synaptic spike, w^{ex} is modified as follows:

$$\begin{aligned} A_- &= A_- + \alpha_- \\ w^{ex} &= hardbound(w^{ex} + A_+, w^{min}, w^{max}) \end{aligned} \tag{14}$$

Besides, when received an inhibitory pre-synaptic spike, g_{in} is revised as follows:

$$g_{in} = g_{in} + w^{in} \tag{15}$$

where w^{in} is the fixed inhibitory synaptic weight.

2) Adaptive Thresholding Rule: As an asynchronous neural network, the proposed SNN framework would bias towards the neurons which fire the spikes first, especially when incorporated lateral inhibition mechanism. As a result, most neurons would remain inactivated throughout the learning procedure.

To resolve this issue, inspired by [57], an adaptive thresholding method is applied in the spiking pattern learning layer. Specifically, we use an adaptive membrane threshold V_t (with the time constant τ_v) to replace the predefined membrane threshold V_{thr}, which can be obtained as follows:

$$\tau_v \frac{d}{dt} V_t = V_{thr} - V_t \tag{16}$$

Since V_t increases by a fixed increment V_i every time a neuron fires a spike, it becomes much harder for the fired neuron to fire a spike again, which induces a stable learning procedure.

$$V_t = V_t + V_i \tag{17}$$

3) Learning Procedure: For better understanding of the proposed learning method, we summarize the whole learning procedure as Fig.8 and briefly explain it as follows:

1) Propagate a video with N consecutive frames $(I_1, I_2, \ldots, I_N)$ into the dynamic movements extracting layer. Given the input video, the dynamic facial movements are represented by a sequence of absolute differential video frames $(D_1, D_2, \ldots, D_{N-1})$, which can be obtained by computing the absolute difference of adjacent frames, as shown in Equation (2).
2) Propagate $(D_1, D_2, \ldots, D_{N-1})$ into the high level feature extracting layer and then apply the procedure in Fig.3 to generate a sequence of global invariant $C2$ feature vectors $(F_1, F_2, \ldots, F_{N-1})$, with each includes 4096 (d in Fig.3) units.

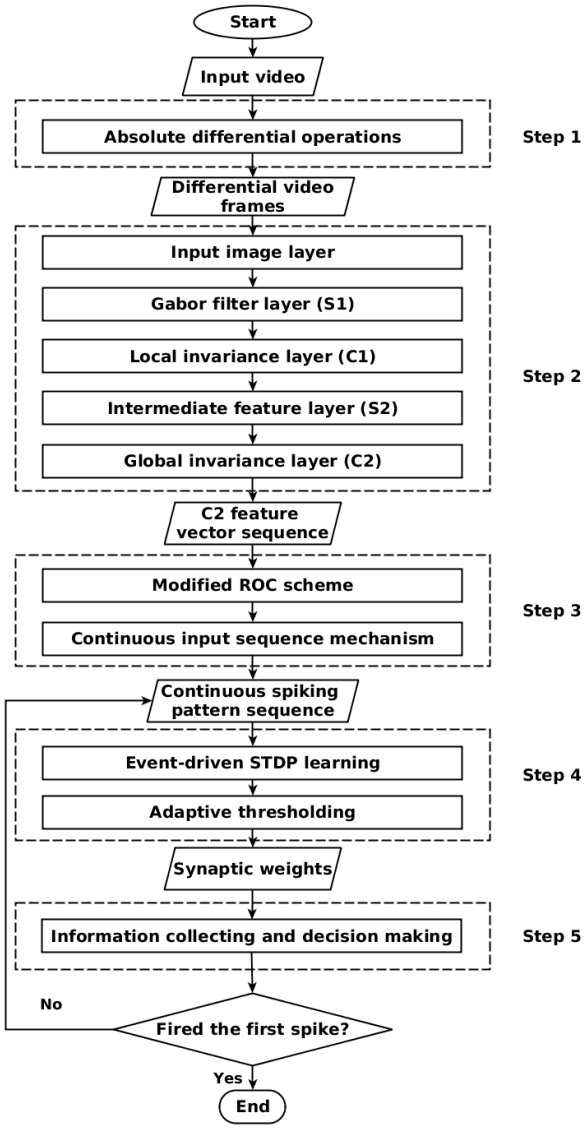


Fig. 8: The flowchart of the proposed VDFR learning method. Step 1-5 in the figure correspond to the five steps of the learning procedure in IV-B3.

