2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A Bio-inspired Approach Modeling Spiking Neural Networks of Visual Cortex for Human Action Recognition

Na Shu, Q Tang and Haihua Liu*

Abstract— **Human visual system is an effective recognition one. Based on information processing mechanism of visual cortex, a bio-inspired approach for the human action recognition from video sequences is proposed in this paper. The approach gives a hierarchical architecture of the feedforward spiking neural network modeling two visual cortical areas: primary visual cortex (V1) and middle temporal area (MT), neurobiologically dedicated to motion processing. We augment the operator of motion information processing with center surround interaction to model the nonclassical receptive field inhibitory effect based on horizontal connection of spiking neurons in each cortical area. The weight function of lateral connection between V1 and MT areas is built based on a previous study that explained direction selectivity in MT area by a linear combination of normalized V1 direction-tuned signals. Moreover, we propose a three-dimensional (3D) Gabor filter to model the spatiotemporal direction and speed tuning properties of time-dependent receptive fields of the V1 cells. The conductance-driven integrate-and-fire (IF) neuron model is used to obtain spike trains generated by the spiking neurons in two cortical areas. Finally, in order to analyze spike trains, we consider a characteristic of the neural code: mean motion map based on the mean firing rates of neurons in MT, called action code, as feature vector representing human actions. The approach is carried out on the Weizmann and KTH action database. Experimental results show that our approach has higher recognition performance and computational efficiency than other bio-inspired ones.**

I. Introduction

Visual scene understanding is of interest in many computer vision applications such as automated surveillance, elderly behavior monitoring and human-computer interaction. Recognizing human actions performed by an individual from videos is a fundamental task in this paper. During the past decade an extensive amount of research on human action recognition has been carried out with the goal to create a robust system, a variety of different tools and techniques have been employed. But it still remains a challenging problem due to the large variations in human appearance, posture and body size within the same class. It also suffers from various factors such as cluttered background, occlusion, camera movement and illumination change. In recent years, with progress of our understanding of the brain mechanisms responsible for the action recognition [1], the various approaches based on vision are reported in literature [2], but it is questionable whether these approaches are suitable for more complex motions.

The visual cortex for processing sensory information contains two pathways: the ventral stream and the dorsal stream. The ventral stream which is usually thought of as dealing mainly with the shape information, contains the visual cortical primary visual cortex (V1), V2 and V4 areas, while the dorsal stream involved with the analysis of motion information, includes the visual cortical V1 and the middle temporal (MT) areas. In dorsal pathway, the external visual stimuli is first accepted into retina, then passes the lateral geniculate nuclei (LGN) and V1, finally arrives at MT, the posterior parietal cortex (PPC). Based on dual-channel theory of vision [3], the researchers do extensive research and propose different bio-inspired models. Firstly, Giese and Poggio [4] evaluated the ventral stream and the dorsal stream in biological motion recognition. Afterwards, using only the information of the dorsal stream, Sigala [5] proposed a biological motion recognition system using a neurally plausible memory trace learning rule. Later, Jhuang [6] proposed a feedforward hierarchical template matching model. This model consists of a hierarchical architecture of spatiotemporal feature detectors of increasing complexity: an input sequence is analyzed by an array of motion-direction sensitive units, thereby extracting spatial and temporal features of video objects and obtaining a better recognition result, but this approach requires heavy computation and lacks biological plausibility. Similarly, Escobar [7][8] simulated dorsal pathway to create a computational model of human action recognition, called V1_MT model based on feedforward spiking neural network, in which motion information was processed in the visual areas V1 and MT, modeling more complex visual characteristics. Although this approach satisfies biology plausibility, there is some problems solved, such as how to focus on human action, why not to use the properties of surround suppression. To solve these problems, we propose a new simplified bio-inspired approach by simulating the mechanism of motion information processing in the visual cortical V1 and MT areas in order to achieve rapid and accurate identification of human actions.

The rest of this paper is organized as follows. Section II describes spiking neural networks and spiking neuron model in details. Section III represents the bio-inspired approach based on spiking neural networks in the visual cortical V1 and MT areas. Feature extraction from the spike trains of neurons in MT and recognizing human action is performed in Section IV. We compare the performance of action recogni-

Na Shu, Q Tang and Haihua Liu are with the School of Biomedical Engineering, South-central University for Nationalities, Wuhan 430074, China; *Haihua Liu is also with the Key Laboratory of Cognitive Science of State Ethnic Affairs Commission, corresponding author (email: lhh@mail.scuec.edu.cn).

This work was supported by the National Natural Science Foundation of China under Grant No. 91320102 and 60972158.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

tion with proposed approach to other bio-inspired ones, such as Jhuang and Escobar, on the Weizmann and KTH database in Section V. In Section VI, the advantages and disadvantages of our approach are discussed and some perspectives are proposed.

II. Spiking Neural Networks

A. Spiking Neuron and Spike Train

The basic units of the visual nervous system are neurons. A typical neuron can be divided into three functionally distinct parts, called dendrites, soma and axon. Roughly speaking, the dendrites play the role of the 'input device' that collects signals from other neurons and transmits them to the soma. The soma is the 'central processing unit' that performs an important non-linear processing step. The output signal is taken over by the 'output device', the axon, which delivers the signal to other neurons.

