



A spatiotemporal energy model based on spiking neurons for human motion perception

Hayat Yedjour¹ · Dounia Yedjour¹

Received: 14 August 2023 / Revised: 30 November 2023 / Accepted: 9 January 2024 / Published online: 7 February 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

Abstract

Inspired by the motion processing pathway, this paper proposes a bio-inspired feedforward spiking network model based on Hodgkin–Huxley neurons for human motion perception. The proposed network mimics the mechanisms of direction selectivity found in simple and complex cells of the primary visual cortex. Simple cells' receptive fields are modeled using Gabor energy filters, while complex cells' receptive fields are constructed by integrating the responses of simple cells in an energy model. To generate the motion map, the spiking output of the network integrates motion information encoded by the responses of complex cells with various preferred directions. Simulation results demonstrate that the spiking neuron-based network effectively replicates the directional selectivity operation of the visual cortex when presented with a sequence of time-varying images. We evaluate the proposed model against state-of-the-art spiking neuron-based motion detection models using publicly available datasets. The results highlight the model's capability to extract motion energy from diverse video sequences, akin to human visual motion perception models. Additionally, we showcase the application of the proposed model in motion segmentation tasks and compare its performance with state-of-the-art motion-based segmentation models using challenging video segmentation benchmarks. The results indicate competitive performance. The motion maps generated by the proposed model can be utilized for action recognition in input videos.

Keywords Motion perception · Gabor energy filters · Motion energy · Hodgkin–Huxley model · Spiking neural networks · Visual cortex

Introduction

The human visual system plays a crucial role in detecting and analyzing motion in a dynamic visual environment. The processing of images in our brains follows a hierarchical arrangement, starting from lower areas where neurons respond to simple visual features and extending to higher areas where neurons respond to more complex visual features. The areas of the human brain primarily involved in motion perception are the primary visual cortex

(V1) and the Medial Temporal (MT) region (Sekuler et al. 2002). An important discovery in the neurophysiology of the primary visual cortex made about 60 years ago and described by Hubel and Wiesel (1962), revealed the existence of two types of cortical cells in V1: simple cells and complex cells. Simple cells are organized into columns, and they are phase-sensitive, meaning that their responses are modulated based on the direction of motion of a stimulus in their receptive fields. Complex cells, situated as the second stage after simple cells in V1, exhibit phase insensitivity, responding optimally to stimuli regardless of the position of motion within their receptive fields. Each complex cell receives input from various simple cells with similar orientation preferences but varying spatial phases (Hubel and Wiesel 1962, 1968).

Some theoretical models have been proposed to simulate the motion processing observed in the biological receptive fields of both simple and complex cells within the human visual system (Barlow and Levick 1965; Hassenstein and

✉ Hayat Yedjour
hayat.yedjour@univ-usto.dz

Dounia Yedjour
dounia.yedjour@univ-usto.dz

¹ Faculty of Mathematics and Computer Science, Department of Computer Science, Université des Sciences et de la Technologie d'Oran Mohamed Boudiaf, USTO-MB, EL M'naouer, BP 1505, 31000 Oran, Algeria

Reichardt 1956). Notable among these is the energy-based model developed by Adelson and Bergen (1985). Studies have shown that the receptive field properties of simple cells can be effectively represented by linear filters, such as Gabor functions, which align well with their impulse responses (Andrews and Pollen 1979; Marcelja 1980; Daugman 1985; Jones and Palmer 1987). However, addressing the more intricate characteristics of complex cells, which exhibit spatial phase invariance across a diverse range, necessitates the use of multiple linear spatial filters, as highlighted by Lian et al. (2021). Complex cells are constructed from simple cells in alternative ways, requiring a more nuanced modeling approach. A pivotal contribution by Pollen and Ronner (1981) unveiled a quadrature phase relation between pairs of simple cells in the same cortical column. This revelation enabled the characterization of complex cells through an energy model, where their responses are computed as the sum of the squared outputs of two linear filters characterized by a quadrature pair of Gabor functions.

Several spiking neural network models (SNNs) have been proposed in the literature to replicate the behavior of receptive fields in simple and complex cells for motion processing (Azzopardi and Petkov 2012; Ayzenshtat et al. 2016). Escobar and Kornprobst (2012) introduced a bioinspired motion approach by modeling various types of V1 cells, using responses modeled as conductance-driven integrate-and-fire neurons. Inspired by the functional architecture in V1. A bio-inspired neuron-based network, based on the Hodgkin–Huxley neuron model dedicated to edge detection, was proposed to reproduce the mechanism of orientation selectivity observed in the visual cortex V1 (Yedjour et al. 2017). Another model for motion detection emerged from a feedforward design based on spiking neurons, incorporating temporal delays in synaptic transmission from the MT area (Yedjour et al. 2022), MT neurons' activation in this model, featured linear receptive fields, and responses were modeled using the Hodgkin–Huxley neuron model. Using the Discrete Leaky Integrate-and-Fire neuron model, a spiking network was proposed to illustrate pattern direction selectivity in the visual area MT for motion processing (Beyeler et al. 2014). Berlin and John (2022) introduced a bio-inspired approach for human action recognition, using a spiking network to aggregate motion information across frames. Simulating the hierarchical structure of the visual pathway, Lin et al. (2022) replicated the information transfer mechanism of the retina to the lateral geniculate nucleus for edge detection. Heitmann and Ermentrout (2020) suggested a neural mechanism through which the visual cortex generates spatiotemporal responses for direction-selective motion. To describe the mechanism underlying motion processing in the visual cortex, Alan (2021) proposed a model leading to

direction selectivity. Wang et al. (2021) developed a neural network for real-time motion detection based on the correlation of spatiotemporal receptive fields.

In recent research, Convolutional Neural Networks (CNNs) and Deep Learning (DL) architectures have emerged as the primary choices for detecting motion across diverse visual data. Tokmakov et al. (2017) introduced a two-stream neural network focusing on spatial and temporal features for effective video object segmentation. This model was refined to include motion, appearance, and temporal cues for enhanced segmentation (Tokmakov et al. 2019). Kamranian et al. (2020) proposed a strategy combining motion boundaries and CNN-based class maps for video object segmentation. Song et al. (2018) introduced a DL-based video saliency model that improves feature detection and segmentation. Jain et al. (2017) suggested a model segmenting both static and moving objects through a two-stream full CNN. Cheng et al. (2017) presented a unified network predicting object segmentation and optical flow with improved performance. Jang and Kim (2017) introduced a video object segmentation algorithm using user annotations and optical flow. Koh and Kim (2017) presented a method for video object segmentation employing color and motion edge analysis. While CNN and DL approaches offer accuracy, their computational costs and overfitting risks with limited data pose challenges for real-time implementation. However, SNNs provide an effective alternative, blending neuroscience and deep learning principles to emulate biological neural processes more efficiently.

Studying how nerve cells compute motion is still a puzzle, despite identifying the involved cortical areas. Most neurons in the visual cortex are directionally selective, meaning they respond best to stimuli moving in a specific direction. Previous models focused on synaptic transmission delays to explain motion computation, treating spatial and temporal responses separately. Recently, it has been found that neural activity elicited by a visual stimulus generates traveling waves, which can enhance direction selectivity without relying on explicit time delays (Townsend et al. 2017). These findings add complexity to our understanding of motion perception, emphasizing the use of phase differences rather than transmission delays.

Directional selectivity is a feature observed in the visual cortex, often associated with cells in the primary visual cortex V1. In this study, our goal is to propose a new SNN model for motion perception, drawing inspiration from the observed direction selectivity mechanisms in the primary visual cortex V1. To do so, we aim to model the motion processing pathway, focusing on cortical area V1, regarded as the initial cortical area in visual stream processing. Unlike bio-inspired approaches that rely on transmission delays, our proposed SNN employs an energy model to

replicate the responses of motion-sensitive receptive fields in V1. The computation occurs in two successive steps based on the properties of the receptive fields of simple and complex cells in the visual cortex. In the initial stage, receptive fields of simple cells are constructed using Gabor functions adapted to their profiles. In the subsequent stage, the receptive fields of complex cells are calculated by integrating responses from simple cells within an energy model. Many spiking neural network models rely on leaky integrate-and-fire neuron models to detect motion. In this study, the network response is determined by the Hodgkin–Huxley (1952) model, faithfully replicating the biological behavior of neurons.

Overall, our contributions can be summarized in the following four aspects:

1. We introduce a novel feedforward spiking network model for motion perception, built upon the core principles of the motion energy model.
2. We model two primary cell types, namely simple and complex cells in the visual cortex, to illustrate the successive phases in the development of cortical receptive fields, as classified by Hubel and Wiesel (1962, 1968).
3. The response of our network is determined by the conductance-based Hodgkin–Huxley neuron model (Hodgkin and Huxley 1952), regarded as one of the most biologically plausible models in computational neuroscience.
4. We demonstrate the practical utility of our model in identifying moving objects in videos and compare its performance with current state-of-the-art methods, using challenging benchmark datasets for video object segmentation. This underscores the practical value of our approach.