3) Create a new map with 4096 neurons within the spiking encoding layer and obtain the continuous spiking pattern sequence $(S_1, S_2, \ldots, S_{N-1})$ using a simple sequential continuous input mechanism. To sequentially connect the spiking patterns, an interval is inserted into the adjacent spiking patterns, in which both have the same time span. For a global invariant $C2$ feature vector F_j, a modified ROC scheme described in Equation (6) is used to generate the corresponding spiking pattern S_j, where $j = 1, 2, \ldots, N-1$.
4) Create a new map with k neurons for each class in the spiking pattern learning layer. To achieve a competitive learning, each neuron would have lateral inhibition connections with other neurons within the same layer. Table I is used to initialize the proposed deep SNN architecture. Propagate the continuous spiking pattern

sequence $(S_1, S_2, \ldots, S_{N-1})$ into the spiking pattern learning layer and update the synaptic weights using Equations (13) and (14). Meanwhile, the membrane threshold V_t of a neuron would be adaptively modified using Equations (16) and (17).

5) Create a new map with one neuron for each class in the output layer. The neuron in each map only connects with related neurons in the previous layer. When a neuron receives an incoming spike, the post-synaptic potential would only increase by a fixed amount and remain at the same level within the whole testing period. The input video would be identified as the class when the related neuron in the output layer fires a spike firstly.

V. Experimental Results

In this section, we first introduce the newly created video disguise face database and the corresponding experimental parameter settings, and then elaborate the experimental results under different scenarios.

A. The Proposed Video Disguise Face Database

To verify the proposed VDFR method, the training/testing dataset should inherently incorporate disguise variations. However, none of the current video face databases meet the above requirement. To this end, a novel disguise video face database named MakeFace DB* is built in this paper, in which two scenarios are considered, namely, subject without any disguise and subject with disguise, as shown in Fig.9. To obtain the dynamic facial movements, each subject would be asked to speak during recording.

Specifically, MakeFace DB consists of 20 subjects in total and the disguise variations are achieved by wearing varying sunglasses, hats and fake beards. Note, normal glasses are not considered as disguise. For each subject, 5 video clips (each with frame rate of 29.97 frames/sec and a time span of 3 seconds) are recorded under varying experimental scenarios. The video frame size is set to 320×240. The comparisons with other video face databases are demonstrated in Table II. Even though it cannot compete with other complicated datasets like YouTube Faces DB or COX Face DB in terms of the number of experimental subjects/videos, MakeFace DB still stands out from other competitors since it incorporates more challenging disguise variations.

B. Parameter Settings

In this paper, the parameter settings listed in Table I are used to model the proposed SNN framework. For biological reasons, in a realistic neural simulation, the associated parameters are often located within a predefined range. Therefore, some parameters within Table I are set to the same values as [50], while others, like [51], are optimized globally. Moreover, all parameters except d are set to the same values as the base computational model [40] in section III-B. Finally, the information fusion mechanism in section III-E is selected as follows: an input video with m frames would be classified as a specific class if the related neuron in the output layer is activated by at least $0.5 \times m$ frames. In the following experiments, the time resolution is set to 0.1 ms.

*can be accessed at http://www.ciluk.org/makefaceDB

TABLE II: Comparison of different video face databases.

Database	Number of subjects/videos	Variations	Scenarios
CMU MoBo	25/150	w	V2V
First Honda/UCSD	20/75	p	V2V
Second Honda/UCSD	15/30	p	V2V
CMU FIA	214/214	p,l,e	V2V
CamFace	100/1400	p,l	V2V
Faces96	152/152	l,r	V2V
VidTIMIT	43/43	p,e	V2V
YouTube Celebrities	47/1910	p,l,e,r,b,w	V2V
MBGC	821/3764	p,l,e,r,b,w	V2S/V2V
ND-Flip-QO	90/14	l,e,r,b	V2V
YouTube Faces DB	1595/3425	p,l,e,r,b,w	V2V
Chokepoint	29/48	p,l,e,r,b	V2V
ScFace	130/910	p,l,r	V2S/V2V
UT Dallas	284/1016	p,l,e,r,b,w	V2S/V2V
UMD Comcast10	16/12	p,l,r,b,w	V2V
PaSC	265/2802	p,l,e,r,b,w	V2S/V2V
Celebrity-1000	1000/7021	p,l,e,r,b,w	V2V
COX Face DB	1000/3000	p,l,e,r,b,w	V2S/S2V/V2V
MakeFace DB	**20/100**	**l,e,d**	**V2V**

• Note: Variations include pose(p), illumination(l), expression(e), resolution(r), motion blur(b), walking(w) and disguise(d). According to the data types (V and S represent video and still image, respectively) applied in the training and testing periods, three scenarios are used in these datasets: video-to-video (V2V), video-to-still (V2S) and still-to-video (S2V).

