Int J Comput Vis (2009) 82: 284–301
DOI 10.1007/s11263-008-0201-1

Action Recognition Using a Bio-Inspired Feedforward Spiking Network

Maria-Jose Escobar · Guillaume S. Masson · Thierry Vieville · Pierre Kornprobst

Received: 3 December 2007 / Accepted: 11 December 2008 / Published online: 12 February 2009

Abstract We propose a bio-inspired feedforward spiking network modeling two brain areas dedicated to motion (V1 and MT), and we show how the spiking output can be exploited in a computer vision application: action recognition. In order to analyze spike trains, we consider two characteristics of the neural code: mean firing rate of each neuron and synchrony between neurons. Interestingly, we show that they carry some relevant information for the action recognition application. We compare our results to Jhuang et al. (Proceedings of the 11th international conference on computer vision, pp. 1–8, 2007) on the Weizmann database. As a conclusion, we are convinced that spiking networks represent a powerful alternative framework for real vision applications that will benefit from recent advances in computational neuroscience.

Keywords Spiking networks · Bio-inspired model · Motion analysis · V1 · MT · Action recognition

M.-J. Escobar (✉) · T. Vieville · P. Kornprobst
INRIA Sophia-Antipolis, 2004 route des Lucioles,
06902 Sophia-Antipolis, France
e-mail: mjescoba@sophia.inria.fr

P. Kornprobst
e-mail: pkornp@sophia.inria.fr

G.S. Masson
Institut de Neurosciences Cognitives de la Méditerranée, CNRS,
Université d'Aix-Marseille, UMR6193, 31 Chemin Joseph
Aiguier, 13402 Marseille, France
e-mail: guillaume.masson@incm.cnrs-mrs.fr

1 Introduction

The output of a spiking neural network is a set of events, called spikes, defined by their occurrence times, up to some precision. Spikes represent the way that the nervous system choose to encode and transmit the information. But decoding this information, that is understanding the neural code, remains an open question in the neuroscience community.

There are several hypotheses on how neural code is formed, but there is a concensus on the fact that rate, i.e., the average spiking activity, is certainly not the only characteristic analyzed by the nervous system to interpret spike trains (see, e.g., some early ideas in Perkel and Bullock 1968).

For example, rank order coding could explain our performances in ultra-fast categorization. In Thorpe et al. (1996), the authors show that the classification of static images can be performed by the visual cortex within very short latencies: 150 ms and even faster. However, if one consider latency times of the visual stream (Nowak and Bullier 1997), such timings can only be explained by a specific architecture and efficient transmission mechanisms. As an explanation to the extraordinary performance of fast recognition, rank order coding was introduced (Thorpe 1990; Gautrais and Thorpe 1998): So one could interpret the neural code by considering the relative order of spiking times. The idea is that most highly excited neurons fire in average more but also faster. With this idea of rank order coding, the authors in fact developed a complete theory of information processing in the brain by successive waves of spikes (VanRullen and Thorpe 2002). Interestingly, the information carried by this first wave has been confirmed by some recent experiments in Gollisch and Meister (2008), where the authors show that certain retinal ganglion cells encode the spatial structure of a briefly presented image with the relative timing of their first spikes.

Another example of relevant spike train characteristics could be synchronies and correlations. The binding-by-synchronization hypothesis holds that neurons that respond to features of one object fire at the same time, but neurons responding to features of different objects do not necessarily. In vision, neuronal synchrony could thereby bind together all the features of one object and segregate them from features of other objects and the background. Several studies have supported this hypothesis by showing that synchrony between neuronal responses to the same perceptual object is stronger than synchrony between responses to different objects. Among the numerous observations in this direction, let us mention Neuenschwander et al. (1999), Fries et al. (2001), Biederlack et al. (2006).[1]

Back to vision application, there are, up to our knowledge, very few attempts to use spikes in real applications. Moreover, existing work concern static images. For example, let us mention two contributions about image recognition (see, e.g., Thorpe 2002 as an application of rank order coding) or image segmentation (see, e.g., Wang and Terman 1995 modeled by oscillator networks), which refer respectively to the two characteristics mentioned above: rank and synchronies.

But analyzing spikes means being able to correctly generate them, which is a difficult issue. At the retina level, some models exist such as Thorpe (2002), Wohrer and Kornprobst (2008) with different degrees of plausibility. When we go deeper in the visual system, this requires even more simplifications since it is not possible to render the complexity of all the successive areas and neural diversity. Here, we propose a simplified spiking model of the V1/MT architecture with one goal: Can the spiking output be exploited in order to extract some content like the action taking place?

The article is organized as follows. Section 2 describes the state-of-the-art in action recognition, from computer vision approaches to bio-inspired ones. Section 3 describes the framework of spiking networks in more details. Section 4 presents our two-stages motion model. Section 5 indicates how the resulting spike trains can be analyzed focusing on two characteristics: the rate and the synchrony between spike trains. In Sect. 6, we clearly leave the bio-inspired modeling to present how our motion maps could be applied in the action recognition application. In this computational part, a supervised classification protocol is proposed and we show how feature vectors can be defined from spike trains (motion maps). We also compare our results with Jhuang et al. (2007) with the same Weizmann database. Finally, the discussion is in Sect. 7, where some perspectives mainly related with the richness of information contained in spike trains are also present.

[1]Note that the link between synchrony and segmentation is still controversial. Results could sometimes be explained by other mechanisms taking over the segmentation by synchrony (see, e.g., Roelfsema et al. 2004).

2 State of the Art in Action Recognition

2.1 How Computer Vision Does?

Action recognition has been addressed in the computer vision community with many ideas and concepts. Proposed approaches often rely on simplified assumptions, scene reconstructions, or motion analysis and representation. For example, some approaches exploit periodicity of motion (Collins et al. 2002; Cutler and Davis 2000; Polana and Nelson 1997; Seitz and Dyer 1997), or model and track body parts (Shah and Jain 1997; Gavrila and Davis 1996; Gavrila 1999), or consider generic human model recovery (Goncalves et al. 1995; Hogg 1983; Rohr 1994), or consider the shape of the silhouette evolution across time (Bobick and Davis 2001; Mokhber et al. 2008; Wang and Suter 2007; Blank et al. 2005).

An important category of approaches in computer vision is based on the motion information. For example, it was shown that a rough description of motion (Efros et al. 2003) or the global motion distribution (Zelnik-Manor and Irani 2001) can be successfully used to recognize actions. Local motion cues are also widely used. For example, in Laptev et al. (2007), the authors propose to use event-based local motion representations (here, spatial-temporal chunks of a video corresponding to $2D + t$ edges) and template matching. This idea extracting spatial-temporal features was proposed in several contributions such as Dollar et al. (2005), and then Niebles et al. (2006), Wong et al. (2007), using the notion of cuboids. Another stream of approaches was inspired by the work by Serre (2006), first applied to object recognition (Serre et al. 2005; Mutch and Lowe 2006) and then extended to action recognition (Sigala et al. 2005; Jhuang et al. 2007).

2.2 How the Brain Does?

Action recognition has been addressed in psychophysics where remarkable advances have been made in the understanding of human action perception (Blake and Shiffrar 2007). The perception of human action is a complex task that combines not only the visual information, but additional aspects as social interactions or motor system contributions. From several studies in psychophysics, it has been shown that our ability to recognize human actions does not need necessarily a real moving scene as input. In fact, we are also able to recognize actions when we watch some point-light stimuli corresponding to joint positions for example. This kind of simplified stimuli, known as *biological motion*, was

highly used in the psychophysics community in order to obtain a better understanding of the underlying mechanism involved. The neural mechanisms, processing *form* or *motion* taking part of biological motion recognition, remain unclear. On the one hand, Beintema and Lappe (2002) suggests that biological motion can be derived from dynamic *form* information of body postures and without local image motion. On the other hand, Casile and Giese (2003) proposes a new type of point-light stimulus suggesting, in this case, that only the *motion* information is enough and the detection of specific spatial arrangements of *opponent-motion features* can explain our ability to recognize actions. Finally, Casile and Giese (2005) showed that biological motion recognition can be done with a coarse spatial location of the mid-level optic flow features.

This dichotomy between motion and form finds some neural basis in the brain architecture and it has been confirmed by fMRI studies (Michels et al. 2005). A simplified representation of the visual processing is that there exists two distinct pathways: the *dorsal* stream (motion pathway) with areas such as V1, MT, MST, and the *ventral* stream (form pathway) with areas such as V1, V2, V4. Both of them seem to be involved in the biological motion analysis.

2.3 Towards a Bio-Inspired System

Based on the *dorsal* and *ventral* streams of visual processing, the seminal work done by Giese and Poggio (2003) evaluates both pathways in biological motion recognition. The analysis is done separately for each pathway and never combined. Afterwards, using only the information of the *dorsal* stream, Sigala et al. (2005) proposed a biological motion recognition system using a neurally plausible memory-trace learning rule.

Starting also from the work done by Giese and Poggio (2003) and Serre et al. (2005), Mutch and Lowe (2006), the recent work presented by Jhuang et al. (2007) shows a hierarchical feedforward architecture, that the authors mapped to the cortical architecture, essentially V1 (with simple and complex cells). Their approach is composed by a sequel of local operations, pooling, max operators, and finally features comparisons. Thanks to this analogy, the authors claimed that their approach was bio-inspired.

In this article, our goal is to propose a bio-inspired model for real video analysis (see the block diagram in Fig. 3). By bio-inspired, it is meant here that our model will communicate through discrete events (i.e., spikes) and its architecture will be inspired by motion-related brain areas V1 and MT. As far as categorization is concerned, we will use some standard algorithms with no link to biology.

By considering motion only, our model is related to several other computer vision models which are only motion-based and do not consider form, see for example Efros et al. (2003), Zelnik-Manor and Irani (2001), Laptev et al. (2007).

But, by considering motion only, the bio-inspiration of the model is clearly limited and, in term of performance, we can expect that we won't be able to deal with any kind of videos (including scale and rotation invariance, complex backgrounds, multiple persons, etc.). As we mentioned in Sect. 2.2, we do not consider the other brain areas involved in human motion analysis, specially interactions with the form pathway but also other motion processing areas. Another simplification comes from the structure of the proposed architecture which is a feedforward architecture, similarly to Jhuang et al. (2007). Finally, attention mechanisms are also here ignored. These simplifications certainly account for the limitations of what pure motion-based models can handle.

Having in mind those limitations, our goal is to propose a competitive model based on a bio-inspired spiking motion model.

3 Spiking Networks?

3.1 Spikes

The elementary units of the central nervous system are neurons. Neurons are highly connected to each other forming networks of spiking neurons. The neurons collect signals from other neurons connected to it (presynaptic neurons), do some non-linear processing, and if the total input exceeds a threshold, an output signal is generated. The output signal generated by the neuron is what is known as *spike* or *action-potential*: it is a short electrical pulse that can be physically measured and has an amplitude of about 100 mV and a typical duration of 1–2 ms. A chain of spikes emitted by one neuron is called *spike train*. The neural code corresponds to the pattern of neuronal impulses (see also Gerstner and Kistler 2002).

Although spikes can have different amplitudes, durations or shapes they are typically treated as discrete events. By discrete events, we mean that in order to describe a spike train, one only needs to know the succession of emission times:

$$\mathcal{F}_i = \{\ldots, t_i^n, \ldots\}, \quad \text{with } t_i^1 < t_i^2 < \cdots < t_i^n < \cdots, \tag{1}$$

where t_i^n corresponds to the nth spike of the neuron of index i.

