Computational Model Based on Neural Network of Visual Cortex for Human Action Recognition

Haihua Liu, Na Shu, Qiling Tang, and Wensheng Zhang

***Abstract*— In this paper, we propose a bioinspired model for human action recognition through modeling neural mechanisms of information processing in two visual cortical areas: the primary visual cortex (V1) and the middle temporal cortex (MT) dedicated to motion. This model, named V1-MT, is composed of V1 and MT models (layers) corresponding to their cortical areas, which are built with layered spiking neural networks (SNNs). Some neuron properties in V1 and MT, such as direction and speed selectivity, spatiotemporal inseparability, and center surround suppression, are integrated into SNNs. Based on speed and direction selectivity, V1 and MT models contain multiple SNN channels, each of which processes motion information in sequences with spatiotemporal tunings of neurons at a certain speed and different directions. Therefore, we propose two operations, input signal perceiving with 3-D Gabor filters and surround inhibition processing with 3-D differences of Gaussian functions, to perform this task according to the spatiotemporal inseparability and center surround suppression of neurons. Then, neurons are modeled with our simplified integrate-and-fire model and motion information is transformed into spike trains. Afterward, we define a new feature vector: a mean motion map computed from spike trains in all channels to represent human actions. Finally, a support vector machine is trained to classify actions represented by the feature vectors. We conducted extensive experiments on public action databases, and the results show that our model outperforms other bioinspired models and rivals the state-of-the-art approaches.**

***Index Terms*— Action recognition, classical receptive field (RF), spiking neural networks (SNNs), surround suppression, visual cortex.**

I. Introduction

HUMAN action recognition is a challenging task in computer vision, and has numerous applications, including automated surveillance, elderly behavior monitoring, and human–computer interaction. Over the past few decades, extensive amounts of research on human action recognition have been conducted with the goal to create a robust system [1]–[7]. However, human action recognition remains a challenging task due to the large variations in human appearance and movement speed. It also suffers from various uncertain challenges, such as cluttered backgrounds, occlusion, and illumination changes.

Recently, with enhanced understanding of the brain mechanisms responsible for action recognition [8], a large number of action recognition approaches based on vision have been proposed. For example, Zhang and Tao [3] introduced a novel method of action recognition based on slow feature analysis (SFA), which extracted slowly varying features from a quickly varying signal. Afterward, Sun *et al.* [9] proposed a two-layered SFA learning method to capture abstract and structural features from the video for human action recognition by combining SFA with deep learning techniques. Additionally, artificial neural networks (ANNs) [10], [11], as a family of statistical learning algorithms inspired by biological neural networks, were proposed to recognize human action from video sequences. The method in [10] highlighted the strength of ANNs in representing and classifying visual information based on learning spatially related body posture prototypes with self-organizing maps. Furthermore, Ji *et al.* [11] proposed a novel 3-D convolutional neural network model for action recognition that extracted features from both spatial and temporal dimensions by performing 3-D convolutions.

The human visual system is effective for action recognition in terms of its complete constructions. In visual cortex, billions of neurons (the nerve cells) form a huge network by connecting to each other, in which information processing (through the ventral and dorsal streams [12]) is regular, parallel, and hierarchical. Wherein, the ventral stream mainly deals with shape information and involves V1, V2, and V4 visual cortical areas. On the other hand, the dorsal stream mainly analyzes motion information and involves V1, the middle temporal (MT), and the posterior parietal cortex. Based on this dual-channel theory of vision [12], researchers proposed different bioinspired models for biological motion recognition. These ventral and dorsal streams are separately evaluated in biological motion recognition [13]. By using only dorsal stream information, Sigala *et al.* [14] proposed a biological motion recognition system with a neurally plausible memory trace learning rule. However, it has never seen practical applications in action recognition.

Later, Jhuang *et al.* [15] proposed a feedforward hierarchical model (HMAX model) in which spatiotemporal feature detectors of increasing complexity extract features of human action. Unfortunately, this model not only required heavy computation but also lacked biological plausibility [16].

Manuscript received June 20, 2016; revised October 15, 2016 and January 26, 2017; accepted February 7, 2017. This work was supported by the National Natural Science Foundation of China under Grant 91320102, Grant 60972158, Grant 61432008, and Grant 61532006. *(Corresponding authors: Haihua Liu; Wensheng Zhang.)*

H. Liu is with the School of Biomedical Engineering, South Central University for Nationalities, Wuhan 430074, China, and the Key Laboratory of Cognitive Science of State Ethnic Affairs Commission and Hubei Key Laboratory of Medical Information Analysis and Tumor Diagnosis & Treatment, Wuhan 430074, China (e-mail: lhh@mail.scuec.edu.cn).

N. Shu and Q. Tang are with the School of Biomedical Engineering, South-central University for Nationalities, Wuhan 430074, China.

W. Zhang is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wensheng.zhang@ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2669522

Schindler and van Gool [17] extended Jhuang's approach by combining both shape and motion responses. Due to a collection of shape and motion features independently obtained in the matching stage, this approach also required complex computation even though fewer frames were used for feature extraction. Similar work was also implemented with the hierarchical space-time model [18].

Escobar *et al.* [19] created a hierarchical computational model (V1/MT model) for human action recognition by simulating two brain areas (V1 and MT) in the dorsal pathway. This model is built with a feedforward spiking neural network (SNN) by mimicking behaviors of spiking neurons in the human visual system, such as direction-selectivity and center-surround interactions of MT neurons, to extract key information from video sequences. Thereafter, Escobar and Kornprobst [20] also proposed a model which went beyond the classical models and further introduced different surround geometries of MT cells receptive fields (RFs) into the SNN following the biological observations. Although these works are biologically plausible and applicable to action recognition, ignoring some of the neuron properties, such as speed selectivity for biological pattern motion and center-surround interactions in both V1 and MT, causes poor performance.

The bioinspired SNN is capable of processing complex information due to its ability to represent and integrate different information dimensions, including time, space, frequency, and phase. Therefore, it has been widely applied to various fields, such as image processing and recognition [21], learning, and understanding of various spatiotemporal and spectrotemporal brain data [22]. With the development of SNN-based techniques to capture spatial and temporal data, new opportunities arose in action recognition. However, these bioinspired approaches share a common shortcoming: high computation cost. Fortunately, engineers, and especially the neuromorphic engineering community, aimed to mimic the human systems, such as spike-based dynamic vision retinas [23] and neuro-inspired spike-based motion [24]. Moreover, it is possible to integrate several thousand of artificial neurons into the same electronic device (neuromorphic devices), which further promotes the development and application of these bioinspired action recognition approaches.

Here, we present a bioinspired computational model for action recognition based on the feedforward SNN. In this model, we integrate more V1 and MT neuron properties than previous methods [19], [20], such as direction and speed selectivity [25], spatiotemporal inseparability [26], and center surround suppression [27], into a two-layered SNN. Unlike Escobar's model [19], our model (V1-MT model) is built with V1 and MT models (layers) on a multichannel parallel architecture by emulating the neural mechanisms of information processing in their corresponding cortical areas. Considering speed selectivity of neurons to adapt to speed changes in actions, each channel in V1 and MT models processes information with the spatiotemporal tuning properties of neurons at a certain speed and different directions. Moreover, according to spatiotemporal inseparability and center surround interactions, information processing in each channel is divided into two stages: information detection and spike conversion based on the conductance-driven integrate-and-fire (IF) model simplified with some constraints.

In the first stage, input signals in V1 and MT layers are perceived by a set of RFs, and then distributed to multiple neurons through synaptic connections for further processing with surround suppression between neurons in a cortical area. Hence, two operations: input signal perceiving and inhibition processing are proposed in this stage. In general, the perceptual functions of V1 neuron are modeled based on the spatial properties of the RF organization [28], [29]. However, further research reveals that the RF structure is inseparable in space time [26] and prefers movement in one direction [30]. Unlike other models, we use the 3-D Gabor filters modeling inseparability of the spatiotemporal tunings of V1 neurons to process input signals in our V1 model.

The perceptual functions of MT neuron are to pool input signals from V1 through synaptic connections between V1 and MT neurons. However, it is unclear in neuroscience how an MT neuron connects to V1 neurons by synapses. Fortunately, we know that, at any given eccentricity, an MT neuron classical RF (cRF) is about four times larger than that of a V1 neuron [31]. An MT neuron must summate input from many V1 neurons with RFs distributed across visual space. Therefore, based on previous work, we propose a mapping method from V1 to MT to model the spatiotemporal frequency tunings of MT neurons for perceiving motion information.

The motion information perceived by MT and V1 neurons is further processed by inhibition processing operations originated from surround suppressive interactions between neurons within V1 and MT [27], [32]. This can detect motion discontinuities or motion boundaries. In contrast to Escobar's model, our model considers the surround suppression not only in MT [32] but also in V1 [27]. Since its functional role depends on the surround's spatial and temporal organization, the inhibition processing operation is modeled with a 3-D difference of Gaussian (DoG) function.

In the second stage of information processing, each neuron in the V1 and MT layers, regarded as a spiking neuron with the IF model, receives information obtained in the previous stage and converts this information into spikes. To facilitate spike conversion, the IF model is simplified based on some constrains. Finally, spike trains are analyzed by considering the mean firing rate of each neuron. An action is represented by the mean motion maps built with the mean firing rates of all MT neurons, and recognized with a support vector machine (SVM) classifier.

Based on the above analysis, the key contributions of this paper are summarized as follows.

1) We propose a multichannel 3-D SNN architecture based on the direction and speed selectivity of V1 and MT neurons to adapt to speed changes in actions.
2) We integrate neuron properties into the 3-D SNNs, and propose two operations so to effectively detect motion information in video streams for action recognition.
3) We also propose a new feature vector to represent human action. This feature vector is built with the mean motion maps by combining spike trains from all channels.