Fig. 9: Comparison between training samples with and without disguises in the proposed MakeFace DB database. 9 subjects are shown in this image.

C. Experiments

To our knowledge, the proposed identification paradigm is the first one to combine the deep learning models with the dynamic facial movements to resolve the VDFR applications. To achieve a fair comparison, a simple and efficient CNN method [40] is used as the baseline in this section. It uses the differential video frames as the input and shares a similar architecture as the proposed VDFR method. For those two methods, the first two layers are almost identical. The difference is that the CNN method directly propagates the output feature vectors obtained from the first two layers to the final output layer, while the proposed VDFR method transforms these feature vectors into spiking patterns to be used by the last three layers in Figure.1. Specifically, to verify the universality of the proposed VDFR method, we test its

TABLE III: Correct recognition performance with various training/testing ratio (5 videos clips in total).

Number of training video clips	Number of testing video clips	Performance
1	4	92.5%
2	3	100%
3	2	100%

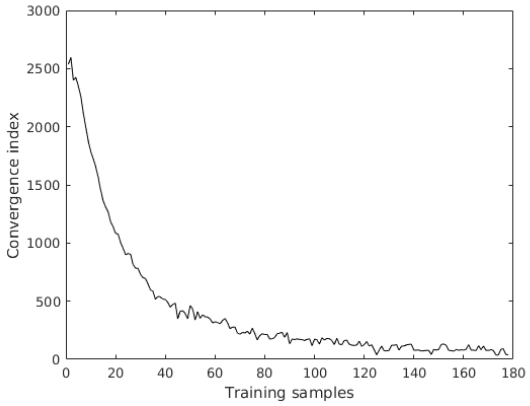


Fig. 10: Convergence status of the experiment without disguise samples. Here, 2 video clips with no disguise are applied within the experiment and training samples are differential video frames extracted from the video clips. Note, the proposed method converges after feeding around 100 differential video frames.

classification performances under the following four different experimental conditions:

1) Experiment without disguise samples: Before exploring the experiments with disguises, a basic requirement for the proposed method is to successfully distinguish different individuals when no disguise is applied. 10 subjects from MakeFace DB are selected to verify the above requirement. Specifically, different training/testing ratios are considered in this experiment and the subjects in the related video clips have no disguise. As shown in Table III, the proposed method can flawlessly resolve the above requirement with only 2 training video clips. Such convergence property would certainly be beneficial for the following experiments with disguises.

Moreover, for the proposed method, a convergence index f is applied to investigate whether it can converge to optimum over the learning procedure, as well as the corresponding convergence rate. It is known that the spiking timings of the activated neurons tend to remain stable (only with minor fluctuations) after convergence. Motivated by the above observation, within a spiking pattern period (consists of the spiking pattern and the adjacent interval), f can be computed by summing the spiking timings of all n fired neurons:

$$f = \sum_{i=1}^{n} (st_i - init) \tag{18}$$

where $init$ represents the starting time of the related spiking pattern, st_i the spiking timing of a possible fired neuron. Both use ms as the units. As demonstrated in Fig.10, the proposed method converges by only using around 100 differential video frames (or equivalently 2 video clips). This kind of efficiency is quite essential for VDFR applications.

TABLE IV: Classification performances of two different methods on testing video clips with disguise (%).

Method	Correct rate	Wrong rate	Unknown rate
CNN [40]	93.1 ± 1.35	6.9 ± 1.35	0
Proposed VDFR method	95.2 ± 2.65	4.8 ± 2.65	0

• Note: The classification rates are obtained by averaging 10 random tests. For the above two methods, a Wilcoxon signed-rank test is conducted on the correct classification performances and a resulting significance level $p \approx 0.03429$ is computed. Since $p < 0.05$, it means these classification performances are statistically independent.

2) Experiment with disguise samples: As demonstrated above, when no disguise is applied, the proposed method could flawlessly detect the individuals within the testing video clips by only using 2 training video clips. However, the above result is far from accomplishing the VDFR application. In this subsection, we would investigate whether the above selectivity could still detect the disguised individuals within the testing period. All 20 subjects from MakeFace DB are applied in this experiment.

In Table IV, two comparison methods - CNN [40] and the proposed VDFR method - are applied to test the classification performance on disguise video clips. 10 random tests are conducted in this experiment. It can be seen that the proposed VDFR method outperforms the CNN method [40] by about 2.1%. Furthermore, to prove the correct classification performances obtained from the above methods are statistically independent, we conduct a Wilcoxon signed-rank test. The resulting significance level $p \approx 0.03429$. Since $p < 0.05$, the generated correct classification performances are statistically independent.