The remainder of this paper is organized as follows: Section “Spatiotemporal-energy SNN for motion perception” provides a detailed description of the framework of the proposed spiking network for motion perception, including the spiking neuron model and the receptive fields. Section “Simulation results and discussion” presents the experimental results, while Section “Performance Evaluation Of The Proposed Model” compares the performance of the proposed model. Finally, Section “Conclusion and future work” concludes the paper and outlines future work.

Spatiotemporal-energy SNN for motion perception

Motion processing in the visual cortex is widely recognized as a complex task. It involves the dynamic interaction of spiking neurons, working together to construct a comprehensive representation of motion by integrating information from various receptive fields. Inspired by the processing pathway of the human visual cortex, a feed-forward SNN model composed of Hodgkin–Huxley neurons is proposed for motion detection (see Fig. 1). The structure of the SNN with various receptive fields, is explained in detail in this section.

Neuron model for the SNN

The Hodgkin–Huxley model (Hodgkin and Huxley 1952), is a successful mathematical representation for understanding how action potentials are generated in the giant squid axon. It uses four nonlinear differential equations to describe this complex biological process.

$$C \frac{dV}{dt} = I - G_K n^4 (V - E_K) - G_{Na} m^3 h (V - E_{Na}) - G_L (V - E_L) \quad (1)$$

$$\frac{dn}{dt} = \alpha_n(V)(1 - n) - \beta_n(V)n \quad (2)$$

$$\frac{dm}{dt} = \alpha_m(V)(1 - m) - \beta_m(V)m \quad (3)$$

$$\frac{dh}{dt} = \alpha_h(V)(1 - h) - \beta_h(V)h \quad (4)$$

where V is the membrane potential, C represents the specific membrane capacitance per unit area, I is the total membrane current per unit area, G_K and G_{Na} are the potassium and sodium conductances per unit area, respectively, E_K and E_{Na} are the potassium and sodium reversal potentials, respectively, and G and E_L are the leak conductance per unit area and leak reversal potential, respectively. The time-dependent elements of Eq. 1 are V , G_K and G_{Na} , where the last two conductances depend explicitly on voltage as well. α_i and β_i are rate constants for the i th ion channel, that depend on the voltage but not on time. n , m and h are dimensionless quantities between 0 and 1 that are associated with potassium channel activation, sodium channel activation and sodium channel inactivation, respectively. The synaptic current I represents the sum of all conductances received by a neuron, it is given by the following equation:

$$I = g(V - E) \quad (5)$$

where g is the time-dependent conductance. E is its reversal potential. The process of neuronal action potential

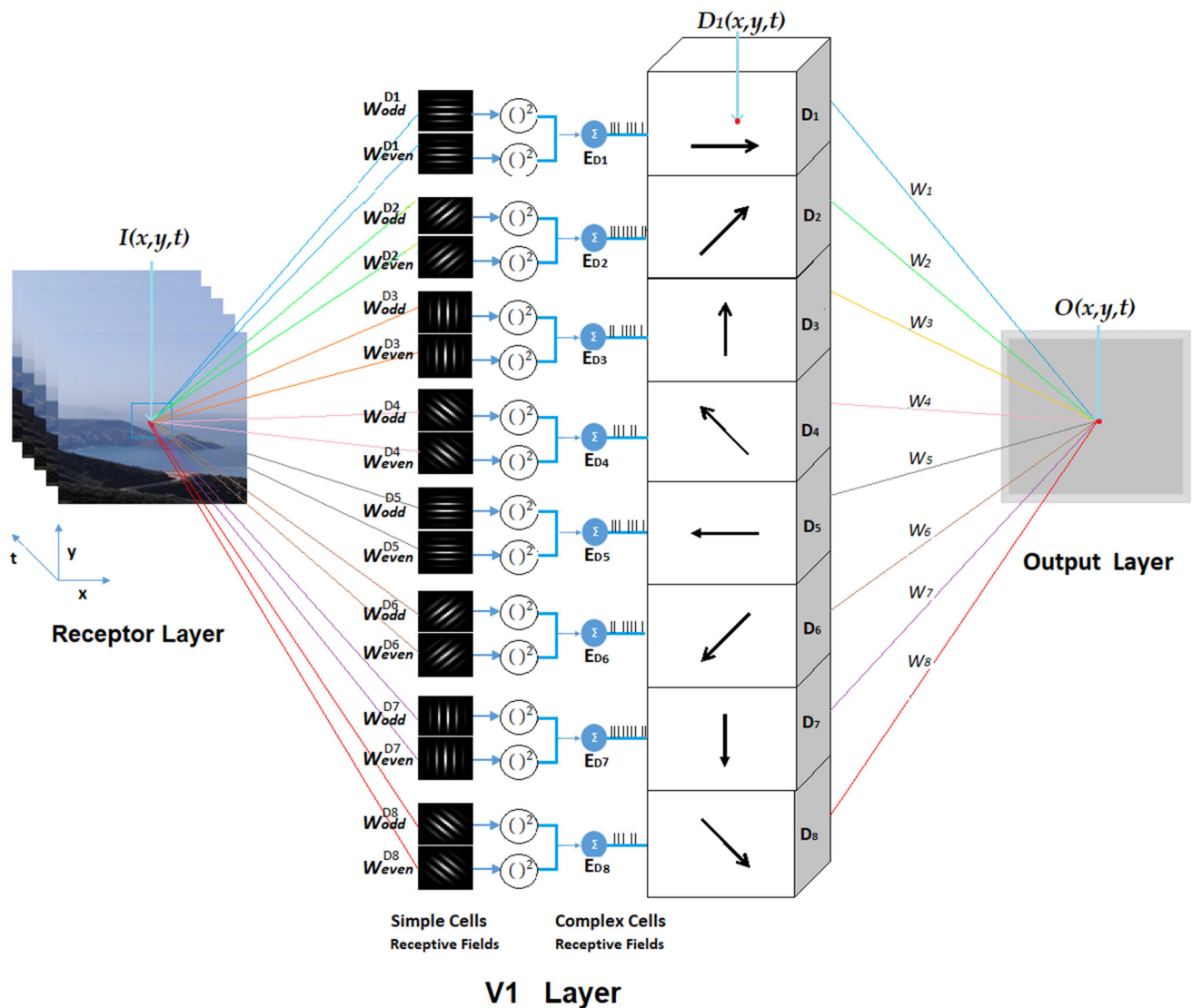


Fig. 1 Overview of the proposed SNN architecture: The SNN consists of three layers: The receptor layer, the V1 layer, and the output layer. Each layer represents a homogeneous population of Hodgkin-Huxley neurons. The Receptor layer, composed of an array with dimensions (854×480) pixels, transforms the pixels of the input image sequence into external currents, which are then fed into the V1 layer. Neurons in the V1 layer mimic the direction selectivity mechanisms found in the primary visual cortex area V1 neurons. They are organized into eight parallel neuron arrays with different direction selectivity, each measuring (854×480) pixels, resulting in a total V1 layer size of $(8 \times 854 \times 480)$ pixels. Each neuron array in the V1 layer has a direction-selective receptive field connected to the receptor array. The construction of these receptive fields is divided into two main stages. In the first stage, Gabor filters are employed to establish

synaptic weights for the receptive fields of simple cells involved in direction detection. These Gabor filters encompass both Sine Gabor (for odd phases) and Cosine Gabor (for even phases) filters. Each simple cell processes information related to a specific motion direction, tuning to eight directions of motion (up, down, left, right, and four diagonals). In the second stage, receptive fields of complex cells are formed by integrating responses from two simple cells using a pair of Gabor filters in quadrature phase, followed by squaring to compute local motion energy. The Output layer, represented by a neuron array of the same size as the receptor layer, integrates information from the V1 layer's spike trains in both space and time to generate a global motion-energy map corresponding to the input image sequence

generation is approximated as a threshold process, when the membrane potential V of neuron reaches the spike threshold V_{th} at time t , a spike S is generated and the membrane potential V is reset to V_{reset} . Any further inputs to the neuron are ignored during a time τ_{ref} called

“absolute refractory period”. S represents a spike train array encoded by binary ones and zeros, generated by the neuron.

$$\text{if}(V(t) > V_{th}) \text{ then } S(t) = 1 \text{ and } V(t) = V_{reset} \\ \text{else } S(t) = 0 \quad (6)$$

SNN architecture and receptive fields construction

The structure of the proposed SNN consists of three layers as depicted in Fig. 1. The first layer, named “receptor”, receives input from the image sequence. The intermediate layer, named “V1”, mimics directional selectivity in the primary visual cortex, connecting to the receptor layer through weight matrices. The output layer pools information from the V1 output in space and time. The network’s response propagates from left to right using the conductance-based Hodgkin-Huxley model to faithfully replicate the biological behavior of neurons.