The neuronal signals consist of short electrical pulses. The pulses, so-called action potentials or spikes, have an amplitude of about 100*mV* and typically a duration of 1-2*ms*. The form of the pulse does not change as the action potential propagates along the axon. A chain of action potentials emitted by a single neuron is called spike train. Since all spikes of a given neuron look alike, the form of the action potential does not carry any information. Rather, it is the number and the timing of spikes which matter. All spikes in spike train are typically treated as discrete events. Therefore, in order to describe a spike train, we only need to know the succession of emission times:

$$T_i = \{..., t_i^n, ...\}, t_i^1 < t_i^2 < ... < t_i^n < ... \qquad (1)$$

where t_i^n denotes the nth spike of the neuron i.

B. Spiking Neural Networks

Connection between any two neurons in the visual cortex is established by a synapse. A neuron sends a signal across a synapse. A single neuron in visual cortex often connects to more postsynaptic neurons. Many of its axonal branches end in the direct neighborhood of the neuron, but the axon can also stretch over several centimeters so as to reach to neurons in other areas of the brain. Based on these biological facts in visual cortex, an architecture of spiking neural network is proposed, shown in Fig. 1. The architecture proposed here is almost the similar as a traditional neural network, in which the neurons of the preceding layer fully connect to the neurons of the next layer. The first layer is composed by the receptive fields neurons, and the neurons of the previous layer will act as the receptive fields of the next layer [9]. A difference from the traditional neural networks is that all output neurons are laterally connected to each other by means of a strongly inhibitory synapse, which is biologically plausible [16].

Moreover, the network built temporally and spatially separates incoming spikes into different regions of the network, that in turn generate corresponding new spikes. Spatially,

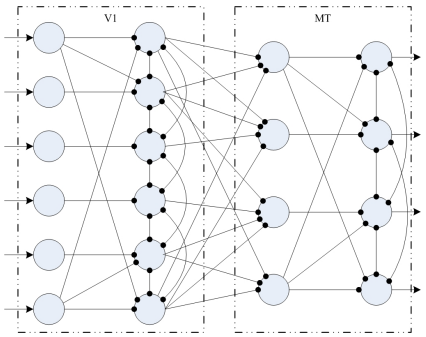

Fig. 1. V1 and MT spiking neural network.

though the distribution of receptor cells on the retina, which map into visual cortex, is like a Gaussian with a small variance in biological vision systems [10], we suppose that the distribution of the cells is uniform in V1 and MT areas, as shown Fig. 2. A black spot in the distribution map represents a single cell and the color circle indicates its classic receptive field (cRF).

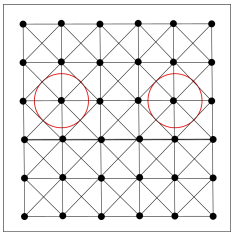

Fig. 2. Distribution schematics of spiking neurons in V1 and MT areas. A black dot is the position of a cell and the color circle indicates the range of the cRF. All cells are interconnected each other.

C. The Neuron Model

To describe neural activity as a simple homogeneous unit, many spiking neuron models have been proposed at a high level of abstraction in the literature. They differ by their biological plausibility and their computational efficiency [11]. For example, using biophysically accurate Hodgkin-Huxley-type models is computationally prohibitive [12], since we can simulate only a handful of neurons in real time. In contrast, using an integrate-and-fire model is computationally effective, but the model is unrealistically simple and incapable of producing rich spiking and bursting dynamics exhibited by cortical neurons [13]. Then, Wielaar [14] and Destexhe [15] respectively proposed a spiking neuron model by modeling a spiking neuron as a conductance-driven integrate-and-fire neuron, which depends on real cortical architecture and matches physiological data. In this paper, we use the conductance-driven integrate-and-fire model. Considering a neuron *i*, the integrate-and-fire equation is given by:

$$\begin{aligned}\frac{du_i(t)}{dt} =& G_i^{exc}(t)(E^{exc} - u_i(t)) + g^L(E^L - u_i(t)) \\ &+ G_i^{inh}(t)(E^{inh} - u_i(t)) + I_i(t).\end{aligned} \qquad (2)$$

This equation represents the spike emission process: the neuron *i* will emit a spike when the normalized membrane

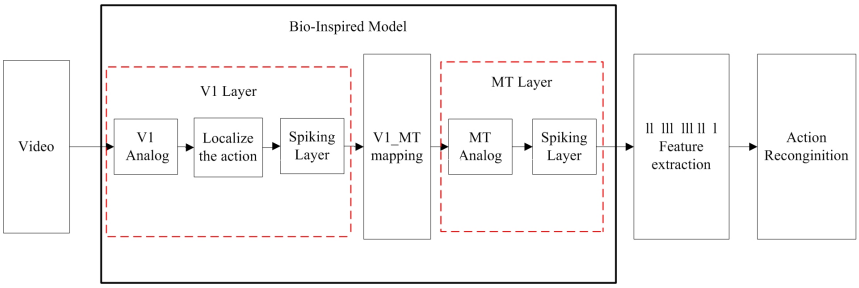

Fig. 3. The architecture of the proposed approach. It is consisted of four parts: V1 Layer model and MT Layer model, feature extraction and action recognition. In V1 Layer, motion energy is detected with non-linear combination of perception information of simple cells modeled by 3D Gabor spatiotemporal filters, and is transformed to the spike train by the spiking neuron model. In MT Layer, MT cells pool the information coming from the V1 cells according to the mapping connection between V1 and MT. The new spike trains are generated by MT spiking neurons. Features are extracted from spike trains of MT spiking neurons and used for action recognition with SVM classifier.