3) Experiment with mixed samples: In the above subsections, only video clips without any disguise are used to learn the selectivity. However, such ideal training samples may be hard to obtain in some real-world scenarios. For instance, females frequently wear light/heavy makeup or criminals often tend to wear masks. In this subsection, to verify the classification performance of the proposed method under the above scenarios, mixed samples (some with disguise while others have no disguise) are incorporated within both training and testing periods.

Furthermore, the above convergence index f is applied to investigate the convergence status of the proposed method in this specific experiment. As shown in Fig.11, the proposed method converges after feeding around 120 differential video frames. This means 2 mixed video clips are sufficient for learning the selectivity. Therefore, in this experiment, we use 2 mixed video clips for training and apply the remaining 3 mixed video clips for testing. All 20 subjects from MakeFace DB are selected for this experiment.

As demonstrated in Table V, the proposed VDFR method reaches 100% correct classification rate and outperforms the CNN method [40] by about 3.3%. This experiment further indicates the superiority of the proposed VDFR method

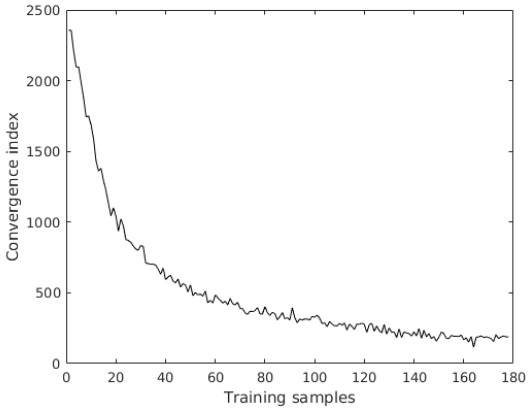


Fig. 11: Convergence status of the experiment with mixed samples. Here, 2 mixed video clips are used within the experiment and training samples are differential video frame extracted from the video clips. Note, the proposed method converges after feeding around 120 differential video frames.

TABLE V: Classification performances of two different methods on testing mixed video clips (%).

Method	Correct rate	Wrong rate	Unknown rate
CNN [40]	96.7 ± 0	3.3 ± 0	0
Proposed VDFR method	100 ± 0	0 ± 0	0

• Note: The classification rate has been computed by averaging 10 random tests.

TABLE VI: Classification performances of two different methods on testing unknown video clips (%).

Method	Correct rate	Wrong rate	Unknown rate
CNN [40]	0	0	100 ± 0
Proposed VDFR method	0	0	100 ± 0

• Note: The classification rate has been computed by averaging 10 random tests.

when dealing with relatively complicated scenario like training/testing the mixed samples.

4) Experiment with unknown samples: In some real-world scenarios, the testing samples can only be gradually obtained over the testing period. For those newly created unknown testing samples, the proposed method should be robust enough to recognize them as unknown rather than some known classes. To this end, we select 10 subjects from MakeFace DB as known training samples and use the remaining 10 subjects as unknown testing samples in this subsection. Specifically, we choose 2 video clips with no disguise from the known training samples to learn the synaptic weights, and select 3 video clips with disguises from the unknown testing samples for evaluation. As demonstrated in Table VI, both methods are capable of flawlessly (with unknown rates of 100%) classifying the testing samples as unknown.

VI. Discussions

In this section, the main contributions of the paper are enumerated in the first subsection, followed by the robustness analysis and the ablation analysis. The comparisons with other state-of-the-art methods are introduced in the last subsection.

A. Contributions

To demonstrate the novelty of our research, we have enumerated its main contributions in this section:

1) A novel identification paradigm is proposed to recognize the disguised individuals, which combines the deep learning techniques with the dynamic facial movements. To our knowledge, neither the conventional CNN models nor the deep spiking learning architectures have been used to accomplish the above paradigm. The proposed VDFR method and the CNN method used in comparison can all be considered as the first ones to implement the above paradigm. Moreover, the proposed paradigm relies on the global dynamic facial movements instead of the specific facial context information around certain key-points, which can still identify the disguised individuals when some facial key-points are occluded.
2) An event-driven continuous STDP learning method with adaptive thresholding is applied to learn the associated spiking patterns generated from the input differential video frames. To achieve a balanced competitive learning, a soft winner-take-all strategy is applied within the learning procedure. Such learning methodology is efficient and physiologically realistic as it can update the associated synaptic weight when received a pre-synaptic spike or fired a post-synaptic spike.
3) A new video-based disguise face recognition dataset - MakeFace DB - has been built in this paper, which provides both visual and audio information for the researchers. Such multi-modal dataset can be further explored to induce more advanced identification methods.

