Computer Vision and Image Understanding 116 (2012) 593–605

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Action recognition via bio-inspired features: The richness of center–surround interaction ☆

María-José Escobar [a,*], Pierre Kornprobst [b,1]

[a] Universidad Técnica Federico Santa María, Department of Electronics Engineering, Avda. España 1680, Valparaíso, Chile
[b] Neuromathcomp Project Team, INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 Sophia-Antipolis, France

ARTICLE INFO

Article history:
Received 9 September 2010
Accepted 7 January 2012
Available online 20 January 2012

Keywords:
Action recognition
Bio-inspired models
V1
MT
Motion analysis
Center–surround interaction

ABSTRACT

Motion is a key feature for a wide class of computer vision approaches to recognize actions. In this article, we show how to define bio-inspired features for action recognition. To do so, we start from a well-established bio-inspired motion model of cortical areas V1 and MT. The primary visual cortex, designated as V1, is the first cortical area encountered in the visual stream processing and early responses of V1 cells consist in tiled sets of selective spatiotemporal filters. The second cortical area of interest in this article is area MT where MT cells pool incoming information from V1 according to the shape and characteristic of their receptive field. To go beyond the classical models and following the observations from Xiao et al. [61], we propose here to model different surround geometries for MT cells receptive fields. Then, we define the so-called bio-inspired features associated to an input video, based on the average activity of MT cells. Finally, we show how these features can be used in a standard classification method to perform action recognition. Results are given for the Weizmann and KTH databases. Interestingly, we show that the diversity of motion representation at the MT level (different surround geometries), is a major advantage for action recognition. On the Weizmann database, the inclusion of different MT surround geometries improved the recognition rate from 63.01 ± 2.07% up to 99.26 ± 1.66% in the best case. Similarly, on the KTH database, the recognition rate was significantly improved with the inclusion of MT different surround geometries (from 47.82 ± 2.71% up to 92.44 ± 0.01% in the best case). We also discussed the limitations of the current approach which are closely related to the input video duration. These promising results encourage us to further develop bio-inspired models incorporating other brain mechanisms and cortical areas in order to deal with more complex videos.

1. Introduction

Motion is a key feature for a wide class of computer vision approaches to recognize actions: Existing approaches consider a variety of motion representations and motion-based features (see, e.g., [45,56,46] for recent surveys). The question is how to define motion features containing sufficient information to perform action recognition? Let us mention some recent examples from the literature. In Efros et al. [18], the authors defined a spatiotemporal descriptor for low resolution videos which is based on optical flow measurements inside a small window. In Zelnik-Manor and Irani [64], the authors defined spatiotemporal features from the local intensity gradients extracted at different temporal scales. Actions were then characterized by the global histograms of image gradients at multiple time scales. In Dollar et al. [17], the authors defined spatiotemporal features based on the generalized Harris detector, which extracts spatiotemporal corners represented by cuboids (see also [28]). The spatiotemporal corners are quite rare and therefore a sparse representation of the motion activity in a video sequence. Similarly to Dollar et al. [17], in Jhuang et al. [24], the authors also proposed the concept of 3D cuboids for the spatiotemporal feature extraction, relating their algorithm to a neurobiological model of motion-form processing in the visual cortex (essentially at the level of the V1-cortical area with simple and complex cells).

As far as the visual system is concerned, one can make a similar observation asking whether motion is also essential to recognize actions (though other features such as form are also useful). This was observed in psychophysics and confirmed in fMRI. In psychophysics, Casile and Giese [13] showed that biological motion recognition can be performed with a coarse spatial location of the mid-level optic flow features. Using fMRI, in Michels et al. [35], the authors located human brain areas involved in biological motion recognition identifying the activation of both motion-

☆ This paper has been recommended for acceptance by J.K. Tsotsos.

* Corresponding author. Fax: +56 32 2 79 74 69.
E-mail addresses: mariajose.escobar@usm.cl (M.-J. Escobar), pierre.kornprobst@inria.fr (P. Kornprobst).
[1] Fax: +33 4 92 38 78 45.

1077-3142/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2012.01.002

processing (dorsal) and form-processing (ventral) pathways of the visual system.

In this article our goal is to propose new motion-based features for action recognition, inspired by visual system processing. To do so, we propose modeling the motion-processing pathway, focusing on cortical areas V1 and MT. The primary visual cortex, designated as V1, is the first cortical area encountered in the visual stream processing. The purpose of V1 can be thought of as similar to many spatially local and complex Fourier transforms. The second cortical area of interest in this article is area MT. MT cells pool incoming information from V1 according to the shape and characteristic of their receptive field and a large portion of the cells are tuned to the speed and direction of moving visual stimuli, so that MT plays a significant role in the processing of visual motion.

The basis of our model will be mostly classical and it will rely on well-established results. Indeed, several bio-inspired motion processing models have been proposed in the computational neuroscience literature, and they can be a useful source of inspiration. Some examples include Nowlan and Sejnowski [36], Rust et al. [49], Simoncelli and Heeger [52], Grzywacz and Yuille [21]. However, the goal of these models was essentially to reproduce certain properties of primate visual systems and to make predictions for neuroscience. Recently, more elaborated models were also proposed, combining several motion-related areas with other cortical areas (see, e.g., [5,3,8,54]. These model can handle more complex stimuli and tend to bridge the gap between computational neuroscience and computer vision. Considering all this past work, our goal is provide a simple model capturing important biological properties and applicable for action recognition.

The bio-inspired features proposed in this article will be directly defined from MT cells' characteristics. Focusing on MT cells, an important observation is that most cells in MT are sensitive to motion contrasts: Different kinds of surrounding geometries of MT receptive fields are observed in the computation of motion structure (see [32,4]), introducing anisotropies in the processing of spatial information. More precisely, in Xiao et al. [61] the authors examined a population of MT cells, revealing that 50% of MT cells have asymmetric receptive fields, and the other 50% is divided into circular symmetric surrounds (20%) and bilaterally symmetric surrounds (25%).

This article is organized as follows. Section 2 presents a classical functional bio-inspired model of the motion-processing pathway, based on existing literature. This model has two main stages: (i) focus on the action; and (ii) motion estimation through the modeling of V1 and MT cortical areas. Main justifications will be given in terms of general biological findings regarding V1 and MT neuron properties. Exact matching of these models with neurophysiological recordings is beyond the scope of this article. Note that an open-source C++ library to simulate V1 cells is proposed. Section 3 focuses on the variety of the MT receptive fields modeled in this article. Section 4 addresses the problem of action recognition. We show how MT cell responses can be used to define bio-inspired features and how these features are relevant in the context of action recognition. Comparisons with state-of-the-art methods and databases are provided. Section 5 summarizes the contribution and propose perspectives of this work. Finally, for those interested, technical details about the model are included in Appendix A.

2. A bio-inspired vision architecture

2.1. Focus on the action

Recognizing human action in real life requires that, if a person is moving across the visual field, our eyes follow the motion. The importance of smooth eye pursuit has been studied in biological motion recognition. Specifically, Orban de Xivry et al. [38] found a strong correlation between the action recognition rate and the level of smooth eye velocity measured after 100 ms stimulus onset. In fact, to follow moving persons with the eyes is absolutely necessary for recognition if the motion is performed in cluttered backgrounds or crowded scenarios. To focus on the action, attentional mechanisms are important to select a part of the sensory information, and thus, to process only a bounded area of the visual scene. Recently, Safford et al. [50] reported that the action recognition performance in biological motion is highly modulated by attention: the best performance is reached when the moving subject is in the focus of attention.

Following these ideas, we preprocess videos containing human motion in order to obtain a self-centered representation of the action, as illustrated in Fig. 1a and b. In practice, focusing on the action can be done in different ways: For example, one can use models of visual attention where salient events are extracted based on combinations of image features and given a certain context (see, e.g., [23,9,10]).

The main consequence of this preprocessing stage is that the overall system will naturally be invariant to position (in the sense "where does the action take place in the scene?") but of course the problem of invariance with respect to viewing angles is still open. Concerning scale-invariance, since actions are rescaled to the same dimension, an action at a distance will be zoomed in but with a degraded resolution. However, the benefit of having all actions represented in the same size is that it will allow us to compare features between nearby and distant actions.

2.2. V1 layer: local motion detectors

The primary visual cortex, designated as V1, is the first cortical area encountered in the visual stream processing. It is probably the most studied visual area but its complexity still makes it very hard to accurately model [37]. However, the current consensus is that early responses of V1 cells consist in tiled sets of selective spatiotemporal filters. The purpose of V1 can be thought of as similar to many spatially local, complex Fourier transforms. For example, in Grzywacz and Yuille [21], the authors showed that several properties of simple/complex cells in V1 can be described by energy filters and in particular by Gabor patches. More recently, Mante and Carandini [33] showed which properties of V1 cells can be explained using an energy model. Theoretically, these filters together can carry out neuronal processing of spatial frequency, orientation, motion, direction, speed (thus temporal frequency), and many other spatiotemporal features.

In this article, we implemented a classical model for V1 complex cells (Fig. 1c). So, let us consider the ith V1 complex cell, located at $\mathbf{x}_i = (x_i, y_i)$, tuned to the spatial orientation θ_i and spatiotemporal orientation $f_i = (\bar{\xi}_i, \bar{\omega}_i)$. Its mean firing rate, denoted by $r_i^{V1}(t)$, is estimated in two stages:

(i) Motion energy detectors: Motion computation is performed by energy motion detectors according to Adelson and Bergen [1]. Each motion energy detector emulates a V1 complex cell, in the sense that it is invariant to contrast polarity, and it is built as a nonlinear combination of linear V1 simple cells (see [12] for V1 cells classification). Each complex cell is thus selective to a certain motion direction inside a spatiotemporal frequency bandwidth. We refer the interested reader to Appendix A where more details are given on the estimation of the V1 complex cells activation, denoted by $C_i(t)$.