potential $u_i(t)$ reaches threshold u_0, then $u_i(t)$ return to resting potential E^L. $G_i^{exc}(t)$ is the excitatory conductance directly associated with the presynaptic neurons connected neuron i. G_i^{inh} is an inhibitory normalized conductance dependent on lateral connections or feedbacks from upper layers. For simplicity, we ignore $G_i^{exc}(t)$ and G_i^{inh}.

III. Bio-Inspired Approach

Visual information processing in the brain has evolved to the highly perfect stage. Numerous studies conducted visual motion analysis in the V1 and MT. The cells sensitive to motion for a specific speed and direction have been found in the two areas and the complex motion information processing needs collaboration of different cortical areas of the brain. In this paper, we propose a bio-inspired approach for human action recognition by modeling properties of V1 and MT areas in biological visual perception system. The approach consists of four parts, shown in Fig. 3, including V1 Layer, MT Layer, extracting feature vector and action recognition based on the spiking neural networks shown in Fig. 1.

A. Surround Interaction

Surround interactions are observed in different cortical regions such as V1 [16], middle temporal (MT/V5) [17]and lateral medical superior temporal(MST) [18], due to lateral connection between neurons. The response of such a neuron is suppressed when moving stimulis are presented in the region surrounding its cRF. In general, The surround suppression is modeled by a 2D difference of Gaussian (DoG) functions [19]. Our surround inhibition operator takes into account the influence of the surround at each spatial location and time instant. We suppose that the cRF of a simple cell is defined as the area by a 3D Gabor function $g_{v,\theta,\varphi}(x,y,t)$. The surround weighting function $w_{v,\theta}(x,y,t)$ to be zero inside the cRF and positive outside it and to decay with the distance to the cRF. Similar to DoG, we take as a surround weighting function with the half-wave-rectified difference of two concentric Gaussian envelopes, defined as follows:

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

where $\|\cdot\|_1$ denotes the L1 norm and

$$I_{v,\theta,k_1,k_2}(x,y,t) = |G_{v,\theta,k_2}(x,y,t) - G_{v,\theta,k_1}(x,y,t)|^+$$

here

$$\begin{aligned} G_{v,\theta,k}(x,y,t) =& \frac{\gamma}{2\pi(k\sigma)^2}\exp\left(\frac{-((\bar{x}+v_c t)^2+\gamma^2\bar{y}^2)}{2(k\sigma)^2}\right) \\ & \times \frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{(t-u_t)^2}{2\tau^2}\right)\times\varepsilon(t) \end{aligned} \tag{4}$$

where $\bar{x} = x\cos(\theta) + y\sin(\theta), \bar{y} = -xsin(\theta) + y\cos(\theta)$, $\varepsilon(t)$ is step function. The parameters v, θ are respectively the preferred speed and the preferred direction of motion, and k is a factor constant which determines the size of center surrounding.

For each point (x,y,t), an inhibition term $S_{v,\theta}(x,y,t)$ can be computed by weighted summation of the motion energy $E_{v,\theta}(x,y,t)$ in the surroundings of that point using the surround weighting function $w_{v,\theta}(x,y,t)$. It is performed by convolution:

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

We next use this inhibition term to define and compute a surround suppressed motion energy $\tilde{E}_{v,\theta}(x,y,t)$ as follows:

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

where the factor α controls the strength with which surround suppression is taken into account, and where $|\cdot|^+$ is an operator with half-wave rectification [20].

B. V1 Layer and Input Current

V1 corresponding to the first area involved on the visual processing in the brain, contains mainly two types of cells: simple cell and complex cell. The RF of simple cell is relatively small, and doesn't respond to a large area of diffuse light, but has strong reaction toward the strip stimulus with a certain position and width. Therefore simple cell is suitable for the detection of the contrast polarity of edges and strict to select the position and orientation of the edges. In contrast, the RF of complex cell is larger than simple cell, there are no obvious excitatory and inhibitory regions, the best stimulus is still the light bar with a certain position and width, but it isn't strict to the position of stimulus in the RF [21].

1) Simple and Complex Cell Model: To model the spatiotemporal properties of simple cell in V1, 3D Gabor spatial-temporal filter [20] is used to simulate its RF. The formula is given by:

$$\begin{aligned} g_{v,\theta,\varphi}(x,y,t) =& \frac{\gamma}{2\pi\sigma^2}\exp(\frac{-((\bar{x}+v_c t)^2+\gamma^2\bar{y}^2)}{2\sigma^2}) \\ & \times\cos(\frac{2\pi}{\lambda}(\bar{x}+vt)+\varphi) \\ & \times\frac{1}{\sqrt{2\pi\tau}}\exp(-\frac{(t-u_t)^2}{2\tau^2})\cdot\varepsilon(t). \end{aligned} \tag{7}$$

3D Gabor function consists of a spatial Gaussian function, a cosine function and a temporal Gaussian function.The parameter γ is the spatial aspect ratio of the spatial Gaussian function, σ determines the size of the receptive field and λ is the spatial period. The ratio σ/λ determines the spatial bandwidth and the number of excitatory and inhibitory stripe zones in the receptive field, set $\sigma/\lambda = 0.56$. φ is a parameter that determines the spatial symmetry of the function. The following relationship exists between the wavelength λ and the speed v.

$$\lambda = \lambda_0\sqrt{1+v^2}. \tag{8}$$

where the constant λ_0 is the spatiotemporal period of the filter. The Equ. (8) shows that the higher the preferred speed v is, the larger the RF of cell is.