3.2 The Neural Code

The set of all spikes from a set of neurons in a period of time is generally represented in a graph called *raster plot*, as illustrated in Fig. 1. Many hypothesis were proposed on the way that this pattern of neuronal impulses is analyzed by the nervous system. The most intuitive is to estimate the mean

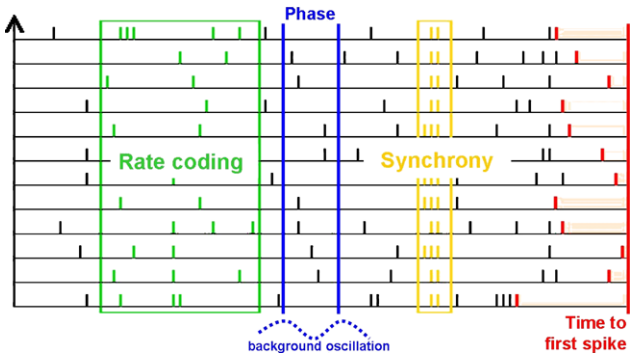

Fig. 1 Example of a raster plot and illustration of some different methods to analyze the neural code (see text for more details). Each *horizontal line* can be interpreted as an axon in which we see spikes traveling (from *left* to *right*)

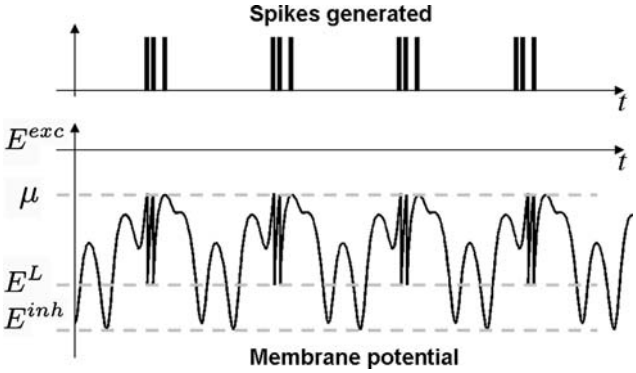

Fig. 2 Temporal evolution of the membrane potential of a neuron and its corresponding spikes generated. A spike is generated when the membrane potential exceeds the threshold μ ($E^L < \mu < E^{exc}$). When the spike is emitted membrane potential returns to its resting value E^L

firing rate over time, which is the average number of spikes inside a temporal window (rate coding).

But what makes the richness of such a representation is the many other ways to analyze spiking networks activity, and that is the idea we wish to push forward for this framework. Methods include rate coding over several trials or over population of neurons, time to first spike, phase, synchronization and correlations, interspike intervals distribution, repetition of temporal patterns, etc.

In spite of these numerous hypotheses, "decoding" the neural code remains an open question in neuroscience (Victor and Purpura 1996; Rieke et al. 1997; Fellous et al. 2004), which is far beyond the scope of this work. Different metrics or weaker similarity measures between two spike trains have also been proposed (see Cessac et al. 2008 for a review).

Here our goal will be to illustrate how the analysis of simulated spike trains can be successfully used in a given vision application. To do this, we will use the mean firing rate and a measure of the synchrony between spike trains (see Sect. 5).

3.3 Spiking Neuron Model

Many spiking neuron models have been proposed in the literature. They differ by their biological plausibility and their computational efficiency (see Izhikevich 2004 for a review).

In this article, a spiking neuron will be modeled as a conductance-driven integrate-and-fire neuron (Wielaard et al. 2001; Destexhe et al. 2003). Considering a neuron i, defined by its membrane potential $u_i(t)$, the integrate-and-fire equation is given by:

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

with the spike emission process: the neuron i will emit a spike when the normalized membrane potential of the cell reaches threshold $u_i(t) = \mu$, then $u_i(t)$ is reinitialized to its resting potential E^L. The neuron membrane potential $u_i(t)$ will evolve according to inputs through either conductances ($G_i^{exc}(t)$ or $G_i^{inh}(t)$) or external currents ($I_i(t)$).

Each variable has indeed some biological interpretation (see Wielaard et al. 2001 for details). $G_i^{exc}(t)$ is the normalized excitatory conductance directly associated with the presynaptic neurons connected neuron i. The conductance g^L is the passive leaks in the cell's membrane. $I(t)$ is an external input current. Finally, $G_i^{inh}(t)$ is an inhibitory normalized conductance dependent on, e.g., lateral connections or feedbacks from upper layers. The typical values for the reverse potentials E^{exc}, E^{inh} and E^L are 0 mV, -80 mV and -70 mV, respectively (see Fig. 2 for an illustration).

4 Bio-Inspired Motion Analysis Model

Several bio-inspired motion processing models have been proposed in the literature (Tsotsos et al. 2005; Nowlan and Sejnowski 1995; Rust et al. 2006; Simoncelli and Heeger 1998; Grzywacz and Yuille 1990), those models were validated considering certain properties of primate visual systems, but none of them has been tested in a real application such as AR. More complex motion processing models combining not only motion information but also connections from different brain areas can be found in, e.g., Berzhanskaya et al. (2007), Bayerl and Neumann (2007).

Visual motion analysis has been studied during many years in several fields such as physiology and psychophysics. Many of those studies tried to relate our perception with the activation of the primary visual cortex V1 and extrastriate visual areas as MT/MST. It seems that the area most involved in motion processing is MT, who receives input motion afferent mainly from V1 (Felleman and Essen 1991).

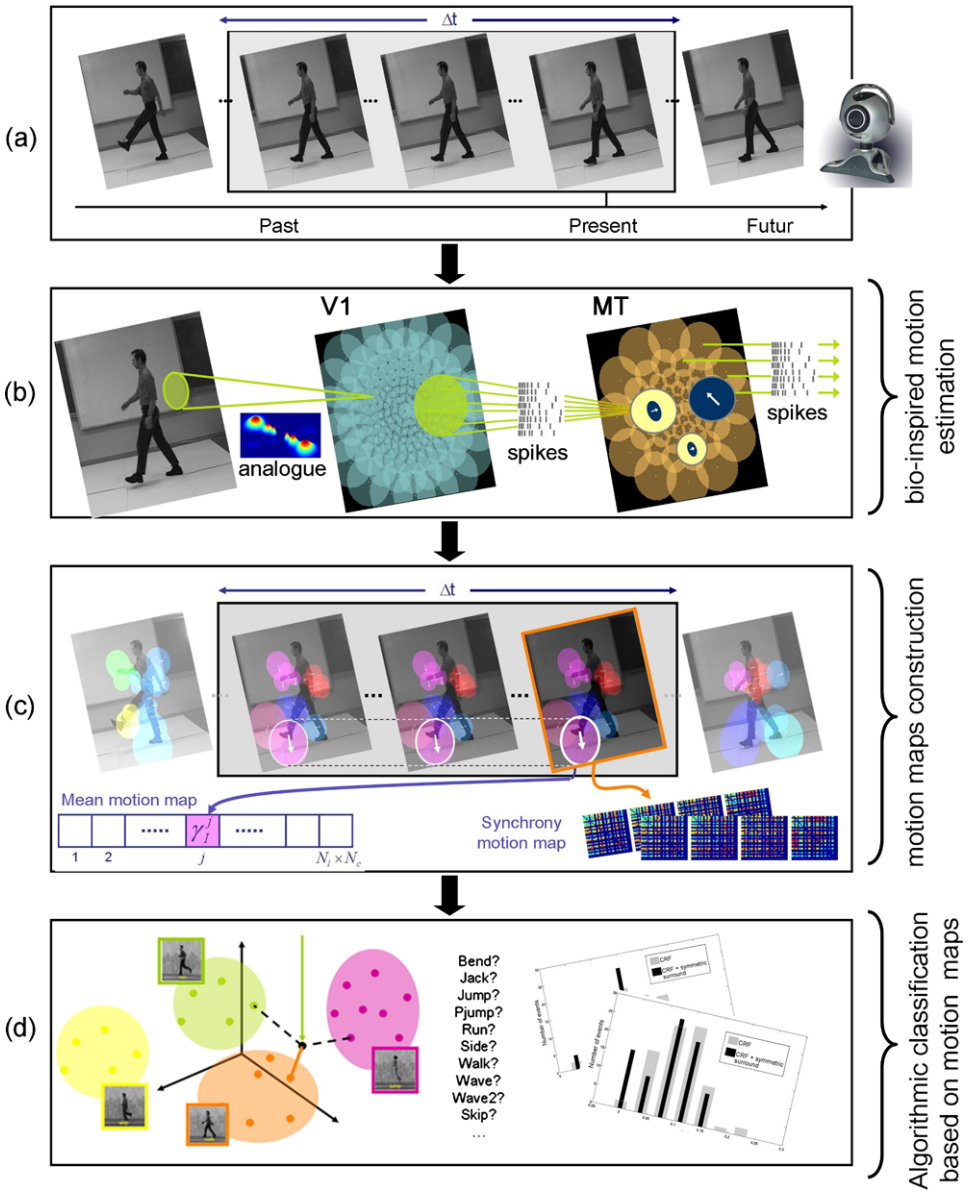

Fig. 3 Block diagram showing the different steps of our approach from the input image sequence as stimulus until its final classification. (**a**) We use real video sequence as input, the input sequences are preprocessed in order to have contrast normalization and centered moving stimuli. To compute the motion maps representing the input image we consider a sliding temporal window of length Δt. (**b**) Directional-selectivity filters are applied over each frame of the input sequence in a log-polar distribution grid obtaining spike trains as V1 output. These spike train feed the spiking MT which integrates the information in space and time. (**c**) The motion maps (*mean motion map* and *synchrony motion map*) are constructed calculating either the mean firing rates of MT spike trains or a synchrony map of the spikes trains generated by MT cells. Both motion maps are created considering the spike trains inside the sliding temporal window of length Δt. (**d**) Classification stage where, starting from the motion maps and the training set content, a final action is assigned to the input image sequence

Several work such as (Conway and Livingstone 2003; De Valois et al. 2000) have established experimentally the spatial-temporal behavior of simple/complex V1 cells and MT cells, in the form of activation maps (see Fig. 4(a)). With different methods, both have found directionally selective cells sensitive to motion for a certain speed and direction. More properties about MT can be found in the survey (Born and Bradley 2005).

Here we propose spiking neuron models for V1 and MT cells, defining also the connections between the cells of these two areas. The model for the spiking V1 neurons is inspired by Adelson and Bergen (1985) (Sect. 4.1). Then, our contribution comes with the spiking MT cells simulator, their interactions and their connections with the underlying V1 level (Sect. 4.2).

4.1 V1 layer: Local Motion Detectors

The primary visual cortex V1 corresponds to the first area involved on the visual processing in the brain. This area contains specialized cells for motion processing (motion detectors).