(ii) Nonlinearity: Passing from activation to mean firing rate is classically modeled by a nonlinearity, since the mean firing rate shows several nonlinearities due to response saturation, response rectification or contrast gain control mechanism

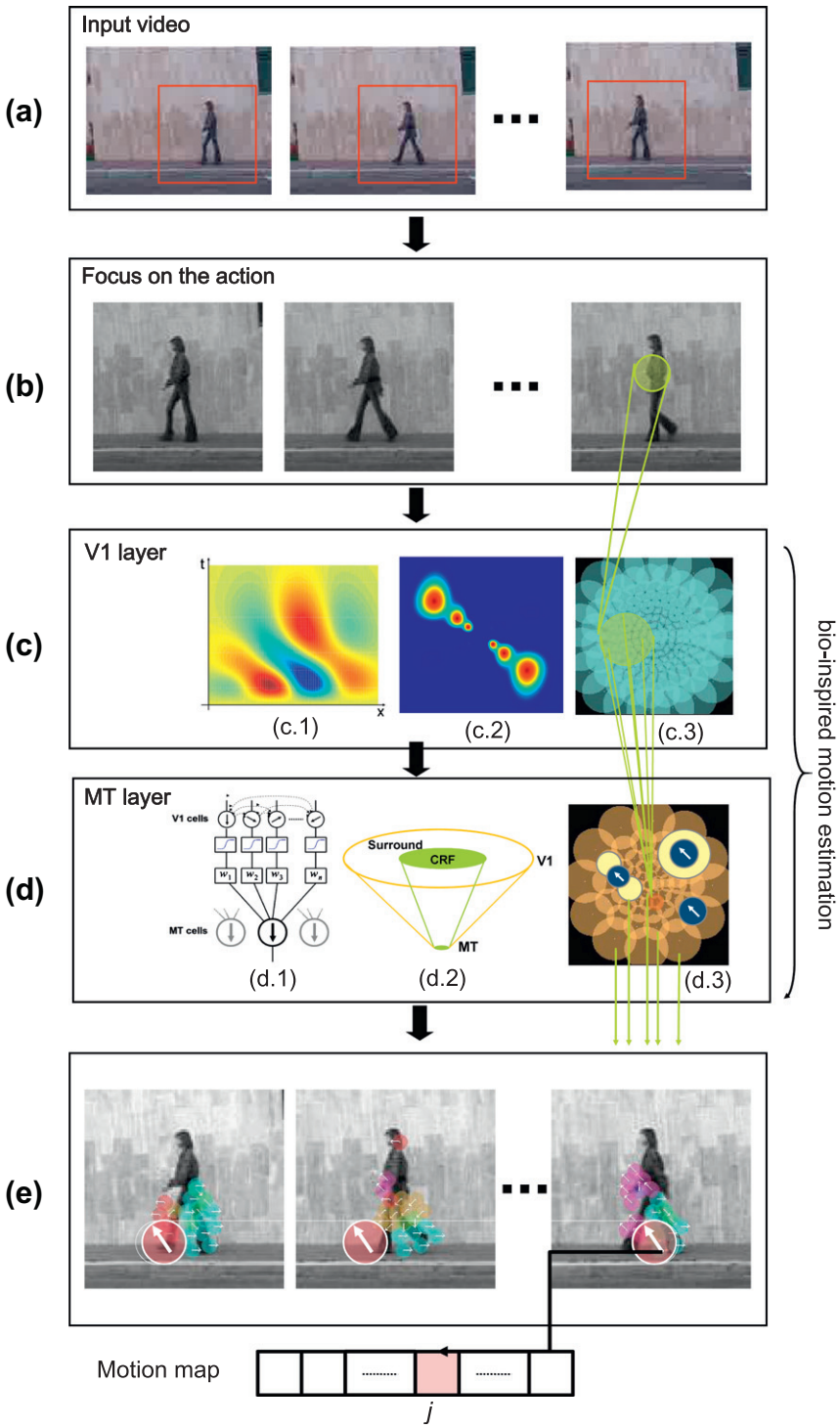

Fig. 1. Block diagram of the proposed approach to estimate our bio-inspired features. (a) Input is a real video sequence. (b) Videos are preprocessed in order to have a focus on the action to be labeled. (c) V1 cortical map is modeled via directional-selective filters (c.1 and c.2) applied to each frame of the input sequence. Cells are organized in a log-polar distribution (c.3). (d) MT cortical map is pooling the information coming from the V1 cortical map (d.1 and d.2). MT cells receptive fields are also organized in a log-polar distribution with different receptive field configurations (d.3). (e) The resulting bio-inspired feature corresponds to a *motion map* which is defined by the average activation of MT cells in time. The motion map has a length of $N_L \times N_c$ elements, where N_L is the number of MT layers and N_c is the number of MT cells per layer. It is this feature which is then used to label videos containing human motion.

([2]). So, it is classical to define the mean firing rate $r_i^{V1}(t)$ by $r_i^{V1} = S(C_i(t))$, where S is a nonlinear function (e.g., a Sigmoid or a Heaviside function).

Although this is a classical model, implementing motion energy detectors is not straightforward and requires some attention, in particular for proper cell tuning. For this reason, we propose the open-source library called ABfilters to simulate V1 complex cells (see Appendix A.3).

2.3. MT layer: higher order motion analysis

The second cortical area of interest in this article is MT area. MT cells pool incoming information from V1 according to the shape

and characteristic of their receptive field. Every V1 cell within the MT receptive field contributes to the MT cell activation, either incrementing or decreasing its activity. As far as modeling is concerned, for a given MT cell, each connected V1 cell has a respective connection weight; the set of all connection weights define the shape of the MT cell receptive field (Fig. 1d.1 and d.2).

As a general neuron model to describe MT cells, we chose the simplified conductance-based neuron model described by Destexhe et al. [16] for in vivo cell recordings. This conductance-based neuron model estimates the contribution of excitatory and inhibitory conductances in non-anesthetized animal cortical cells. Following this model, the membrane potential of the ith MT cell, denoted by $u_i^{MT}(t)$, is defined by

$$u_i^{MT}(t) = \frac{G_i^{exc}(t)E^{exc} + G_i^{inh}(t)E^{inh} + g^L E^L}{G_i^{exc}(t) + G_i^{inh}(t) + g^L}, \tag{1}$$

where E^{exc}, E^{inh} and E^L are constants with typical values of 70 mV, -10 mV and 0 mV, respectively; g^L is the leak reversal potential here considered as a constant; $G_i^{exc}(t)$ and $G_i^{inh}(t)$ are the excitatory and inhibitory conductances which will directly depend on V1 cells firing rate.

More precisely, the conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ are obtained by pooling the activity of all the pre-synaptic cells connected to it. Each MT cell has a receptive field built from the convergence of pre-synaptic afferent V1 complex cells (Fig. 1c.1). The excitatory conductance $G_i^{exc}(t)$ is related to the activation of V1 cells lying inside the center region of the MT cell (also called classical receptive field, denoted by CRF). The inhibitory conductance $G_i^{inh}(t)$ is related to the activation of V1 cells lying inside the surround region of the MT cell (Fig. 1c.2). Note that an important difference between the center and surround areas is that V1 cells in the surround cannot elicit a response to the MT cell if the V1 cells in the center are not activated: V1 cells in the surround can only modulate the activation of the MT cell when V1 cells in the center are activated. So, the input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ of the post-synaptic MT neuron i can be defined by

$$G_i^{exc}(t) = \max\left(0, \sum_{j\in\Omega_i} w_{ij} r_j^{V1}(t) + \sum_{j\in\Omega_i'} w_{ij} r_j^{V1}(t)\right), \tag{2}$$

and

$$G_i^{inh}(t) = \sum_{j\in\Phi_i} w_{ij} r_j^{V1}(t), \tag{3}$$

where

$$\begin{cases} \Omega_i = \{j \in \text{CRF} | \varphi_{ij} < \pi/2\}, \\ \Omega_i' = \{j \in \text{CRF} | \varphi_{ij} > \pi/2\}, \\ \Phi_i = \{j \in \text{Surround} | \varphi_{ij} < \pi/2\}, \end{cases}$$

and where φ_{ij} is the absolute angle between the preferred cell direction of the MT cell i and the preferred cell direction of the V1 cell j; w_{ij} is the efficacy of the synapse from neuron j to neuron i which depends on φ_{ij} and the relative positions (centers of the receptive fields) between cells. The precise definitions of w_{ij} will be given in Section 3. Finally, note that the values of the conductances will be always greater or equal to zero so that their positive or negative contribution to $u_i^{MT}(t)$ is due to the signs of E^{exc} and E^{inh}.

2.4. Retinotopic organization

In the V1–MT cortical maps, cells form a map of the visual field, also called a retinotopic map: This means that adjacent cells have receptive fields that include slightly different but overlapping portions of the visual field. Also, cell density and receptive fields vary according to their eccentricity (distance to the center of fixation). In this article, we reproduced these biological properties of the visual system: The centers of the receptive fields of both V1 and MT cells are arranged along a radial retinotopic scheme including a foveal uniform zone. More precisely, the density of cells, denoted by d, is defined by

$$d(r) = \begin{cases} d_0 & \text{if } r \leqslant R_0, \\ d_0 R_0 / r & \text{if } R_0 < r \leqslant R_{max}, \end{cases} \tag{4}$$

where r is the eccentricity. Cells with an eccentricity r less than R_0 belong to the fovea and have small receptive fields. Cells with an eccentricity greater than R_0 are outside the fovea and have receptive fields with a size increasing with r. This retinotopic organization of the V1–MT cells is illustrated in Fig. 1c.3 and d.3.