Receptor layer

The receptor layer represents the input image sequence in grayscale. It is assumed that each receptor can transform a grayscale value of a pixel image, denoted as $I(x, y, t)$ at time t , to peak conductance q using the following formula:

$$q(x, y, t) = g_{norm} I(x, y, t) \quad (7)$$

Here, g_{norm} is a normalization constant, specifically $g_{norm} = \frac{1}{I_{max}}$, where I_{max} represents the maximum pixel value in a grayscale image at time t .

V1 layer

Hubel and Wiesel (1962) distinguished between two distinct types of cells in the primary visual cortex V1: simple cells and complex cells. They observed that both simple and complex cells exhibit direction selectivity, meaning they respond most effectively to oriented stimuli moving in a particular direction. To illustrate motion direction selectivity in the primary visual cortex, we incorporated a V1 layer with eight subpopulations of spiking neurons, denoted as $D_{i=1..8}$, each corresponding to a specific receptive field finely tuned to one of eight directions: leftward, rightward, upward, downward, and the four diagonals. These subpopulations, responsible for processing specific motion directions, are connected to the receptor layer through distinct weight matrices. The receptive fields in the V1 layer are constructed in two consecutive stages to replicate the behavior of simple and complex cells. For example, neuron $D_1(x, y)$, selective for rightward motion, connects to the receptor layer through its receptive fields, built up with both simple cells (W_{odd}^{D1} , W_{even}^{D1}) and complex cells $E_{(x,y,\sigma,f)}^{D1}$.

Simple cells’ receptive fields Simple cells are the first class of neurons in the visual cortex. These cells are made

up of excitatory and inhibitory regions that form oriented receptive fields (Hubel and Wiesel 1962, 1968). These receptive fields consist of even and odd symmetric components, that closely resemble Gabor functions (Daugman 1985). The even-symmetric parts capture the overall spatial structure, while the odd-symmetric parts enable these cells to distinguish between different orientation patterns. This dual organization contributes to the ability of simple cells to detect and respond to specific visual orientations. To accurately model synaptic connections between neurons in the receptor layer and those in the V1 layer, we represent the receptive fields of simple cells using separable components of the Gabor energy function, which include both even and odd phases. These components are expressed as spatio-temporal Gaussian filters multiplied by cosine waves (for even phases) and sine waves (for odd phases). Figure 2 illustrates the distribution of synapses for odd and even connections, tuned for detecting 8 directions.

Let’s consider a neuron $D_i(x, y)$ with the receptive field centered at position (x_0, y_0) and the center frequencies (f_{x0}, f_{y0}, f_{t0}) . The weight distribution for synapse connections W_{odd}^{Di} and W_{even}^{Di} for a neuron $D_i(x, y)$ is computed using the following expressions:

$$W_{odd}^{Di}(x, y, t, \sigma, f) = \frac{1}{(2\pi)^{\frac{3}{2}} \sigma_x^2 \sigma_y^2 \sigma_t^2} \exp\left(-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} + \frac{t^2}{\sigma_t^2}\right)\right) \sin(2\pi f_{x0} + 2\pi f_{y0} + 2\pi f_{t0}) \quad (8)$$

$$W_{even}^{Di}(x, y, t, \sigma, f) = \frac{1}{(2\pi)^{\frac{3}{2}} \sigma_x^2 \sigma_y^2 \sigma_t^2} \exp\left(-\frac{1}{2} \left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} + \frac{t^2}{\sigma_t^2}\right)\right) \cos(2\pi f_{x0} + 2\pi f_{y0} + 2\pi f_{t0}) \quad (9)$$

Here, $(\sigma_x, \sigma_y, \sigma_t)$ represent the standard deviations of the spatio-temporal Gaussian envelope. For each neuron $D_{i=1..8}$, the conductances $g_{odd(x,y)}^{Di}(t)$ and $g_{even(x,y)}^{Di}(t)$ are governed by the following equations:

$$\frac{g_{odd(x,y,\sigma,f)}^{Di}(t)}{dt} = -\frac{1}{\tau} g_{odd(x,y,\sigma,f)}^{Di}(t) + \sum_{r=-R}^R \sum_{s=-R}^R W_{odd}^{Di}(r, s, \sigma, f) q(x-r, y-s) \quad (10)$$

$$\frac{g_{even(x,y,\sigma,f)}^{Di}(t)}{dt} = -\frac{1}{\tau} g_{even(x,y,\sigma,f)}^{Di}(t) + \sum_{r=-R}^R \sum_{s=-R}^R W_{even}^{Di}(r, s, \sigma, f) q(x-r, y-s) \quad (11)$$

where τ is synaptic decay times, W_{odd}^{Di} and W_{even}^{Di} represent synaptic strength matrices with a synapse radius R centered at a pixel image $I(x, y, t)$, and q represent the peak synaptic conductance.

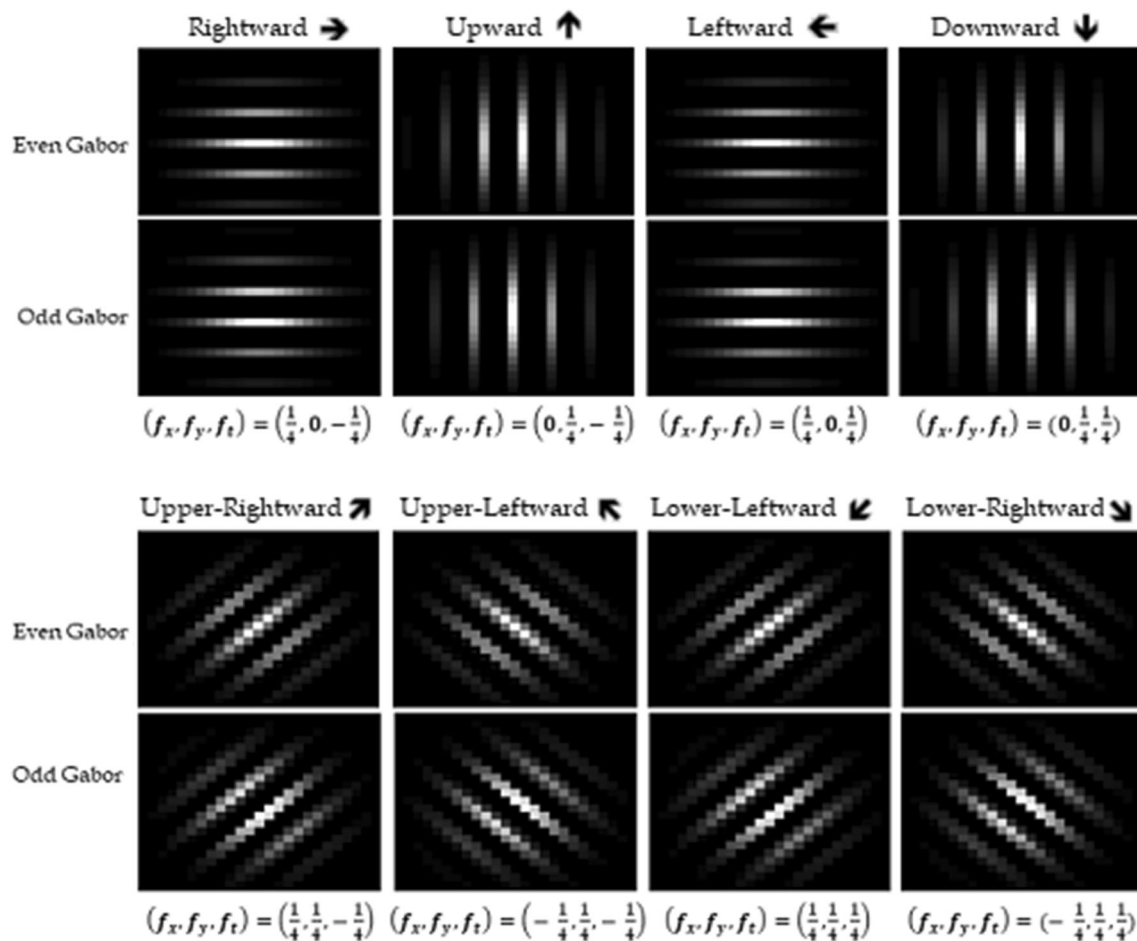


Fig. 2 Receptive Fields in V1 Layer: Slices through the even (cosine) and odd (sine) components of the receptive fields of simple cells, utilizing separable components of the Gabor energy functions to model 8 motion-sensitive simple cell receptive fields, each tuned to center frequencies corresponding to 8 directions. The first row depicts responses to the 4 cardinal directions, while the second row illustrates responses to the 4 diagonals. Notably, the first two columns in each row mirror the last two columns, differing only in the sign of f_t .