Our spiking V1 model is built with a bank of energy motion detectors as a local motion estimation. The model is divided in two stages: The analog processing, where the motion information is extracted, and the spiking layer, where each neuron is modeled as a spiking entity whose inputs are the information obtained in the previous stage. The analog processing is done through energy filters which is a reliable and biologically plausible method for motion information analysis (Adelson and Bergen 1985). Each energy motion

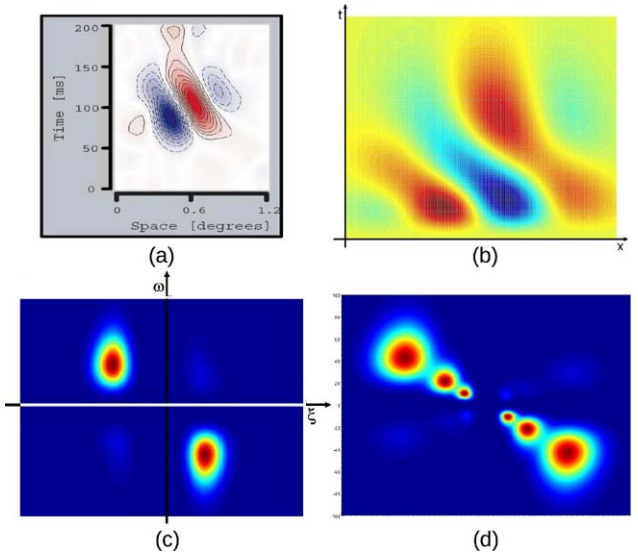

Fig. 4 (**a**) Example of spatial-temporal map of one directionally-selective V1 simple cell (De Valois et al. 2000). (**b**)–(**c**) Space-time diagrams for $F^a(x,t)$ and its power spectrum $|\tilde{F}^a(\xi,\omega)|^2$. Both graphs were constructed considering just one spatial dimension x. (**b**) It is possible to see directionality-selective obtained after the linear combination of cells. It is important also to observe the similarities with the biological activation maps measured by (De Valois et al. 2000) (**a**). (**c**) Spatio-temporal energy spectrum of the directional-selective filter $F^a(x,t)$. The slope formed by the peak of the two blobs is the speed tuning of the filter. (**d**) Different filters tuned at the same speed used to tile the spatial-temporal frequency space

detector will emulate a complex cell, which is formed by a nonlinear combination of V1 simple cells (see Hubel and Wiesel 1962 for V1 cells classification).

4.1.1 V1 Cells Model

In Grzywacz and Yuille (1990), the authors showed that several properties of simple/complex cells in V1 can be described with energy filters and in particular using Gabor filters. The individual energy filters are not velocity tuned, however it is possible to use a combination of them in order to have a velocity estimation.

Simple cells are characterized with linear receptive fields where the neuron response is a weighted linear combination of the input stimulus inside its receptive field. By combining two simple cells in a linear manner it is possible to get direction-selective neurons, that is, simple cells selective for stimulus orientation and spatial frequency.

The direction-selectivity (DS) refer to the property of a neuron to respond selectively to the direction of the motion of a stimulus. The way to model this selectivity is to obtain receptive fields oriented in space and time. Let us define the following spatial-temporal oriented simple cells

$$\begin{aligned} F^a_{\theta,f}(x,y,t) &= F^{odd}_{\theta}(x,y)H_{fast}(t) \\ &\quad - F^{even}_{\theta}(x,y)H_{slow}(t), \\ F^b_{\theta,f}(x,y,t) &= F^{odd}_{\theta}(x,y)H_{slow}(t) \\ &\quad + F^{even}_{\theta}(x,y)H_{fast}(t), \end{aligned} \tag{3}$$

where simples cell defined in (3) are spatially oriented in the direction θ, and spatio-temporal oriented to $f = (\bar{\xi},\bar{\omega})$, where $\bar{\xi}$ and $\bar{\omega}$ are the spatial and temporal maximal responses, respectively (see Fig. 4(b)). The spatial parts $F^{odd}_{\theta}(x,y)$ and $F^{even}_{\theta}(x,y)$ of each conforming simple cell are formed using the first and second derivative of a Gabor function spatially oriented in θ. The temporal contributions $H_{fast}(t)$ and $H_{slow}(t)$ come from the substraction of two Gamma functions with a difference of two in their respective orders.

$$\begin{aligned} H_{fast}(t) &= T_{3,\tau}(t) - T_{5,\tau}(t), \\ H_{slow}(t) &= T_{5,\tau}(t) - T_{7,\tau}(t), \end{aligned} \tag{4}$$

and $T_{\eta,\tau}(t)$ is defined by

$$T_{\eta,\tau}(t) = \frac{t^{\eta}}{\tau^{\eta+1}\eta!}\exp\left(-\frac{t}{\tau}\right), \tag{5}$$

which models the series of synaptic and cellular delays in signal transmission, from retinal photoreceptors to V1 afferent serving as a plausible approximation of biological findings (Robson 1966). The biphasic shape of $H_{fast}(t)$ and $H_{slow}(t)$ could be a consequence of the combination of cells of M and P pathways (De Valois et al. 2000; Saul et al. 2005) or be related to the delayed inhibitions in the retina and LGN (Conway and Livingstone 2003).

Thinking about the design of our filter bank, we are interested in the estimation of the spatial-temporal bandwidths of our V1 simple cell model. For simplicity and without loss of generality, we will use just one spatial dimension x and focus on the function $F^a_{\theta,f}(x)$ (from now on noted as $F^a(x)$). The power spectrum $\tilde{F}^a(\xi,\omega)$ of $F^a(x)$ is shown in Fig. 4(c). The quotient between the highest temporal frequency activation and the highest spatial frequency is the speed of the filter. It is also possible to see a small activation for the same speed but in the opposite motion direction. The activation in the anti-preferred direction tuning is an effect also seen in real V1-MT cells data (Snowden et al. 1991), where V1 cells have a weak suppression in anti-preferred direction (30%) compared with MT cells (92%).

As we can see, for a given speed, the filter covers a specified region of the spatial-temporal frequency domain. So, the filter will be able to see the motion for a stimulus whose spatial frequency is inside the energy spectrum of the filter. To pave all the space in a homogeneous way, it is necessary

to take more than one filter for the same spatial-temporal frequency orientation. A diagram with the filter bank tuned at the same speed can be seen in Fig. 4(d).

In our case, the causality of $H_{fast}(t)$ and $H_{slow}(t)$ generates a more realistic model than the one proposed by Simoncelli and Heeger (1998), where a Gaussian is proposed as a temporal profile which is non-causal and inconsistent with V1 physiology. Using the temporal profiles defined in (4)—unlike Simoncelli and Heeger (1998) where the choice of a Gaussian as temporal profile is computationally convenient—the search of an analytic expression for $|\tilde{F}^a(\xi,\omega)|^2$ is not an easy task, specially due to the non-separability of $F^a(x,t)$.

Complex cells are also direction-selective neurons, however they include other characteristics that cannot be explained by a linear combination of the input stimulus. Their responses are relatively independent of the precise stimulus position inside the receptive field, which suggest a combination of a set of V1 simple cells responses. The complex cells are also invariant to contrast polarity which indicates a kind of rectification of their ON-OFF receptive field responses.

Based on Adelson and Bergen (1985), we define the ith V1 complex cells, located at $\mathbf{x}_i = (x_i, y_i)$, with spatial orientation θ_i and spatio-temporal orientation $f_i = (\bar{\xi}_i, \bar{\omega}_i)$ as

$$C_{\mathbf{x}_i,\theta_i,f_i}(t) = \left[\left(F^a_{\theta_i,f_i} * L\right)(\mathbf{x}_i,t)\right]^2 + \left[\left(F^b_{\theta_i,f_i} * L\right)(\mathbf{x}_i,t)\right]^2 \quad (6)$$

where the symbol $*$ represents the spatio-temporal convolution, and $F^a_{\theta_i,f_i}$ and $F^b_{\theta_i,f_i}$ are the simple cells defined in (3). This definition gives a cell response independent of stimulus contrast sign and constant in time for a drifting sinusoidal as stimulus.

4.1.2 Foveated Organization of V1

Given V1 cells modeled by (6), we consider N_L layers of V1 cells (see Fig. 5). Each layer is built with V1 cells with the same spatial-temporal frequency tuning $f_i = (\bar{\xi}_i, \bar{\omega}_i)$ and N_{or} different orientations. The related spatial-temporal frequency, and the physical position of the cell inside V1 define its receptive field. All the V1 cells belonging to one layer, with receptive fields centered in the position (x_i, y_i), form what we call a *column*. One *column* has as many elements as the number of orientations defined N_{or}. See Fig. 5 for an illustration.

The centers of the receptive fields are distributed along a radial log-polar scheme with a foveal uniform zone. The related one-dimensional density $d(r)$, depending of the eccentricity r, is taken as

$$d(r) = \begin{cases} d_0 & \text{if } r \le R_0, \\ d_0 R_0 / r & \text{if } r > R_0, \end{cases} \quad (7)$$

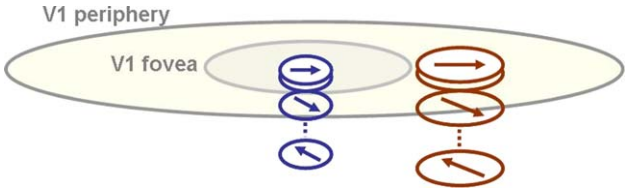

Fig. 5 Diagram with the architecture of one V1 layer. There are two different regions in V1, the fovea and periphery. Each element of the V1 layer is a column of N_{or} V1 cells, where N_{or} corresponds to the number of orientations

The cells with an eccentricity r less than R_0 have an homogeneous density and their receptive fields refer to the retina fovea (*V1 fovea*). The cells with an eccentricity greater than R_0 have a density depending on r and receptive fields lying outside the retina fovea (*V1 periphery*).

4.1.3 Analogous to Spike Conversion

The response of the V1 complex cell—formed as a combination of the V1 simple cells defined in (3)—is analogous. To transform the analogous response to a spiking response, the cell will be modeled as the conductance-driven integrate-and-fire neuron described in (2).

So, let us consider a spiking V1 complex cell i whose center is located in $\mathbf{x}_i = (x_i, y_i)$ of the visual space. For this neuron, $G_i^{exc}(t)$ is the normalized excitatory conductance directly associated with the pre-synaptic neurons connected to V1 cells. The external input current $I_i(t)$ is here associated with the analogous V1 complex cell response. Finally, $G_i^{inh}(t)$ is an inhibitory normalized conductance dependent on the spikes of neighboring cells of the same V1 layer.

We model the external *input current* $I_i(t)$ of the ith cell in (2) as the analog response

$$I_i(t) = k_{exc}\Lambda_i(t)C_{\mathbf{x}_i,\theta_i,f_i}(t), \quad (8)$$

where k_{exc} is an amplification factor, $C_{\mathbf{x}_i,\theta_i,f_i}$ refers to the complex cell response defined in (6) and Λ_i groups the interactions within V1 cells provoked by *local and global divisive* inhibitions (Simoncelli and Heeger 1998).

The *inhibitory conductance* $G_i^{inh}(t)$ and the *excitatory conductance* $G_i^{exc}(t)$ of (2) are not considered in this first approach, equally than the leak conductance g^L.

4.2 MT Layer: Global Motion Analysis

4.2.1 MT Cells Model

Our model is a feedforward spiking network where each entity or node is a MT cell. Each MT cell i can be modeled as conductance-driven integrate-and-fire neuron described in (2).