The simple cell perceives the spatiotemporal information of a video can be simulated by the response of the 3D Gabor filter. The response $r_{v,\theta,\varphi}(x,y,t)$ of 3D Gabor filter with function $g_{v,\theta,\varphi}(x,y,t)$ to image sequence $l(x,y,t)$, is computed by convolution, as follows:

$$r_{v,\theta,\varphi}(x,y,t) = l(x,y,t) * g_{v,\theta,\varphi}(x,y,t). \tag{9}$$

Based on the theory proposed by Adelson and Bergen [22], we use non-linear combination of two mutually orthogonal simple cells to simulate the response of complex cells responding to the input stimuli. The formula is defined as follows:

$$E_{v,\theta}(x,y,t) = \sqrt{r^2_{v,\theta,0}(x,y,t) + r^2_{v,\theta,\pi/2}(x,y,t)}. \tag{10}$$

This quantity, called motion energy, can be used as a model of the response of a complex cell.

In this paper, the motion energy of the video images sensitive to the specific speed and direction is achieved with 3D Gabor filters and their nonlinear combination, and surround suppressed motion energy is also obtained with surround suppression. To characterize tuning properties of V1 cells, we consider N_v different speeds and N_θ directions. $N_v N_\theta$ levels of V1 cells are built. The V1 cells in each level is with the same speed and direction tuning.

Fig. 4 shows the results which we apply this idea of motion detection to an input video from KTH database, including spatiotemporal motion energy and surround suppressed motion energy filters at speed $v = 1ppF$ (pixel per Frame) and direction $\theta = 0$.

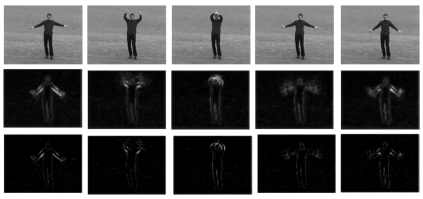

Fig. 4. Detection of motion energy. From the first to final row: the snapshots from a video sequence in KTH database, motion energy corresponding to snapshots, surround suppressed motion energy.

2) Focus on the action: In the real scene, if a person is moving across the visual field, our eyes follow the motion. This process, called smooth pursuit, is regulated by visual attention mechanism. Smooth pursuit is also an important psychological visual adjustment mechanism. Studies have shown that there is a strong correlation between the action recognition rate and the level of smooth pursuit velocity [23]. For example, Safford [30] reported that the action recognition performance in biological motion is highly modulated by attention: the best performance is reached when the moving subject is in the focus of attention. Furthermore, the role of attention mechanism is often based on initial perception of moving objects.

In this paper, we form initial visual perception with a series of local motion detector in V1 layer and use focusing on the action to simulate the visual attention mechanism. The objects of human action are detected based on the saliency maps proposed by Meur [31]. Due to scale variability of a object in a video and different videos, it introduces different number of cells to perceive it, resulting in different coding of motion information. Concerning scale-invariance, an action object is re-scaled to the same dimension, similar to freely adjustment of the eye focal length, shown in Fig. 5.

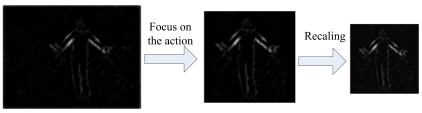

Fig. 5. Focus on the action. The first is sptiotemporal information perceived for a frame image, the second is one of action object localized, action object re-scaled with the same aspect ratio is shown in the last.

3) Input current: Objective of the spiking neuron model described above is transforming the analogous response of the cell defined in Equ. (6) to the spiking response so as to characterize the activity of a neuron. From Equ. (2), the activity of a neuron is determined by external input current $I_i(t)$ of the spiking neuron and the membrane potential threshold.

First, let us consider input of a spiking neuron *i* in V1 whose center is located in (x_i, y_i). Its external input current $I_i(t)$ associates with the analogous response of V1 cell defined in Equ. (6). However, the activation of the cell is in range of cRF. The computational operator over RF in a layer (e.g. same preferred motion direction and speed) is needed [32]. Thus, the external *input current* $I_i(t)$ of the *i*th neuron

is modeled in Equ. (11) as follows:

$$I_i(t) = K_{exc} \max_i \{\widetilde{E}_{v,(\theta)}(\mathbf{x}, t)\} \tag{11}$$

where K_{exc} is an amplification factor, $\widetilde{E}_{v,(\theta)}(\mathbf{x}, t)$ refers to V1 cell response defined in Equ. (6) with $k_2 = 4$ and $k_1 = 1$, and $\max_i$ is a operator of local maximum [33].