The remainder of this paper is organized as follows. Section II describes SNNs and spiking neuron models in detail. Section III describes our bioinspired approach based on SNNs for human action recognition. In Section IV, we compare our approach with other bioinspired models and current state-of-the-art approaches on the Weizmann, KTH, and UFC Sport databases. Finally, we present both advantages and disadvantages as well as some perspectives of our approach in Section V.

II. Spiking Neural Networks

A. Spiking Neuron Model and Spike Train

Neurons are the basic unit of the nervous system. A neuron receives and transmits information from/to other neurons by synapses. If the sum of input signals into a neuron surpasses a certain threshold, the relevant neuron will send an electrical pulse. This behavior of biological neurons has been modeled with various spiking neuron models [33]–[35].

In this paper, we adopt the conductance-driven IF model [33], [35] to analyze neuron responses to external stimuli. Considering a neuron i, the model is

$$\frac{du_i(t)}{dt} = G_i^{\text{exc}}(t)(E^{\text{exc}} - u_i(t)) + g^L(E^L - u_i(t)) + G_i^{\text{inh}}(t)(E^{\text{inh}} - u_i(t)) + I_i(t). \quad (1)$$

where $I_i(t)$ is input current, and $G_i^{\text{exc}}(t)$ and G_i^{inh} are the excitatory conductance and inhibitory normalized conductance, respectively. This equation represents the spike emission process. The neuron i will first emit a spike when the normalized membrane potential $u_i(t)$ reaches threshold u_{th}. Then, $u_i(t)$ returns to resting potential E^L. Since the shapes of all spikes of a given neuron do not carry any information, regarding spike number and timing, we can treat all spikes as discrete time events forming a spike train $\eta_i = \{\ldots, t_i^n, \ldots\}$, where t_i^n denotes the nth spike of neuron i.

B. Spiking Neural Networks

In the visual cortex, millions of neurons are interconnected through synapses to build biological neural networks. To model biological neural networks, some SNN models are proposed based on biological behavior.

We also propose an SNN to model biological neural networks in V1 and MT, as shown in Fig. 1(a). This SNN is a two-layered feedforward architecture and similar to that of a traditional ANN. The processing unit, however, is a spiking neuron, which is typically modeled by an IF model in (1). In this SNN, the computation of each layer is similar, but inputs to the V1 layer are the spikes from lateral geniculate nucleus (LGN) while those to the MT layer come from V1 outputs.

In (1), $u_i(t)$ is determined by three parameters: $G_i^{\text{exc}}(t)$, $G_i^{\text{inh}}(t)$, and $I_i(t)$. However, $G_i^{\text{exc}}(t)$ and G_i^{inh} depend on the connections between neurons. Based on biological findings of information transmission in the visual cortex [27], we only consider the contribution of connections of intracortical neurons to $G_i^{\text{inh}}(t)$, and ignore connections of intercortical neurons. On the contrary, only connections of intercortical neurons are considered to compute $G_i^{\text{exc}}(t)$.

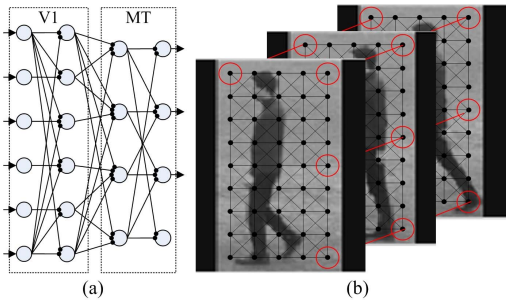

Fig. 1. SNN. (a) Network architecture modeling V1 and MT areas. (b) Connections of all neurons in space and time dimensions.

According to the above assumption, an excitatory V1 or MT neuron in (1) does not have any external input current $I_i(t)$, but receives spikes from the LGN and V1 neurons through delayed synaptic connections. Thus, a neuron has an excitatory current $I_i^{(e)}(t) = G_i^{\text{exc}}(t)(E^{\text{exc}} - u_i(t))$, which is computed by RF. Simultaneously, the neuron receives inhibitory current $I_i^{(i)}(t) = G_i^{\text{inh}}(t)(E^{\text{inh}} - u_i(t))$ due to lateral connections within a cortical area, which is computed by the surround inhibitor (details in Section III). As a result, (1) is simplified using the total current $I_i(t) = I_i^{(i)}(t) + I_i^{(e)}(t)$ as follows:

$$\frac{du_i(t)}{dt} = g^L(E^L - u_i(t)) + I_i(t). \quad (2)$$

Here, each neuron is a coincidence detection unit allowed to emit only one spike for a sequence image.

Moreover, our SNN based on visual cortical stream is 3-D in space. For simplicity, we assume that spiking neurons in the V1 and MT layers are uniformly distributed in more contiguous sequence images to capture motion information encoded in the video sequence, as shown in Fig. 1(b). A black spot represents a cRF center of single neuron and the colored circle indicates its size. Each neuron receives spikes from other neurons in space and time through their synaptic connections.

III. Bioinspired Approach

Numerous studies have conducted visual motion analysis in the V1 and MT. In this paper, we propose a bioinspired approach based on SNNs for human action recognition by modeling some properties of V1 and MT neurons. This approach consists of three parts: bioinspired computational models for motion analysis, feature extraction, and action recognition, as shown in Fig. 2. The core of our approach is the bioinspired model (V1-MT model) with a two-layered feedforward architecture composed of both V1 and MT models (layers). Neurons in LGN first encode the video sequence $s(\mathbf{x}, t), \mathbf{x} = (x, y)$ using a family of their equally spaced overlapping Gaussian RFs, and generate spike trains η^L with the IF model. Then, these spike trains are distributed to multiple V1 input neurons through synaptic connections and subsequently processed by V1 and MT models to generate new spike trains η^V and η^M. Next, we analyze the spike trains η^M coming from the V1-MT model to extract motion features with the average firing rates of all neurons. Finally, these features are used to train an SVM classifier for action recognition.

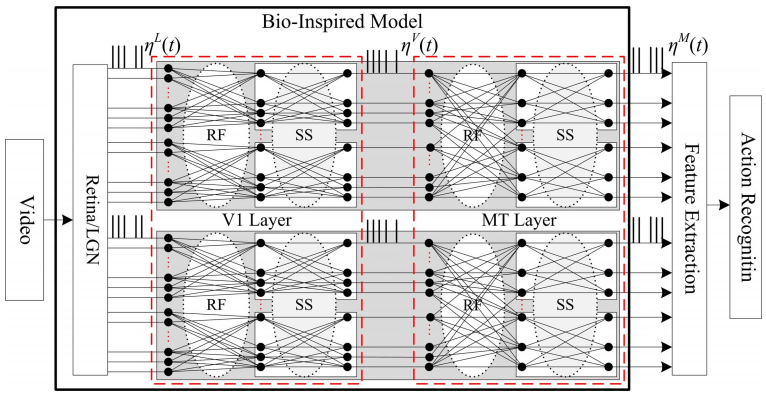

Fig. 2. Architecture of the proposed approach. It consists of three parts: V1 and MT layer models, feature extraction, and action recognition.

Each V1 and MT model is a 3-D SNN and is divided into two stages: 1) information detection, where the input spatiotemporal information is detected through modeling the properties of RF and surround suppression and 2) spike conversion, where a neuron as a spike entity converts information obtained in the previous stage into spike trains. To model the direction and speed selectivity of spiking neurons in V1 and MT, our 3-D SNN is built with multiple channels determined by the number of different tuning speeds (Fig. 2). The complexity of each channel is influenced by the number of selected tuning directions. These are discussed in detail in Section III: A-D. Additionally, because surround suppression is an important property of V1 and MT neurons, its computation for V1 and MT models is discussed separately.

A. Surround Suppression

Neurophysiological studies show that surround suppression is observed in different cortical areas, such as V1 [27] and MT [32]. That is to say, once a neuron is activated by input spikes in its cRF, another in the same cortical area outside that field can have an effect on the neural response. Although the suppressive effect is generally modeled by a 2-D DoG function [36], it does not consider the influence of the surround at that time instant. Considering delayed synaptic connections between intracortical neurons, we define an inhibition processing operation with a 3-D weighting function in the time-space domain $w_{v,\theta}(\mathbf{x},t)$, $\mathbf{x}=(x,y)$ as follows:

$$w_{v,\theta,k_1,k_2}(\mathbf{x},t)=\frac{I_{v,\theta,k_1,k_2}(\mathbf{x},t)}{\|I_{v,\theta,k_1,k_2}(\mathbf{x},t)\|_1} \tag{3}$$

where $\|\cdot\|_1$ denotes the L1 norm, and $I(\mathbf{x},t)$ is a 3-D DoG function $D(\mathbf{x},t)$ with half-wave rectification $|\cdot|^+$ [37], namely

$$D_{v,\theta,k_1,k_2}(\mathbf{x},t)=G_{v,\theta,k_2}(\mathbf{x},t)-G_{v,\theta,k_1}(\mathbf{x},t) \tag{4}$$

here

$$G_{v,\theta,k}(\mathbf{x},t)=\frac{\gamma}{2\pi(k\sigma)^2}\exp\left(\frac{-((\bar{x}+v_c t)^2+\gamma^2\bar{y}^2)}{2(k\sigma)^2}\right)\times\frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{(t-u_t)^2}{2\tau^2}\right)\cdot\varepsilon(t)$$

where $\bar{x}=x\cos(\theta)+y\sin(\theta)$, $\bar{y}=-x\sin(\theta)+y\cos(\theta)$, and the step function $\varepsilon(t)$ ensures 3-D DoG function causality.

Fig. 3. Spatiotemporal behavior of a surround weighting function $w_{v,\theta}(\mathbf{x},t)$.

The parameters v and θ are the preferred motion speed and direction, respectively, while σ and γ are the standard deviation and spatial aspect ratio of the spatial Gaussian factor, respectively. The parameter σ also represents the spatial size of the neuron cRF and is determined by the preferred speed v: $\sigma=\lambda_0\sqrt{1+v^2}$, where λ_0 is a constant.