B. Robustness Analysis

In the visual cortex, interferences such as the time jitter or the background neural noise are ubiquitous. Such interferences would affect the visual pathway and subsequently deteriorate the identification performance. To avoid the impact of the interferences, the proposed method should be robust enough to ignore the interferences and only concentrate on learning the input visual stimuli.

To achieve the above robust learning procedure, two important facts are incorporated in this paper: 1) Interferences tend to be independent between each others; 2) The event-driven continuous STDP learning method is essentially an inhibitive model since the integral area of LTD is larger than the integral area of LTP. Those two facts would guarantee the proposed method to ignore the interferences and only concentrate on learning the input visual stimuli. As an inhibitive model, the proposed learning method would only activate the neurons which constantly receive similar spiking patterns. It would inhibit the neurons to fire spikes if independent spiking patterns are received. In other words, the neurons would gradually

TABLE VII: Classification performances of four different configurations on testing video clips with disguise (%).

Configuration	Correct rate	Wrong rate	Unknown rate
Differential frames + SNN (**VDFR**)	95.2 ± 2.65	4.8 ± 2.65	0
Original frames + SNN	71.9 ± 2.48	28.1 ± 2.48	0
Differential frames + CNN [40]	93.1 ± 1.35	6.9 ± 1.35	0
Original frames + CNN [40]	70.2 ± 1.27	29.8 ± 1.27	0

• Note: The classification rate has been computed by averaging 10 random tests. Differential frames are used to model the dynamic facial movements.

become inactive when they repeatedly receive independent spiking patterns generated by the input interferences.

C. Ablation Analysis

As a multi-stage algorithm, the identification performance of the proposed VDFR method relies on various modules. Ablation analysis is essential to differentiate the importance of different modules. Moreover, it could help us finding the optimal modules. To verify whether the current modules are the optimal ones for the proposed method and differentiate the importance of different modules, two ablation studies are investigated in this section.

Firstly, for the proposed method, the feature extracting layer is an important module since it can greatly reduce the data dimensionality of the input stimuli. To pursue an optimal preprocessing strategy, we need to consider the following two factors: performance and efficiency. The ideal preprocessing strategy should achieve a considerable performance within a limited training time period. Even though the complicated CNN models such as VGG or ResNet can achieve superior performances, the time cost of these models is unbearable for our methodology. For the disguise face recognition application, it is crucial to process the input data in a timely manner. The feature extracting methodology used in this paper can reach a considerable performance using much less training time than the above models. Moreover, it is more biologically plausible since it is essentially an unsupervised learning method.

Secondly, the dynamic facial movements and the SNN learning model are two essential modules of the proposed VDFR method. The first three configurations in Table VII are devised to differentiate the importance of those two modules. Specifically, the first configuration (the proposed VDFR method) is used as the baseline since it achieves the best classification performance. The following two configurations are devised to evaluate the importances of the dynamic facial movement module and the SNN learning module, respectively. Given the baseline, we can calculate the performance gaps for the following two configurations. The bigger the performance gap, the more important the corresponding module. It can be seen that the performance gap of the second configuration is much larger than that of the third configuration, which implies the dynamic facial movement module is more important than the SNN learning module. Moreover, it can be seen from Table VII that the differential frames, compared with original frames, can help both SNN and CNN learning modules to substantially boost the recognition performances.

D. Comparisons with State-of-the-art Methods

With the increasing performances, the deep learning models are becoming ubiquitous. For instance, [16] introduces a novel CNN model to efficiently implement a multi-scale and sliding window approach. To improve the generalization capability with a much deeper network, [17] incorporates a micro neural network with more complex structures and [18] extends the knowledge distillation approach to allow the training of a student that is deeper and thinner than the teacher. However, none of the previous deep learning models are designed for the video-based disguise face recognition application.

To our knowledge, the current disguise face recognition methods [1], [2], [3], [4] are only applied for the still images. Most of them use the context information around certain facial key-points to identify the disguised faces. However, such methods would not work or perform poorly if some of the key contextual information are missing. For instance, to perform disguise face identification, [4] needs to first detect 14 facial key-points. However, the 10 key-points required around eyes can be easily covered by wearing a black sun-glass. Even though it is possible to apply additional compensation strategy to approximate the key-points, the recognition performance would be greatly devastated.