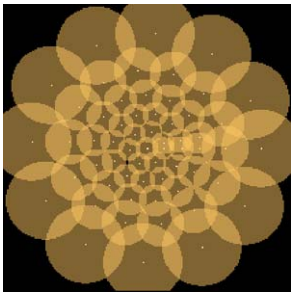

Fig. 6 Sample of log-polar architecture used for a MT layer. The cell distribution law is divided into two zones, a homogeneous distribution in the center with a certain radius and then a periphery where the density of cells decays with the eccentricity

The neuron i is a part of a spiking network where the input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ are obtained considering the activity of all the pre-synaptic neurons connected to it. For example, if a pre-synaptic neuron j has fired a spike at time $t_j^{(f)}$, this spike reflects an input conductance to the post-synaptic neuron i during a time course $\alpha(t - t_j^{(f)})$. In our case the pre-synaptic neurons refer to the V1 outputs (see Fig. 7). According to this, the total input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ of the post-synaptic neuron i are expressed as

$$\begin{aligned} G_i^{exc}(t) &= \sum_j w_{ij}^{+} \sum_f \alpha(t - t_j^{(f)}), \\ G_i^{inh}(t) &= \sum_j w_{ij}^{-} \sum_f \alpha(t - t_j^{(f)}) \end{aligned} \tag{9}$$

where the factor w_{ij}^{+} (w_{ij}^{-}) is the efficacy of the positive (negative) synapse from neuron j to neuron i (see Gerstner and Kistler 2002 for more details). The time course $\alpha(s)$ of the post-synaptic current in (9) can be modeled as an exponential decay with time constant τ_s as follows

$$\alpha(s; \tau_s) = \left(\frac{s}{\tau_s}\right) \exp\left(-\frac{s}{\tau_s}\right). \tag{10}$$

Each MT cell has a receptive field made from the convergence of afferent V1 complex cells. The V1 afferent are the pre-synaptic neurons j in (9). Those inputs will be excitatory or inhibitory depending on the characteristic and shape of the corresponding MT receptive fields (Xiao et al. 1997, 1995). Half of MT surface is assigned to process the information coming from the central 15° of the visual field, which receptive field size of a MT cell inside this region is about 4–6 times bigger than the V1 receptive field (Mestre et al. 2001).

The MT cells are distributed in a log-polar architecture, with a homogeneous area of cells in the center and a periphery where the density decreases with the distance to the center of focus. While the density of cells decreases with the eccentricity, the size of the receptive fields increases preserving its original shape. Figure 6 shows an example of the log-polar distribution of MT cells.

Different layers of MT cells conform our model. Each layer is built with MT cells of the same characteristics, same

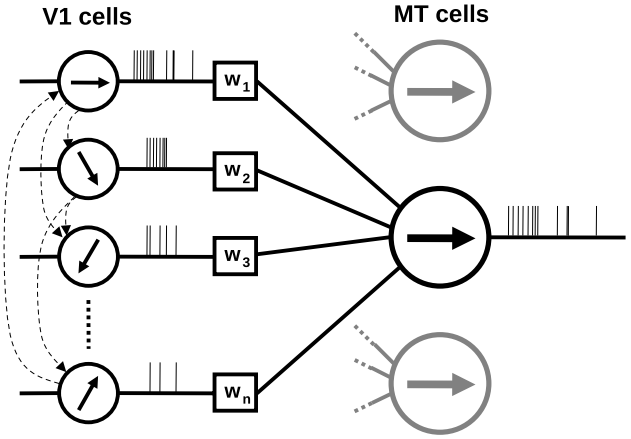

Fig. 7 Architecture of the feedforward spiking network to model MT. Each MT cell receives as input the afferent V1 cells

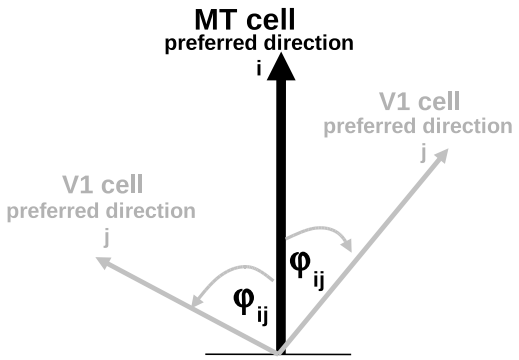

Fig. 8 The connection weights between V1 and MT cells are modulated by the cosine of the angle φ_{ij} between the preferred direction of ith MT cell and the preferred direction of jth V1 cell

speed and direction tuning. The group of V1 cells connected with a MT cell and their respective connection weights depend on the tuning values desired for the MT cell. The criteria of selection is to consider all the V1 cells inside the MT receptive field with an absolute difference of motion direction-selectivity respect MT cell no more than $\pi/2$ radians. The weight associated to the connection between pre-synaptic neuron j and post-synaptic neuron i is proportional to the angle φ_{ij} between the two preferred motion direction-selectivity (see Fig. 8). The connection weight w_{ij} between the jth V1 cell and the ith MT cell is given by

$$w_{ij} = \begin{cases} k_c w_{cs}(\mathbf{x}_i - \mathbf{x}_j)\cos(\varphi_{ij}) & \text{if } 0 \le \varphi_{ij} \le \frac{\pi}{2}, \\ 0 & \text{if } \frac{\pi}{2} < \varphi_{ij} < \pi, \end{cases} \tag{11}$$

where k_c is an amplification factor, α_{ij} is the absolute angle between the preferred ith MT cell direction and the preferred jth V1 cell direction. $w_{cs}(\cdot)$ is the weight associated to the difference between the center of MT cell $\mathbf{x}_i = (x_i, y_i)$ and the V1 cell center position $\mathbf{x}_j = (x_j, y_j)$. The value of $w_{cs}(\cdot)$ depends on the shape of the receptive field associated

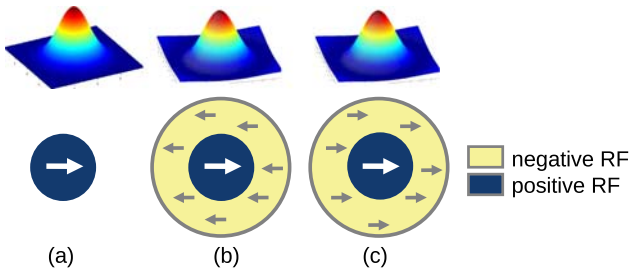

Fig. 9 Center-surround interactions modeled in the MT cells. The classical receptive field (CRF) is modeled through a Gaussian (**a**). The two receptive fields with inhibitory surround (**b**), (**c**) are modeled with a Difference-of-Gaussians. The cells with inhibitory surround have either antagonistic direction tuning between the center and surround or the same direction tuning

to the MT cell (see Sect. 4.2.2). The sign of w_{cs} will set the values of w_{ij}^{+} (if $w_{cs} > 0$) and w_{ij}^{-} (if $w_{cs} < 0$).

4.2.2 Receptive Fields: Geometry and Interactions

The geometry and interactions of the MT receptive fields is far from being completely understood. Half of MT neurons have asymmetric surrounds introducing anisotropies in the processing of the spatial information (Lui et al. 2007). The neurons with asymmetric surrounds seem to be involved in the encoding of important surfaces features, such as slant and tilt or curvature (Buracas and Albright 1996; Xiao et al. 1997). The surround geometry and its interactions with the classical receptive field could be the main responsible of dynamic effects seen in MT cells, as e.g., switching from component to pattern behavior (Smith et al. 2005) or showing a direction reversal from preferred to antipreferred direction tuning (Perge et al. 2005).

Regarding organization and center-surround interactions, Born (2000) shows two different types of cells, the pure integrative cell, where only the activation of the classical receptive field (CRF) is taken into account, and the cell with an antagonistic surround who modulates the activation of the CRF. The direction tuning of the surround is always broader than the center. The direction tuning of the surround compared with the center tends to be either the same or opposite, but rarely orthogonal. The antagonistic surrounds are insensitive to wide-field motion but sensitive to local motion contrast. By the other hand, the cells with only CRF are best sensitive to wide-field motion.

Considering the results found by Born (2000), we include three types of MT center-surround interactions in our model. Our claim is that the antagonistic surrounds contain key information about the motion characterization, which could highly help the motion recognition task. We propose a cell with only the activation of its classical receptive field (CRF) and two cells with inhibitory surrounds as shown in Fig. 9.

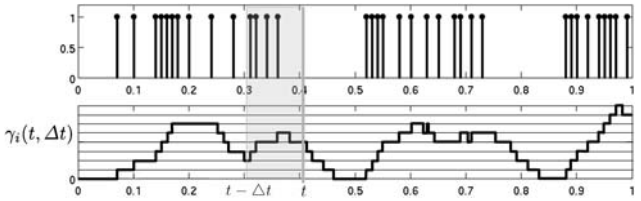

Fig. 10 Mean firing rate

5 Spike Train Analysis

Given the spiking output from the network presented in Sect. 4, we present in this section two methods to describe its activity: the mean firing rate of a spike train and a synchrony measure between pairs of spike trains. These two quantities will be then directly used in the action recognition application described in Sect. 6.

Remark 1 Note that we do not consider high-level statistics of spike trains (Rieke et al. 1997), since this requires large ergodic spike sequences, whereas we are interested here in recognition tasks from non-stationary spike trains generated by some dynamic input. Also, we do not considered spike-train metrics in the strict sense (Victor and Purpura 1996), since we do not have enough knowledge from the biology to predict the firing times in a deterministic way. For the same reason, we do not compare, here, spike patterns (Fellous et al. 2004). In fact, these aspects are perspectives of the present work.

5.1 Mean Firing Rate of a Neuron

Let us consider a spiking neuron i. The spike train $\mathcal{F}_i$ associated to this neuron is defined in (1). We defined the *windowed firing rate* $\gamma_i(\cdot)$ by

$$\gamma_i(t, \Delta t) = \frac{\eta_i(t - \Delta t, t)}{\Delta t}, \tag{12}$$

where $\eta_i(\cdot)$ counts the number of spikes emitted by neuron i inside the sliding time window $[t - \Delta t, t]$ (see Fig. 10 and e.g., Dayan and Abbott 2001).

5.2 Synchrony between Two Spike Trains

Let us consider the recent approach proposed by Kreuz et al. (2007) to estimate the similarity between two spike trains, as a measure of synchrony. The authors proposed to compute first the interspike interval (ISI) instead of the spike as a basic element of comparison. The use of ISI has the main advantage to be parameter-free and self-adaptive, so that there is no need to fix a time scale beforehand ("binless") or to fit any parameter.

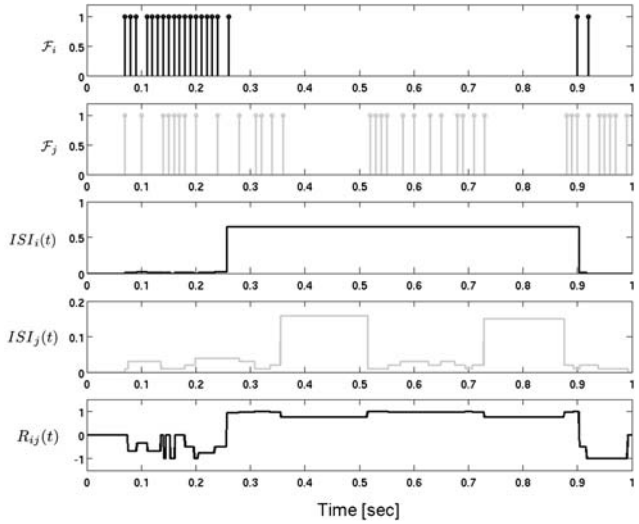

Fig. 11 Synchrony between the spike trains of a pair of neurons. $\mathcal{F}_i$ and $\mathcal{F}_j$ are the spike trains of MT neurons i and j, respectively. The respective *ISI* representations defined in (13) are shown as $ISI_i(t)$ and $ISI_j(t)$. Finally, the ratio between $ISI_i(t)$ and $ISI_j(t)$ is shown as $R_{ij}(t)$

So, for the neuron i the ISI representation $ISI_i(t)$ is given by

$$ISI_i(t) = \min(t_i^{(f)} \mid t_i^{(f)} > t) - \max(t_i^{(f)} \mid t_i^{(f)} < t), \qquad (13)$$

for $t_i^{(f)} < t$. Considering the ISI representation of two neurons i and j, the next step is to calculate the ratio $R_{ij}(t)$ defined as

$$R_{ij}(t) = \begin{cases} \frac{ISI_i(t)}{ISI_j(t)} - 1 & \text{if } ISI_i(t) \leq ISI_j(t), \\ -(\frac{ISI_j(t)}{ISI_i(t)} - 1) & \text{otherwise.} \end{cases} \qquad (14)$$

$R_{ij}(t)$ will be zero in case of completely synchrony between $ISI_i(t)$ and $ISI_j(t)$. In the cases of a big difference between the two ISI representation, $R_{ij}(t)$ will tend to ± 1 (see Fig. 11).