The value of surround weighting function is 0 inside the cRF, positive outside the cRF, and decays with distance from the cRF. Here, k_1 and k_2 are set to 1 and 4, respectively, determining the size of the surrounding area. $v_c=v$ is the speed with which the center of the spatial Gaussian envelope moves along the $\overline{x}$ axis. This behavior is shown in Fig. 3, where the speed v is set as 1 ppF (pixel per frame) and the direction θ is 0.

For each neuron at $(\mathbf{x},t)$, its inhibition term $S_{v,\theta}(\mathbf{x},t)$ is computed by weighted summation of the responses $E_{v,\theta}(\mathbf{x},t)$ of other neurons in the surrounding area as follows:

$$S_{v,\theta}(\mathbf{x},t)=E_{v,\theta}(\mathbf{x},t)*w_{v,\theta}(\mathbf{x},t). \tag{5}$$

Thus, its surround suppressed response $\widehat{E}_{v,\theta}(\mathbf{x},t)$ (or surround suppressive motion energy) is given by

$$\widehat{E}_{v,\theta}(\mathbf{x},t)=|E_{v,\theta}(\mathbf{x},t)-\alpha S_{v,\theta}(\mathbf{x},t)|^+ \tag{6}$$

where factor α controls the suppressive strength, and is set to 2 in our experiments.

B. Input Spike Trains

First, a sequence $s(\mathbf{x},t)$ is perceived by LGN neurons with a cRF function $D(\mathbf{x},t)$. Their response to the sequence is computed with convolution following half-wave rectification:

$$L(\mathbf{x},t)=|s(\mathbf{x},t)*D(\mathbf{x},t)|^+. \tag{7}$$

The physiological data [25] demonstrates that the cRF profile of an LGN neuron is a separable spatiotemporal function and is described as a product of a spatial term and a temporal

term. Hence, we define the function $D(\mathbf{x}, t)$ with a 3-D DoG similar to (4), but reset the parameters. Here, the parameters θ, υ_c, and γ in (4) are set to 0, 0, and 1, respectively. The spatial term becomes a 2-D DoG independent of time and the temporal term is a time Gaussian function. Other parameters are set according to the literature [38].

After the response of an LGN neuron is estimated, it is transformed to spikes using the IF model in (1). Here, only considering short-range connections among neurons within LGN, the surround suppression interactions are ignored and the inhibitory current is 0. Moreover, as there are not the presynaptic neurons connected to an LGN neuron, its excitatory current is also 0. Hence, the IF model is also simplified to (2) and the membrane potential of an LGN neuron is determined by its external input current $I_i(t)$ associated with the response $L(\mathbf{x}, t)$. For a spiking neuron i located in $\mathbf{x}$, the current $I_i(t)$ is computed as follows:

$$I_i(t) = K_L L(\mathbf{x}, t) \tag{8}$$

where K_L is a factor. The activity of an LGN neuron is determined by the membrane potential in (2) and its threshold which is defined with biological data. Thus, a neuron in each pixel location $\mathbf{x}$ corresponding to video stimuli generates a spike train $\eta_i^L(t)$, and an image sequence $s(\mathbf{x}, t)$ is converted into a spiking sequence $l(\mathbf{x}, t)$.

C. V1 Model

The spikes are propagated from the excitatory LGN neurons to the V1 model in a feedforward manner and processed by the V1 model to generate new spikes.

1) Simple and Complex Cell Models: Based on their cRF structures, V1 neurons are classically divided into two types: simple cells and complex cells [39]. In the V1 model, input information is processed by simple cells. Studies have shown that the cRF profiles of many simple cells are inseparable functions of space and time [37]. Thus, similar to [25] and [38], we construct the cRF with a 3-D Gabor spatiotemporal filter to model their properties, defined as follows:

$$\begin{aligned} g_{\upsilon,\theta,\varphi}(\mathbf{x}, t) = & \frac{\gamma}{2\pi\sigma^2} \exp\left(\frac{-((\bar{x} + \upsilon t)^2 + \gamma^2 \bar{y}^2)}{2\sigma^2}\right) \\ & \times \cos\left(\frac{2\pi}{\lambda}(\bar{x} + \upsilon t) + \varphi\right) \\ & \times \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(t - u_t)^2}{2\tau^2}\right) \cdot \varepsilon(t) \end{aligned} \tag{9}$$

where the inseparability of the filter $g_{\upsilon,\theta,\varphi}(\mathbf{x}, t)$ in space and time reflects 2 aspects: that a spatial Gaussian factor is relative to time and speed, and that the spatial cRF size σ is determined by the preferred speed υ. Thus, the motion information perceived by a model simple cell is computed with the response $r_{\upsilon,\theta,\varphi}(\mathbf{x}, t)$ of a 3-D Gabor filter to a sequence $l(\mathbf{x}, t)$, which is defined as follows:

$$r_{\upsilon,\theta,\varphi}(\mathbf{x}, t) = |l(\mathbf{x}, t) * g_{\upsilon,\theta,\varphi}(\mathbf{x}, t)|^+. \tag{10}$$

Next, the responses of simple cells are further processed by complex cells. Currently, there are three types of complex cell models [40]–[42]: the energy model, MAX model, and learn model. We use the energy model to compute the response of a complex cell (motion energy) as follows:

$$E_{\upsilon,\theta}(\mathbf{x}, t) = \sqrt{r^2_{\upsilon,\theta,0}(\mathbf{x}, t) + r^2_{\upsilon,\theta,\pi/2}(\mathbf{x}, t)}. \tag{11}$$

Then, the motion energy $E_{\upsilon,\theta}(\mathbf{x}, t)$ is used to compute surround suppressive energy $\widehat{E}_{\upsilon,\theta}(\mathbf{x}, t)$ with (6). Fig. 4 shows the intermediate results generated by the V1 model on a sequence in KTH, which include the total motion energy and surround suppressed motion energy in all directions at speed $\upsilon = 1$ ppF.

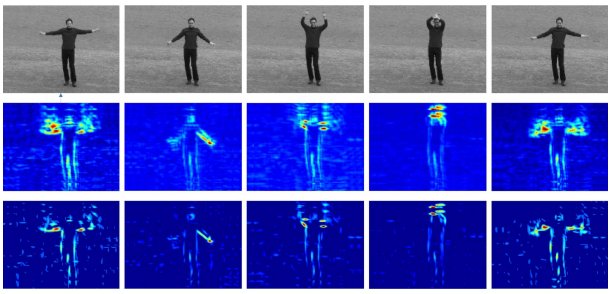

Fig. 4. Spatiotemporal information detected by V1 model. From the first to final row: snapshots of a sequence, motion energy, and surround suppressed motion energy.

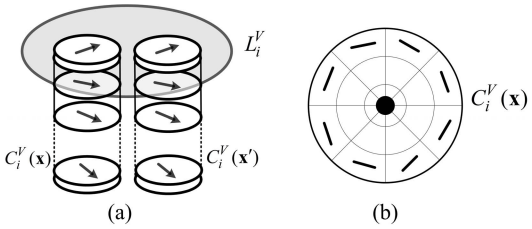

Fig. 5. Architecture of a V1 channel. (a) Column of N_θ V1 cells. (b) Pinwheel.

Based on the V1 cell model in (11), we consider V1 cells at N_υ tuning speeds. This network of V1 model consists of N_υ channels (or layers) (see Fig. 2), each of which is built with V1 cells with the same speed tuning but N_θ different direction tunings. Since all cells belonging to a layer $L_i^V = \{C_i^V(\mathbf{x})|\mathbf{x}\}$ and centered at position $\mathbf{x}$ form a functional *column*, $C_i^V(\mathbf{x}) = \{E_{\upsilon_i,\theta_j}(\mathbf{x})|j\}$, or direction *pinwheel* as shown in Fig. 5, each channel contains N_θ subchannels, as shown in Fig. 2. In this way, all V1 neurons in each subchannel share the same speed and direction tuning.

2) Action Detection and Rescaling: In the real scene, our eyes follow the motion when a person moves across our visual field. This smooth pursuit is regulated by an attention mechanism. Studies showed that the action recognition performance in biological motion is highly modulated by attention [43].

Based on the attention mechanism of initial perception, we first form initial visual perception with a series of 3-D Gabor filters at different speeds and directions. Action objects are then detected based on saliency maps [44] with only slight human intervention. Meanwhile, we also focus on the scale variability of an object in large range. When the scale variability of an object exceeds the maximal spatial frequency of V1 neurons in consecutive frames, it causes "induced motion,"

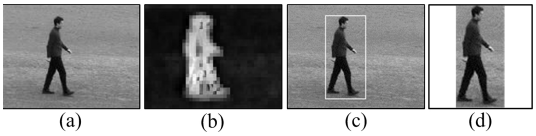

Fig. 6. Action detection and rescaling. (a) Original snapshot. (b) Saliency map. (c) Original image with ground truth rectangle. (d) Rescaled image.

resulting in error coding of motion information. Therefore, once action objects with the large scale variations are detected from a video, rescale them so that the high dimension is the same value, similar to focal length adjustment, as shown in Fig. 6. To avoid distortion of the detected object, expansions or contractions are done with the same spatial aspect ratio.

3) Input Current: The objective of the spiking neuron model is to transform the analog response of a neuron obtained by cRF and surround inhibition processing to a spiking response, and characterize its activity. From (2), the activity of a neuron is determined by its input current $I_i(t)$ and membrane potential threshold.