Complex cells' receptive fields The complex cells in our network constitute the second class of neurons in the V1 layer. We model complex cells based on the motion energy model (Adelson and Bergen 1985). Each complex cell combines energy responses from two simple cells characterized by Gabor functions of the same frequency but in quadrature pair, i.e., odd-Gabors and even-Gabors. The energy response of a complex cell for a neuron $D_i(x, y)$ is computed as the squared sum of the energy responses of odd-phase and even-phase Gabor filters, expressed as:

$$E_{(x,y,\sigma,f)}^{D_i}(t) = \left(g_{odd(x,y,\sigma,f)}^{D_i}(t)\right)^2 + \left(g_{even(x,y,\sigma,f)}^{D_i}(t)\right)^2 \quad (12)$$

In accordance with (Eq. 1), the dynamics of the membrane potential V for a neuron $D_i(x, y)$ is described by the following equation:

Reversing the sign of f_t results in a 180-degree flip in the filter direction. The size of the receptive field is determined by the fixed standard deviation of the Gaussian factor, set at at $(\sigma_x, \sigma_y, \sigma_t) = (4, 4, 1)$. The selection of center frequencies (f_x, f_y, f_t) maintains a constant $(f_x^2 + f_y^2)$, with the temporal frequency $f_t = (\frac{1}{4})$ (Heeger 1988)

$$C \frac{dV}{dt} = E_{(x,y,\sigma,f)}^{D_i}(V - E) - G_K n^4 (V - E_K) - G_{Na} m^3 h (V - E_{Na}) - G_L (V - E_L) \quad (13)$$

Let $S_{D_i=1..8}(t)$ represent spike trains generated by (Eq. 6) for neurons $D_{i=1..8}$ respectively. The firing rates of the V1 neuron population, selective for motion direction, serve as input to the spiking neurons in the output layer.

Output layer

The output layer receives feedforward connections from the V1 layer. Each neuron in the output layer, denoted as $O(x, y)$, integrates the outputs of corresponding neurons from the V1 layer and can respond to motion in any

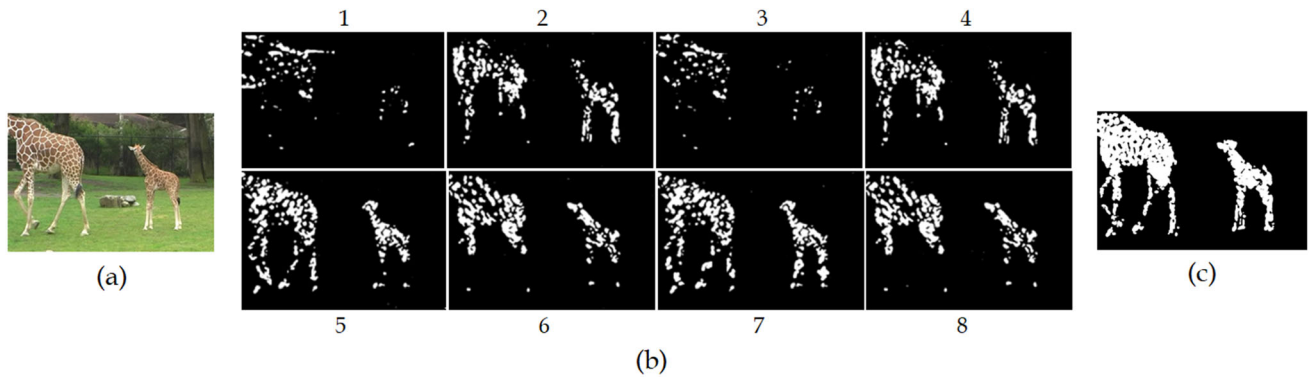


Fig. 3 Direction-selective motion energy responses on an input image from the FBMS dataset (Ochs et al. 2014): **a** Original frame of the image sequence; **b** Spatio-temporal energy responses for different

motion directions: (b-1) Rightward. (b-2) Upward. (b-3) Leftward. (b-4) Downward. (b-5) Upper-rightward. (b-6) Upper-leftward. (b-7) Lower-leftward. (b-8) Lower-rightward; **c** global motion energy

direction within its receptive field. According to the conductance-based Hodgkin-Huxley model, the dynamics of a neuron $O(x, y)$ in relation to $D_i(x, y)$ are described by the following equations:

$$\frac{g_{O(x,y)}(t)}{dt} = -\frac{1}{\tau} g_{O(x,y)}(t) + \sum_{i=1}^8 W_i S_{D_i(x,y)}(t) \quad (14)$$

$$C \frac{dV}{dt} = g_{O(x,y)}(V - E) - G_K n^4 (V - E_K) - G_{Na} m^3 h (V - E_{Na}) - G_L (V - E_L) \quad (15)$$

The spike condition (Eq. 6) is applied to neuron $O(x, y)$ to generate a spike train array denoted by $S_{O(x,y)}(t)$. To compute the firing rate $F_{O(x,y)}$ for neuron $O(x, y)$ during the simulation time T , we use the following expression:

$$F_{O(x,y)} = \frac{\sum_{t=1}^T S_{O(x,y)}(t)}{T} \quad (16)$$

Simulation results and discussion

The proposed SNN model for visual motion perception was implemented in MATLAB using a set of parameters that were consistent with the biophysical neuron model of (Hodgkin and Huxley 1952). The parameters are summarized as follows: $C = 1 \mu F/cm^2$, $E_L = 10,613 mV$, $E_{Na} = 115 mV$, $E_K = -12 mV$, $G_L = 0.3 mS/cm^2$, $G_{Na} = 120 mS/cm^2$, $G_K = 36 mS/cm^2$, $V_{th} = -65 mV$, $V_{reset} = -70 mV$, $E = 0 mV$, $\tau = 3 ms$, $T = 100 ms$, $\tau_{ref} = 6 ms$. Synapse weights $W_{i=1..8}$ are constants, all set to 1. (n , m and h) for activation and inactivation, are quantities between 0 and 1. (n_0 , m_0 , and h_0) are their initial values, calculated using the resting states defined by: $\frac{dn}{dt} = 0$, $\frac{dm}{dt} = 0$ and $\frac{dh}{dt} = 0$, of Eqs. (2)–(4) respectively. With the value $V = 0$ of membrane potential, the following initial point is found: ($n_0 = 0.1765$, $m_0 = 0.0529$ and $h_0 = 0.5961$).

We constructed eight Gabor energy filters to capture the energy responses of moving objects in a video sequence, considering all possible directions of motion. Each directional filter approximates the profile receptive field of simple cells, with the size of the receptive field determined by the standard deviation (σ_x , σ_y , σ_t) of the Gaussian factor. Experimental results indicate that using a smaller size of receptive fields results in lower energy for a given direction, while larger receptive fields produce higher energy. To minimize ambiguity between directional filters, we selected standard deviations (σ_x , σ_y) greater than σ_t to accurately mimic the human visual system. In this paper, we fixed the standard deviation values to $\sigma_x = \sigma_y = 4$ and $\sigma_t = 1$, as suggested in (Heeger 1988). Additionally, empirical findings suggest that setting $\sigma_t = 1$ provides effective coverage of the frequency-temporal domain, including temporal frequencies $f_t = -1/4$, $f_t = 0$ and $f_t = 1/4$ (Heeger 1988; Daugman 1985; Spinéi et al. 2000).

In the first experiment, we aim to evaluate the ability of the proposed SNN model to replicate directional selectivity responses. The original frame of an image sequence featuring a giraffe from the FBMS dataset (Ochs et al. 2014), as shown in Fig. 3, undergoes processing with the receptive fields of the simple and complex cells in the V1 layer. The firing rates of V1 neurons generate motion energy responses across various directions, showing strong responses to specific motion directions and comparatively weaker responses to others. The output layer integrates energy responses from the V1 layer, providing a comprehensive motion energy depiction in the current frame. These results align with the established properties of the biological visual system.

Another experiment was conducted using three image sequences sourced from diverse benchmark datasets, each presenting distinct challenges and featuring complex background images. The first sequence is from the YMB (YouTube Motion Boundaries) dataset (Weinzaepfel et al.