Having the ratio $R_{ij}(t)$ it is possible to calculate, for a finite time Δt, a measure of spike train distance $\zeta_{ij}(t; \Delta t)$, which is an estimator of the spike train synchrony between neurons i and j

$$\zeta_{ij}(t; \Delta t) = \frac{1}{\Delta t} \int_{t-\Delta t}^{t} |R_{ij}(s)| ds. \qquad (15)$$

Remark 2 Completely synchrony $\zeta_{ij}(\cdot) = 0$ was assigned for two cells not emitting spikes, while the maximal desynchronization $\zeta_{ij}(\cdot) = 1$ was assigned to the case where only one cell fired spikes.

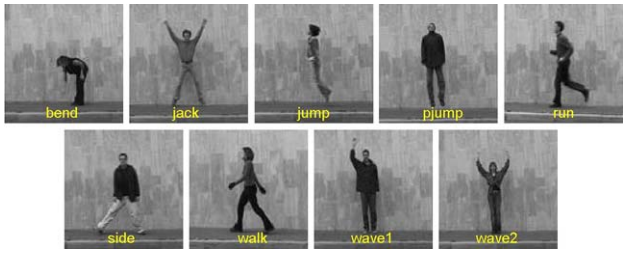

Fig. 12 Sample frames of each of the nine actions conforming the Weizmann database. The actions are: bending (bend), jumping-jack (jack), jumping-forward-on-two-legs (jump), jumping-in-place-on-two-legs (pjump), running (run), galloping-sideways (side), walking (walk), waving-one-hand (wave1) and waving-two-hands (wave2)

6 From Spikes to Action Recognition

The system that we described (see Fig. 3) takes as input a sequence of images $L(x, y, t)$ where a human action is performed. The directionally-selective V1 filters are applied over each frame of the input sequence in a log-polar distribution grid. The spike trains generated feed the MT layers where the activation of each MT cell depends on the activation of the V1 stage. The MT cells are arranged in a log-polar grid as well, working jointly with V1 cells as a spiking network.

In this section, we show how the activation of MT cells is used to define motion maps, and we also show a notion of distance between these maps. Based on this vectorial representation of a piece of sequence, we consider here the action recognition application with a standard supervised classification framework.

6.1 Database and Settings

We ran the experiment using Weizmann[2] database. Weizmann database consists in 9 different subjects performing 9 different actions. A representative frame of each action is shown in Fig. 12. The number of frames per sequence is variable and the original video streams were resized and centered to have sequences of 210×210 pixels.

General V1 and MT settings are shown in Table 1. V1 has a total of 72 layers, formed by 8 orientations and 9 different spatial-temporal frequencies, giving a total of 3302 cells per layer. Following the biological result mentioned in (Watson and Ahumada 1983) the value of σ_{V1} is $1.324/(4\pi f)$. The 72 layers of V1 cells are distributed in the frequency space in order to tile the whole space of interest. We considered a maximal spatial frequency of 0.5 pixels/s and a maximal temporal frequency of 12 cycles/s. In the case of MT, 8 (1×8 orientations) or 24 (3×8 orientations) layers were used depending on the center-surround configuration defined in Fig. 9.

[2] http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html.

Table 1 Parameters used for V1 and MT layers

	V1	MT
Fovea radius	80[pixels]	40[pixels]
Layer radius	100[pixels]	100[pixels]
Cell density in fovea	0.4[cells/pixel]	0.1[cells/pixel]
Eccentricity decay	0.02	0.02
Radius receptive field in fovea	$2\sigma_{V1}$[pixels]	9[pixels]
Number orientations	8	8
Number cells per layer	3302	161

6.2 Defining Motion Maps as Feature Vectors

6.2.1 The Mean Motion Map

Starting from the idea proposed in Escobar et al. (2006), we define the *mean motion map* $H_L(\cdot)$ representing the input stimulus $L(x, y, t)$ by

$$H_L(t, \Delta t) = \left\{\gamma_j^L(t, \Delta t)\right\}_{j=1,\ldots,N_l \times N_c}, \tag{16}$$

where N_l is the number of MT layers and N_c is the number of MT cells per layer. Each element γ_j^L with $j = 1, \ldots, N_l \times N_c$ is the windowed firing rate defined in (12). One illustration is given in Fig. 3(c).

The representation (16) has several advantages. It is invariant to the sequence length and its starting point (for Δt high enough depending on the scene). It is also includes information regarding the temporal evolution of the activation of MT cells, respecting the causality in the order of events. The use of a sliding window allows us to include motion changes inside the sequence.

The comparison between two mean motion maps $H_L(t, \Delta t)$ and $H_J(t, \Delta t)$, can be defined by

$$\mathcal{D}(H_L(t, \Delta t), H_J(t', \Delta t')) = \frac{1}{N_l \times N_c} \sum_{l=1}^{N_l \times N_c} \frac{(\gamma_l^L(t, \Delta t) - \gamma_l^J(t', \Delta t'))^2}{\gamma_l^L(t, \Delta t) + \gamma_l^J(t', \Delta t')}. \tag{17}$$

This measure refers to the *triangular discrimination* introduced by Topsoe (2000). Note that another measures derivated from statistics, such as *Kullback and Leiber* (KL) could also be considered. However, we didn't find any significant improvement with the KL measure for example.

6.2.2 The Synchrony Motion Map

As it is shown in Sect. 5.2, for each pair of cells $\{i, j\}$ it is possible to obtain a measure of synchrony using $\zeta_{ij}(\cdot)$ defined in (15).

The whole population of MT cells is divided into N_l subpopulations. Inside each subpopulation we created a map with the values of $\zeta_{ij}(t; \Delta t)$ obtained to every cell in the subpopulation within a sliding time window of size Δt. So, each sequence L will be represented by a *synchrony motion map* $\tilde{H}_L(t, \Delta t)$ defined as

$$\tilde{H}_L(t, \Delta t) = \{\mathbf{D}_k^L(t; \Delta t)\}_{k=1..N_l}, \tag{18}$$

where $\mathbf{D}_k^L(\cdot) = \{\zeta_{mn}(\cdot)\}_{m=1..N_c, n=1..N_c}$ is a matrix of $N_c \times N_c$ elements containing the measures $\zeta_{mn}(\cdot)$ between the mth and nth neurons of the kth layer of MT cells defined in (15). The $\tilde{H}_L$ construction can be summarized in Fig. 3(c).

The comparison between two synchrony motion maps $\tilde{H}_L(t, \Delta t)$ and $\tilde{H}_J(t', \Delta t')$. can be defined by the Euclidean distance

$$\tilde{\mathcal{D}}(\tilde{H}_L(t, \Delta t), \tilde{H}_J(t', \Delta t')) = \sqrt{\sum_{N_l} \|\mathbf{D}_k^L - \mathbf{D}_k^J\|^2}. \tag{19}$$

A real example showing the *synchrony motion maps* obtained for two different sequences ($N_l = 8$) can be seen in Fig. 13(b).

6.3 Results

6.3.1 Action Recognition Performance

To evaluate recognition performance of our approach using the motion maps defined in Sects. 6.2.1 and 6.2.2, we followed a similar experimental protocol than the one proposed by Jhuang et al. (2007). The *mean motion maps* and *synchrony motion maps* of all the 81 sequences forming Weizmann database were calculated, removing in both cases the first 5 frames containing initialization information (see Fig. 13(a)). With these motion maps, the training set and testing set were then constructed. The *training set* was built considering actions of 6 different subjects (6 subjects × 9 actions = 54 motion maps). The *testing set* was built with the remaining 3 subjects (3 subjects × 9 actions = 27 motion maps). Unlike Jhuang et al., we ran all the possible training sets (84) and not only 5 random trials. Each motion map is compared to every motion map in the training set. The match class will be the class associated to the motion map with the lowest distance.

For each training set, the experiment was performed twice: one time considering 8 layers of MT cells ($N_l = 8$) with the activation of the CRFs for the 8 different orientations, and a second time with 24 layers of MT cells ($N_l = 24$) using, for each orientation, all the surround interactions shown in Fig. 9. We constructed a histogram with the different recognition error rates obtained by our approach (see Fig. 14) using *mean motion maps* and *synchrony motion maps*. As we can see in Fig. 14, the values have a strong