For a spiking neuron i in V1 located in $\mathbf{x}_i$, the input current $I_i(t)$ associates with its surround suppressive motion energy $\widehat{E}_{\upsilon,\theta}(\mathbf{x}, t)$, as defined in (6). Thus, the *input current* $I_i(t)$ of the ith neuron is modeled as follows:

$$I_i(t) = K_{\text{exc}} \widehat{E}_{\upsilon,\theta}(\mathbf{x}, t) \tag{12}$$

where K_{exc} is a control factor and its value is obtained from experiments. The membrane potential is computed with (2). According to the above computation, a V1 neuron i with preferred direction θ and speed υ generates a spike train $\eta_i^V(t)$ corresponding to input spike trains $\eta^L(t)$ coming from LGN.

D. MT Model and Input Current

The spikes generated by V1 model neurons continue to propagate to the MT model as input. The MT model is responsible for integrating these spikes and generating new spikes. However, how MT neurons integrate motion visual signals still remains a conundrum. Therefore, we aim to simulate partial biological properties of MT neurons to realize effective feature extraction of human action.

1) Mapping Between V1 and MT: In general, a single MT neuron connects to multiple V1 neurons. However, how MT neurons receive afferent connections from V1 neurons is anatomically unclear. Therefore, similar to the V1 model, we consider MT neurons at N_υ preferred speeds and build the network of MT model with N_υ channels (or layers) (Fig. 2). In each channel, the neurons have the same speed tuning but different direction tunings. In position $\mathbf{x}$ of a layer $L_i^M = \{C_i^M(\mathbf{x})|\mathbf{x}\}$, N_θ sublayers $C_i^M(\mathbf{x}) = \{E^M_{\upsilon_i,\theta_j}(\mathbf{x})|j\}$ are set. In this way, we assume that an MT neuron in a channel only connects to V1 neurons in the corresponding channel. In other words, the inputs of an MT neuron in channel L_i^M only depend on outputs of V1 neurons in channel L_i^V, and are independent of other V1 neurons.

In addition, neurophysiological research indicates that the cRF size of an MT neuron is four times bigger than that of the V1 cRF [31]. Therefore, we only consider the connections

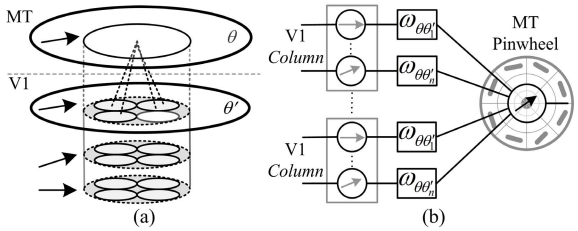

Fig. 7. Mapping connection between V1 and MT neurons. (a) Spatial projection of an MT cRF into V1 area. (b) Connection of an MT neuron to many V1 neurons.

to an MT neuron from V1 neurons spatially localized within the MT neuron's cRF, as shown in Fig. 7. Connections of other V1 neurons outside a projecting region of the MT cRF are neglected.

According to the above assumptions, an MT neuron in a layer (a channel in Fig. 2) receives spike trains from V1 neurons in the corresponding layer (channel) and within the cRF region of this MT neuron [Fig. 7(a)]. Therefore, the motion energy of an MT neuron depends on connection weights between MT and V1 neurons, and spike trains from V1 neurons. As shown in Fig. 7(b), we compute the motion energy of an MT neuron as follows:

$$E^M_{\upsilon,\theta}(\mathbf{x}_i, t) = \sum_{\theta'} \sum_{\mathbf{x}_j} \sum_{t^f} w_{\theta\theta'}(\mathbf{x}_i - \mathbf{x}_j)\alpha(t - t^f)\eta^V_{j,\upsilon,\theta'} \tag{13}$$

where $\eta^V_{j,\upsilon,\theta'}$ represents the spike train generated by jth V1 neuron with speed υ and direction θ' tuning, $w_{\theta\theta'}(\mathbf{x}_i - \mathbf{x}_j)$ is the weight factor between the MT neuron i (at $\mathbf{x}_i$ position and with θ direction) and V1 neuron j (at $\mathbf{x}_j$ position and with θ' direction), and t^f denotes the time of the fth spike of the V1 excitatory neuron. The temporal kernel $\alpha(s)$ models the time course of the postsynaptic current responding to spikes arriving from V1 neurons. The temporal kernel is

$$\alpha(s, \tau_s) = \frac{s}{\tau_s} \exp\left(-\frac{s}{\tau_s}\right). \tag{14}$$

The weight $w_{\theta\theta'}(\mathbf{x}_i - \mathbf{x}_j)$ represents the strength of spatial coupling between V1 neuron j and MT neuron i and depends on their spatial and direction-selective relationship. Since the direction selectivity of MT neurons could be inherited from direction-selective inputs, we can assume that the weight between neuron j and neuron i is proportional to the angle φ_{ij} between the two preferred directions, as shown in Fig. 8. Here, the connection weight $w_{\theta\theta'}(\mathbf{x}_i - \mathbf{x}_j)$ is given by

$$w_{\theta\theta'}(\mathbf{x}_i - \mathbf{x}_j) = k_c w_d(\mathbf{x}_i - \mathbf{x}_j)\cos(\varphi_{ij}) \tag{15}$$

where k_c is a factor, $\varphi_{ij} = |\theta - \theta'|$ is the absolute angle between two preferred directions, and $w_d(\mathbf{x}_i - \mathbf{x}_j)$ is the weight associated with the difference between their center positions. Because the local connections appear spatially isotropic based on neuroanatomy, we take the weight of these local connections as Gaussian, specifically described by

$$w_d(\mathbf{x}_i - \mathbf{x}_j) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\mathbf{x}_i - \mathbf{x}_j)^2}{2\sigma^2}\right] \tag{16}$$

where σ is determined by the cRF size of the MT neuron.

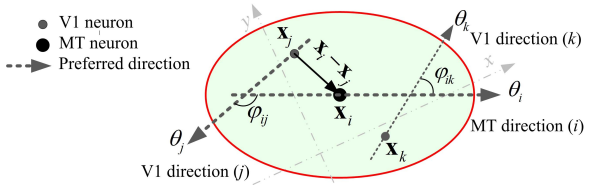

Fig. 8. Connection weights between V1 and MT neurons are modulated by the cosine of the absolute difference φ_{ij} and φ_{ik} of the preferred directions and spatial distance between the ith MT neuron and the jth and kth V1 neuron.

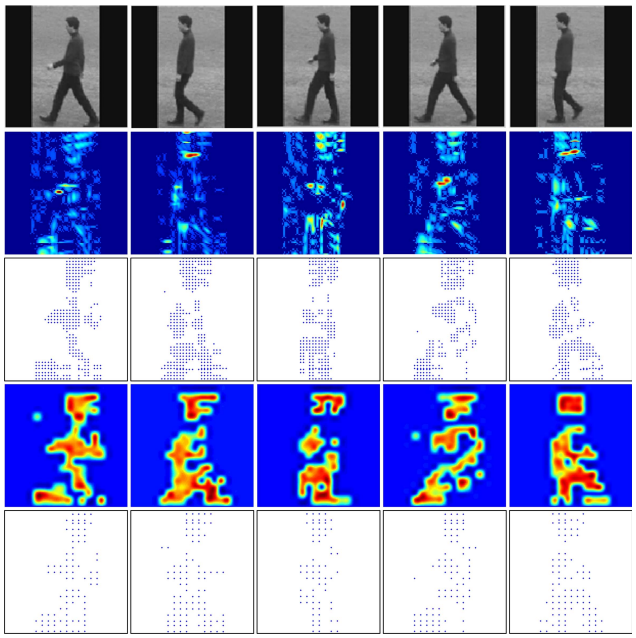

Fig. 9. Spike maps of V1 and MT neurons at 1 ppF speed on an action. *From first to final row*: snapshots of a sequence, surround suppression motion energy in V1, spike maps of V1 neurons, surround suppression motion energy in MT, and spike maps of MT neurons.

2) Input Current and Spike Trains: As aforementioned, we suggest that a model MT neuron only pools spikes from V1 neurons mapping in its cRF region, while other V1 neurons outside that region are ignored. This is because the contribution of V1 neurons mapping outside the MT cRF is indirectly affected by the surround interaction between MT neurons. Thus, the surround suppressed motion energy of an MT neuron $\widehat{E}^M_{v,\theta}(\mathbf{x},t)$ is computed by (6) with $E^M_{v,\theta}$ replacing $E_{v,\theta}$. Parameter σ in (4) is the spatial scale of an MT neuron.

Similar to V1 neurons, the input current $I^M_i(t)$ of the ith MT neuron in (2), associated with its surround suppressed motion energy $\widehat{E}^M_{v,\theta}(\mathbf{x},t)$, is computed by (12), and then is transformed into spikes by the modified IF model described in (2), forming its spike train $\eta^M_{i,v,\theta}$.

Fig. 9 shows spike maps of our models on action *walking* in KTH. It includes surround suppressive motion energy of V1 and MT models computed with (6) at 1 ppF speed. The corresponding spike maps obtained are also shown. Fig. 10 shows the spike trains of V1 and MT neurons responding to KTH(s1) 2 actions: *handclapping* and *walking* (see Section IV-A). The raster plots consider all V1 and MT neurons with only cRF of a given direction (0°) and speed (1 ppF).

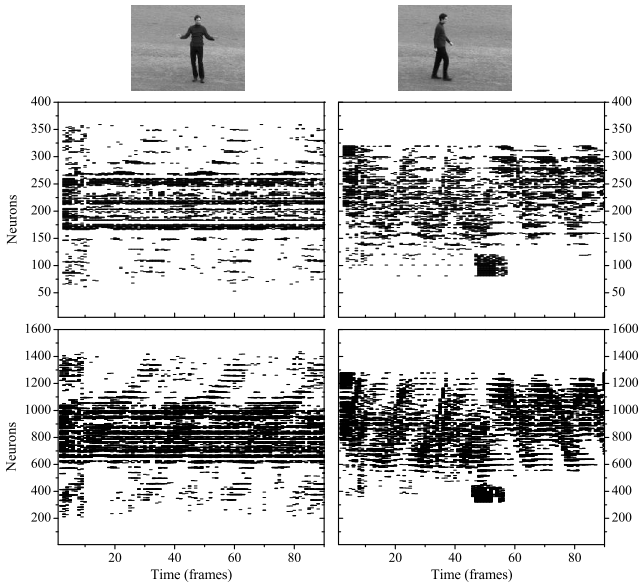

Fig. 10. Spike trains of V1 and MT neurons corresponding to two actions: *handclapping* and *walking*. *First row*: original snapshots of two actions. *Second row*: spike trains of MT cells. *Third row*: spike trains of V1 cells.