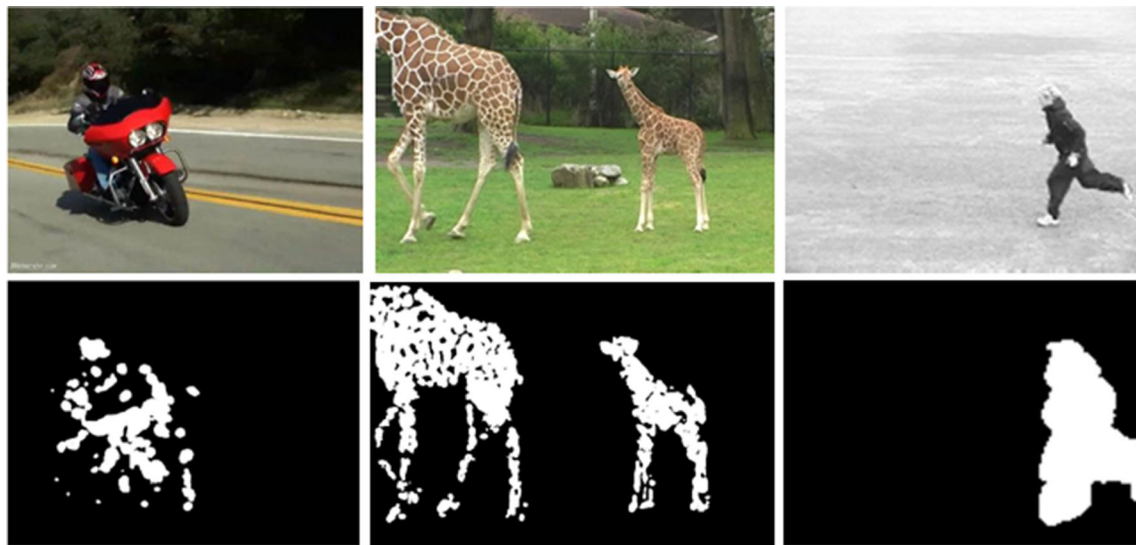


Fig. 4 Sample results of our scheme on three video segmentation datasets. First row shows one frame of the video sequence taken from (Left to right): YMB dataset (Weinzaepfel et al. 2015), FBMS dataset

(Ochs et al. 2014) and WEIZ dataset (Alpert et al. 2012). The second row shows the corresponding energy motion results

2015). The second comes from the FBMS (Freiburg–Berkeley Motion Segmentation) dataset (Ochs et al. 2014), and the last one is from the Weizmann dataset (Alpert et al. 2012). The experimental results, as depicted in Fig. 4, showcase the proposed SNN’s capability to effectively extract the energy of moving objects from image sequences.

Performance evaluation of the proposed model

To demonstrate the effectiveness of our approach, we conducted a comparative analysis involving SNN-based approaches for motion detection. The evaluation encompasses two distinct datasets: BMC (Sobral and Vacant 2014) and CDNet2014 (Wang et al. 2014). We then illustrate the utility of the identified motion energy for motion segmentation tasks. Initially, we evaluate the model using conventional motion segmentation approaches and subsequently explore advanced DL motion segmentation methods. This evaluation covers four datasets: BMC dataset (Sobral and Vacant 2014), MoSeg dataset (Brox and Malik 2010), Davis2016 dataset (Perazzi et al. 2016), and SegTrackv2 dataset (Li et al. 2013).

Datasets and evaluation metrics

The MoSeg dataset (Brox and Malik 2010) offers short video sequences (20 to 60 frames) capturing the movement of objects like cars and people, with ground truth annotations. The videos have a resolution of (640×480) pixels.

The BMC dataset (Sobral and Vacant 2014) serves as a benchmark for motion detection and moving object segmentation. It includes 10 synthetic video sequences depicting diverse scenes and weather conditions, with annotations for 1500 frames at a resolution of (640×480) pixels.

The CDNet2014 dataset (Wang et al. 2014) is a large benchmark for real-life motion detection. It consists of 11 categories with 4 to 6 sequences each, featuring challenges like changing light, diverse weather conditions, and complex objects, providing annotated ground truth and segmentation masks.

The Davis2016 dataset (Perazzi et al. 2016) contains 50 high-quality video sequences (3455 frames) presenting challenges in video object segmentation, including object transitions, occlusions, and motion blur, with detailed pixel-wise annotations.

The SegTrackv2 dataset (Li et al. 2013) is a benchmark for video object segmentation, featuring 14 diverse sequences (1066 frames) with challenges like object-background similarity, non-rigid deformations, fast camera motion, varying lighting, and occlusions. It provides detailed pixel-wise annotations.

The evaluation metrics employed include common metrics: precision (P), recall (R), and F-score (F), and Intersection Over Union (IOU) index. These metrics, calculated using ground truth annotations, are defined as follows:

$$P = \frac{tp}{tp + fp}, R = \frac{tp}{tp + fn}, F = 2 * \frac{P * R}{R + P} \text{ and } IOU = \frac{tp}{tp + fp + fn} \quad (17)$$

Here, (tp) represents the number of pixels correctly detected as moving, (fp) is the number of incorrectly detected pixels as moving, and (fn) is the number of moving pixels not detected. A higher F-score and *IOU* indicate better segmentation performance.

Energy detection performance against SNN-based approaches

To demonstrate the effectiveness of our approach, we conducted a comparative analysis involving two SNN-based methods for motion detection. The first one, based on temporal difference (denoted as TD) (Yedjour et al. 2019), extracts moving objects between consecutive frames in a dynamic scene and feeds them into a network of Hodgkin-Huxley neurons for boundary detection (Yedjour et al. 2017). The second one, based on axonal delays (denoted as AD) (Yedjour et al. 2022), uses temporal delays in synapses to construct a hierarchical feedforward model for the MT area, introducing distinct MT receptive fields for leftward and rightward motion.

The BMC dataset (Sobral and Vacant 2014) includes 10 sequences showcasing two scenes: a street and a rotary scene. Each sequence demonstrates various weather conditions and noise levels. Specifically: Sequences 112 and 122 have normal conditions without acquisition noise. Sequences 212 and 222 are under cloudy conditions with salt and pepper noise. Sequences 312 and 322 are sunny with noise, including moving cast shadows. Sequences 412 and 422 are foggy with added noise, making analysis challenging. Sequences 512 and 522 are windy with noise and a moving background. In our evaluation shown in Fig. 5, we present qualitative results comparing our approach to SNN-based methods for five street scene sequences, each representing different weather conditions (sequences 112, 212, 312, 412, and 512).

The SNN takes an image from a sequence of images as input, performing motion detection at each time step. We use performance parameters precision (P), recall (R), and F-score (F) to measure the similarity between the motion map obtained by a given model and the ground truth. To provide a comprehensive assessment, we conducted experiments on all ten video sequences from the BMC dataset (Sobral and Vacant 2014) and calculated the average over them. We then compared the performance of our approach to TD (Yedjour et al. 2019) and AD (Yedjour et al. 2022) approaches. Quantitative results are presented in Table 1.

The results presented in Table 1 and Fig. 5 demonstrate that our proposed approach achieved the highest F-score compared the TD (Yedjour et al. 2019) and the AD (Yedjour et al. 2022) methods, showcasing its ability to

segregate moving objects in an image sequence. The second-highest F-score was obtained by the AD method (Yedjour et al. 2022). While this method effectively extracts motion objects in dynamic environments, it struggles to accurately delineate them. Despite its faster processing time, the TD method (Yedjour et al. 2019) obtained the lowest F-score. This suggests that the TD method is less suitable for sequences characterized by illumination variation and cluttered environments. In summary, our approach significantly outperforms the TD (Yedjour et al. 2019) and AD (Yedjour et al. 2022) methods, as evidenced by Table 1 and Fig. 5. This highlights a more precise alignment with the ground truth, facilitating the optimized extraction of motion energy. These results affirm the effectiveness of our proposed approach in motion detection tasks.

We extended our evaluation to include the CDNet2014 dataset (Wang et al. 2014), encompassing various scenarios such as highways, offices, pedestrians, and PETS2006. We compared the motion maps generated by our approach with those of the TD (Yedjour et al. 2019) and AD (Yedjour et al. 2022) methods against their respective ground truth annotations on the CDNet2014 dataset. This comparison utilized standard CDNet2014 challenge metrics: precision (P), recall (R), and F-score (F), averaged across all video frames within each scenario.

The results presented in Table 2 clearly demonstrate the effectiveness of the proposed approach in motion detection across different video sequences. The proposed method shows a significant improvement in F-score compared to the SNN approaches, TD and AD, in all evaluated scenarios, particularly pronounced in the video sequences of Highway and PETS2006, where motion is more evident. This improvement is particularly noteworthy, with an F-score of 0.91 in the Highway scenario, representing a 22% increase compared to AD (0.75) and a 54% increase compared to TD (0.59). Similarly, in the PETS2006 scenario, the proposed approach outperforms the SNN approaches, AD and TD, with F-score improvements of 13% and 69%, respectively. The average F-score improvement of 13% compared to the TD (Yedjour et al. 2019) and 71% compared to the AD (Yedjour et al. 2022) confirms the robustness of the proposed approach in capturing motion nuances in various environmental contexts. These superior results in diverse scenarios highlight the effectiveness of our proposed method, positioning it as a promising solution for motion detection tasks.