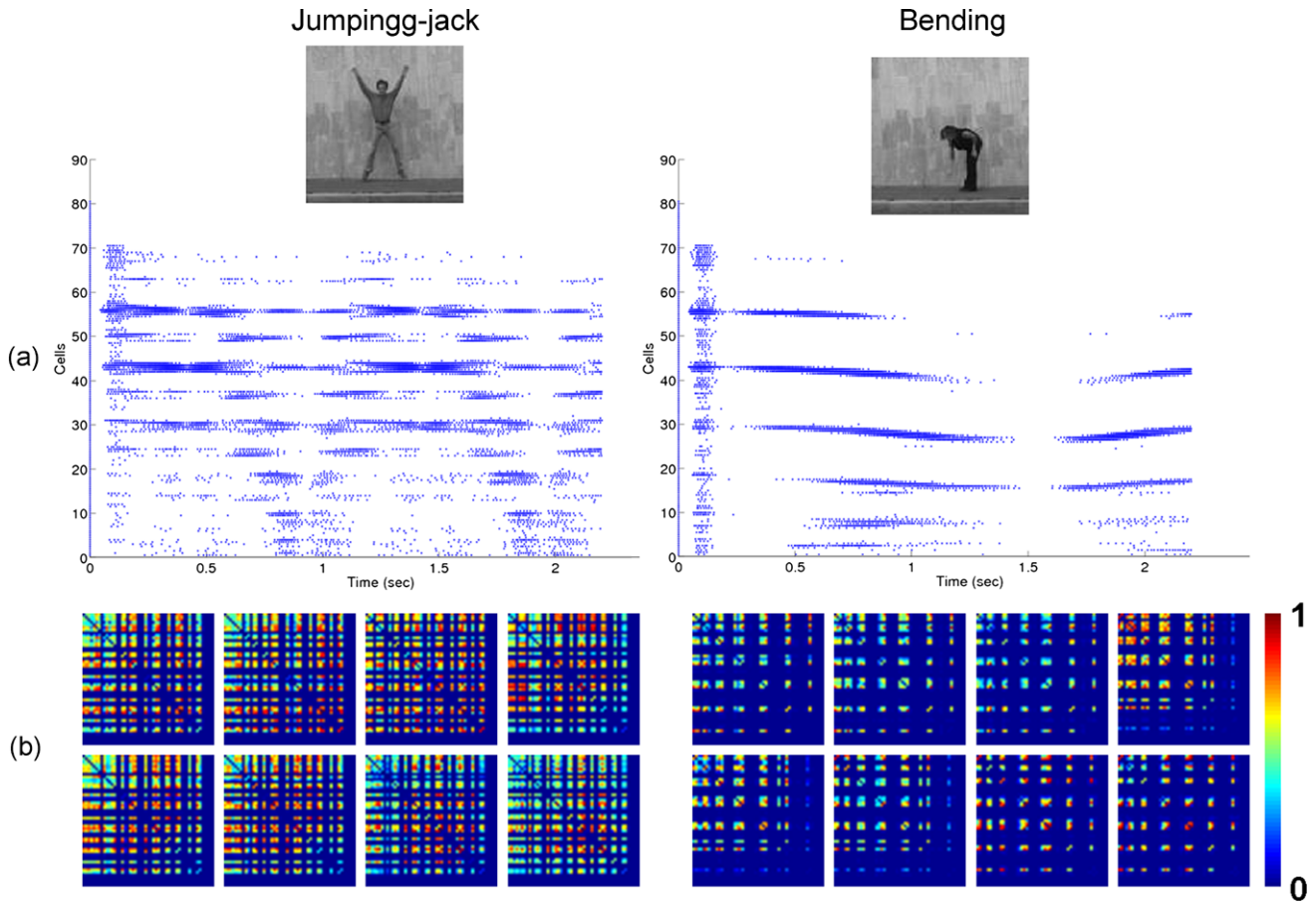

Fig. 13 (**a**) Raster plots obtained considering the 161 MT cells with only CRF of a given orientation in two different actions: *jumping-jack* and *bending*. Looking at the raster plots obtained, is evident that the information contained into the spike trains is much richer than a simplified mean firing rate. The frame rate is 25 frames per seconds. (**b**) Matrices conforming the *synchrony motion maps* defined in (18), each matrix shows the synchronization (see (15)) between the spike trains of iso-oriented cells members of the same MT population. It is possible to see the big differences between the synchrony maps of *jumping-jack* action and *bending* action

variability and the recognition performance highly depends on the sequences used to construct the training set, reaching in most of the cases 100% of correct recognition.

A comparison with the results obtained by Jhuang et al. (2007) is shown in Table 2. It is important to remark that our results were obtained using the 84 training sets built with 6 subjects (i.e., all possible combinations) and not only 5 trials as in Jhuang et al. (2007). As remarked previously, because of the high variability of classification performance depending on the training set chosen, results in Jhuang et al. (2007) are hard to interpret.

6.3.2 *Confusion Matrices*

In order to have a qualitative comparison between the quality of the human action representation using the two motion maps defined in Sect. 6.2, we estimated the *confusion matrices* for the 81 sequences conforming Weizmann database (see Fig. 15). The sequences were grouped according to the action performed (total of 9 actions), and for each pair of actions the mean distance value was obtained. The matrices are 9×9 and they were built using $N_l = 8$ (just MT CRF) and $N_l = 8 \times 3$ (using the three MT center-surround interactions of Fig. 9). Interestingly, despite of the lower recognition performance of *synchrony motion maps* compared with *mean motion maps*, *spiking motion maps* better separates the data belonging to different classes, specially for actions were only a limited part of the body performs the action (*waving-one-hand*, *waving-two-hands*, *bending*).

In order to quantify the inter-class separability we applied a simple statistical analysis (t-student test). Applying the t-student test on the obtained distances matrices we numerically observe for intra-cass distances a range of t-value $\in [0.20; 0.26]$ for *mean motion maps* and t-value $\in [0.29; 0.31]$ for *synchrony motion maps*, which in term of probabilities means that the probability to have distances different of zero is $P < 0.60$ and $P < 0.61$, respectively. A significant difference is seen in the inter-class distances,

where the range of t-values for *running/all-other-sequences* is t-value $\in [1.40; 2.93]$ (*synchrony motion maps*) and t-value $\in [0.44; 0.55]$ (*mean motion maps*). This can be interpreted, for instance, that for *jumping/walking* the distances are different from 0 with a probability of $P < 0.69$ for *mean motion maps* and $P < 0.90$ for *synchrony motion maps*. Although t-test values obtained for *mean motion maps* are numerically higher for inter-class than intra-class distances, it appears that they are not "significantly" higher compared to the ones obtained with the *synchrony motion maps*.

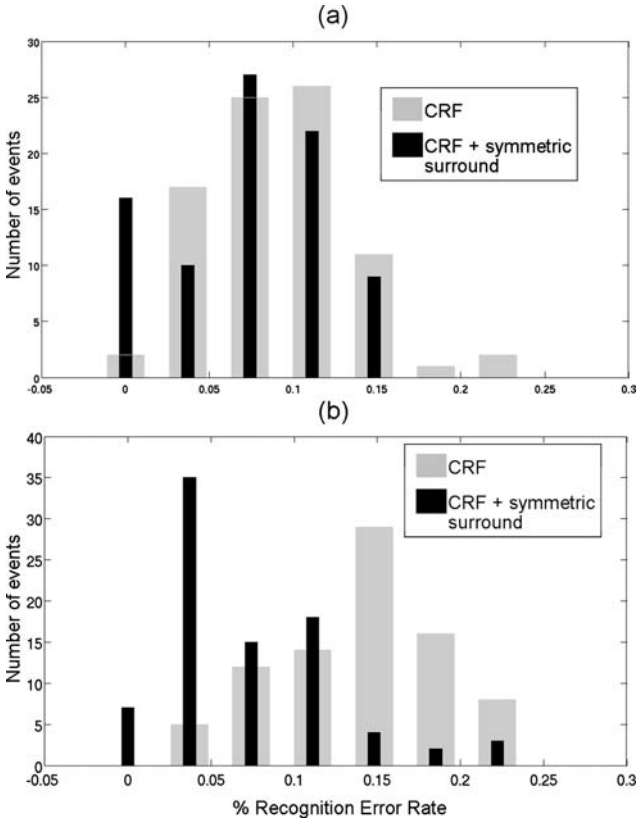

Fig. 14 Histograms representing the recognition error rates obtained by our approach in Weizmann database, using: MT CRFs (*gray bars*) and MT center-surround interactions shown in Fig. 9 (*black bars*). The results were obtained using the 84 possible training sets built with 6 different subjects. (**a**) Histogram obtained for *mean motion maps*. (**b**) Histogram obtained using *synchrony motion maps*

6.3.3 Robustness

To evaluate some kind of robustness of the approach, we considered input sequences with perturbations. Snapshots of the sequences considered to measure the robustness of the model are shown in Fig. 16. We considered three kinds of perturbations: *noisy* sequence (Fig. 16(2)), *legs-occluded* sequence (Fig. 16(3)) and *moving-background* sequence (Fig. 16(4)). Both *noisy* and *legs-occluded* sequences were created starting from the sequence shown in Fig. 16(1), which was extracted from the training set for the robustness experiments. The *legs-occluded* sequence was created placing a black box on the original sequence before the centered cropping. The *noisy* sequence was created adding

Table 2 Mean recognition error rates and standard deviation obtained by our approach and by Jhuang et al. (2007)

	Mean error rate ± STD	#trials
Juang et al. Grc_2 dense C_2 features	8.9%/ ± 5.9	5
Juang et al. Grc_2 dense C_2 features	3.0%/ ± 3.0	5
Mean motion maps CRF	9.08% ± 4.40	84
Mean motion maps CRF + symmetric surrounds	7.32% ± 4.62	84
Synchrony motion maps CRF	13.89% ± 4.95	84
Synchrony motion maps CRF + symmetric surrounds	7.19% ± 5.15	84

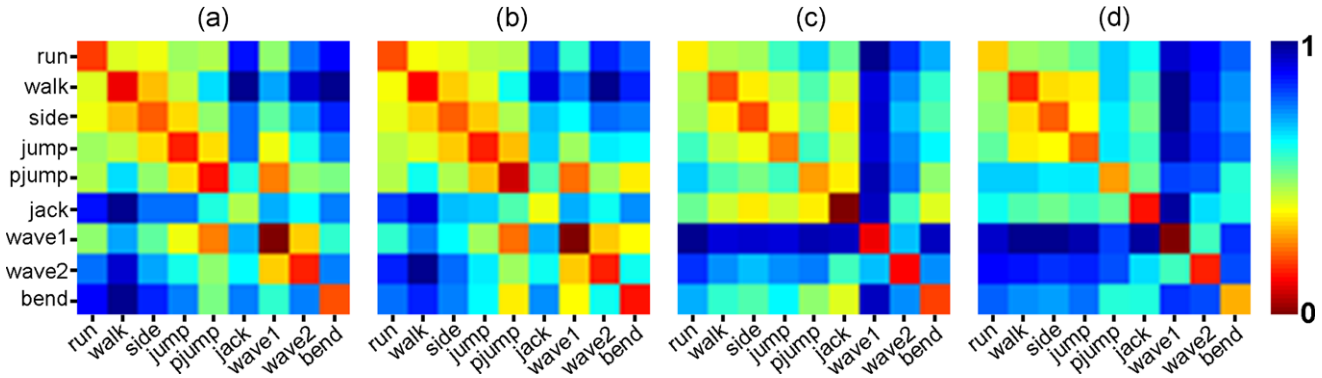

Fig. 15 Confusion matrices obtained using two different readouts: (**a**)–(**b**) *mean motion maps* defined in (16) and (**c**)–(**d**) *synchrony motion maps* defined in (18). We also here compare: (**a**)–(**c**) considering only MT CRFs and (**b**)–(**d**) considering all the MT center-surround interactions defined in Fig. 9

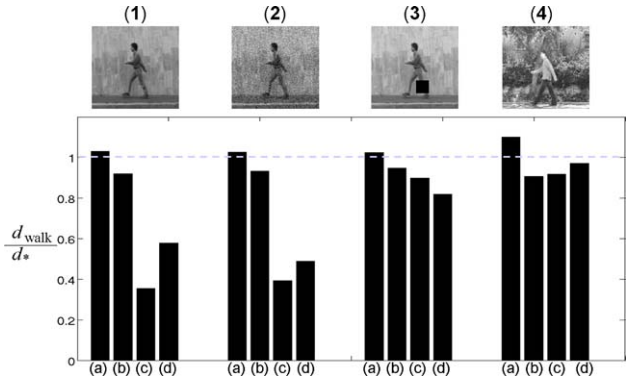

Fig. 16 Results obtained in the robustness experiments for the four input sequences represented by the snapshots at the *top* of the image. From *left* to *right*: (**1**) *normalwalker*, (**2**) *noisy* sequence, (**3**) *occluded-legs* sequence and (**4**) *moving-background* sequence. For each input sequence the action recognition experiment was performed 4 times: (a) *synchrony motion maps* with MT CRF, (b) *synchrony motion maps* with MT CRF + symmetric surrounds, (**c**) *mean motion maps* with MT CRF and (d) *mean motion maps* with MT CRF + symmetric surrounds. The *black bars* indicate the ratio between the distance to *walking* class d_{walk} and the distance to the second closest class d_* (*galloping-sideways*, *bending* or *jumping-forward-in-two-legs*)

a Gaussian noise. The *moving-background* sequence was taken from Blank et al. (2005). For the original sequence and the three modified input sequences the recognition was correctly performed as *walking*.