3. Modeling the richness of surround modulations

3.1. What biology tells us

The activation of a MT neuron inside its classical receptive field can be modulated by the activation of a surround area. This has been widely studied in the neuroscience community but is usually ignored in most MT-like models. Interestingly, in Born [7] the authors found two different types of MT cells:

(i) Cells purely integrative where only the response to the classical receptive field (CRF) stimulation is considered. These neurons without surround interactions strongly respond to wide-field motion.
(ii) Neurons with antagonistic surrounds modulate the response to CRF stimulation, normally suppressing it. These neurons are unresponsive to wide-field motion but they strongly respond to local motion contrasts.

Regarding direction selectivity, the direction tuning of the MT surround is broader than that of the center and it tends to be either the same or opposite, but rarely orthogonal [7]. This characteristic was modeled and implemented by Beck et al. [4] for an object segmentation application. In Beck et al. [4] the authors implemented symmetric MT surrounds with different direction selectivities to detect motion boundaries, combining them with temporal occlusion, and thus, to improve the detection of kinetic boundaries in artificial and real scenarios. Apparently, inhibitory surrounds contain key information about motion characterization, such as motion contrasts.

But the geometries of MT surrounds are far from just being symmetric. Half of MT neurons have asymmetric receptive fields introducing anisotropies in the processing of spatial information [61,32]. Within the half of MT neurons with symmetric surrounds, Xiao et al. [61] reported two types of interactions: circular symmetric surrounds (20% of the whole population) and bilaterally symmetric surrounds. The latter is formed by a pair of surrounding regions on opposite sides (25% of the whole population). Neurons with asymmetric receptive fields seem to be involved in the encoding of important surface features, such as slant and tilt or curvature. Different surround geometries could be the main actor of dynamic effects observed in MT neuron response, such as, component-to-pattern behavior [53] or changes in motion direction selectivity [41,40]. Finally, note that in most of the cases MT surround suppresses the response of the MT classical receptive field. But the surround could also facilitate neuron response depending on the input stimulus [22] and contrast [7].

Considering all this information, and more specifically the work by Xiao et al. [61], we propose to model four types of MT cells as shown in Fig. 2: One basic type of cell only activated by its CRF,

and three other types with inhibitory surrounds. The tuning direction of the surround is always the same as the CRFs, but their spatial geometry changes, from symmetric-isotropic to symmetric-anisotropic and asymmetric-anisotropic surround interactions. In the next section, we present the analytical expression of the connection weights corresponding to these geometries.

3.2. Connecting V1 to MT

From Eqs. (2) and (3), it remains to define the weights w_{ij} for both excitatory and inhibitory conductances. Let us consider an MT cell i with a classical receptive field of radius ρ, located at $\mathbf{x}_i$, with a motion direction selectivity of α_i radians. According to the biological properties mentioned above, we define

$$w_{ij} = \begin{cases} k_c w_d(\mathbf{x}_i - \mathbf{x}_j)\cos(\varphi_{ij}) & \text{if } j \in \Omega_i \text{ or } j \in \Phi_i, \\ -k_c w_d(\mathbf{x}_i - \mathbf{x}_j)\cos(\varphi_{ij}) & \text{if } j \in \Omega_i' \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where k_c is an amplification factor and w_d is a function of the difference between the positions (centers of the receptive fields) of the cells i and j. The specification of the function w_d will then completely define the connection weights, and it is the term defining the geometries described in Section 3.1. Note that negative weights in Eq. (5) are included in order to improve the direction selectivity of MT cells eliminating the two-blob shape that characterizes V1 cells [43]. The result of this pooling mechanism (without any other interaction or dynamic) improves the direction selectivity of MT cells, obtaining strong motion direction selectivity.

The analytical expression of the function w_d will depend on the type of MT cell that is considered. The four kinds of center–surround interactions treated in this article are represented in Fig. 2. Analytically, if we denote $\boldsymbol{\delta x} = \mathbf{x_j} - \mathbf{x_i}$, the function w_d can be decomposed as

$$w_d(\boldsymbol{\delta x}) = w_c(\boldsymbol{\delta x}) - w_s(\boldsymbol{\delta x}). \tag{6}$$

In case (a) where no surround modulation occurs, we have:

$$w_c(\boldsymbol{\delta x}) = \frac{\exp\left(-\boldsymbol{\delta x}^2/2\sigma_c^2\right)}{\sigma_c\sqrt{2\pi}} \quad \text{and} \quad w_s(\boldsymbol{\delta x}) = 0.$$

In case (b) with a symmetric and isotropic surround, we have

$$w_c(\boldsymbol{\delta x}) = \frac{\exp\left(-\boldsymbol{\delta x}^2/2\sigma_c^2\right)}{\sigma_c\sqrt{2\pi}} \quad \text{and} \quad w_s(\boldsymbol{\delta x}) = \frac{\exp\left(-\boldsymbol{\delta x}^2/2\sigma_s^2\right)}{\sigma_s\sqrt{2\pi}}$$

In case (c) with a symmetric and anisotropic surround, we have

$$w_c(\boldsymbol{\delta x}) = \frac{\exp\left(-\boldsymbol{\delta x}^2/2\sigma_c^2\right)}{\sigma_c\sqrt{2\pi}} \quad \text{and}$$

$$w_s(\boldsymbol{\delta x}) = \frac{\exp\left(-(\boldsymbol{\delta x}-\boldsymbol{\mu})(\boldsymbol{\delta x}-\boldsymbol{\mu})^T/2\sigma_{ss}^2\right)}{\sigma_{ss}\sqrt{2\pi}} + \frac{\exp\left(-(\boldsymbol{\delta x}-\boldsymbol{\nu})(\boldsymbol{\delta x}-\boldsymbol{\nu})^T/2\sigma_{ss}^2\right)}{\sigma_{ss}\sqrt{2\pi}}$$

In case (d) with a asymmetric and anisotropic surround, we have

$$w_c(\mathbf{x}) = \frac{\exp\left(-\boldsymbol{\delta x}^2/2\sigma_c^2\right)}{\sigma_c\sqrt{2\pi}} \quad \text{and}$$

$$w_s(\mathbf{x}) = \frac{\exp\left(-(\boldsymbol{\delta x}-\boldsymbol{\mu})(\boldsymbol{\delta x}-\boldsymbol{\mu})^T/2\sigma_{ss}^2\right)}{\sigma_{ss}\sqrt{2\pi}}.$$

For all these expressions, we chose

$$\boldsymbol{\mu} = \begin{bmatrix} \rho/2 + 3\sigma_{ss}\cos(\alpha) \\ \rho/2 + 3\sigma_{ss}\sin(\alpha) \end{bmatrix} \quad \text{and} \quad \boldsymbol{\nu} = \begin{bmatrix} \rho/2 + 3\sigma_{ss}\cos(\alpha+\pi) \\ \rho/2 + 3\sigma_{ss}\sin(\alpha+\pi) \end{bmatrix},$$

where the values of σ_c, σ_s and σ_{ss} are set as $\rho/3$, $2.2\rho/3$ and $\rho/3$, respectively. Details about parameter settings are described in Section 4.1.

4. Action recognition via bio-inspired features

4.1. Specification of the bio-inspired architecture

4.1.1. Focus on the action

Based on what we mentioned in Section 2.1, the goal of the preprocessing stage is to focus on the action. To do so, we used the approach developed by Kornprobst et al. [25]. In Kornprobst et al. [25], the authors proposed a variational approach to restore and segment in coupled way noisy videos with a static background (i.e., no camera motion and no change of focus). Using this approach to detect the foreground, the videos can be easily cropped leaving the action at the center. After cropping the actions in the original videos, all images were resized to be 210×210 pixels.

4.1.2. V1 and MT cortical maps

The parameters used for the V1 and MT cortical maps and neurons are shown in Table 1. They were defined following functional properties reported in the neurophysiology literature. This was done *qualitatively* because: (i) most of the neurophysiological data available in the literature is obtained using artificial input stimulus, such as, gratings, plaids or random dots; and (ii) neurons do not respond equally when the sensory visual system is stimulated by natural images instead of artificial stimuli [34,59].

The functional properties of interest for this article are the following:

- Connection weights between V1 and MT neurons were chosen in order to reproduce two main properties: (i) MT neurons are selective to motion direction independently of the motion speed [27,47,39]; (ii) MT neurons have a strong inhibition in the anti-preferred direction [27,43].
- MT surround characterization was chosen in order to reproduce three main properties: (i) Half of MT neurons have asymmetric surrounds [32,60,61,11]; (ii) The motion direction selectivity of the surround, with respect to that of the center, tends to be

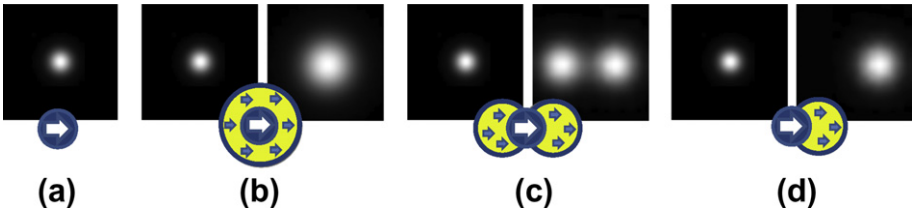

Fig. 2. MT center–surround interactions modeled in our approach. In (a) no surround modulation occurs. In (b) the surround is symmetric and isotropic. In (c) the surround is symmetric and anisotropic. In (d) the surround is asymmetric and anisotropic.

Table 1
Parameters used for V1 and MT cortical maps and neurons.