E. Action Recognition

To effectively recognize human actions from video sequences, we must analyze spike trains generated by neurons and extract the features truly representing these actions. Researchers have proposed different analytical methods on the formation of neural code, such as rank order coding [45], synchronization [46], and mean firing rate [47]. Although this paper does not focus on neural code, an effective method is beneficial for recognition performance. Here, we use mean firing rate as a simple and effective method. Considering a spiking neuron i, the mean firing rate is computed as follows:

$$r_i(t,\triangle t)=\frac{n_i(t-\triangle t,t)}{\triangle t} \tag{17}$$

where $n_i(t-\triangle t,t)$ is the number of all spikes within a sliding time window $[t-\triangle t,t]$.

The mean firing rate, which is the average number of spikes between times $t-\triangle t$ and t, has many advantages. As a function of time, it is independent of sequence length and can reduce the effect of noise during feature extraction. However, due to the small time window, the mean firing rate only represents spike density at time t generated by a neuron responding to a sequence image, and cannot express the process of neuronal activity excited by an action sequence over time. Human action is a process of acting or doing, and excites a series of continuous sequence images in the visual system. Thus, recognizing human action must consider the mean firing rates of a neuronal array in a long time internal responding to the video sequence. In this way, the number of features with mean firing rates is increased. To reduce the feature dimension and build a feature vector independent of time, we define a new feature vector $H_I(\cdot)$, the mean motion maps, through averaging over mean firing rates in the interval

TABLE I
PARAMETERS FOR 3-D GABOR AND IF MODEL

λ_0	γ	θ	u_t	τ	φ	E^L	u_{th}	g^L
1	0.5	π/n	1.75	2.75	$0, \pi/2$	-70	-50	0.1

T as follows:

$$H_I = \{h_i\}_{i=1,\ldots,N\times N_L}, \text{ and } h_i = \frac{\sum_{i=1}^{T} r_i(t, \triangle t)}{T} \quad (18)$$

where N is the number of neurons per subchannel (or sublayer), and N_L is a product of N_v and N_θ.

As above, we only model V1 and MT areas for feature extraction and do not consider other brain areas involved in motion analysis. Therefore, the final step is to train the classifier without linking to biology using the features. SVM as a classifier is directly used for human action recognition.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Database and Settings

To demonstrate the ability of the proposed approach in classifying actions performed by different people under different environments, we use three public human action databases: Weizmann,[1] KTH,[2] and UCF Sports.[3]

The Weizmann database consists of nine different subjects performing nine different actions: *bend, jack, jump, pjump, run, side, walk, wave1, and wave2*. Conversely, the KTH database contains six types of human actions: *walking, jogging, running, boxing, handwaving, and handclapping*. These actions are executed multiple times by 25 subjects in 4 different conditions: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), and indoors with lighting variation (s4). For both databases, we automatically obtain subimage sequences focusing on human actions by using an existing approach based on visual attention. For computation sake, all subimages are rescaled to 120×120 pixels.

To evaluate the performance of our approach, all experiments are executed with a cross validation protocol similar to Escobar's [19]. For the Weizmann database, we use actions of six subjects for the training set, and the remaining three subjects for the testing set. We run all training sets to compute the average recognition rate (ARR). For KTH, we use the actions from 9 subjects for training, and the remaining 16 subjects for testing. However, since there are many possible training sets for KTH (C_{25}^9), we choose 84 trials to compute the performance (ARR).

B. Parameter Settings

Many parameters are involved in our model. To achieve better recognition performance, we specify parameter values. Considering biological experimental data [33], [48] and previous related work [37], the 3-D Gabor filters and IF neuron model parameters are set, as shown in Table I. For both V1 and MT models in our proposed V1-MT model, parameterization considers some restrictions found in experimental data [31] and computation cost. General V1 and MT settings are shown in Table II. For example, both V1 and MT models are built with N_v channels, each of which is formed with four sublayers corresponding to four directions. However, the number of neurons per sublayer differs: a V1 sublayer contains 1600 neurons, while a total of 400 MT neurons are sparsely distributed in an MT sublayer. The values of K_{exc} in (12) for V1 and MT layers are set to 600 and 100, respectively.

TABLE II
PARAMETERS USED FOR V1 AND MT LAYERS

	V1	MT
Number of speeds	N_L	N_L
Number of orientations	4	4
Size of receptive field	σ_{V1}	$4\sigma_{V1}$
Cell density	0.33[cells/pixel]	0.17[cells/pixel]
Number of cells per sublayer	1600	400

C. Performance With Different Parameters

In the proposed approach, the feature vector H_I in (18) obtained by our model, which represents human action in a sequence, depends on many parameters, including the frame length of a subsequence T, the size of the sliding time window $\triangle t$, and the number (N_v) of the preferred speeds and their values (v). Thus, we perform some test experiments with different parameters and evaluate their influence on action recognition performance. To ensure simplicity and consistency of experiments, all tests are conducted under the same conditions, unless otherwise stated.

1) Frame Length of Test Sequences: From (18), we can observe that the longer the frame length T, the more information encoded. When T matches a duration of a whole action, the complete motion information of an action is captured. However, the same actions performed by different people show variations in execution speed and style, and choosing T with an action duration is impractical. Therefore, we must assess the impact of different frame lengths of the subsequences cut out from original sequences on recognition performance. Some aforementioned tests with cross validation are performed on subsequences with frame lengths ranging from 20 to 50 frames. Note that the sliding time window $\triangle$ is set to 3, and speeds v are, respectively, set to 1, 2, and 3 ppF in all tests. Classification results on the KTH database with different frame lengths are shown in Fig. 11.

As shown in Fig. 11, ARRs of the proposed approach increase with the increase of frame length T even if small variations occur in a small range. This performance improvement validates that additional encoded information is helpful for action recognition. However, the longer the subsequence is, the heavier the computational load becomes. As a compromise between performance and computational load, the frame length is finally set to 50 for our model.

2) Size of Sliding Time Window: To evaluate the influence of the size of a sliding time window Δt on recognition results,

[1] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html
[2] http://www.nada.kth.se/cvap/actions/
[3] http://crcv.ucf.edu/data/UCF_Sports_Action.php

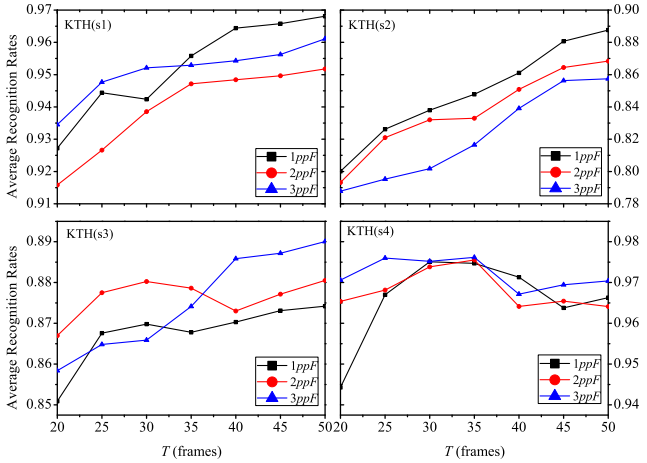

Fig. 11. ARRs of our model on the KTH database with different frame lengths T and different speeds, where $\triangle t = 3$.

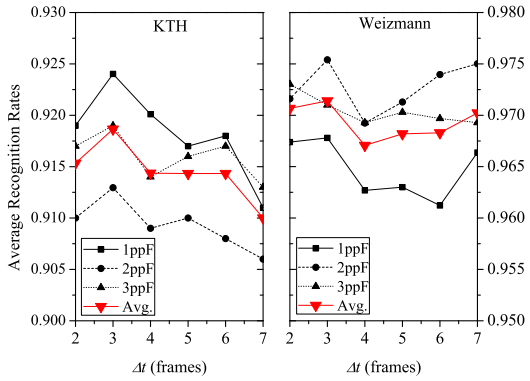

Fig. 12. ARRs of our model with different sizes $\triangle t$ and different speeds on the KTH (*left*) and Weizmann (*right*) databases.

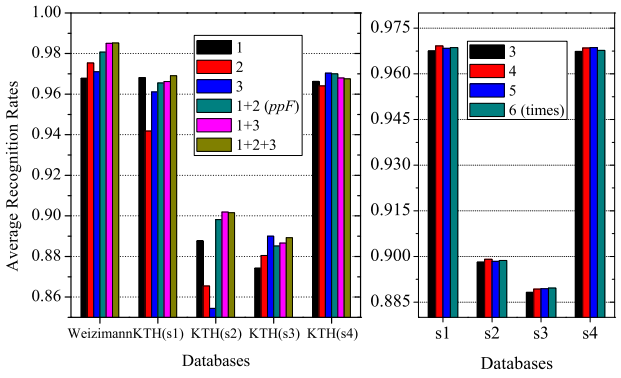

Fig. 13. Average recognition performance under different conditions. *Left*: different speeds and their combination. *Right*: different sizes of MT cRF.