The bars of Fig. 16 represent the ratio between the shortest distance to *walking* (d_{walk}) class and the distance to the second closest class (d_*), which can vary from *galloping-sideways* to *bending* or *jumping-forward-in-two-legs*. Note that in most of the cases the action was correctly recognized as *walking*, giving a ratio $d_{walk}/d_* < 1$. The recognition failed in the case of *synchrony motion maps* (a) who consider only the CRF activation. In those cases the action was always misclassified as *bending* ($d_{walk}/d_{bend} > 1$). This performance is considerably improved if the information of the surround interactions is added to the *synchrony motion maps* (case (b)), confirming its important role in motion representation.

7 Discussion

We proposed a generic spiking V1-MT model that can be used for high level tasks such as human action recognition. Our model takes as input a sequence of images. V1 cells implement a linear spatial-temporal filtering stage followed by both local and global nonlinearities known as normalizations. Activities in V1 neurons are then transformed into spike trains using a LIF neuron model, adding implicit nonlinearities contained in the spike generation process. These spike trains feed a second layer of spiking neurons, which was designed using biological findings of motion processing in primates, area MT (Born and Bradley 2005). From the activation of the MT cells, we defined two kinds motion maps (*mean motion maps* and *synchrony motion maps*) which represent the activation of the different MT spiking network in a temporal window. Finally, we showed that these motion maps can be used in a classical supervised classification technique, and we gave some recognition results on a classical database.

Of course, more validation would be needed. We tested the model with Weizmann database, obtaining the results shown in Sect. 6.3. The good recognition performance obtained with our spike-to-spike model reinforces our hypothesis about the representability of our motion maps. Weizmann database was also used by, e.g., Blank et al. (2005) and Jhuang et al. (2007) to validate their model. However, test conditions and experimental protocol are not the same than the ones considered in our experiments, and therefore most of the times recognition performances cannot be compared. We only compared our recognition performance with the results obtained by Jhuang et al. (2007), showing that due to the high variability of the results, the recognition percentages of Jhuang et al. (2007) are not so representative. Another concern is that it is not possible to claim that our system will work in any condition. But that concern is in fact general as remarked by Pinto et al. (2008): It is an overclaim to declare that the whole action recognition problem is solved only based on some results obtained with a given database. So, more validation with other database such as KTH[3] database would be needed.

Recognition results obtained using *synchrony motion maps* are slightly inferior than the ones obtained using *mean motion maps*, specially if we only consider the activation of MT CRFs. This difference is enhanced in the robustness experiments. As an explanation, we think that because the synchrony analysis largely forgets about the rate, it lacks a fundamental information about network activity. Nevertheless, by considering synchronies only, satisfying recognition performance can be achieved. Also, note that the use of the synchrony to encode the input motion information improves the inter-class separability obtaining a better class clustering (see Fig. 15 and Table 3). These results are consistent with neuroscience findings about the complementarity of rate and synchrony codes: There are evidence from motor and visual cortex that both, rate and synchrony code, are conjointly used to extract complementary information (Maldonado et al. 2008; Riehle et al. 1997). As a future work, we plan to combine these two motion maps in order to have a better representation of the input motion information.

Earlier models have suggested that biological motion perception depends on strong interactions between motion and

[3] http://www.nada.kth.se/cvap/actions/.

Table 3 The Null hypothesis rejection probability associated with the t-test values obtained from the distance matrices built using *mean motion maps* and *synchrony motion maps* (case CRF + symmetric surrounds). The corresponding action for each value is the same than the ones shown in Fig. 15

Mean motion map

0.59	0.70	0.71	0.69	0.68	0.62	0.67	0.68	0.72
0.70	0.59	0.69	0.68	0.72	0.65	0.70	0.70	0.74
0.71	0.69	0.60	0.66	0.68	0.63	0.68	0.69	0.72
0.69	0.68	0.66	0.60	0.72	0.62	0.68	0.69	0.75
0.68	0.72	0.68	0.72	0.60	0.59	0.64	0.66	0.72
0.62	0.65	0.63	0.62	0.59	0.59	0.61	0.64	0.65
0.67	0.70	0.69	0.68	0.64	0.61	0.58	0.64	0.69
0.68	0.70	0.69	0.69	0.66	0.64	0.64	0.58	0.68
0.72	0.74	0.72	0.75	0.72	0.65	0.69	0.68	0.59

Synchrony motion map

0.61	0.86	0.88	0.90	0.97	0.98	1.00	0.99	0.99
0.86	0.62	0.89	0.90	0.97	0.94	0.99	0.98	0.97
0.88	0.89	0.62	0.86	0.98	0.97	1.00	0.96	0.98
0.90	0.91	0.86	0.62	0.99	0.96	1.00	0.99	0.99
0.97	0.97	0.98	0.99	0.61	0.85	0.93	1.00	0.91
0.98	0.94	0.97	0.96	0.85	0.62	0.96	0.93	0.94
1.00	0.99	1.00	1.00	0.93	0.96	0.60	0.76	0.86
0.99	0.98	0.96	0.99	0.99	0.93	0.75	0.60	0.96
0.99	0.97	0.98	0.99	0.91	0.94	0.86	0.96	0.61

form pathways (see Blake and Shiffrar 2007 for a review). In the model proposed by Giese and Poggio (2003), both form and motion pathways learn sequences or "snapshots" of human shapes and optic flow patterns, respectively. Several models have been proposed to dynamically constrain such motion model using local information about shapes, features and contours (e.g., Bayerl and Neumann 2007). Since configural information are important for biological motion recognition (e.g., Hiris et al. 2005) future work will investigate how local form information can be dynamically merged and integrated with the motion pathway to improve the representability of motion maps, specially in the case of complex backgrounds where motion integration could play an important role (see Fig. 16(d)).

Specifically, Giese and Poggio (2003) proposed a neurophysiological model for the visual information processing in the dorsal (*motion*) and ventral (*form*) pathways. The model is validated in the action recognition task using as input stimulus stick figures constructed from real sequences. Assuming no interaction between the two pathways, they found that both motion and form pathways are capable to do action recognition. One of the main difference with our approach is the fact that new parameters need to be fitted if a new action must be considered. In our case, no parameters must be adjusted and only the new motion maps must be inserted into the training set. Moreover, their model exhibit several interesting properties for biological pattern motion recognition such as spatial and temporal scale invariance, robustness to noise added to point-like motion stimuli and so on. More recent work from Jhuang et al. (2007) implemented this invariance for spatial and temporal scales (i.e. stimulus size and execution time, respectively). Their approach uses a bio-inspired model for the action recognition task based in Giese and Poggio (2003) and Serre et al. (2005). The invariance to spatial and temporal scale is achieved considering as many motion detector layers as the number of spatial and temporal scales to be detected. This can be easily implemented in our approach adding more layers with different spatial and temporal scales and therefore apply the *max* operator between the different layers coding the same motion direction.

Finally, contrarily to the bio-inspired model of Giese and Poggio (2003), our model relies on a general purpose motion processing based upon the known properties of the two-stages biological motion pathway where V1 and MT neurons implement detection and integration stage, respectively. The architecture is rooted on the linear-nonlinear ("L-N") model, of a kind that is increasingly used in sensory neuroscience (see Simoncelli and Heeger 1998; Rust et al. 2006 for instance). Recent version of this L-N models propose that complex motion analysis can be done through a cascade of L-N steps, followed by a Poisson spiking generation process (Rust et al. 2006). Our generic motion model departs from this cascaded L-N model in several important way.

- For early local motion detection, Simoncelli and Heeger (1998) proposed local units modeled through spatial-temporal energy filters. However, those filters have a temporal profile that is non-causal and inconsistent with V1 cell physiology. Our approach, on the other hand uses temporal profiles consistent with V1 cell physiology. These biologically plausible temporal profiles bring out not trivial calculation for the tuning of the spatial-temporal frequency orientation. As a consequence, motion orientation tuning must be computed using numerical approximations.
- Each L-N stage is followed by spike generation process, using LIF neurons. Each spiking process introduces additional nonlinearities due to spike generation process. Moreover, the responses of MT neurons now operates on spike trains from an afferent population of nonlinear V1 cells. Given the others nonlinearities found in the MT layer we add more complexity to the system, making it more suitable for natural images analysis.
- Our model implements different MT non classical receptive fields by having different classes of center-surround interactions (e.g. Xiao et al. 1995; Born 2000). The role of different MT receptive field shapes in the action recognition task has not been evaluated before (see Giese and Poggio 2003; Sereno and Sereno 1999). Here we present some results in the action recognition performance using

three different structures of receptive fields as observed in monkey area MT (Born 2000; Xiao et al. 1995, 1997), showing their crucial role in our motion representation (see, e.g., Figs. 14 and 16). Using the same architecture, we can implement more complex center-surround interactions such as oriented, non-isotropic inhibitory surrounds (Xiao et al. 1995, 1997) which was modeled in Escobar and Kornprobst (2008). We have shown already that more complex spatial integration mechanism has a significant impact on the discrimination of motion maps. In future work we will consider how the diversity of center-surround interactions enable a generic motion integration model to process complex synthetic and natural images flows.

– Lastly, the dynamical changes in the receptive field organization and in MT direction tuning reported, e.g., by Pack et al. (2005), Perge et al. (2005), Smith et al. (2005) suggest that the connectivity between V1 and MT cells is highly dynamical, allowing adaptive changes in motion maps. Those changes can be easily implemented in a wholly spiking network as the one proposed in our approach.

Acknowledgements This work was partially supported by the EC IP project FP6-015879, FACETS and CONICYT Chile. We also thank to Olivier Rochel for his Mvaspike simulator, this tools allowed us to create and simulate spiking networks in an easy way.

References

Adelson, E., & Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*, 284–299.

Bayerl, P., & Neumann, H. (2007). Disambiguating visual motion by form–motion interaction—a computational model. *International Journal of Computer Vision*, *72*(1), 27–45.

Beintema, J., & Lappe, M. (2002). Perception of biological motion without local image motion. *Proceedings of the National Academy of Sciences of the USA*, *99*(8), 5661–5663.

Berzhanskaya, J., Grossberg, S., & Mingolla, E. (2007). Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception. *Spatial Vision*, *20*(4), 337–395.

Biederlack, J., Castelo-Branco, M., Neuenschwander, S., Wheeler, D. W., Singer, W., & Nikoli, D. (2006). Brightness induction: rate enhancement and neuronal synchronization as complementary codes. *Neuron*, *52*(6), 1073–1083.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology*, *58*, 12.1–12.27.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space-time shapes. In *Proceedings of the 10th international conference on computer vision* (Vol. 2, pp. 1395–1402).

Bobick, A., & Davis, J. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3), 257–267.

Born, R. T. (2000). Center-surround interactions in the middle temporal visual area of the owl monkey. *Journal of Neurophysiology*, *84*, 2658–2669.

Born, R., & Bradley, D. (2005). Structure and function of visual area MT. *Annual Reviews—Neuroscience*, *28*, 157–189.

Buracas, G. T., & Albright, T. D. (1996). Contribution of area mt to perception of three-dimensional shape: a computational study. *Vision Research*, *36*(6), 869–87.

Casile, A., & Giese, M. (2003). Roles of motion and form in biological motion recognition. In *Lecture notes in computer science: Vol. 2714. Artificial networks and neural information processing* (pp. 854–862). Berlin: Springer.

Casile, A., & Giese, M. (2005). Critical features for the recognition of biological motion. *Journal of Vision*, *5*, 348–360.

Cessac, B., Rostro-Gonzalez, H., Vasquez, J., & Vieville, T. (2008). To which extend is the "neural code" a metric? In *Deuxième conférence française de neurosciences computationnelles*.