C. MT Layer and Input Current

The most important cortical area in the brain that processes visual motion information is the MT area, which is responsible for integrating and segregating visual motion information projected by the visual cortex V1, and forms visual motion perception projection. How to extract and segregate motion visual signals in the MT area is a mystery so far. Achieving complete information processing of MT area has exceeded the scope of this paper. This paper aims to simulate partial properties of V1 and MT neurons and realizes rapidly and highly effective human action recognition. When simulating the characteristics of MT neurons, we use the similar information processing way in V1. It contains computing motion energy using spike trains from V1 layer, obtaining surround suppressed motion energy with surround interaction, finally converting those into spike trains by spiking neuron model in MT layer.

1) The mapping relationship between V1 and MT: In order to build spike transmission from V1 to MT, what we first need study is the mapping connection and afferent projections from V1 cells to MT cells. In a general sense, a single neuron in MT often connects to more presynaptic neurons in V1. Each MT cell has a RF made from the convergence of afferent V1 complex cells. The afferent V1 will be excitatory or inhibitory depending on the characteristic and shape of the corresponding MT RFs. The evidence of neurophysiological research indicates that RF size of a MT cell is about 4-6 times bigger than the V1 RF. Therefore, we take only into account the connection of a MT neuron with 4 V1 neurons in the region of a MT cell RF. The connections of other V1 neurons outside the MT cell RF to the MT neuron are neglected because these can act on the MT cell indirectly by the horizontal connection in MT area.

Fig. 6 shows mapping connection between V1 neurons and MT ones. A MT neuron receives spike inputs from four V1 neurons in different directions. For a MT cell i in (x_i, y_i) space, the motion energy perceived by it depends on connection weights as a weight factor and spike trains of V1 neurons in the RF of MT cells. It is defined as following:

$$E_i^{MT}(t) = \max(0, \sum_{j \in cRF^{MT}} w_{ij} \eta_j^{V1}(t)). \tag{12}$$

where $\eta_j^{V1}(t)$ represents the spike trains of V1 neurons, cRF^{MT} is the RF of MT cell, j is the jth V1 neuron in MT cRF, and w_{ij} is the weight factor between MT neuron i and V1 neuron j.

Similar to V1 architecture, each layer is built with MT neurons of the same characteristics, same speed and direction tuning. The group of V1 neurons connected with a MT neuron and their respective connection weights depend on the tuning values desired for the MT neuron. The weight associated to the connection between presynaptic neuron j and post-synaptic neuron i is proportional to the angle ϕ_{ij} between the two preferred motion direction-selectivity (see Fig. 7). The connection weight w_{ij} between the jth V1 neuron and the ith MT neuron is given by

$$w_{ij} = \begin{cases} k_c w_d(\mathbf{x}_i - \mathbf{x}_j)\cos(\varphi_{ij}), & \text{if } \varphi_{ij} < \pi/2; \\ -k_c w_d(\mathbf{x}_i - \mathbf{x}_j)\cos(\varphi_{ij}), & \text{if } \varphi_{ij} > \pi/2; \\ 0, & \text{if } \varphi_{ij} = \pi/2; \end{cases} \tag{13}$$

where k_c is an amplification factor. $\varphi_{ij} = |\varphi_i - \varphi_j|$ is the absolute angle between the preferred ith MT neuron direction and the preferred jth V1 neuron direction. $w_d(\cdot)$ is the weight associated to the difference between the center position of MT neuron and V1 neuron. Specific description is shown in Equ. (14).

$$w_d = \frac{\exp(-(\mathbf{x}_i - \mathbf{x}_j)^2 / 2\sigma^2)}{\sigma\sqrt{2\pi}}. \tag{14}$$

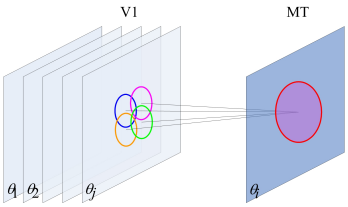

Fig. 6. The mapping connection between V1 and MT. Four V1 neurons are connected to a MT neuron because a RF of MT cell covers four RFs of V1 cells.

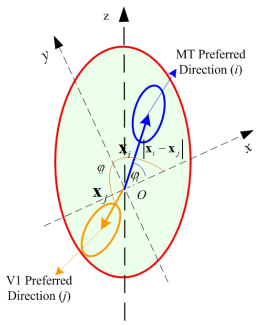

Fig. 7. The connection weights between V1 and MT neurons are modulated by the cosine of the angle φ_{ij} between the preferred direction of ith MT neuron and the preferred direction of jth V1 neuron.

2) Input Current and Spike Train: Due to surround suppression also observed in MT, surround suppressed motion energy $\tilde{E}_{v,\theta}^{MT}(x, y, t)$ is obtained by Equ. (6). However, parameter σ in Equ. (4) need change because RF size of a MT cell is 4 times bigger than the V1 RF.

Because the response of the MT cell, $\tilde{E}_{v,\theta}^{MT}(x, y, t)$, is also analogous. we need transform the analogous response to a spiking response. Similarly, the conductance-driven integrate-and-fire neuron model described in Equ. (2) is used again. The external input current $I_i^{MT}(t)$ of the ith MT cell

in Equ. (2) as the analog response is computed with Equ. (11). Finally, spike trains in MT layers are generated. Fig. 8 shows a example that spike maps output from the V1 and MT neurons on a action in KTH database. Fig. 9 shows the spike trains of all MT neurons corresponding to the action.