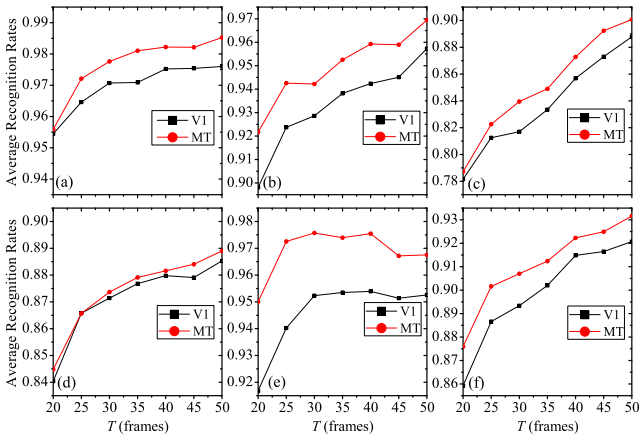

Fig. 14. Performance comparison of V1 model and V1-MT model with different sequence length. (a) Weizmann. (b) KTH(s1). (c) KTH(s2). (d) KTH(s3). (e) KTH(s4). (f) All of KTH.

we perform our model on all videos in the Weizmann and KTH databases. According to the above conclusions, the length of all training and testing subsequences is 50 (if the original sequence is less than 50 frames, it is directly used).

Experimental results with different values of Δt are shown in Fig. 12. Each result at a speed is the average performance of all actions in each database at that speed. ARRs at different speeds vary with Δt, but not regularly. However, for more general sequences in KTH, recognition performance decreases when Δt is greater than 3, as shown in Fig. 12 (*left*). This indicates that although a large window is beneficial to reducing interruptions of undesired stimuli to feature extraction, it also reduces distinction between features representing different actions, leading to performance degradation. After averaging ARRs (*red lines*) at all speeds on each database, we find the optimal value of the sliding time window is 3.

3) Speed Combinations: From the experimental results on the KTH database shown in Fig. 11, we can find that different speeds lead to different ARR performances. The same conclusion is also seen in Fig. 12. For example, the highest ARR on KTH(s2) is obtained at the speed 1 ppF, while the optimal performance on KTH(s3) is at 3 ppF. This is because each action in different sequences operates at different speeds. Thus, it is impossible for effective representation of human actions to extract features from sequences at a single speed. To adapt to speed changes of actions in sequences, we consider multiple speeds in our model, and build a multichannel architecture.

Theoretically, feature extraction at more speeds is beneficial for recognition performance improvement, but increases computation load. Statistical analysis of real video sequences suggests that more than 70% of motion vectors are enclosed in the central 5×5 area [49]. In other words, more than 70% of motion speeds are no larger than 3 ppF. Moreover, considering the biological result of the maximal spatial and temporal frequencies of each visual cortical area [50], we combine different speeds (such as 1, 2, and 3 ppF) to evaluate speed influence on performance of our approach. Partial experimental results at a single speed and combined speeds on the Weizmann and KTH databases are shown in Fig. 13 (*left*). Overall, our model with a combination of increased speeds provides higher action classification accuracy than that with a single speed. Therefore, we perform our approach using the combination of three speeds (1, 2, and 3 ppF) over all experiments.

4) Other Parameters: We also address the performance of our approach with respect to other challenging factors, such as the MT RF size and rescaling operation. First, in the above tests of performance evaluation, the size of the MT RF is

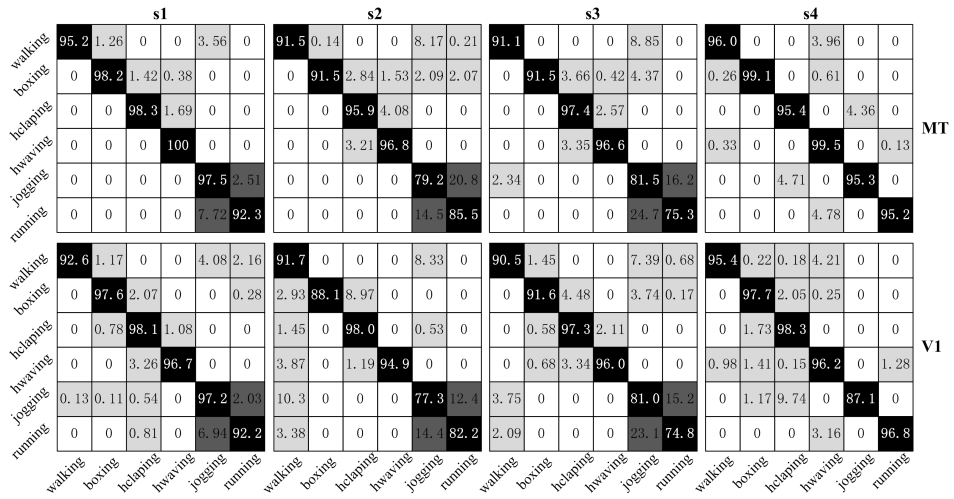

Fig. 15. Confusion matrices of classification on the KTH database obtained by our models. *Top*: MT model. *Bottom*: V1 model. From *left* to *right*: s1, s2, s3, and s4

TABLE III
RECOGNITION PERFORMANCE WITH RESCALING ON KTH

	s1	s2	s3	s4
Rescaling	96.9	90.1	88.9	96.8
No Rscaling	96.9	88.7	88.8	96.7

four times greater than the size of the corresponding V1 RF. To evaluate the impact of MT RF size on performance, we test our approach with different MT RF sizes on the KTH database. Fig. 13 (*right*) shows the experimental results with three to six times greater sizes of the V1 RF. Note that all tests are performed under the same conditions: frame length $T = 50$, sliding time window $\triangle t = 3$, and the combination of three speeds $v \in \{1, 2, 3\}$ ppF. According to the experimental results, four to six times V1 RF makes our approach almost the same recognition performance, while three times V1 RF gives the lowest performance. However, larger MT RF sizes cause more heavy computational cost. Thus, MT RF size is finally set to four times size of V1 RF in our approach.

Additionally, we introduced the rescaling operation in the above tests. To evaluate the effects of this operation on recognition performance, we perform exhaustive comparative experiments on the KTH database. The action recognition experiment is performed twice with and without rescaling operation for each input sequence. Table III compares their results to each other. We clearly see that the rescaling operation improves recognition performance on KTH, but primarily KTH (s2). This means that the rescaling operation is helpful for action recognition on sequences including zoom shots.

D. Performance Comparison

1) Comparison Based on Different Visual Areas: Studies have shown that biological motion patterns are recognized in either MT or V1 [51], [52]. That is to say, motion information processed through only the V1 area, or through V1 and MT areas, can be used to recognize human actions. Therefore, not only our V1-MT computational model but our V1 model built with only V1 layer should be able to recognize human actions from videos. To validate these two different models, features representing human action are extracted from the spike trains coming from the models. Performance is evaluated on all video sequences in the Weizmann and KTH databases with different frame lengths (T).

TABLE IV
PERFORMANCE COMPARISON WITH OTHER BIOINSPIRED METHODS ON WEIZMANN DATABASE

	ARR(%)	std(%)	Trials
Ours	98.52	1.61	84
V1/MT (Mean) [19]	92.68	4.62	84
V1/MT (Synchrony) [19]	92.81	5.15	84
V1/MT (TD) [20]	96.34	0.72	84
V1/MT (SKL) [20]	96.47	0.81	84
HMAX (StC2 dense) [15]	91.10	5.90	5
HMAX (StC2 sparse) [15]	97.00	3.00	5

Experimental results of V1 and V1-MT models in Fig. 14 show the characteristics of two aspects. On the one hand, all ARRs of either V1 or V1-MT models on all databases still increase with growing frame length. On the other hand, regardless of frame length, ARRs of the V1-MT model are higher than those of the V1 model [see Fig. 14(a) and (f)]. Especially in simple scenes, such as Weizmann, KTH(s1), and KTH(s4), performance improvement on human action recognition is significant [see Fig. 14(a), (b), and (e)]. However, for videos in complex scenes, performance improvement is unremarkable [see Fig. 14(d)]. This is mainly because the subjects in KTH(s3) wear different clothes, such as coat and dress, resulting in action occlusion and inaccurate feature extraction. Although our V1-MT model, which only uses a bioinspired feedforward spiking network, improves action recognition performance by using sparse features in the advanced visual cortical area, it cannot achieve the intended effect for action recognition in complex scenes.

Fig. 15 presents the confusion matrices of the classification on the KTH database by V1 and V1-MT models. The column represents the instances to be classified, while each row

TABLE V
PERFORMANCE COMPARISON WITH OTHER BIOINSPIRED METHODS ON KTH DATABASE

ARR(%)/std(%)	S1	S2	S3	S4	Avg	Trials
Ours	96.9/1.6	90.1/2.7	88.9/3.8	96.8/2.0	93.2/2.5	84
V1/MT [20]	83.1/2.0	–	69.8/2.8	83.8/1.9	79.8/2.2	100
V1/MT [20]	92.0/0.01	–	84.4/1.2	92.4/0.01	89.6/0.4	5
HMAX (StC2 dense) [15]	89.8/3.1	81.3/4.2	85.0/5.3	93.2/1.9	87.3/3.6	5
HMAX (StC2 sparse) [15]	96.0/2.1	86.1/4.6	88.7/3.2	95.7/2.1	91.6/3.0	5

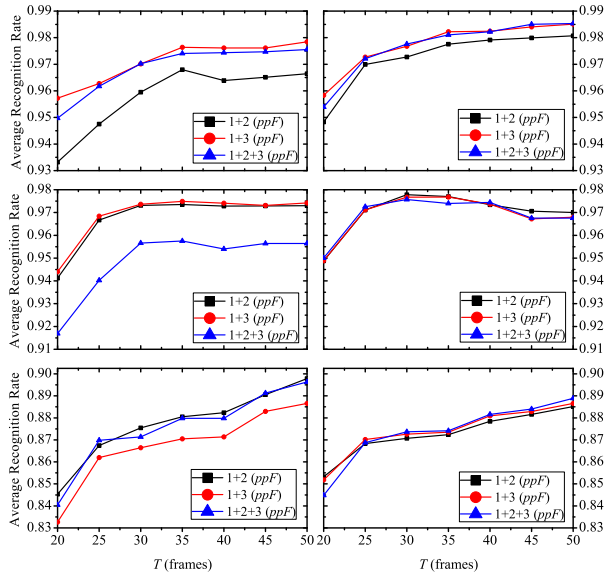

Fig. 16. Performance comparison of V1 model (*left column*) and V1-MT model (*right column*) with different speed combinations. From the *first* to *third* row: Weizmann, KTH(s4), and KTH(s3).

represents the corresponding classification results. The main confusion occurs between *jogging* and *running*. It is very challenging to distinguish the two actions, because they are similar when performed by some subjects. Furthermore, it is seen that actions confused with other actions in V1-MT model are fewer than those in V1 model.