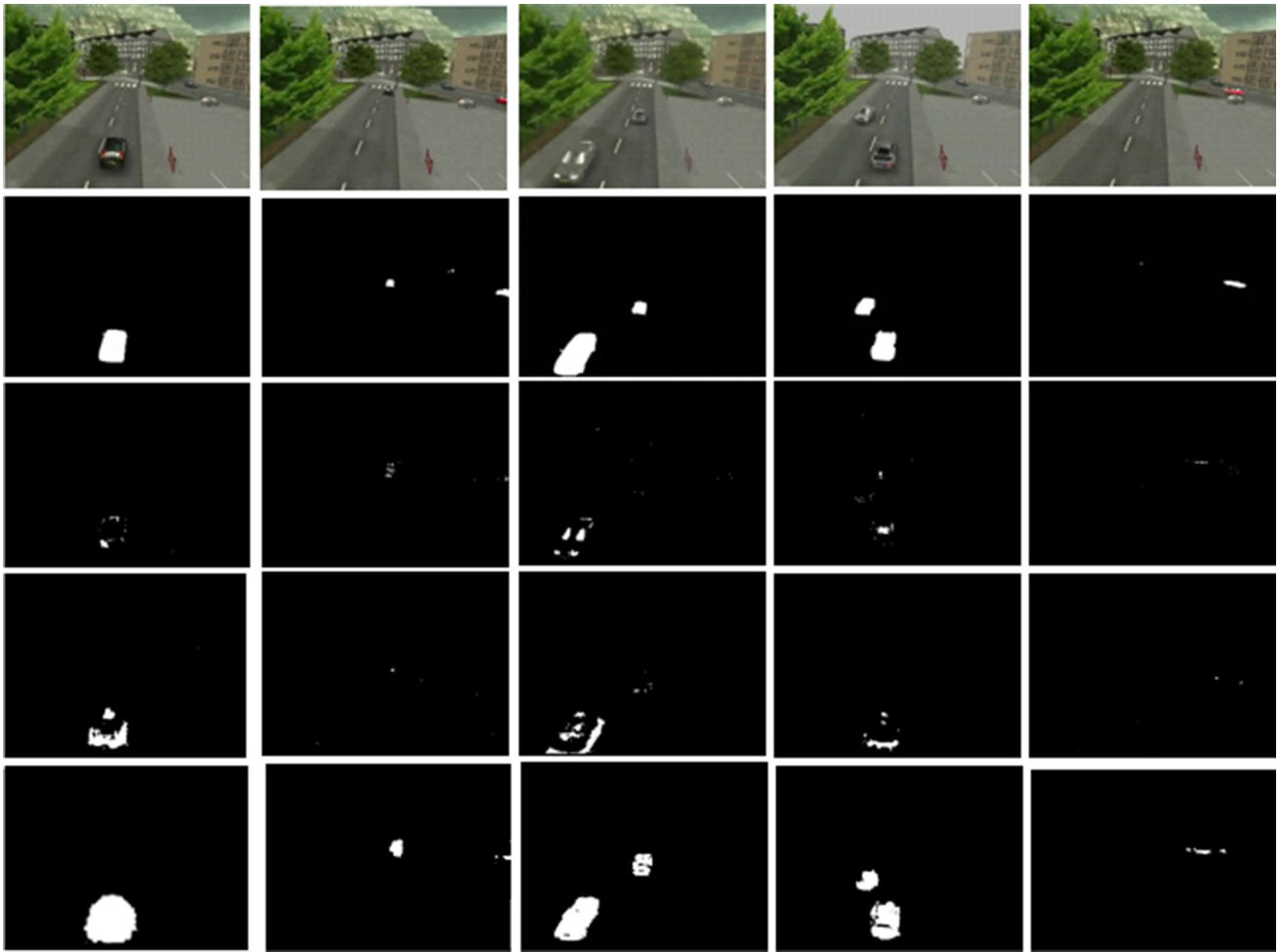


Fig. 5 Qualitative results for the BMC dataset (Sobral and Vacant 2014). From top to bottom: image sequence, ground truth, TD (Yedjour et al. 2019), AD (Yedjour et al. 2022), Proposed approach. Left to right: 112, 212, 312, 412, 512 video sequences

Table 1 Comparison results of the proposed approach to TD (Yedjour et al. 2019) and AD (Yedjour et al. 2022) approaches on the BMC dataset (Sobral and Vacant 2014)

BMC Sequences	TD (Yedjour et al. 2019)			AD (Yedjour et al. 2022)			Proposed Approach		
	P	R	F	P	R	F	P	R	F
112	0.73	0.53	0.61	0.88	0.76	0.82	0.94	0.93	0.93
122	0.73	0.56	0.63	0.86	0.78	0.82	0.95	0.92	0.93
212	0.58	0.45	0.51	0.81	0.75	0.78	0.93	0.91	0.92
222	0.56	0.50	0.53	0.83	0.79	0.81	0.93	0.93	0.93
312	0.67	0.57	0.62	0.90	0.75	0.82	0.94	0.94	0.94
322	0.63	0.56	0.59	0.87	0.74	0.80	0.88	0.95	0.91
412	0.72	0.57	0.64	0.81	0.78	0.79	0.90	0.94	0.92
422	0.70	0.48	0.57	0.79	0.77	0.78	0.93	0.91	0.92
512	0.68	0.51	0.58	0.79	0.68	0.73	0.93	0.95	0.94
522	0.62	0.49	0.56	0.77	0.66	0.71	0.89	0.94	0.91
AVG	0.66	0.52	0.58	0.83	0.75	0.79	0.92	0.93	0.93

Table 2 Comparison results of the proposed approach to TD (Yedjour et al. 2019) and AD (Yedjour et al. 2022) approaches on the CDNet2014 dataset (Wang et al. 2014)

CDNET2014 baseline sequences	TD (Yedjour et al. 2019)			AD (Yedjour et al. 2022)			Proposed approach		
	P	R	F	P	R	F	P	R	F
HIGHWAY	0.63	0.55	0.59	0.79	0.72	0.75	0.95	0.87	0.91
OFFICE	0.55	0.43	0.48	0.79	0.79	0.79	0.89	0.84	0.86
PEDESTRIANS	0.52	0.43	0.47	0.72	0.69	0.70	0.89	0.88	0.88
PETS2006	0.58	0.49	0.53	0.82	0.78	0.80	0.91	0.89	0.90
AVG	0.57	0.48	0.52	0.78	0.75	0.76	0.91	0.65	0.89

Energy segmentation performance

Comparison with conventional approaches

We demonstrate that the detected motion energy can be effectively utilized in motion segmentation tasks. To assess the robustness of our approach, we compare it with two dedicated motion segmentation methods. The first method is based on video object segmentation to establish occlusion relations robustly (Taylor et al. 2015). The second method employs the Region-of-Interest (denoted as RoiSeg) approach (Zhang et al. 2022). Quantitative evaluation results for Taylor et al. (2015) were conducted on the MoSeg dataset (Brox and Malik 2010), while evaluation results for Zhang et al. (2022) were carried out on the BMC dataset (Sobral 2014). Building on these comparisons, we assess the performance of our energy-based model on two datasets: MoSeg (Brox and Malik 2010) and BMC (Sobral 2014) datasets. For the performance evaluation of segmentation, we use the metrics precision (P), recall (R), and F-score (F), which are also adopted by the comparative approaches. These metrics are reported based on the provided ground truth annotations.

The approach by Taylor et al. (2015) has been compared to two methods, Lee et al. (2011) and Papazoglou and Ferrari (2013), which employ individual component models. These models generate binary segmentation algorithms, including the (BASIC) model, its temporal extension (TE), foreground–background (FG), and the full model (FULL) as a combination of all components. Results are reported for various algorithms, such as (BASIC), (TE), (FG), Lee et al. (2011), and Papazoglou and Ferrari (2013). Our model is specifically compared to Taylor et al. (2015) method for motion segmentation. The comparative analysis of the performance of different methods is presented in Table 3. The highest F-score indicates a perfect match with the ground truth. According to the results in Table 3, our approach outperforms the comparative approaches in F-score on MoSeg dataset (Brox and Malik 2010).

The RoiSeg (Zhang et al. 2022) has been compared to three of the best traditional foreground segmentation algorithms, all evaluated on the BMC dataset (Sobral

Table 3 Quantitative comparison of the proposed approach with individual and Full components of the Taylor et al. (2015) method, Lee et al. (2011), and Papazoglou and Ferrari (2013) methods on the MoSeg dataset (Brox and Malik 2010)

	P	R	F
BASIC	0.93	0.34	0.49
TE	0.78	0.55	0.64
FG	0.77	0.56	0.65
FULL	0.87	0.63	0.73
Lee et al. (2011)	0.62	0.56	0.59
Papazoglou and Ferrari (2013)	0.76	0.63	0.69
Proposed approach	0.88	0.65	0.75

2014). The first one, named DPWrenGABGS (DRG), is an enhanced version of the Gaussian Mixture Model designed for a specific color space. The second, Mixture Of GaussianVIBGS (MOG), represents an improved version that adapts the system to illumination variation. The third, MultiLayer BGS (BGS), is a background subtraction algorithm based on color and texture. We evaluated our approach across 10 frame sequences, as utilized in RoiSeg (Zhang et al. 2022). The quantitative results are then averaged for the 10 frame sequences covering the entire BMC dataset (Sobral 2014) and are summarized in Table 4.