	Weizmann		KTH	
	V1	MT	V1	MT
Fovea radius (R_0)	80 (pixels)	40 (pixels)	80 (pixels)	40 (pixels)
Layer radius (R_{max})	100 (pixels)	100 (pixels)	100 (pixels)	100 (pixels)
Cell density in fovea (d_0)	0.3 (cells/pixel)	0.08 (cells/pixel)	0.2 (cells/pixel)	0.08 (cells/pixel)
Receptive field size ($\varnothing$) [fovea]	8, 17, 34 (pixels)	20 (pixels)	8, 17, 34 (pixels)	24 (pixels)
Number of motion directions	8	8	8	8
Number of layers (N_l)	72	8, 16, 32	72	8, 16, 32
Number of cells per layer (N_c)	2604	106	1147	106
Leak (g_L)	–	0.1	–	0.1

either the same or opposite but rarely orthogonal [7]; (iii) The extent of the MT surrounds has a characteristic size two times bigger than the classical receptive field [61].

In our model, V1 neurons are arranged into layers. Each V1 layer contains V1 complex cells sharing the same motion direction selectivity and the same spatiotemporal frequency tuning (see Section 2.2). V1 layers are distributed in the frequency space in order to pave the spatiotemporal frequency space of interest in a homogeneous manner. Following the work done by Mante and Carandini [33], the frequency space of interest is limited to a spatial frequency range of 0.05–0.2 cycles/pixel, and temporal frequency range of 2–8 cycles/s (in a digitalized version, 0.08 and 0.32 cycles/frame, respectively). Within this frequency space we considered three different spatial frequencies (0.05, 0.1 and 0.2 cycles/pixel) and three temporal frequencies (2, 4, and 8 cycles/s). So, with eight possible motion directions, there is a total of $N_l = 72$ V1 layers. Using these values, the power spectrum for a given motion direction θ is shown in Fig. 3.

Similarly, MT neurons are also arranged into layers but here, each layer contains MT neurons sharing the same center–surround interaction and motion direction selectivity. So, with eight possible directions, we will have $N_l = 8$, 16 or 32 layers depending on the MT center–surround interactions chosen in the model (Fig. 2).

Concerning the size of receptive fields ($\varnothing$), there are two cases. Inside the fovea, the size of receptive fields for V1 neurons is given by $\varnothing = 3\sigma$ where σ depends on the spatial frequency (Appendix A, Eq. (A.3)). In Table 1, the three possible values are given. For MT neurons, the size was chosen to be fixed. Outside the fovea, the receptive field sizes, for both V1 and MT neurons, are scaled by a factor equal to the inverse of the density $d(r)$ (Eq. (4)).

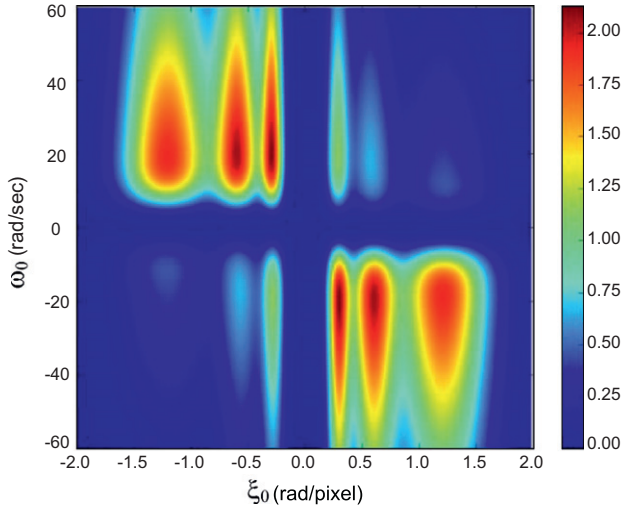

Fig. 3. Power spectrum of all V1 complex cells used in this article. This image was obtained combining nine V1 complex cells with spatial frequency tuning $\xi_0 = \{0.05, 0.1, 0.2\}$ (cycles/pixel) and temporal frequency tuning $\omega_0 = \{2, 4, 8\}$ (Hz).

Using this configuration, Fig. 4 shows the 20 most activated MT cells, considering only the interaction of its classical receptive field, for a video from the Weizmann database (*jumping-jack denis*). Receptive fields are represented with a color code corresponding to the motion direction-tuning of the cells. This example illustrates the behavior of our system to properly detect motion direction.

4.2. Classification approach and feature space

In previous sections, we have proposed a model of MT cells so that, given an input video, we can estimate an activity at each time step for each cell. Now the goal is to investigate whether this activity can be used to categorize the action present in the input video.

To do so, in this article we propose (i) to define features from the MT cells activity as well as a distance between features; (ii) to use a standard supervised classification method (which has no biological inspiration). Here we considered the simplest case of supervised classification with training sets (category is assumed to be known for some videos). Then, the recognition performance is evaluated on the testing set, defined by remaining videos.

Given a video stream $L(\mathbf{x}, t)$ and the corresponding membrane potential for each MT cell, we define the feature vector by a vector $H^L \in R^{N_l \times N_c}$ whose components are the time-average membrane potential of the MT cells during the length of the video (Fig. 1e). Thus, H^L is defined by:

$$H^L = \left\{\gamma_j^L\right\}_{j=1,\ldots,N_l \times N_c} \quad \text{with} \quad \gamma_j^L = \frac{1}{T}\int_T u_j^{MT}(s)ds, \tag{7}$$

where N_l is the number of MT layers, N_c is the number of MT cells per layer, and, T is the total length of the video. In the sequel, H^L will be designated as the *motion map*.

From this definition, there is one major remark to be made. By averaging over all the duration of the video, it is implicitly assumed that the video contains only one action with a constant viewpoint. If in the same video, a person was running and then jumping, or running and changing direction, the averaged activation in time would not be appropriate. This is a technical assumption that could be removed by considering sliding time windows to estimate dynamical motion maps (and not one motion map for the entire video). This is currently under research but beyond the scope of this article.

To compute the similarity between the *testing motion map* and the *training motion maps*, we used two measures: triangular discrimination (TD, see [19]) and symmetric Kullback–Liebler (SKL) divergence.

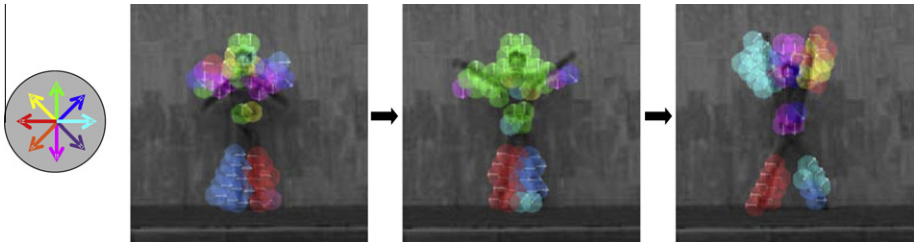

Fig. 4. Representation of the 20 most activated MT cells for a video from the Weizmann database (*jumping-jack denis*). Receptive fields are represented here with the color code corresponding to the motion direction-tuning of the cells, given on the left hand-side.

Table 2
Recognition performance for Weizmann database with the triangular discrimination (TD) and symmetric Kullback–Liebler (SKL) divergence.

Distance → Size of training set →	TD All 84 trials (%)	TD Five random trials (%)	SKL All 84 trials (%)	SKL Five random trials (%)
Case (i)	62.48 ± 2.08	–	63.01 ± 2.07	–
Case (ii)	87.52 ± 2.23	–	87.87 ± 2.03	–
Case (iii)	**96.34** ± 0.72	**98.53** ± 2.02	**96.47** ± 0.81	**99.26** ± 1.66

4.3. Results on Weizmann database

Weizmann database[2] contains nine subjects performing nine actions: 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*). The number of frames per sequence varies and depends on the action.

Following the same experimental protocol described in Jhuang et al. [24], we select six random subjects as a training set ($6 \times 9 = 54$ videos) and we use the remaining three subjects as a testing set ($3 \times 9 = 27$ videos). Motions maps are first estimated for every video from the training set. Then, for each video from the testing set, a motion map is estimated together with the distances to each motion map from the training set. The class of the video from the training set with the shortest distance is then kept and classification errors are estimated.

Our main result is to evaluate how the system benefits from the richness of the center–surround interaction defined in the feature space. To do so, we ran experiments with the different configurations of center–surround interaction. We considered three cases:

- Case (i): CRF only (Fig. 2a)
- Case (ii): CRF and the isotropic surround interaction (Fig. 2a and b)
- Case (iii): CRF and both isotropic and anisotropic surrounds (Fig. 2a–d)

The recognition performance is shown in Table 2. Fig. 5 shows the confusion matrices obtained for each type of motion map using symmetric Kullback–Leibler divergence. A significant improvement (paired t-test $P < 0.0001$) in the recognition performance is obtained when all the surround geometries described in Fig. 2 are considered, case (i) versus case (iii).

In Table 2 we can also observe a high variability in the recognition performance depending on which subjects formed the training set. So, when results are shown based only on five random training sets (such as in, e.g., [24]), it is *a priori* not thorough enough to draw strong conclusions about the performance. To overcome this difficulty, we estimated the recognition performance over all the possible training sets that can be built with six subjects, giving a total of 84 training sets. Table 2 also shows the resulting recognition performance obtained considering triangular discrimination and only five random training sets. The variability observed in the recognition performance is represented by the histograms shown in Fig. 6, where the abscissa is the number of mismatched video from the testing set, and the ordinate is the number of times, over the 84 trials, that the number of mismatched sequences was obtained. In other words, the histograms shown in Fig. 6 approximate the probability distribution of the recognition performance obtained with the architecture proposed in this article.