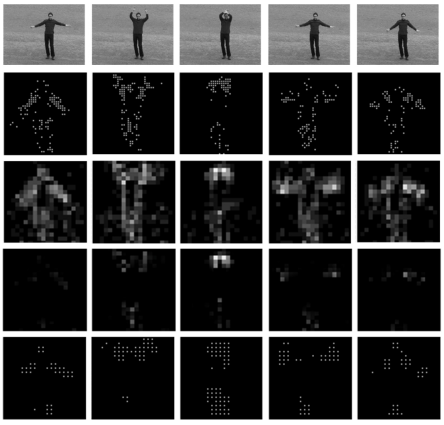

Fig. 8. The spike maps of V1 and MT neurons on an action. *First row* shows the snapshots from a sequence in KTH database. *Second row* shows the spike maps in a V1 layer with 0° orientation at 1*ppF* speed. Motion energy of MT cells is shown in the *third row* and the corresponding surround suppression motion energy is given in *fourth row*. *Final row* shows the spike maps in a MT layer with 0° orientation at 1*ppF* speed. Note that a light dot in the spike maps indicate a spike generated by the neuron in this position.

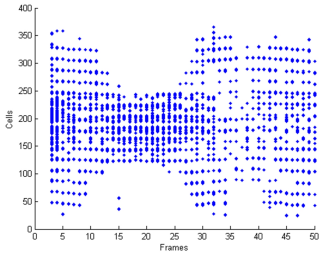

Fig. 9. The spike trains of all MT neurons corresponding to the action. Raster plots are obtained considering the 400 MT cells with a given orientation and speed for the actions shown in Fig. 8.

IV. Action Recognition

Researchers have proposed different analytical methods on how neural code is formed, methods include rank order coding [24], synchronization and correlations [25] [26] [27] [28], mean firing rate [29], etc. Mean firing rate is considered as the most general and effective method. Let us consider a spiking neuron *i*, mean firing rate is computed in Equ. (15), defined as following:

$$r_i(t, \triangle t) = \frac{n_i(t - \triangle t, t)}{\triangle t}. \tag{15}$$

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

Using mean firing rate as a feature vector has many advantages: It doesn't depend on the sequence length and its starting point. It represents the process of cell excitability over time and distinguishes different actions in a video sequence. Therefore, it can be taken as a motion feature of a action. But this time-related feature vector is not conducive to characteristic expression. In order to construct the feature vector not depending on the time, the paper will average over mean firing rate as the final feature vector $H_I(\cdot)$.

$$h_i = \frac{\sum_{i=1}^{T} r_i(t, \triangle t)}{T}. \tag{16}$$

$$H_I = \{h_i\}_{i=1,\ldots,N\times N_L}. \tag{17}$$

Where N is the number of MT cells per layer, N_L is the number of layers with same characteristics.

The final step in the system, feature vectors extracted are sent to the classifier for classification. In this article, in order to reduce calculation, the SVM is used as classifier with no linking to biology.

V. Experimental Results and Analysis

A. Database and Settings

In order to test the effectiveness of the approach, we use a public Weizmann and KTH human action database as experimental subjects. We follow a similar experimental protocol than Jhuang's [6].

The Weizmann human action database consists of 9 different subjects performing 9 different actions: bend, jack, jump, pjump, run, side, walk, wave1 and wave2. The training set is built considering actions of 6 different subjects (6 subjects $\times$ 9 actions=54 videos). The testing set is built with the remaining 3 subjects (3 subjects $\times$ 9 actions=27 videos). The KTH human action database contains six types of human actions: walking, jogging, running, boxing, handwaving and handclapping. These actions are performed several times by twenty-five subjects in four different conditions: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors with lighting variation (s4). For training, we use the actions from 9 random subjects (9 $\times$ 6=54 videos), the actions of the remaining 16 subjects (16 $\times$ 6=96 videos) are then used for testing. Unlike Jhuang et al, we run all the possible training sets (84) and not only 5 random trials. Since the KTH database contains more blank frames and a lot of useless information, in order to reduce the impact of useless information on the recognition results, so we remove blank frames. Moreover, the number of frames per sequence in the database are very long, in order to reduce computing time, we design the maximum frame f_s to control the input video frames and fix f_s=50.

B. Parameter settings

In order to get better recognition results, the approach must set the corresponding parameters. First, for 3D Gabor filter, the parameters are listed in Table I. Then, based on the biological data, the parameters about IF neuron model are set in Table II.

TABLE I

PARAMETERS FOR 3D GABOR

λ_0	γ	θ	u_t	τ	φ
1	0.5	π/n	1.75	2.75	$0, \pi/2$

TABLE II

PARAMETERS FOR IF NEURON MODEL

	K_{exc}	E^L	u_0	g^L
V1	200	-70	-50	0.1
MT	100	-70	-50	0.1

General V1 and MT settings are shown in Table III, V1 has a total of 12 layers, formed by 3 speeds (1, 2 ,3 ppF) and 4 orientations, giving a total of 1600 cells per layer. MT contains the same speeds and orientations with V1, but cell distribution is sparse and the total cells per layer are a quarter of V1's. We use publicly binary image in Weizmann database to focus on the action. In KTH database, there is not publicly binary image, we use the approach based on visual attention mechanism to focus on the action, which is helpful to remove interference from background information and maximize the extraction of the interesting object. After the original video streams are centered, they are resized in 80×80 pixels, forming new sequences.