Collins, R., Gross, R., & Shi, J. (2002). Silhouette-based human identification from body shape and gait. In *5th intl. conf. on automatic face and gesture recognition* (p. 366).

Conway, B., & Livingstone, M. (2003). Space-time maps and two-bar interactions of different classes of direction-selective cells in macaque V1. *Journal of Neurophysiology*, *89*, 2726–2742.

Cutler, R., & Davis, L. (2000). Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8)

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Cambridge: MIT Press.

De Valois, R., Cottaris, N., et al. (2000). Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity. *Vision Research*, *40*, 3685–3702.

Destexhe, A., Rudolph, M., & Paré, D. (2003). The high-conductance state of neocortical neurons in vivo. *Nature Reviews Neuroscience*, *4*, 739–751.

Dollar, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *VS-PETS* (pp. 65–72).

Efros, A., Berg, A., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Proceedings of the 9th international conference on computer vision* (Vol. 2, pp. 726–734).

Escobar, M. J., & Kornprobst, P. (2008). Action recognition with a bio–inspired feedforward motion processing model: The richness of center-surround interactions. In *Lecture notes in computer science. Proceedings of the 10th European conference on computer vision*. Berlin: Springer.

Escobar, M. J., Wohrer, A., Kornprobst, P., & Vieville, T. (2006). Biological motion recognition using an mt-like model. In *Proceedings of 3rd Latin American robotic symposium*.

Felleman, D., & Essen, D. V. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex*, *1*, 1–47.

Fellous, J. M., Tiesinga, P. H. E., Thomas, P. J., & Sejnowski, T. J. (2004). Discovering spike patterns in neural responses. *The Journal of Neuroscience*, *24*(12), 2989–3001.

Fries, P., Neuenschwander, S., Engel, A. K., Goebel, R., & Singer, W. (2001). Rapid feature selective neuronal synchronization through correlated latency shifting. *Nature Neuroscience*, *4*(2), 194–200.

Gautrais, J., & Thorpe, S. (1998). Rate coding vs temporal order coding: a theoretical approach. *Biosystems*, *48*, 57–65.

Gavrila, D. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, *73*(1), 82–98.

Gavrila, D., & Davis, L. (1996). 3-D model-based tracking of humans in action: a multi-view approach. In *Proceedings of the international conference on computer vision and pattern recognition*. San Francisco: IEEE.

Gerstner, W., & Kistler, W. (2002). *Spiking neuron models*. Cambridge: Cambridge University Press.

Giese, M., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements and actions. *Nature Reviews Neuroscience*, *4*, 179–192.

Gollisch, T., & Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science*, *319*, 1108–1111.

Goncalves, L., DiBernardo, E., Ursella, E., & Perona, P. (1995). Monocular tracking of the human arm in 3D. In *Proceedings of the 5th international conference on computer vision* (pp. 764–770).

Grzywacz, N., & Yuille, A. (1990). A model for the estimate of local image velocity by cells on the visual cortex. *Proceedings of the Royal Society London B: Biological Sciences*, *239*(1295), 129–161.

Hiris, E., Humphrey, D., & Stout, A. (2005). Temporal properties in masking biological motion. *Perception and Psychophysics*, *67*(3), 435–443.

Hogg, D. (1983). Model-based vision: a paradigm to see a walking person. *Image and Vision Computing*, *1*(1), 5–20.

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat visual cortex. *Journal of Physiology*, *160*, 106–154.

Izhikevich, E. (2004). Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks*, *15*(5), 1063–1070.

Jhuang, H., Serre, T., Wolf, L., & Poggio, T. (2007). A biologically inspired system for action recognition. In *Proceedings of the 11th international conference on computer vision* (pp. 1–8).

Kreuz, T., Haas, J. S., Morelli, A., Abarbanel, H. D., & Politi, A. (2007). Measuring spike train synchrony. *Journal of Neuroscience Methods*, *165*, 151–161.

Laptev, I., Capuo, B., Schultz, C., & Lindeberg, T. (2007). Local velocity-adapted motion events for spatio-temporal recognition. *Computer Vision and Image Understanding*, *108*(3), 207–229.

Lui, L. L., Bourne, J. A., & Rosa, M. G. P. (2007). Spatial summation, end inhibition and side inhibition in the middle temporal visual area MT. *Journal of Neurophysiology*, *97*(2), 1135.

Maldonado, P., Babul, C., Singer, W., Rodriguez, E., Berger, D., & Grün, S. (2008). Synchronization of neuronal responses in primarily visual cortex of monkeys viewing natural images. *Journal of Neurophysiology*, *100*, 1523–1532.

Mestre, D. R., Masson, G. S., & Stone, L. S. (2001). Spatial scale of motion segmentation from speed cues. *Vision Research*, *41*(21), 2697–2713.

Michels, L., Lappe, M., & Vaina, L. (2005). Visual areas involved in the perception of human movement from dynamic analysis. *Brain Imaging*, *16*(10), 1037–1041.

Mokhber, A., Achard, C., & Milgram, M. (2008). Recognition of human behavior by space-time silhouette characterization. *Pattern Recognition Letters*, *29*(1), 81–89.

Mutch, J., & Lowe, D. G. (2006). Multiclass object recognition with sparse, localized features. In *Proceedings of the international conference on computer vision and pattern recognition* (pp. 11–18).

Neuenschwander, S., Castelo-Branco, M., & Singer, W. (1999). Synchronous oscillations in the cat retina. *Vision Research*, *39*(15), 2485–2497.

Niebles, J. C., Wang, H., & Fei-Fei, L. (2006). Unsupervised learning of human action categories using spatial-temporal words. In *British machine vision conference*.

Nowak, L., & Bullier, J. (1997). The timing of information transfer in the visual system. In *Cerebral cortex* (Vol. 12, pp. 205–241). New York: Plenum Press. Chap. 5.

Nowlan, S., & Sejnowski, T. (1995). A selection model for motion processing in area MT of primates. *Journal of Neuroscience*, *15*, 1195–1214.

Pack, C. C., Hunter, J. N., & Born, R. T. (2005). Contrast dependence of suppressive influences in cortical area mt of alert macaque. *Journal of Neurophysiology*, *93*(3), 1809–1815.

Perge, J., Borghuis, B., Bours, R., Lankheet, M., & van Wezel, R. (2005). Temporal dynamics of direction tuning in motion-sensitive macaque area mt. *Journal of Neurophysiology*, *93*, 2194–2116.

Perkel, D. H., & Bullock, T. H. (1968). Neural coding. *Neurosciences Research Program Bulletin*, *6*, 221–348.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology*, *4*(1), e27.

Polana, R., & Nelson, R. (1997). Detection and recognition of periodic, non-rigid motion. *International Journal of Computer Vision*, *23*(3), 261–282.

Riehle, A., Grün, S., Diesmann, M., & Aertsen, A. (1997). Spike synchronization and rate modulation differentially involved in motor cortical function. *Science*, *278*, 1950–1953.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge: Bradford Books.

Robson, J. (1966). Spatial and temporal contrast-sensitivity functions of the visual system. *Journal of Optical Society of America*, *69*, 1141–1142.

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (2004). Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nature Neuroscience*, *7*(9), 982–991.

Rohr, K. (1994). Toward model-based recognition of human movements in image sequences. *CVGIP, Image Understanding*, *1*, 94–115.

Rust, N., Mante, V., Simoncelli, E., & Movshon, J. (2006). How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, *11*, 1421–1431.

Saul, A., Carras, P., & Humphrey, A. (2005). Temporal properties of inputs to direction-selective neurons in monkey v1. *Journal of Neurophysiology*, *94*, 282–294.

Seitz, S., & Dyer, C. (1997). View-invariant analysis of cyclic motion. *The International Journal of Computer Vision*, *25*(3), 231–251.

Sereno, M. E., & Sereno, M. L. (1999). 2-d center-surround effects on 3-d structure-from-motion. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1834–1854.

Serre, T. (2006). *Learning a dictionary of shape-components in visual cortex: Comparison with neurons, humans and machines*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Serre, T., Wolf, L., & Poggio, T. (2005). Object recognition with features inspired by visual cortex. In *Proceedings of the international conference on computer vision and pattern recognition* (pp. 994–1000).

Shah, M., & Jain, R. (1997). *Motion-based recognition. Computational imaging and vision series*. Dordrecht: Kluwer Academic.

Sigala, R., Serre, T., Poggio, T., & Giese, M. (2005). Learning features of intermediate complexity for the recognition of biological motion. In *LNCS: Vol. 3696. ICANN 2005* (pp. 241–246). Berlin: Springer.

Simoncelli, E. P., & Heeger, D. (1998). A model of neuronal responses in visual area MT. *Vision Research*, *38*, 743–761.

Smith, M., Majaj, N., & Movshon, A. (2005). Dynamics of motion signaling by neurons in macaque area mt. *Nature Neuroscience*, *8*(2), 220–228.

Snowden, R. J., Treue, S., Erickson, R. G., & Andersen, R. A. (1991). The response of area mt and v1 neurons to transparent motion. *The Journal of Neuroscience*, *11*(9), 2768–2785.

Thorpe, S. (1990). Spike arrival times: A highly efficient coding scheme for neural networks. In *Parallel processing in neural systems and computers* (pp. 91–94).

Thorpe, S. (2002). Ultra-rapid scene categorization with a wave of spikes. In *Lecture notes in computer science: Vol. 2525. Biologically motivated computer vision* (pp. 1–15). Berlin: Springer.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.

Topsoe, F. (2000). Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, *46*(4), 1602–1609.

Tsotsos, J., Liu, Y., Martinez-Trujillo, J., Pomplun, M., Simine, E., & Zhou, K. (2005). Attending to visual motion. *Computer Vision and Image Understanding*, *100*, 3–40.

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research*, *42*, 2593–2615.

Victor, J., & Purpura, K. (1996). Nature and precision of temporal coding in visual cortex: a metric-space analysis. *Journal of Neurophysiology*, *76*, 1310–1326.

Wang, L., & Suter, D. (2007). Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In *Proceedings CVPR*.

Wang, D. L., & Terman, D. (1995). Locally excitatory globally inhibitory oscillator networks. *IEEE Transactions on Neural Networks*, *6*, 283–286.

Watson, A., & Ahumada, A. (1983). *A look at motion in the frequency domain* (NASA Tech. Memo).

Wielaard, D. J., Shelley, M., McLaughlin, D., & Shapley, R. (2001). How simple cells are made in a nonlinear network model of the visual cortex. *The Journal of Neuroscience*, *21*(14), 5203–5211.

Wohrer, A., & Kornprobst, P. (2008). Virtual Retina: A biological retina model and simulator, with contrast gain control. *Journal of Computational Neuroscience*. doi:10.1007/s10827-008-0108-4.

Wong, S. F., Kim, T. K., & Cipolla, R. (2007). Learning motion categories using both semantic and structural information. In *Proceedings of the international conference on computer vision and pattern recognition* (pp. 1–6).

Xiao, D., Raiguel, S., Marcar, V., Koenderink, J., & Orban, G. A. (1995). Spatial heterogeneity of inhibitory surrounds in the middle temporal visual area. *Proceedings of the National Academy of Sciences*, *92*(24), 11303–11306.

Xiao, D. K., Raiguel, S., Marcar, V., & Orban, G. A. (1997). The spatial distribution of the antagonistic surround of MT/V5 neurons. *Cereb Cortex*, *7*(7), 662–677.

Zelnik-Manor, L., & Irani, M. (2001). Event-based analysis of video. In *Proceedings of CVPR'01* (Vol. 2, pp. 123–128).