Finally, we tested the robustness of our approach by considering videos with different kinds of perturbations (Fig. 7): (a) Gaussian noise of 12%, (b) Gaussian noise of 20%, (c) occlusion and (d) moving textured background. Both *noisy* (a) and (b) and *legs-occluded* (c) videos were created starting from the *walking-denis* video, which was extracted from the training set for the robustness experiments. The *legs-occluded* video sequence was created placing a black box on the original video before the centered cropping. The *noisy* videos were created adding Gaussian noise. The *moving-background* video was taken from Blank et al. [6]. For the four videos tested, the recognition was correctly performed as *walking*. More precisely, a graph with the ratio between the shortest distance to *walking* class and the distance to the second closest class (*running* or *galloping-sideways* in our tests) is shown in Fig. 7. This result shows another benefit of using rich center–surround interaction: taking into account anisotropic surround interaction makes the model less sensitive to occlusions or noise.

4.4. Results on KTH database

KTH database[3] contains 25 subjects performing 6 actions: *handclapping*, *handwaving*, *boxing*, *jogging*, *running* and *walking*. The sequences are separated in four different scenarios: outdoors (d1), outdoors with scale variations (d2), outdoors with different vestment (d3) and indoor with lighting variations (d4). Note that for some videos, the action is repeated several times (e.g., for running, one video may contain a person crossing the scene several times). Because of the different scenarios, this database is more challenging than the Weizman database.

[2] The Weizmann database can be downloaded from http://www.wisdom.weizmann.ac.il/vision/SpaceTimeActions.html.

[3] The KTH database can be downloaded from http://www.nada.kth.se/cvap/actions.

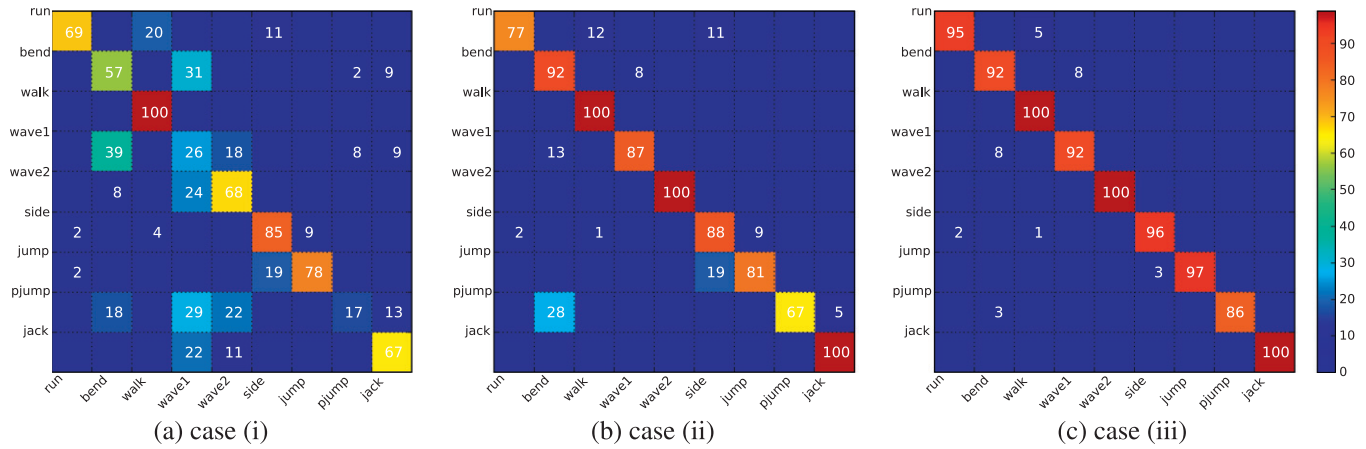

Fig. 5. Confusion matrices obtained for Weizmann database and symmetric Kullback–Leibler divergence representing the recognition performance obtained for the 84 different training sets. Numbers inside the boxes indicate the percentage of correct (diagonal) or mismatched (non-diagonal) classification.

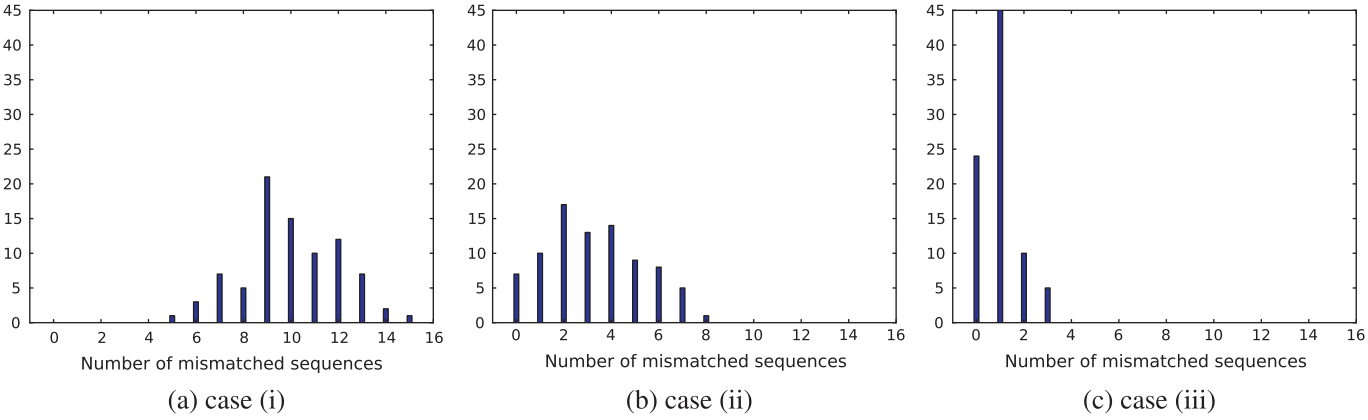

Fig. 6. Histograms obtained for Weizmann database and the symmetric Kullback–Leibler divergence. The abscissa represents the number of mismatched sequences in the training set, and the ordinate, the number of times (of a total of 84 trials) that this number of mismatched sequences was obtained. The histograms are shown for the three MT center–surround interactions considered in this article: (a) CRF, (b) CRF + isotropic surrounds, and (c) CRF + isotropic/anisotropic surrounds.

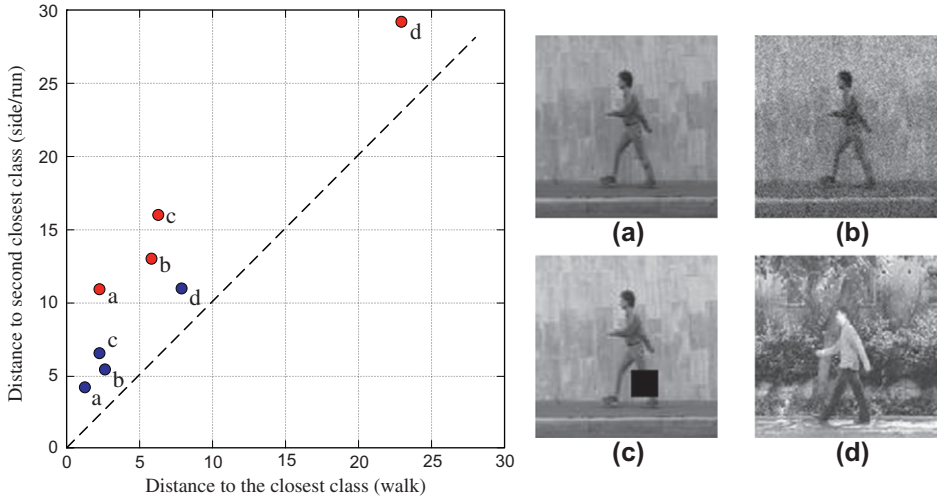

Fig. 7. Graph showing the robustness of our approach with four kinds of perturbations: (a) *walking-denis* modified with a gaussian noise of 12%; (b) *walking-denis* modified with a gaussian noise of 20%; (c) *walking-denis* modified with an occlusion over the legs; (d) video from Blank et al. [6] with a moving background. Each point represents the ratio between the action recognized (*walking*) and the second closest action (*running* or *galloping-sideways*). *Blue points* represent the performance obtained for motion maps considering only the activation of MT classical receptive field (case (i)). *Red points* represent the performance obtained for motion maps considering all center–surround interaction shown in Fig. 2 (case (iii)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Recognition performance for KTH human database obtained with triangular discrimination.

Size of training set → considering *running* →	100 Trials No (%)	100 Trials Yes (%)	Five random trials No (%)
d1-case (i)	55.24 ± 2.69	–	–
d1-case (iii)	**83.09** ± 1.95	74.63 ± 2.82	**92.00** ± 0.01
d3-case (i)	47.40 ± 2.22	–	–
d3-case (iii)	**69.75** ± 2.81	65.48 ± 2.81	**84.44** ± 1.22
d4-case (i)	47.82 ± 2.71	–	–
d4-case (iii)	**83.84** ± 1.90	71.19 ± 2.66	**92.44** ± 0.01

As before, our approach starts with preprocessing the data so that the input will be videos of 210 × 210 pixels, with the action which is self-centered and represented at a given scale (see Section 2.1). But there are two main differences with respect to the Weizmann database:

- For the KTH videos where actions are repeated, besides cropping and resizing the actions, we also selected only the first instance of the action (and we did not consider the other repetitions of the same action). Note that it is not always clear how other approaches deal with this concern. In our case, this choice had to be made simply because in our framework, we represent a video by the time-average activity of MT cells, so that estimating an average activity corresponding to different instances of an action would be ill-posed. But by doing so, starting from initial videos, our preprocessing may sometimes only keep very short pieces of videos. For example, this is the case for some *running* videos (one action will contain only a few frames), causing problems related to the required time to correctly estimate