Unlike the above algorithms, the current VFR methods [5], [6], [7], [8] use both temporal and spatial features in the recognition procedure. However, such methods are not suitable for disguised face recognition tasks. This is because the required context information will be difficult to capture with the key areas being occluded. For instance, in [7], within the proposed individual expression recognition (IER) block, to obtain the behavioral map (BM) containing facial evolutions of microexpressions in each frame, at least an eye, a brow or a cheek need to be detected within the video frames. Unfortunately, within VDFR applications, such critical requirement cannot be satisfied.

Instead of using local context information, the proposed identification paradigm relies on the global dynamic facial movements, which could still identify the disguised individuals even some of the key context information are not available. It is the first one to identify disguised faces by combining the deep learning models (conventional CNN or deep spiking networks) with the dynamic facial movements. Specifically, two deep learning models - the CNN model [40] and the proposed VDFR architecture - are applied to learn the input dynamic facial movements in this paper. Unlike the CNN model [40], the proposed VDFR method adds an additional competitive learning layer. It combines an event-driven continuous STDP learning algorithm with a soft winner-take-all strategy, which can be considered as an expectation-maximization (EM) algorithm [27]. Through such competitive learning, the proposed VDFR method improves its generalization capability and thus achieves a better identification performance.

VII. Conclusion

To resolve VDFR task, we propose a new learning paradigm in this paper, which combines deep learning architectures with dynamic facial movements. To balance the identification

performance and the computational complexity, a novel deep SNN framework is further proposed, which takes dynamic facial movements as the sole input and apply an event-driven continuous STDP learning rule to train the associated synaptic weights. It can still identify the individuals even certain key facial areas are disguised. Systematic experiments on the newly created MakeFace DB dataset show the proposed method can achieve very satisfying classification performances - from 95% to 100% correct classification rates under various experimental scenarios. Future works include extending the current learning scheme to resolve the malicious spoofing attacks, as well as pursuing more advanced VDFR methods for video samples with complicated moving background.

ACKNOWLEDGMENT

The authors have been supported by EU Horizon 2020 projects STEP2DYNA (691154), ENRICHME (643691) and ULTRACEPT (778062).

REFERENCES

[1] R. Singh, M. Vatsa, and A. Noore, "Face recognition with disguise and single gallery images," *Image and Vision Computing*, vol. 27, no. 3, pp. 245–257, Feb. 2009.

[2] T. I. Dhamecha, A. Nigam, R. Singh, and M. Vatsa, "Disguise detection and face recognition in visible and thermal spectrums," in *2013 International Conference on Biometrics (ICB)*, Jun. 2013, pp. 1–8.

[3] T. I. Dhamecha, R. Singh, M. Vatsa, and A. Kumar, "Recognizing Disguised Faces: Human and Machine Evaluation," *PLOS ONE*, vol. 9, no. 7, p. e99212, Jul. 2014.

[4] A. Singh, D. Patil, G. M. Reddy, and S. Omkar, "Disguised face identification (dfi) with facial keypoints using spatial fusion convolutional network," in *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1648–1655.

[5] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "A Benchmark and Comparative Study of Video-Based Face Recognition on COX Face Database," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5967–5981, Dec. 2015.

[6] J. Yu, "Super-resolution and Facial Expression for Face Recognition in Video," Ph.D. dissertation, University of California, Riverside, USA, 2007, aAI3298271.

[7] M. Gavrilescu, "Study on using individual differences in facial expressions for a face recognition system immune to spoofing attacks," *IET Biometrics*, vol. 5, no. 3, pp. 236–242, 2016.

[8] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning Euclidean-to-Riemannian Metric for Point-to-Set Classification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1677–1684.

[9] R. Gross and J. Shi, "The cmu motion of body (mobo) database," Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-01-18, June 2001.

[10] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," *IEEE Conf. On Computer Vision and Pattern Recognition*, vol. 1, pp. 313–320, 2003.

[11] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 529–534.

[12] J. F. Cohn, K. Schmidt, R. Gross, and P. Ekman, "Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification," in *Fourth IEEE International Conference on Multimodal Interfaces, 2002. Proceedings*, 2002, pp. 491–496.

[13] S. Yue and F. Rind, "Collision detection in complex dynamic scenes using an LGMD-based visual neural network with feature enhancement," *IEEE Transactions on Neural Networks*, vol. 17, no. 3, pp. 705–716, 2006.

[14] S. Yue and F. Rind, "Redundant Neural Vision Systems -Competing for Collision Recognition Roles," *IEEE Transactions on Autonomous Mental Development*, vol. 5, no. 2, pp. 173–186, 2013.