TABLE III

PARAMETER USED FOR V1 AND MT LAYERS

	V1	MT
Number of prefered speeds	3	3
Number of prefered orientations	4	4
Size of receptive field	σ	2σ
Number cells per layer	800	200

C. Recognition Performance

In order to evaluate the effectiveness of our approach proposed in this paper, we compare the performance of our approach with others. Experimental results are shown in Table IV and Table IV.

TABLE IV

PERFORMANCE COMPARISON ON WEIZMANN DATABASE

	ARR(%)	std(%)	trials
Ours	98.63	2.74	84
Escobar (Mean) [7]	92.68	4.62	84
Escobar (Synchrony) [7]	92.81	5.15	84
Escobar (TD) [8]	96.34	0.72	84
Escobar (SKL) [8]	96.47	0.81	84
Jhuang (StC2 dense)	91.10	5.90	5
Jhuang (StC2 sparse)	97.00	3.00	5

To be more meaningful and fair, the performance comparison of different approaches is made on the same database. Firstly, we compare our approach with bio-inspired ones described by Escobar in [7] and [8], and Jhuang in [6]. As we can see in Table IV, the best recognition rate of 96.47% in [8] is achieved using symmetric Kullback-Liebler divergence on Weizmann database over 84 trails, while the best performance of [6] on Weizmann database achieves 97% using SVM even if only random 5 training sets are used. However, our approach achieves the performance of 98.63%, which is higher than Escobar's and Jhuang's.

TABLE VI

PERFORMANCE COMPARISON WITH OTHERS ON KTH DATABASE

	ARR(%)	year
Ours	92.30	-
Shi [34]	92.70	2013
Zhang&Tao [35]	93.50	2012
Wang[36]	94.20	2011

From Table V, we can find the average performance on KTH database is superior to other approaches. Although the result of our approach under s1 and s4 on KTH database is slightly lower than Jhuang approach, but is much higher under s2 and s3. It is seen that our approach can overcome the effects coming from scale variation and lighting variation. It is worth noting that the recognition result of Jhuang method on KTH database is obtained over random five trials, but our result using the average of random the 84 results, is more stable than Jhuang's.

Finally, we also compare the performance of our approach with others except bio-inspired ones on KTH database, shown in Table VI, From it, we can see that our result is low but comparable to others with respect to recognition rates.

VI. Conclusion

We propose a biological model of motion processing to recognize human actions. The difference between our approach and Escobar's is mainly reflected in the following aspects: Firstly, in the V1 layer, Escobar used a combination of spatial and temporal filter to simulate simple cells, which cause the omission of unilateral information, such as spatial information or temporal information. Instead, we simulate simple cells with 3D Gabor filter, which can detect the complete spatial and temporal information. Furthermore, Escobar takes localization of action objects as a preprocessing for video with lack of biological plausible. Biological studies show vision carries on visual attention and depth perception. In our model, the initial perception obtained can be used to localize the action objects by simulating visual attention mechanism, which is biological plausible. Finally, Escobar considered only surround inhibition in the MT stage. However, the study found that V1 and MT stage existed surround inhibition, the proposed approach with both introducing surround inhibition in V1 and MT, more accurately restores visual information processing and improves action recognition performance.

Compared to the human visual system, the approach herein is relatively simple, only considering motion information processing of the dorsal pathway. Early studies have shown that the biological motion recognition depends on the com-

TABLE V

PERFORMANCE COMPARISON WITH OTHER BIO-INSPIRED METHODS ON KTH DATABASE

$ARR(\%)/std(\%)$	S1	S2	S3	S4	Avg	trials
Ours	$95.3/2.8$	$89.5/3.5$	$89.2/4.0$	$95.2/3.5$	$92.3/3.1$	84
Escobar [8]	$83.1/2.0$	-	$69.8/2.8$	$83.8/1.9$	$79.8/2.2$	100
Escobar [8]	$92.0/0.01$	-	$84.4/1.22$	$92.4/0.01$	$89.6/0.4$	5
Jhuang (StC2 dense) [6]	$89.8/3.1$	$81.3/4.2$	$85.0/5.3$	$93.2/1.9$	$87.3/3.6$	5
Jhuang (StC2 sparse) [6]	$96.0/2.1$	$86.1/4.6$	$88.7/3.2$	$95.7/2.1$	$91.6/3.0$	5

bination of the dorsal pathway and the ventral pathway, the ventral stream of the visual cortex, involved with the analysis of shape may also be important for the recognition of motion. The shape information of the ventral pathway added the system will become one of our future studies.

References

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

[2] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976-990, 2010.

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

[4] M. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movement and action," *Nature Reviews Neuroscience*, vol. 4, pp. 179-192, 2003.