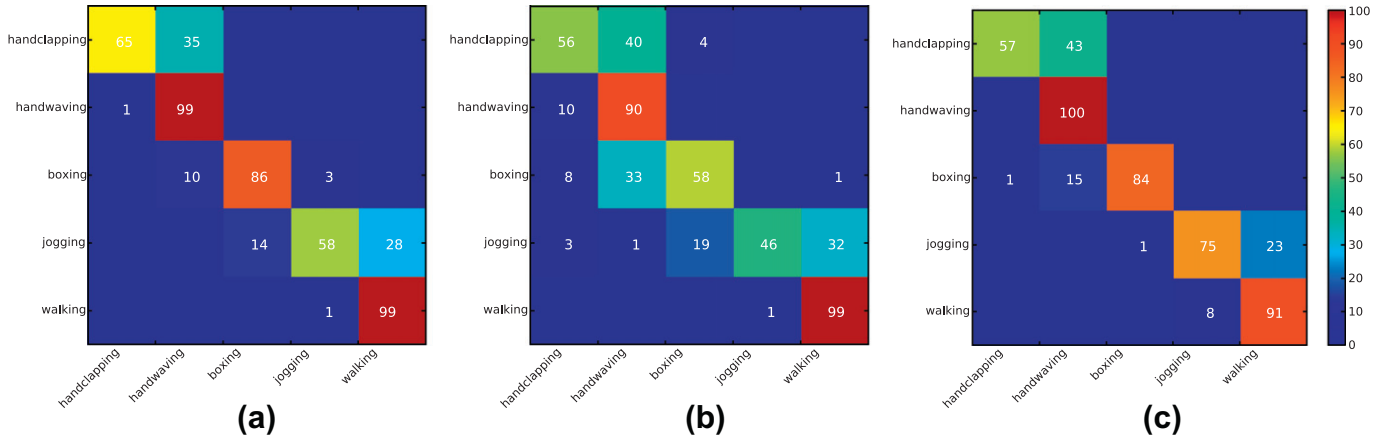

Fig. 8. Confusion matrices, obtained for KTH human database and triangular discrimination divergence, representing the recognition performance obtained for 100 different training sets. Numbers inside the boxes indicate the percentage of correct (diagonal) or mismatched (non-diagonal) classification. These matrices display the results obtained considering case (iii) for (a) d1, (b) d3 and, (c) d4.

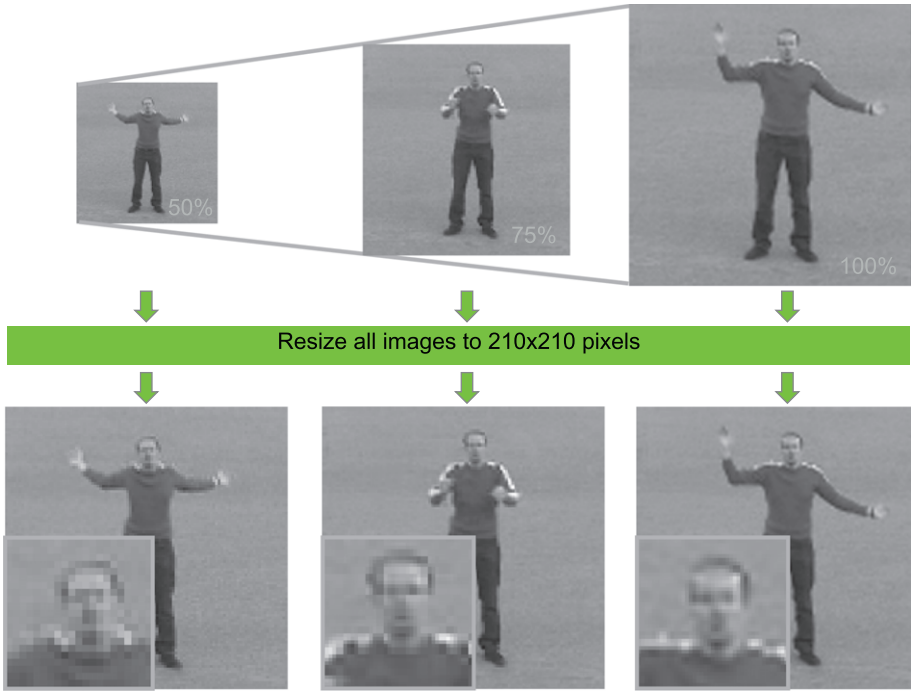

Fig. 9. Synthesized videos where zoom level is changed manually. The top row shows images zoomed by a factor of 0.5–1 (50–100%, respectively). The bottom row shows zoomed images rescaled to the original dimensions, resulting in images with different resolutions.

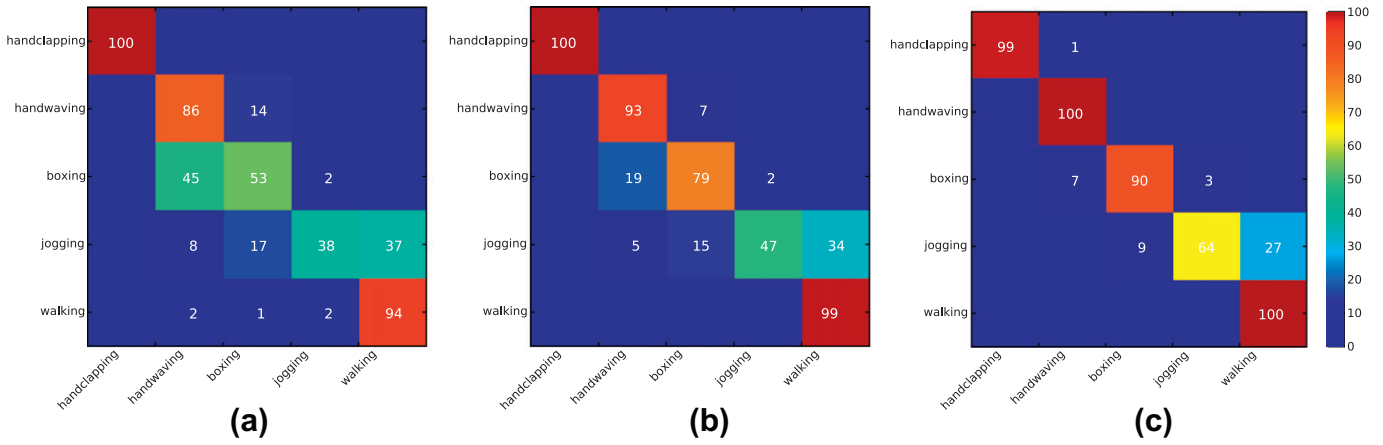

Fig. 10. Confusion matrices, obtained for synthesized zoomed KTH human database and triangular discrimination divergence. The confusion matrices represent the recognition performance obtained for 100 different training sets. Numbers inside the boxes indicate the percentage of correct (diagonal) or mismatched (non-diagonal) classification. These matrices display the results obtained considering case (iii) for (a) d1, (b) d3 and, (c) d4.

the temporal filtering associated to V1 motion detectors. In the sequel, we will show both results: with and without *running* sequences in the dataset.

- In the results presented therein, we did not consider the scenario d2 because it requires more development at the preprocessing level. The scenario d2 presents two main difficulties: the background is not always fixed (change of focus), and also, large shadows are sometimes present in the videos. As mentioned in Section 4.1, the approach Kornprobst et al. [25] can only handle videos with static backgrounds. Therefore, more sophisticated approaches are needed for tracking humans. This is still an active research field in the computer vision community (e.g., [48,65,62,30,26]. Note that the fact that the actor may vary in size is not a problem in our approach since actions are resized before being processed through the V1–MT cortical maps (see also Section 4.5).

Similar to what has been done for the Weizmann database, we first followed the experimental protocol described by Jhuang et al. [24]: We selected 16 random subjects as a training set ($16 \times 25 = 400$ sequences) and the remaining 9 subjects as a testing set ($9 \times 25 = 225$). In Jhuang et al. [24], the authors presented results averaged over five randomly selected training sets. As remarked in the previous section, this may induce strong biases. But considering all the possible combinations of 16 subjects as a training set gives a total of 2,042,975 possible training sets which would be numerically very expensive to do. So, as a compromise in order to add more representability to our results, we estimated results from 100 randomly selected training sets. The recognition performance for the 100 training sets, with and without *running* sequences, is shown in Table 3. The respective confusion matrices are shown in Fig. 8.

Results confirm the improvement in the recognition rate obtained when the different MT center–surround configurations are used to build the *motion maps*. For the three subsets (d1,d3,d4) the results obtained in the case (iii) are significantly better than the ones obtained in the case (i) (paired t-test $P < 0.0001$). The effect of *running* sequences in the recognition performance is also significant. The inclusion of *running* sequences decreases the recognition rate of the system, as we previously mentioned, mainly because running sequences do not have enough frames to make the response of V1 complex cells converge.

4.5. About scale invariance

In this section, we show the scale invariance property of our model by considering synthesized videos where zooming level is changed manually. To do so, we started from the *d1* subset of the KTH database and introduced a zooming factor, as shown in Fig. 9: The first frame of the video was resized to 50% of its original size, and then the zooming factor was gradually increased to one. These frames represent the cropped actions that are then resized to the original dimension (210×210 pixels), resulting in images with different resolutions.

We tested our approach on these synthesized videos with the same procedure as described in Section 4.4 (100 random training sets). The confusion matrices are shown in Fig. 10, and the recognition rates are: **74.27%** ± 1.94 for case (i), **83.73%** ± 1.90 for case (ii), and, **90.56%** ± 1.43 for case (iii). Interestingly, we observe that the action recognition rate improved in comparison to the original videos results, reaching up to 90.56% success. This suggests that the parameters used to define the motion energy filters may not be the most suitable, given the characteristics of the videos. An interesting perspective will be to investigate how to choose the spatiotemporal frequency bandwidth of the energy filters better, which should lead to further improvements in recognition rates.