[15] T. Poggio and E. Bizzi, "Generalization in vision and motor control," *Nature*, vol. 431, no. 7010, pp. 768–774, Oct. 2004.

[16] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[17] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[18] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[19] T. Masquelier and S. J. Thorpe, "Unsupervised learning of visual features through spike timing dependent plasticity," *PLoS Computational Biology*, vol. 3, no. 2, p. e31, 2007.

[20] M. Beyeler, N. D. Dutt, and J. L. Krichmar, "Categorization and decision-making in a neurobiologically plausible spiking network using a STDP-like learning rule," *Neural Networks*, vol. 48, pp. 109–124, Dec. 2013.

[21] E. Neftci, S. Das, B. Pedroni, K. Kreutz-Delgado, and G. Cauwenberghs, "Event-driven contrastive divergence for spiking neuromorphic systems," *Neuromorphic Engineering*, vol. 7, p. 272, 2014.

[22] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in Computational Neuroscience*, p. 99, 2015.

[23] B. Zhao, R. Ding, S. Chen, B. Linares-Barranco, and H. Tang, "Feedforward Categorization on AER Motion Events Using Cortex-Like Features in a Spiking Neural Network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 9, pp. 1963–1978, Sep. 2015.

[24] P. U. Diehl, D. Neil, J. Binas, M. Cook, S. C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *2015 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2015, pp. 1–8.

[25] P. J. Sjstrm, E. A. Rancz, A. Roth, and M. Husser, "Dendritic excitability and synaptic plasticity," *Physiological Reviews*, vol. 88, no. 2, pp. 769–840, Apr. 2008.

[26] G. Q. Bi and M. M. Poo, "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type," *J Neurosci*, vol. 18, pp. 10 464–10 472, 1998.

[27] B. Nessler, M. Pfeiffer, and W. Maass, "Stdp enables spiking neurons to detect hidden causes of their inputs," in *Advances in neural information processing systems*, 2009, pp. 1357–1365.

[28] M. Beyeler, E. Rounds, K. Carlson, N. Dutt, and J. L. Krichmar, "Sparse coding and dimensionality reduction in cortex," *bioRxiv*, p. 149880, 2017.

[29] D. Liu and S. Yue, "Visual pattern recognition using unsupervised spike timing dependent plasticity learning," in *Neural Networks (IJCNN), 2016 International Joint Conference on*, 2016, pp. 285–292.

[30] D. Liu and S. Yue, "Fast unsupervised learning for visual pattern recognition using spike timing dependent plasticity," *Neurocomputing*, vol. 249, pp. 212–224, 2017.

[31] D. Liu and S. Yue, "Event-driven continuous stdp learning with deep structure for visual pattern recognition," *IEEE Transactions on Cybernetics*, vol. 49, no. 4, pp. 1377–1390, 2019.

[32] D. Liu and S. Yue, "Video-based disguise face recognition based on deep spiking neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 01–08.

[33] P. Rogister, R. Benosman, S.-H. Ieng, P. Lichtsteiner, and T. Delbruck, "Asynchronous event-based binocular stereo matching," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 2, pp. 347–353, 2012.

[34] J. A. Pérez-Carrasco, B. Zhao, C. Serrano, B. Acha, T. Serrano-Gotarredona, S. Chen, and B. Linares-Barranco, "Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing–application to feedforward convnets," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2706–2719, 2013.

[35] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, "A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex," 2005.

[36] Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[37] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, Nov. 1999.

[38] S. Thomas, W. Lior, B. Stanley, R. Maximilian, and P. Tomaso, "Robust object recognition with cortex-like mechanisms," *IEEE Trans. Pattern Anal. Mach. Intelli.*, vol. 29, no. 3, pp. 411–426, 2007.

[39] T. Serre, G. Kreiman, M. Kouh, C. Cadieu, U. Knoblich, and T. Poggio, "A quantitative theory of immediate visual recognition," *Progress in Brain Research*, vol. 165, pp. 33–56, 2007.

[40] J. Mutch and D. G. Lowe, "Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields," *International Journal of Computer Vision*, vol. 80, no. 1, pp. 45–57, Jan. 2008.

[41] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, Oct. 1959.

[42] J. Mutch, U. Knoblich, and T. Poggio, "CNS: a GPU-based framework for simulating cortically-organized networks," Massachusetts Institute of Technology, Cambridge, MA, Tech. Rep. MIT-CSAIL-TR-2010-013 / CBCL-286, February 2010.

[43] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, no. 6582, pp. 520–522, 1996.