Moreover, to further examine the robustness of our model, we compare the performance of our V1-MT and V1 models at different speed combinations. Experimental results in Fig. 16 show that the V1-MT model provides almost consistent high performance at different speed combinations, whereas the V1 model achieves different performances at different speed combinations. For example, in the V1 model, the speed combinations of $1+3$ ppF and $1+2+3$ ppF provide more accurate ARRs than the combination of $1+2$ ppF on the Weizmann database with the corresponding frame length, but these two combinations, respectively, provide the least accurate ARRs on KTH(s3) and KTH(s4). This indicates that the V1 model is easily influenced by the moving speed of human action, while the V1-MT model maintains its robustness. This is consistent with neurophysiological research on the visual cortex: V1 neurons can only perceive motion information in a very limited area due to a small cRF, while MT neurons with larger cRFs can effectively integrate motion information.

2) Comparison With Other Bioinspired Approaches: To evaluate the effectiveness of our V1-MT model, we compare its performance to other bioinspired approaches.

TABLE VI
PERFORMANCE COMPARISON WITH OTHER APPROACHES ON WEIZMANN AND KTH DATABASES

Approaches	ARR(%)		Year
	KTH	Weizmann	
Ours (V1-MT Model)	93.16	98.52	-
Multi-ch. Gabor Poly [6]	92.90	-	2014
Multi-ch. Gabor Sphe [6]	93.80	-	2014
Multi-ch. Gabor SOD [6]	94.80	-	2014
CGDLA (Class independent) [7]	92.72	98.92	2014
DL-SFA [9]	93.10	-	2014
3D CNN [11]	90.20	-	2013
SMT [5]	97.10	96.80	2013
RSS [4]	92.70	-	2013
SFA [3]	93.50	93.87	2012
DT [2]	94.20	-	2011

To guarantee fairness, the performance comparison between different approaches is made on the same databases. Experimental results are shown in Tables IV–VI.

First, we compare the performance on the Weizmann with bioinspired approaches described in [15], [19], and [20]. As seen in Table IV, our V1-MT model approach achieves 98.52% higher performance than the best results of Escobar's and Jhuang's approaches. Although Escobar's and Jhuang's approaches also achieve the acceptable performance (96.47% in [20] and 97% in [15]), they have defects. For example, the asymmetric and the anisotropic surround operation for information processing in [20] increases computational complexity greatly, and S2 maps in [15] also need heavy computational cost.

Similarly, performance comparisons on the KTH database are listed in Table V. The average performance (93.2%) of our approach is superior to Escobar's (89.6%) and Jhuang's (91.6%). Moreover, the ARR of our approach in each KTH condition is also higher than that of Escobar's and Jhuang's approaches, especially (s2) (outdoors with scale variation) and (s4) (indoors with lighting variation). This indicates that our approach effectively overcomes the influence of scale (s2) and lighting variation (s4). It is also notable that the results of our approach are more reliable than those in [15] and [19], because they are obtained using 84 training sets and not only 5 trials as in [15] and [19].

3) Comparison With the State-of-the-Art Approaches: We also compare the proposed approach on the Weizmann and KTH databases with the state-of-the-art approaches. To ensure consistency and comparability, we list some representative studies in terms of the same databases in Table VI. These approaches are the newest and highest achieving solutions in human motion or action recognition. From Table VI, we can see that the performance of our approach demonstrated here

TABLE VII
AVERAGE RECOGNITION RATES ON UCF SPORTS ACTION DATABASE

Approaches	ARR(%)	Year
Ours (V1-MT Model)	88.6	-
Ours (V1 Model)	87.3	-
Multi-ch. Gabor Poly [6]	85.2	2014
Multi-ch. Gabor Sphe [6]	86.3	2014
Multi-ch. Gabor SOD [6]	87.5	2014
DL-SFA [9]	86.6	2014
SFA (One Layer) [9]	79.8	2014
Multi-ch. Gabor+HOG3D [53]	85.6	2013
ST-SIFT + HOG3D [54]	80.5	2012
Hierarchical ISA [55]	80.5	2011

is comparable with that of other approaches. Although the performance of our approach is slightly lower than some, such as [2], [3], [5], and [6] on KTH, and [7] on Weizmann, we uses only four directions for motion information processing in each channel of our model. Furthermore, because of the interest here in the bioinspired computational model for recognition tasks, we do not consider high-level statistics of spike trains for feature extraction.

4) Experimental Results on UCF Sports Action Database: Experiments on both KTH and Weizmann databases already validate the effectiveness of our approach for simple human action recognition. However, we further conduct experiments on recognizing sport actions in the UCF Sports action database. The UCF Sports action database consists of ten different human actions: *swinging*, *diving*, *kicking*, *weight-lifting*, *horse riding*, *running*, *skateboarding*, *swinging* (at the high bar), *golf swinging*, and *walking*. This database contains 150 video sequences that show large intraclass variabilities.

We follow the aforementioned protocol to use our V1 and V1-MT models for action recognition. The computed classification accuracies are listed in Table VII. Results show that the recognition performance of our V1-MT model (88.6%) is higher than that of the V1 model (87.3%), once again confirming that the features further processed by the MT model benefits action recognition. Second, our approach with the V1-MT model outperforms other state-of-the-art approaches. For example, the performance of the DL-SFA method [9] is 86.6%, which is approximately equivalent to that of our V1 model approach. Finally, based on the above experimental data on all databases, we conclude that our V1-MT model achieves consistent good performance not only for simple actions, but also for complex actions.

5) Computational Cost: Our approach runs in MATLAB and takes a little over 30 s per test sequence (on KTH, 50 frames, Xeon 2.4 GHz). This shows that the computational cost of our model is relatively high, but lower than Jhuang's and Escobar's approaches. This is a common shortcoming of bioinspired models for human action recognition. However, the system implemented with our model has a hierarchical architecture. In each layer, the basic computational operations are similar, mainly focusing on (5), (10), and (13), which contain many 3-D convolution operations. These operations can be implemented with parallel computation. Therefore, the GPU (GTX TITAN Z) is used to accelerate calculations for real-time tasks (frame rate at 30 frames/s). However, the system is quasi-real time, because the feature extraction from a subsequence of approximately 50 frames induces a delay.

V. CONCLUSION

In this paper, we propose a bioinspired computational model to recognize human actions by stimulating the neural networks of visual areas V1 and MT. A neuron as a spiking entity in both visual cortical areas V1 and MT is first modeled with a simplified IF model. Each spiking neuron converts spatiotemporal information into spikes. A neuron in the V1-MT model detects and processes spatiotemporal information by modeling its speed- and direction-tuned properties and surround suppressive property.

Finally, we analyze spike trains from our V1-MT model and define a mean motion map based on the mean firing rates of neurons to represent a human action, and then use these maps to recognize different actions in real videos with a classical SVM classifier.

To validate our model, we conduct extensive experiments. Two sets of experiments on the KTH and Weizmann human action databases suggest that our model can extract motion pattern effectively and outperform other bioinspired models in action recognition tasks. Furthermore, two sets of experiments on the KTH and UFC Sport action databases demonstrate that our model is competitive with the state-of-the-art methods. The good performance obtained with our model also reinforces the representability of the mean motion maps for human action.

Moreover, our experimental results show that despite stable recognition performance given by the proposed model based on the feedforward architecture, performance improvements for human actions in complex scenes are limited. This indicates that theory based on feedback architecture is required for complex human action recognition, which is consistent with neuroscience findings [46]. In a future work, we plan to structure a feedback path from MT to V1 to achieve a better representation of motion information.

Compared with the human visual system, the model herein is relatively simple, only considering motion information processed by the dorsal pathway and neglecting form information processed by the ventral pathway. Previous studies have shown that biological motion recognition depends on the combination of information in both pathways. Therefore, integrating form information integrated into the bioinspired model for action recognition is a direction for future studies. Furthermore, because an SNN has the advantages of easy realization and low power consumption, designing a neuromorphic device with parallel computation to address computational cost is another task for future research.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, Jun. 2010.
[2] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. CVPR*, Jun. 2011, pp. 3169–3176.
[3] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 436–450, Mar. 2012.
[4] F. Shi, E. Petriu, and R. Laganière, "Sampling strategies for real-time action recognition," in *Proc. CVPR*, Jun. 2013, pp. 2595–2602.
[5] I. Jargalsaikhan, S. Little, C. Direkoglu, and N. E. O'Connor, "Action recognition based on sparse motion trajectories," in *Proc. ICIP*, Sep. 2013, pp. 3982–3985.

[6] H. Zhang, W. Zhou, C. Reardon, and L. E. Parker, "Simplex-based 3D spatio-temporal feature description for action recognition," in *Proc. CVPR*, 2014, pp. 2067–2074.

[7] G. Luo, S. Yang, G. Tian, C. Yuan, W. Hu, and S. J. Maybank, "Learning human actions by combining global dynamics and local appearance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2466–2482, Dec. 2014.