5. Conclusion

In this article we proposed bio-inspired motion features for action recognition. Our approach is based on a state-of-the-art model of the visual stream to process motion. The model has two main stages. The first stage consists of focusing on the action which is an important condition for the visual system to recognize actions. The second stage consists of developing a model of the V1 and MT cortical areas. The V1 cells are modeled following classical ideas from the literature, and we provide open source code to simulate these cells. Our main novelty is to take into account the richness of center–surround interactions at the level of MT: We modeled different kinds of MT cells, corresponding to different center–surround configurations. Then, based on the average activity of MT cells, bio-inspired features were defined and used to perform action recognition via a standard classification procedure (not related to biological facts).

Our method has been tested on two classical databases (Weizman and KTH) and comparisons were made to the state-of-art literature. But comparing recognition rates is not as simple as it seems and needs to be done carefully. In general, there are three concerns. The first concern is the definition experimental protocol, that is not only the size of the training sets but also the composition or number of training sets used to obtain the rates. For instance, Jhuang et al. [24] considered a very small number of training sets (five subjects) which may be not representative with

respect to all possible training sets. Similarly, in KTH dataset Le et al. [29] and Wang et al. [57] considered a single fixed training set to evaluate their algorithms. In our opinion it is misleading to compare recognition rates obtained in different testing conditions. Indeed, we have observed a high variability in the recognition rate depending on the conforming subjects in the training sets. That is why we reported here histograms of recognition rates for many different training sets (all of them for the Weizmann dataset), and that our average is estimated over this ensemble. The second concern is the classifier. In general, sophisticated classifiers are used, such as, linear SVM [63] or nonlinear SVM [24,29,57]). In our case, since our goal was to compare the information conveyed by the different bio-inspired features here proposed, we did not use any sophisticated classifier to improve further the recognition rates and we used the simplest one. Finally, the third concern is the preprocessing of the videos. In general, details are missing concerning the preprocessing done on each database, which is a very critical point since it significantly influences the level of performance that can be reached.

Our main result is that the bio-inspired features introduced therein achieve a high level of performance, being also robust to perturbations such as noise, small occlusions or different backgrounds. More importantly, we also showed that taking into account the diversity of MT center–surround interaction enriches motion representation by improving the action recognition rate.

More precisely, results concerning the two databases can be commented as follows. Concerning the Weizmann database, the recognition rates reported in this article were obtained considering all the possible training sets that can be built using six subjects. In order to compare our results with the ones reported by Jhuang et al. [24], Table 2 also shows the recognition rates obtained for five random trials. The results here shown here are better, but statistically they are not significantly better (t-student $P = 0.1787$). Concerning the KTH database, some technical difficulties appeared: Due to the preprocessing needed by our approach to be *focused on the action*, some sequences were cropped and centered leaving only very few frames. But our V1 motion detectors need time to correctly compute the temporal convolutions, and thus, for some actions the response obtained by our system is not accurate. This effect is critical in the KTH-*running* sequence, were there are not enough frames to make the response of V1 motion detectors converge. In order to isolate this effect, we calculated the performance of our model including and excluding *running* sequences from the database.

One clear advantage of our model is that it is generic: Unlike Giese and Poggio [20], there is no need to tune the properties of local motion given the specific application of action recognition. And unlike optical-flow based models, where a single velocity is assigned to each point, our model reproduces to some extent the richness of center–surround interactions, giving different kinds of motion contrasts for several orientations at every point. Our interpretation is that cells with inhibitory surrounds bring information related to velocity opponency or singularities in the velocity field of the input stimulus (see also [4]).

Of course this approach could be extended in several manners and three main perspectives seem promising. The first perspective is essentially technical and consists of improving the system at two levels by (i) using a more sophisticated human tracking approach as mentioned in Section 4.4 and (ii) defining dynamical motion maps by averaging the activity of MT cells over sliding time windows (instead of the entire video). This would allow us to deal with more complex videos. The second perspective is to enrich the model adding other brain functions or cortical maps. Of course, the motion pathway is not the only actor in action recognition in the visual system. Like every motion-based approach for action recognition, our approach is likely to be limited [20,51]. It will fail in complex situations such as those with large occlusions, complex backgrounds or multiple persons. To do this, one has to consider more complex processing corresponding to additional brain areas (e.g., V2, V4 or IT) and top-down mechanisms such as attention (e.g., [55]. The third perspective is to investigate how speed selectivity could be achieved. Speed coding relies on complex and unclear mechanisms. Many studies on MT focus on motion direction selectivity (DS), but very few on speed selectivity (see, e.g., [47,44,31]), showing that speed coding relies on complex and unclear mechanisms. Based on this, here we only considered the motion direction and not the motion speed, as can be seen in Eq. (2): Our MT cells pool V1 cells considering only their motion direction selectivity, and not their spatiotemporal tuning. However, note that it is also possible to pool different V1 cells in order to extract some speed information, as proposed for example in Simoncelli and Heeger [52], Grzywacz and Yuille [21], Perrone [42]. As a result, one could obtain a velocity field qualitatively similar to an optical flow (i.e., one velocity per position), which could be used for action recognition but also compared to the literature in optical flow.

Acknowledgments

This work was partially supported by the EC IP project FP6–015879, FACETS, UTFSM by DGIP–Grant 231009 and CONICYT Chile by the Programa de Inserción de Capital Humano Avanzado en la Academia Nro. 79100014.

Appendix A. Model of the V1 cells and software library

A.1. V1 simple cells

Simple cells are characterized by 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 cells. The direction-selectivity refers to the property of a neuron to respond to the direction of the stimulus' movement. The way to model this selectivity is obtaining receptive fields oriented in space and time (Fig. 1c.1).

Given an input stimulus $L(\mathbf{x},t)$, the response of a spatiotemporal oriented V1 simple cell $F^s(x,t)$ is obtained by the convolution

$$L(x,t) * F^s(x,t), \tag{A.1}$$

where $F^s(x,t)$ can be defined by one of the following filters

$$\begin{aligned} F^a(x,t) &= F^{odd}(x)H_{fast}(t) - F^{even}(x)H_{slow}(t),\\ F^b(x,t) &= F^{odd}(x)H_{slow}(t) + F^{even}(x)H_{fast}(t), \end{aligned} \tag{A.2}$$

which are spatially located at $x = (x,y)$. The spatial components $F^{odd}_{\theta}(x)$ and $F^{even}_{\theta}(x)$ of each conforming simple cell are the first and second derivative of a Gabor function spatially oriented in θ, with spatial frequency f and a standard deviation of (Watson and Ahumada [58]):

$$\sigma = 0.5622/f. \tag{A.3}$$

The temporal contributions $H_{fast}(t)$ and $H_{slow}(t)$ are defined by

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

where $T_{\eta,\tau}(t)$ is a Gamma function defined by

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

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 [15] or to be related

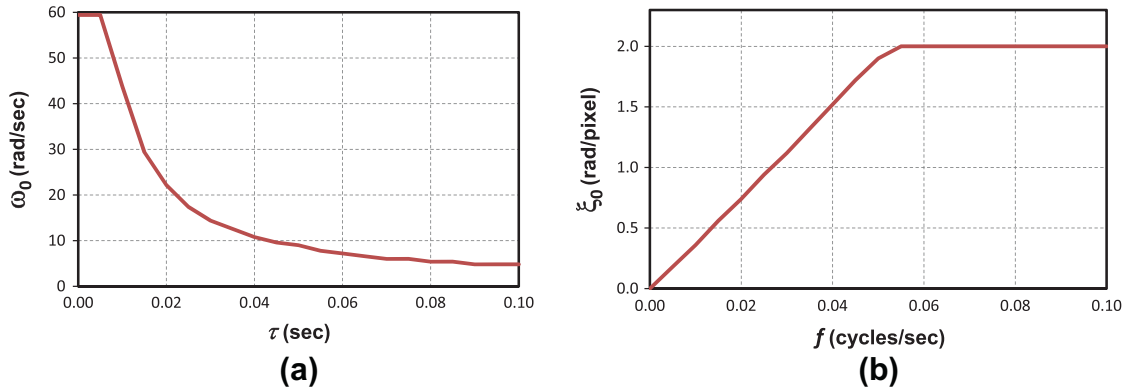

Fig. A.11. Values of ω_0 and ξ_0 as a function of the input parameters f and τ. (a) ω_0 as a function of τ for $|\tilde{F}^a(\xi,\omega)|^2$ (blue) and $|\tilde{F}^b(\xi,\omega)|^2$ (red). (b) ξ_0 as a function of f for $|\tilde{F}^a(\xi,\omega)|^2$ (blue) and $|\tilde{F}^b(\xi,\omega)|^2$ (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to the delayed inhibition in the retina and LGN (Conway and Livingstone [14]). Fig. 1c.1 shows the respective spatial and temporal contribution for a V1 simple cell defined by $F^a(x,t)$.

Note that the causality of $H_{fast}(t)$ and $H_{slow}(t)$ generates a more realistic model than the one proposed by Simoncelli and Heeger [52] (see also [24]), where a Gaussian is proposed as temporal profile, which is non-causal and inconsistent with V1 physiology.

The spatial parameters of the Gabor function: θ, f and σ; and the temporal parameter τ of the Gamma function (Eq. (A.5)) define the spatiotemporal orientation of V1 simple cells $F^a(x,t)$ and $F^b(x,t)$.