According to the results in Table 4, we observe that the proposed approach achieved the highest F-scores in video sequences 312 and 512, despite the dynamic and cluttered environment in sequence 312 and the windy weather in sequence 512. This can be explained by the fact that significant motion leads to a higher energy response. The second-highest F-scores for our method were for sequences 112 and 122, which experienced cloudy conditions without noise, and sequence 222, which included salt and pepper noise. The F-score is slightly higher for the RoiSeg (Zhang et al. 2022) in sequence 322 compared to ours. For sequence 212, the (BGS) method achieved the highest F-score despite the presence of noise in this sequence. A suboptimal F-score obtained by our model for videos 212 and 322 can be attributed to the reduced brightness contrast between moving objects and the background in these

Table 4 Comparative performance metrics on the BMC dataset (Sobral 2014) for the proposed approach, RoiSeg (Zhang et al. 2022), and the three traditional segmentation algorithms

BMC Sequences	DRG Method			MOG Method			BGS Method			RoiSeg Approach			Proposed Approach		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
112	0.87	0.87	0.87	0.96	0.74	0.84	0.92	0.95	0.93	0.89	0.93	0.91	0.94	0.93	0.93
122	0.91	0.87	0.89	0.96	0.70	0.80	0.91	0.94	0.93	0.91	0.94	0.92	0.95	0.92	0.93
212	0.92	0.86	0.89	0.97	0.74	0.84	0.94	0.94	0.94	0.89	0.93	0.91	0.93	0.91	0.92
222	0.93	0.86	0.90	0.96	0.73	0.81	0.94	0.93	0.93	0.91	0.94	0.92	0.93	0.93	0.93
312	0.65	0.78	0.71	0.98	0.68	0.80	0.96	0.87	0.91	0.89	0.93	0.91	0.94	0.94	0.94
322	0.89	0.78	0.83	0.95	0.65	0.77	0.94	0.85	0.89	0.91	0.94	0.92	0.88	0.95	0.91
412	0.53	0.76	0.62	0.98	0.69	0.81	0.71	0.84	0.77	0.89	0.93	0.91	0.90	0.94	0.92
422	0.53	0.75	0.62	0.97	0.64	0.77	0.77	0.79	0.78	0.91	0.94	0.92	0.93	0.91	0.92
512	0.63	0.86	0.73	0.82	0.74	0.78	0.65	0.93	0.76	0.81	0.91	0.86	0.93	0.95	0.94
522	0.80	0.86	0.83	0.91	0.69	0.79	0.88	0.93	0.90	0.89	0.93	0.91	0.89	0.94	0.91
AVG	0.77	0.83	0.79	0.95	0.70	0.80	0.86	0.90	0.87	0.89	0.93	0.91	0.92	0.93	0.93

scenarios, making it challenging for our model to extract motion energy. Note that video 322 includes a sunny rotary scene, which presents difficulties such as general climatic conditions and fast light changes compared to video 312, which depicts a street scene with a dynamic and cluttered environment. Based on the results in Table 4, the proposed approach, on average, demonstrates superior performance compared to the RoiSeg (Zhang et al. 2022) and the three traditional segmentation methods across all ten videos in the BMC dataset (Sobral 2014), with a slight decrease in accuracy observed in specific scenarios.

Comparison with DL approaches

We compare our method with state-of-the-art methods based on DL models (Tokmakov et al. 2017, 2019; Kamranian et al. 2020; Song et al. 2018; Jain et al. 2017; Cheng et al. 2017; Jang and Kim 2017; Koh and Kim 2017). All of these methods are evaluated on all 50 video sequences of Davis2016 dataset using two metrics: the Intersection Over Union (*IOU*) index and the F-score. On SegTrackv2 dataset, the *IOU* metric is measured across all 14 videos for comparison. Table 5 presents the quantitative results for each method, including the proposed approach.

On the Davis2016 dataset, our proposed approach demonstrates superior performance, surpassing all state-of-the-art methods in terms of *IOU* with 78.57% and F-score with 88.0%. Additionally, it is slightly better than (Tokmakov et al. 2019), which achieved the second-highest *IOU* at 78.2% and F-score at 75.9%. Note that the approach by Tokmakov et al. (2019) represents an enhancement of their previous model (Tokmakov et al. 2017) and is

recognized as one of the best-reported techniques, typically relying on a learning mechanism for foreground object identification. In contrast, our method significantly outperforms it without involving any learning processes.

On the SegTrackv2 dataset, our approach achieved an *IOU* of 82.61%, outperforming all methods except Tokmakov et al. (2019) and Jain et al. (2017). While our approach significantly outperforms them on the Davis16 dataset, Tokmakov et al. (2019) and Jain et al. (2017) excel on the SegTrack dataset with *IOU* scores of 83.4% and 83%, respectively. Our lower performance may be attributed to the low quality and poor resolution of the SegTrackv2 videos, where objects often lack a clear distinction from the background. Nevertheless, the proposed approach still significantly improves upon state-of-the-art motion-based segmentation models. This reaffirms its capability to effectively separate moving objects from an image sequence, aligning with the characteristics of the primary visual cortex in the human brain.

Finally, our approach excels in motion segmentation compared to both conventional and DL models, despite facing challenges on the SegTrackv2 dataset. The overall results obtained underscore the robustness and competitiveness of our model, which relies on intrinsic mechanisms such as directional selectivity and spatiotemporal tuning, without resorting to explicit learning mechanisms. This proposed approach proves effective and practical for understanding the underlying mechanisms of human motion perception.

Table 5 Quantitative results of the proposed approach compared to state-of-the-art methods on Davis2016 and SegTrackv2 datasets

Method	Davis16 (<i>IOU</i>)	Davis16 (<i>F</i>)	SegTrackv2 (<i>IOU</i>)
Tokmakov et al. (2017)	75.9	72.1	77.1
Kamranian et al. (2020)	69.3	-	63.8
Song et al. (2018)	77.2	74.5	73.8
Tokmakov et al. (2019)	78.2	75.9	83.4
Jain et al. (2017)	70.7	65.3	83.0
Cheng et al. (2017)	67.4	66.7	81.4
Jang and Kim (2017)	75.5	71.4	-
Koh and Kim (2017)	76.2	70.6	79.2
Proposed approach	78.57	88.0	82.61

The best two results are highlighted in bold

Conclusion and future work

While traditional approaches to motion detection have shown improvement, they still face limitations in capturing the complexity of real-world visual data. Bio-inspired models, particularly Spiking Neural Networks composed of spiking neuron models, replicate the intricate circuitry of visual information processing in the brain, offering powerful functionality for input signal combination and spike generation. This provides a biologically plausible solution that surpasses conventional methods. In this paper, we proposed a novel SNN model based on the Hodgkin–Huxley neuron model and spatiotemporal receptive fields for visual motion perception. Our approach processes motion based on the responses of both simple and complex cells in visual area V1. Synaptic weights, determined by Gabor energy filters, play a crucial role in executing computations, filtering, adaptation, and dynamic properties. We emphasized the benefits of incorporating various directionally selective neurons, whose responses are defined by the Hodgkin–Huxley model of neurons, providing powerful functionality for spike generation and intricate computations. Our model has been tested on diverse video sequences. The obtained results show the model's ability to extract motion energy leading to better agreement with the physiological data. When compared to state-of-the-art motion-based models, including SNN, conventional, and DL approaches, the proposed approach outperforms them across five datasets with different challenges. In future work, we can exploit the motion energy results to recognize the action present in an input video.

Many neural network models have been identified in biological experiments, and the transfer of these models to computational frameworks requires in-depth study. This paper demonstrates the applicability of a biologically inspired network to solve specific problems, such as motion perception. The proposed network model, based on the Hodgkin–Huxley neuron model, offers several key

advantages. Firstly, synapses are able to collaborate with the population of neurons in the network to address various perturbations, including noise, small occlusions, or moving backgrounds. Secondly, our model does not exhibit speed selectivity for motion, instead, it relies on general mechanisms, such as directional selectivity and spatiotemporal tuning, to extract motion energy. Thirdly, our SNN model is designed to detect motion without requiring additional adjustments or an explicit learning process. It leverages built-in mechanisms, such as the integration of inputs in space and time and the generation of spikes, aligning them with observed motion processing principles in the biological visual system. However, the model exhibits certain limitations that warrant further exploration and discussion. Firstly, focusing exclusively on V1, without accounting for other cortical areas involved in motion processing, like the Medial Temporal (MT), the Inferior Temporal (IT), and the Medial Superior Temporal (MST), may limit the model in capturing the full complexity of visual processing. Additionally, challenges presented by complex situations, such as large occlusions or intricate backgrounds, highlight the importance of incorporating attention mechanisms into the model. In future work, we can address these limitations by introducing attention mechanisms in the model.