[8] R. Blake and M. Shiffrar, "Perception of human motion," *Annu. Rev. Psychol.*, vol. 58, pp. 47–73, Jan. 2007.

[9] L. Sun, K. Jia, T.-H. Chan, Y. Fang, G. Wang, and S. Yan, "DL-SFA: Deeply-learned slow feature analysis for action recognition," in *Proc. CVPR*, Jun. 2014, pp. 2625–2632.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 412–424, Mar. 2012.

[11] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[12] L. G. Ungerleider and M. Mishkin, "Two cortical visual systems," in *Analysis of Visual Behavior*, D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, Eds. Cambridge, MA, USA: MIT Press, 1982, pp. 549–586.

[13] M. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movements," *Nature Rev. Neurosci.*, vol. 4, pp. 179–192, Mar. 2003.

[14] R. Sigala, T. Serre, T. Poggio, and M. Giese, "Learning features of intermediate complexity for the recognition of biological motion," Lecture Notes in Computer Science, vol. 3696, pp. 241–246, 2005.

[15] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *Proc. ICCV*, 2007, pp. 1–8.

[16] M.-J. Escobar and P. Kornprobst, "Action recognition with a bio-inspired feedforward motion processing model," in *Proc. ECCV*, 2008, pp. 186–199.

[17] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require?" in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[18] H. Ning, T. X. Han, D. B. Walther, M. Liu, and T. S. Huang, "Hierarchical space-time model enabling efficient search for human actions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 6, pp. 808–820, Jun. 2009.

[19] M.-J. Escobar, G. S. Masson, T. Vieville, and P. Kornprobst, "Action recognition using a bio-inspired feedforward spiking network," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 284–301, May 2009.

[20] M.-J. Escobar and P. Kornprobst, "Action recognition via bio-inspired features: The richness of center–surround interaction," *Comput. Vis. Image Understand.*, vol. 116, no. 5, pp. 593–605, May 2012.

[21] Z. Zhang, Q. Wu, Z. Zhuo, X. Wang, and L. Huang, "Wavelet transform and texture recognition based on spiking neural network for visual images," *Neurocomputing*, vol. 151, no. 3, pp. 985–995, Mar. 2015.

[22] N. K. Kasabov, "NeuCube: A spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data," *Neural Netw.*, vol. 52, pp. 62–76, Apr. 2014.

[23] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 $\times$ 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.

[24] F. Perez-Peña *et al.*, "Neuro-inspired spike-based motion: From dynamic vision sensor to robot motor open-loop control through spike-VITE," *Sensors*, vol. 13, no. 11, pp. 15805–15832, Nov. 2013.

[25] G.-E. La Cara and M. Ursino, "Direction selectivity of simple cells in the primary visual cortex: Comparison of two alternative mathematical models. II: Velocity tuning and response to moving bars," *Comput. Biol. Med.*, vol. 37, no. 5, pp. 598–610, May 2007.

[26] J. McLean and L. A. Palmer, "Contribution of linear spatiotemporal receptive field structure to velocity selectivity of simple cells in area 17 of cat," *Vis. Res.*, vol. 29, no. 6, pp. 675–679, 1989.

[27] H. E. Jones, K. L. Grieve, W. Wang, and A. M. Sillito, "Surround suppression in primate V1," *J. Neurophysiol.*, vol. 86, no. 4, pp. 2011–2028, 2001.

[28] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Amer. A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[29] S. E. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on Gabor filters," *IEEE Trans. Image Process.*, vol. 11, no. 10, pp. 1160–1167, Oct. 2002.

[30] J. Lei and C. Li, "Spatiotemporal organization of simple-cell receptive fields in area 18 of cat's cortex," *Sci. China Ser. C, Life Sci.*, vol. 41, no. 1, pp. 1–8, Feb. 1998.

[31] A. Casile and M. A. Giese, "Critical features for the recognition of biological motion," *J. Vis.*, vol. 5, no. 6, pp. 348–360, 2005.

[32] S. Raiguel, M. M. van Hulle, D.-K. Xiao, V. L. Marcar, and G. A. Orban, "Shape and spatial distribution of receptive fields and antagonistic motion surrounds in the middle temporal area (V5) of the macaque," *Eur. J. Neurosci.*, vol. 7, no. 10, pp. 2064–2082, Oct. 1995.

[33] D. J. Wielaard, M. Shelley, D. McLaughlin, and R. Shapley, "How simple cells are made in a nonlinear network model of the visual cortex," *J. Neurosci.*, vol. 21, no. 14, pp. 5203–5211, Jul. 2001.

[34] A. Destexhe, M. Rudolph, and D. Paré, "The high-conductance state of neocortical neurons *in vivo*," *Nature Rev. Neurosci.*, vol. 4, pp. 739–751, Sep. 2003.

[35] E. M. Izhikevich, "Which model to use for cortical spiking neurons?" *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1063–1070, Sep. 2004.

[36] P. Kruizinga and N. Petkov, "Computational model of dot-pattern selective cells," *Biol. Cybern.*, vol. 83, no. 4, pp. 313–325, Sep. 2000.

[37] N. Petkov and E. Subramanian, "Motion detection, noise reduction, texture suppression, and contour enhancement by spatiotemporal Gabor filters with surround inhibition," *Biol. Cybern.*, vol. 97, no. 5, pp. 423–439, Dec. 2007.

[38] G. Azzopardi and N. Petkov, "A CORF computational model of a simple cell that relies on LGN input outperforms the Gabor function model," *Biol. Cybern.*, vol. 106, no. 3, pp. 177–189, Mar. 2012.

[39] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, 1962.

[40] K. Sakai and S. Tanaka, "Spatial pooling in the second-order spatial structure of cortical complex cells," *Vis. Res.*, vol. 40, no. 7, pp. 855–871, Mar. 2000.

[41] I. Lampl, M. Riesenhuber, T. Poggio, and D. Ferster, "The MAX operation in cells in the cat visual cortex," *Soc. Neurosci. Abstracts*, vol. 619, p. 30, 2001.

[42] P. O. Hoyer and A. Hyvärinen, "A multi-layer sparse coding network learns contour coding from natural images," *Vis. Res.*, vol. 42, no. 12, pp. 1593–1605, Jun. 2002.

[43] A. S. Safford, E. A. Hussey, R. Parasuraman, and J. C. Thompson, "Object-based attentional modulation of biological motion processing: Spatiotemporal dynamics using functional magnetic resonance imaging and electroencephalography," *J. Neurosci.*, vol. 30, no. 27, pp. 9064–9073, 2010.

[44] Z. Gao, J. Zeng, and H. Liu, "A biologically-inspired model for dynamic saliency detection," in *Proc. IEEE Conf. Multisensor Fusion Inf. Integr. Intell. Syst. (MFI)*, Sep. 2014, pp. 1–7.

[45] S. J. Thorpe, "Spike arrival times: A highly efficient coding scheme for neural networks," in *Parallel Processing in Neural Systems and Computers*. Amsterdam: Elsevier, 1990, pp. 91–94.

[46] J. Biederlack, M. Castelo-Branco, S. Neuenschwander, D. W. Wheeler, W. Singer, and D. Nikolić, "Brightness induction: Rate enhancement and neuronal synchronization as complementary codes," *Neuron*, vol. 52, no. 6, pp. 1073–1083, Dec. 2006.

[47] D. H. Perkel and T. H. Bullock, "Neural coding," *Neurosci. Res. Program Bull.*, vol. 6, no. 3, pp. 221–348, 1968.

[48] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development," *J. Neurophysiol.*, vol. 69, no. 4, pp. 1091–1117, 1993.

[49] B. Zeng, R. Li, and M. L. Liou, "Optimization of fast block motion estimation algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 6, pp. 833–844, Dec. 1997.

[50] V. Mante and M. Carandini, "Mapping of stimulus energy in primary visual cortex," *J. Neurophysiol.*, vol. 94, pp. 788–798, Mar. 2005.

[51] P. Bayerl and H. Neumann, "A fast biologically inspired algorithm for recurrent motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 246–260, Feb. 2007.

[52] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vis. Res.*, vol. 38, no. 5, pp. 743–761, Mar. 1998.

[53] I. Everts, J. C. van Gemert, and T. Gevers, "Evaluation of color STIPs for human action recognition," in *Proc. CVPR*, 2013, pp. 2850–2857.

[54] M. Al Ghamdi, L. Zhang, and Y. Gotoh, "Spatio-temporal SIFT and its application to human action classification," in *Proc. ECCV*, 2012, pp. 301–310.

[55] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. CVPR*, 2011, pp. 3361–3368.

Haihua Liu received the Ph.D. degree in computer system architecture from the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, in 2006.

He is currently a Professor with the College of Biomedical Engineering, South Central University for Nationalities, Wuhan. He has authored over 60 research articles in domestic and foreign academic journals, such as the *Information Science*, the *Pattern Recognition*, the *Neurocomputing*, the *Magnetic Resonance Imaging*, and the *Science China: Information Sciences*. His current research interests include computer vision, medical image analysis, pattern recognition, and cognitive computation.

Na Shu is currently pursuing the M.S. degree in biomedical engineering from the South-central University for Nationalities, Wuhan, China.

Her current research interests include computer vision and computer models of vision system for action recognition.

Qiling Tang received the Ph.D. degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology, Wuhan, China, in 2007.

He was a Post-Doctoral Researcher in computer science with the Huazhong University of Science and Technology. He is currently an Associate Professor with the College of Biomedical Engineering, South Central University for Nationalities, Wuhan. His current research interests include computer vision, medical image analysis, and computational modeling of human vision.

Wensheng Zhang received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, in 2000.

He joined the Institute of Software, CAS, in 2001. He is currently a Professor of Machine Learning and Data Mining and the Director of the Research and Development Department, Institute of Automation, CAS. His current research interests include computer vision, pattern recognition, artificial intelligence, and computer–human interaction.