[44] T. Masquelier, R. Guyonneau, and S. J. Thorpe, "Spike timing dependent plasticity finds the start of repeating patterns in continuous spike trains," *PLoS ONE*, vol. 3, no. 1, p. e1377, 2008.

[45] A. Delorme and S. Thorpe, "Face identification using one spike per neuron: resistance to image degradation," *Neural Networks*, vol. 14, no. 6-7, pp. 795–803, 2001.

[46] A. Delorme, J. Gautrais, R. van Rullen, and S. Thorpe, "Spikenet: a simulator for modeling large large networks of integrate and fire neurons," *Neurocomputing*, vol. 26, pp. 989–996, 1999.

[47] A. Delorme, L. Perrinet, and S. Thorpe, "Networks of integrate-and-fire neurons using rank order coding," *Neurocomputing*, vol. 38-40, pp. 539–545, 2001.

[48] T. Rumbell, S. Denham, and T. Wennekers, "A Spiking Self-Organizing Map Combining STDP, Oscillations, and Continuous Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 894–907, 2014.

[49] D. F. Goodman and R. Brette, "The brian simulator," *Frontiers in neuroscience*, vol. 3, no. 2, p. 192, 2009.

[50] S. Song, K. D. Miller, and L. F. Abbott, "Competitive Hebbian learning through spike-timing-dependent synaptic plasticity," *Nature Neuroscience*, vol. 3, no. 9, pp. 919–926, Sep. 2000.

[51] S. Song and L. F. Abbott, "Cortical development and remapping through spike timing-dependent plasticity," *Neuron*, vol. 32, no. 2, pp. 339–350, Oct. 2001.

[52] D. O. Hebb, *The Organization of Behavior: a neuropsychological theory*. New York: Wiley, 1949.

[53] W. Gerstner, R. Kempter, J. L. van Hemmen, and H. Wagner, "A neuronal learning rule for sub-millisecond temporal coding," *Nature*, vol. 386, pp. 76–78, 1996.

[54] R. Kempter, W. Gerstner, and J. L. van Hemmen, "Hebbian learning and spiking neurons," *Phys. Rev. E*, vol. 59, pp. 4498–4514, 1999.

[55] H. Markram, J. Lubke, M. Frotscher, and B. Sakmann, "Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps," *Science*, vol. 275, pp. 213–5, 1997.

[56] P. J. Sjöström, G. G. Turrigiano, and S. B. Nelson, "Rate, timing, and cooperativity jointly determine cortical synaptic plasticity," *Neuron*, vol. 32, pp. 1149–1164, 2001.

[57] D. Querlioz, O. Bichler, P. Dollfus, and C. Gamrat, "Immunity to Device Variations in a Spiking Neural Network With Memristive Nanodevices," *IEEE Transactions on Nanotechnology*, vol. 12, no. 3, pp. 288–295, May 2013.

Daqi Liu is a Research Fellow at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. This work was done during the author was affiliated with University of Lincoln. His research interests include bio-inspired computational models, bayesian machine learning, computer vision and pattern recognition. He has published several scientific papers in top-ranked journals, including IEEE transactions on Cybernetics, Neurocomputing etc.

Nicola Bellotto is a Reader (Associate Professor) in the School of Computer Science, University of Lincoln, UK, and a member of the Lincoln Centre for Autonomous Systems. His main research interests are in machine perception, especially for human detection, tracking, identification and activity recognition with autonomous mobile robots. He has a Master in Electronic Engineering from the University of Padova, Italy, and a PhD in Computer Science from the University of Essex, UK. Before joining the University of Lincoln, he was a researcher in the Active Vision Lab at the University of Oxford. Dr Bellotto is the recipient of a Google Faculty Research Award and a PI/Co-I in several national and international projects on autonomous mobile robots.

Shigang Yue (M05)(SM'17) is a Professor in the School of Computer Science, University of Lincoln, United Kingdom. He also holds a Professorship in the Machine Life and Intelligence Research Centre in Guangzhou University, in collaborating with Prof. Jigen Peng. He received his PhD and MSc degrees from Beijing University of Technology (BJUT) in 1996 and 1993. He worked in BJUT as a Lecturer (1996-1998) and an Associate Professor(1998-1999). He was an Alexander von Humboldt Research Fellow (2000,2001) at University of Kaiserslautern, Germany. Before joining the University of Lincoln as a Senior Lecturer (2007) and promoted to Professor (2012), he held research positions in the University of Cambridge, Newcastle University and the University College London(UCL) respectively. His research interests are mainly within the field of artificial intelligence, computer vision, robotics, brains and neuroscience. He is a senior member of IEEE, and member of INNS and ISBE.