References

- Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. *J Opt Soc Am* 2:284–299
- Alan WF (2021) A model for the origin of motion direction selectivity in visual cortex. *J Neurosci* 41(1):89–102. <https://doi.org/10.1523/JNEUROSCI.1362-20.2020>
- Alpert S, Galun M, Basri R, Brandt A (2012) Image segmentation by probabilistic bottom-up aggregation and cue integration. *IEEE Trans Pattern Anal Mach Intell* 34(2):315–327
- Andrews BW, Pollen DA (1979) Relationship between spatial frequency selectivity and receptive field profile of simple cells. *J Physiol (london)* 287:163–176

- Ayzenshtat I, Jackson J, Yuste R (2016) Orientation tuning depends on spatial frequency in mouse visual cortex. *eNeuro* 3(5):ENEURO.0217-16.2016. <https://doi.org/10.1523/ENEURO.0217-16.2016>
- Azzopardi G, Petkov N (2012) A CORF computational model of a simple cell that relies on LGN input outperforms the Gabor function model. *Biol Cybern* 106(3):177–189
- Barlow HB, Levick WR (1965) The mechanism of directionally selective units in rabbit's retina. *J Physiol* 178(3):477
- Berlin SJ, John M (2022) Spiking neural network based on joint entropy of optical flow features for human action recognition. *Vis Comput* 38:223–237. <https://doi.org/10.1007/s00371-020-02012-2>
- Beyeler M, Richert M, Dutt ND, Krichmar JL (2014) Efficient spiking neural network model of pattern motion selectivity in visual cortex. *Neuroinformatics* 12(3):435–454
- Brox T, Malik J (2010). Object segmentation by long-term analysis of point trajectories. In: ECCV (2010). Retrieved from <http://lmb.informatik.unifreiburg.de/Publications/2010/Bro10c>
- Cheng J, Tsai Y-H, Wang S, Yang M-H (2017) SegFlow: joint learning for video object segmentation and optical flow. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp 686–695
- Daugman JG (1985) Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J Opt Soc Am A* 2:1160–1169
- Escobar MJ, Kornprobst P (2012) Action recognition via bioinspired features: the richness of center-surround interaction. *Comput vis Image Underst* 116(5):593–605
- Hassenstein V, Reichardt W (1956) System theoretical analysis of time, sequence, and sign analysis of the motion perception of the snout-beetle *Chlorophanus*. *German z Naturforsch* 11:513–524
- Heeger DJ (1988) Optical flow using spatiotemporal filters. *Int J Comput Vision* 1:279–302. <https://doi.org/10.1007/BF00133568>
- Heitmann S, Ermentrout GB (2020) Direction-selective motion discrimination by traveling waves in visual cortex. *PLoS Comput Biol* 16(9):e1008164. <https://doi.org/10.1371/journal.pcbi.1008164>
- Hodgkin A, Huxley A (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* 117:500–544
- Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J Physiol* 160:106–154
- Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. *J Physiol-Lond* 195:215–243
- Jain SD., Xiong B, Grauman K (2017) FusionSeg: learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Jul. 2017, pp 2117–2126
- Jang W-D, Kim C-S (2017) Online video object segmentation via convolutional trident network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp 7474–7483. <https://doi.org/10.1109/CVPR.2017.790>
- Jones JP, Palmer LA (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol* 58:1233–1258
- Kamranian Z, Naghsh Nilchi AR, Sadeghian H, Rahmati M (2020) Joint motion boundary detection and CNN-based feature visualization for video object segmentation. *Neural Comput Appl* 32(32):4073–4091
- Koh YJ, Kim C-S (2017) Primary object segmentation in videos based on region augmentation and reduction. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Jul. 2017, pp 7417–7425
- Lee YJ, Kim J, Grauman K (2011) Key-segments for video object segmentation. In: 2011 International Conference on Computer Vision. IEEE, pp 1995–2002
- Li F, Kim T, Humayun A, Tsai D, Rehg JM (2013) Video segmentation by tracking many figure-ground segments. In: Computer vision (ICCV) international conference, IEEE, pp 2192–2199
- Lian Y, Almasi A, Grayden DB, Kameneva T, Burkitt AN, Meffin H (2021) Learning receptive field properties of complex cells in V1. *PLoS Comput Biol* 17(3):e1007957. <https://doi.org/10.1371/journal.pcbi.1007957>
- Lin C, Zhang Z, Hu Y (2022) Bio-inspired feature enhancement network for edge detection. *Appl Intell* 52:11027–11042. <https://doi.org/10.1007/s10489-022-03202-2>
- Marcelja S (1980) Mathematical description of the responses of simple cortical cells. *J Opt Soc Am* 70(11):1297–1300
- Ochs P, Malik J, Brox T (2014) Segmentation of moving objects by long-term video analysis. *IEEE Trans Pattern Anal Mach Intell* 36(6):1187–1200
- Papazoglou A, Ferrari V (2013) Fast object segmentation in unconstrained video. In: ICCV, 2013
- Perazzi F, Pont-Tuset J, McWilliams B, Van Gool L, Gross M, Sorkine-Hornung A (2016) A benchmark dataset and evaluation methodology for video object segmentation. In: Computer vision and pattern recognition (CVPR) conference. IEEE, pp 724–732.
- Pollen DA, Ronner SF (1981) Phase relationships between adjacent simple cells in the visual cortex. *Science* 212(4501):1409–1411. <https://doi.org/10.1126/science.7233231>
- Sekuler R, Watamaniuk SNJ, Blake R (2002) Motion perception. Steven's Handbook of Experimental Psychology 1:121–176
- Sobral A, Vacavant AA (2014) Comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Comput vis Image Understand (CVIU)* 122:4–21
- Song H, Wang W, Zhao S, Shen J, Lam K-M (2018) Pyramid dilated deeper ConvLSTM for video salient object detection. In Proc Eur Conf Comput vis (ECCV) 2018:715–731
- Spinéi A, Pellerin D, Fernandes D, Héroult J (2000) Fast hardware implementation of Gabor filter-based motion estimation. *Integr Comput Aided Eng* 7:67–77
- Taylor B, Karasev V, Soatto S (2015) Causal video object segmentation from persistence of occlusions. In: IEEE conference on computer vision and pattern recognition (CVPR), pp 4268–4276. <https://doi.org/10.1109/CVPR.2015.7299055>
- Tokmakov P, Schmid C, Alahari K (2019) Learning to segment moving objects. *Int J Comput Vision* 127(3):282–301
- Tokmakov P, Alahari K, Schmid C (2017) Learning video object segmentation with visual memory. In: Proceedings of the IEEE international conference on computer vision (ICCV), Oct. 2017, pp 4481–4490
- Townsend RG, Solomon SS, Martin PR, Solomon SG, Gong P (2017) Visual motion discrimination by propagating patterns in primate cerebral cortex. *J Neurosci* 37(42):10074–10084. <https://doi.org/10.1523/JNEUROSCI.1538-17.2017>
- Wang W, Covi E, Milozzi A, Farronato M, Ricci S, Sbandati C, Pedretti G, Ielmini D (2021) Neuromorphic motion detection and orientation selectivity by volatile resistive switching memories. *Adv Syst* 3:2000224. <https://doi.org/10.1002/aisy.202000224>
- Wang Y, Jodoin P-M, Porikli F, Konrad J, Benezeth Y, Ishwar P (2014) CDnet 2014: an expanded change detection benchmark dataset. In: Proceedings of the IEEE Workshop on Change Detection (CDW-2014) at CVPR-2014, pp 387–394

- Weinzaepfel P, Revaud J, Harchaoui Z, Schmid C (2015) Learning to detect motion boundaries. In : CVPR-IEEE conference on computer vision and pattern recognition. IEEE, pp 2578–2586
- Yedjour H, Meftah B, Lézoray AB (2017) Edge detection based on Hodgkin–Huxley neuron model simulation. *Cogn Process* 18(3):315–323
- Yedjour H, Yedjour D, Meftah B (2019) Motion boundary detection improved by bio-inspired approach. *Int J Commun Antenna Propag* 9(5):343–350
- Yedjour H, Meftah B, Yedjour D et al (2022) The Hodgkin–Huxley neuron model for motion detection in image sequences. *Neural Comput Appl* 34:1123–1133. <https://doi.org/10.1007/s00521-021-06446-0>
- Zhang Z, Pei Z, Tang Z, Gu F (2022) RoiSeg: an effective moving object segmentation approach based on region-of-interest with unsupervised learning. *Appl Sci* 12(5):2674. <https://doi.org/10.3390/app12052674>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.