[5] R. Sigala, T. Serre, T. Poggio and M. Giese, "Learning features of intermediate complexity for the recognition of biological motion," *In LNCS*, vol. 3696, pp. 241-246, ICAN 2005, Berlin:Springer.

[6] H. Jhuang, T. Serre, L. Wolf and T. Poggio, "A biologically inspired system for action recognition," *In Proceedings of the 11th international conference on computer vision*, pp. 1-8, 2007.

[7] M. J. Escobar and P. Kornprobst, "Action recognition with a bio-inspired feedforward motion processing model," *Proc.10th European Conference on Computer Vision*, vol. 4, pp. 186-199, 2008.

[8] M. J. Escobar and P. Kornprobst, "Action recognition via bio-inspired features: The richness of center-surround interaction," *Computer Vision and Image Understanding*, vol. 116, pp. 593-605, 2012.

[9] S. M. Bohte, H. L. Poutr and J. N. Kok, "Unsupervised clustering with spiking neurons by sparse temporal coding and multilayer RBF networks," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 426-435, 2002.

[10] E. Kowler, "Eye movements," *Visual Cognition*, S. M. Kosslyn and D. N. Osherson, Eds. Cambridge, MA: MIT Press, pp. 215-266, 1995.

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

[12] A. L. Hodgkin and A. F. Huxley, "A quantitatively description of membrane current and its application to conduction and excitation in nerve," *Journal of Physiology*, vol. 117, pp. 500-544, 1952.

[13] E. M. Izhikevich, "Simple Model of Spiking Neurons," *IEEE Trans. Neural Netwoks*, vol. 14, no. 6, pp. 1569-1572, 2003.

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

[15] A. Destexhe, M. Rudolph and D. Par, "The high-conductance state of neocortical neurons in vivo," *Nature Reviews Neuroscience*, vol. 4, pp. 739-751, 2003.

[16] H. E. Jones, K. L. Grieve and W. Wang, "Surround suppression in primate V1," *J Neurophysiol*, vol. 86, pp. 2011-2028, 2001.

[17] 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, 1995.

[18] S. Eifuku and R. H. Wurtz, "Response to motion in extrastriate cortex MSTl: Center-surround interactions," *Journal of Neurophysiology* , vol. 80, no. 1, pp. 282-296, 1998.

[19] P. Kruizinga and N. Petkov, "Computational model of dot pattern selective cells," *Biological Cybernetics*, vol. 83, no. 4, pp. 313-325, 2000.

[20] N. Petkov and E. Subramanian, "Motion detection, nosie reduction, texture suppresion and contour enhancement by spatiotemporal Gabor filters with surround inhibition," *Biol Cybern*, vol. 97, pp. 423-439, Jun. 2007.

[21] D. H. Hubel and T. N. Wiesel, "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," *J. Physiology*, vol. 160, pp. 106-154, 1962.

[22] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J Opt Soc Am A*, vol. 2, pp. 284-299, 1985.

[23] J. J. Orban de Xivry, S. Coppe, P. Lefvre and M. Missal, "Biological motion drives perception and action," *Journal of Vision*, vol. 10, no. 2, pp. 1-11, 2010.

[24] S. Thorpe, "Spike arrival times: A highly efficient coding scheme for neural networks," *Parallel processing in neural systems and computers*, pp. 91-94, 1990.

[25] S. Neuenschwander, M. Castelo-Branco and W. Singer, "Synchronous oscillations in the cat retina," *Vision Research*, vol. 39, no. 15, pp. 2485-2497, 1999.

[26] P. Fries, S. Neuenschwander, A. K. Engel, R. Goebel, and W. Singer, "Rapid feature selective neuronal synchronization through correlated latency shifting," *NatNeurosci*, vol. 4, no. 2, pp. 194-200, 2001.

[27] F. Grammont and A. Riehle, "Spike synchronization and firing rate in a population of motor cortical neurons in relation to movement direction and reaction time," *Biological cybernetics*, vol. 88, no. 5, pp. 260-373, 2003.

[28] 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, 2006.

[29] D. H. Perkel and T. H. Bullock, "Neural coding," *Neurosciences Research Program Bulletin*, vol. 6, pp. 221-348, 1968.

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

[31] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model the bottom-up visual attention," *IEEE Trans. PAMI*, vol. 28, no. 5, pp. 802-817, 2006.

[32] I. M. Finn and D. Ferster, "Computational Diversity in Complex Cells of Cat Primary Visual Cortex," *J. Neuroscience*, vol. 27, no. 36, pp. 9638-9648, 2007.

[33] A. J. Yu, M. A. Giese and T. A. Poggio, "Biophysiologically plausible implementations of the maximum operation," *Neural Computation*, vol. 14, no. 12, pp. 2857-2881, 2002.

[34] Feng Shi, E. Petriu and R. Laganiere, "Sampling Strategies for Real-time Action Recognition," *CVPR*, pp. 2595-2602, 2013.

[35] Zhang Zhang and Dacheng Tao, "Slow Feature Analysis for Human Action Recognition," *IEEE Trans. PAMI*, vol. 34, no. 3, pp. 436-450, 2012.

[36] Heng Wang, A. Klaser, C. Schmid, Cheng-Lin Liu, "Action Recognition by Dense Trajectories," in Proc. *CVPR*, pp. 3169-3176, 2011.