The spatiotemporal orientation of a V1 simple cell is better visualized in the Fourier space. In the Fourier space the power spectrum of a V1 simple cell ($|\tilde{F}^a(\xi,\omega)|^2$ for $F^a(x,t)$ and $|\tilde{F}^b(\xi,\omega)|^2$ for $F^b(x,t)$) is described by two blobs centered at $(-\xi_0,\omega_0)$ and $(\xi_0,-\omega_0)$, where $\xi_0 = (\xi_0^x \xi_0^y)$ and ω_0 are the preferred spatial and temporal frequencies, respectively (see Fig. 1c.2). The quotient between the highest temporal frequency activation (ξ_0) and the highest spatial frequency (ω_0) is the speed selectivity of the filter $\mathrm{v} = (v_x, v_y) = (\omega_0/\xi_0^x, \omega_0/\xi_0^y)$ inside the limited frequency bandwidth of the neuron.

Analytic expressions for ξ_0 and ω_0 do not exist and these values must be estimated numerically. The numerical solution shows that, for a fixed value of σ as function of f, the value of the preferred temporal frequency ω_0 depends only on τ, while the maximal spatial frequency ξ_0 depends on θ and f, as it is shown in Fig. A.11.

A.2. V1 complex cells

Some characteristics of V1 complex cells can be explained using a nonlinear combination of V1 simple cells. For instance, V1 complex cells are invariant to contrast polarity, which indicates a kind of rectification on their ON–OFF receptive field regions.

Based on Adelson and Bergen [1], the ith contrast invariant V1 complex cell, located at $\mathbf{x}_i = (x_i, y_i)$, with spatial orientation θ_i and spatiotemporal orientation $f_i = (\bar{\xi}_i, \bar{\omega}_i)$ is defined as

$$C_i(t) = [(F^a * L)(\mathbf{x}_i, t)]^2 + [(F^b * L)(\mathbf{x}_i, t)]^2, \tag{A.6}$$

where the symbol $*$ represents the spatiotemporal convolution, and $F^a(\cdot)$ and $F^b(\cdot)$ are the V1 simple cells defined in Eq. (A.2). This definition gives independence to stimulus contrast and the cell response for a drifting grating is constant in time.

A.3. A documented C++ software library called ABFilter

Since the implementation of V1 cells is probably the most technical part of the architecture presented therein, we propose to the interested reader a documented C++ software library called ABFilter, which implements the Adelson and Bergen energy motion detector filters [1]. The ABFilter library is under a CeCill-C open-source license. It can be downloaded from: http://www-sop.inria.fr/neuromathcomp/public/software/abfilter-1.0.tar.gz.

References

[1] E. Adelson, J. Bergen, Spatiotemporal energy models for the perception of motion, Journal of the Optical Society of America A 2 (1985) 284–299.
[2] D.G. Albrecht, W.S. Geisler, A.M. Crane, Nonlinear properties of visual cortex neurons: Temporal dynamics, stimulus selectivity, neural performance, in: L.M. Challupa, J.S. Werner (Eds.), The Visual Neurosciences, Vol. 1, MIT press, 2004, pp. 747–764.
[3] P. Bayerl, H. Neumann, Disambiguating visual motion by form–motion interaction – a computational model, International Journal of Computer Vision 72 (1) (2007) 27–45.
[4] C. Beck, T. Ognibeni, H. Neumann, Object segmentation from motion discontinuities and temporal occlusions - a biologically inspired model, PLoS ONE 3 (11) (2008) 1–14.
[5] J. Berzhanskaya, S. Grossberg, E. Mingolla, Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception, Spatial Vision 20 (4) (2007) 337–395.
[6] 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.
[7] R.T. Born, Center–surround interactions in the middle temporal visual area of the owl monkey, Journal of Neurophysioly 84 (2000) 2658–2669.
[8] Bouecke, J.D., Tlapale, É., Kornprobst, P., Neumann, H., 2011. Neural mechanisms of motion detection, integration, and segregation: From biology to artificial image processing systems. EURASIP, special issue on Biologically inspired signal processing: Analysis, algorithms, and applications 2011.
[9] N. Bruce, J. Tsotsos, Saliency, attention, and visual search: An information theoretic approach, Journal of Vision 9 (3) (2009) 1–24. 5.
[10] Bruce, N.D.B., Kornprobst, P., 2009. Harris corners in the real world: A principled selection criterion for interest points based on ecological statistics. In: cvp (2009).
[11] G.T. Buracas, T.D. Albright, Contribution of area MT to perception of three-dimensional shape: a computational study, Vision Res 36 (6) (1996) 869–887.
[12] M. Carandini, J.B. Demb, V. Mante, D.J. Tollhurst, Y. Dan, B.A. Olshausen, J.L. Gallant, N.C. Rust, Do we know what the early visual system does?, Journal of Neuroscience 25 (46) (2005) 10577–10597
[13] A. Casile, M. Giese, Critical features for the recognition of biological motion, Journal of Vision 5 (2005) 348–360.
[14] B. Conway, M. Livingstone, Space-time maps and two-bar interactions of different classes of direction-selective cells in macaque V1, Journal of Neurophysiology 89 (2003) 2726–2742.
[15] R. De Valois, N. Cottaris, et al., Spatial and temporal receptive fields of geniculate and cortical cells and directional selectivity, Vision Research 40 (2000) 3685–3702.
[16] A. Destexhe, M. Rudolph, D. Paré, The high-conductance state of neocortical neurons in vivo, Nature Reviews Neuroscience 4 (2003) 739–751.
[17] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features. In: VS-PETS. pp. 65–72.
[18] Efros, A., Berg, A., Mori, G., Malik, J., Oct. 2003. Recognizing action at a distance. In: Proceedings of the 9th International Conference on Computer Vision. Vol. 2. pp. 726–734.
[19] M.-J. Escobar, G.S. Masson, T. Vieville, P. Kornprobst, Action recognition using a bio-inspired feedforward spiking network, International Journal of Computer Vision 82 (3) (2009) 284.

[20] M. Giese, T. Poggio, Neural mechanisms for the recognition of biological movements and actions, Nature Reviews Neuroscience 4 (2003) 179–192.
[21] N. Grzywacz, A. Yuille, A model for the estimate of local image velocity by cells on the visual cortex, Proc R Soc Lond B Biol Sci. 239 (1295) (1990) 129–161.
[22] X. Huang, T.D. Albright, G.R. Stoner, Stimulus dependency and mechanisms of surround modulation in cortical area mt, Journal of Neuroscience 28 (51) (2008) 13889–13906.
[23] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.
[24] 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.
[25] P. Kornprobst, R. Deriche, G. Aubert, Image sequence analysis via partial differential equations, Journal of Mathematical Imaging and Vision 11 (1) (1999) 5–26.
[26] Kuo, C.-H., Huang, C., Nevatia, R., 2010. Multi-target tracking by on-line learned discriminative appearance models. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. IEEE, p. 685+692.
[27] L. Lagae, S. Raiguel, G.A. Orban, Speed and direction selectivity of macaque middle temporal neurons, Journal of Neurophysiology 69 (1) (1993) 19–39.
[28] I. Laptev, B. Caputo, C. Schuldt, T. Linderberg, Local velocity-adapted motion events for spatio-temporal recognition, Computer vision and image understanding 108 (2007) 207–229.
[29] Le, Q., Zou, W., Yeung, S., Ng, A., 2009. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: cvp (2009), pp. 3361–3368.
[30] Li, Y., Huang, C., Nevatia, R., 2009. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In: cvp (2009), p. 2953+2960.
[31] J. Liu, W.T. Newsome, Functional organization of speed tuned neurons in visual area MT, Journal of Neurophysiology 89 (2003) 246–256.
[32] L.L. Lui, J.A. Bourne, M.G.P. Rosa, Spatial summation, end inhibition and side inhibition in the middle temporal visual area MT, Journal of Neurophysiology 97 (2) (2007) 1135.
[33] V. Mante, M. Carandini, Mapping of stimulus energy in primary visual cortex, Journal of Neurophysiology 94 (2005) 788–-798.
[34] R. Masland, P. Martin, The unsolved mystery of vision, Current Biology 17 (15) (2007) R577–R582.
[35] L. Michels, M. Lappe, L. Vaina, Visual areas involved in the perception of human movement from dynamic analysis, Brain Imaging 16 (10) (2005) 1037–1041.
[36] S. Nowlan, T. Sejnowski, A selection model for motion processing in area MT of primates, J. Neuroscience 15 (1995) 1195–1214.
[37] B. Olshausen, D. Field, How close are we to understanding V1?, Neural Computation 17 (8) (2005) 1665–1699
[38] J.-J. Orban de Xivry, S. Coppe, P. Lefèvre, M. Missal, Biological motion drives perception and action, Journal of Vision 10 (2) (2010) 1–11.
[39] C. Pack, B. Conway, R. Born, M. Livingstone, Spatiotemporal structure of nonlinear subunits in macaque visual cortex, Journal of Neuroscience 26 (3) (2006) 893–907.
[40] C. Pack, A. Gartland, R. Born, Integration of contour and terminator signals in visual area MT of alert macaque, The Journal of Neuroscience 24 (13) (2004) 3268–3280.
[41] J. Perge, B. Borghuis, R. Bours, M. Lankheet, R. van Wezel, Temporal dynamics of direction tuning in motion-sensitive macaque area mt, Journal of Neurophysiology 93 (2005). pp. 2194-2116.
[42] J. Perrone, A visual motion sensor based on the properties of V1 and MT neurons, Vision Research 44 (2004) 1733–1755.
[43] J. Perrone, R. Krauzlis, Spatial integration by mt pattern neurons: a closer look at pattern-to-component effects and the role of speed tuning, Journal of Vision 8 (9) (2008) 1–14.
[44] J. Perrone, A. Thiele, Speed skills: measuring the visual speed analyzing properties of primate mt neurons, Nature Neuroscience 4 (5) (2001) 526–532.
[45] R. Poppe, Vision-based human motion analysis: An overview, Computer Vision and Image Understanding 108 (1–2) (2007) 4–18.
[46] R. Poppe, A survey on vision-based human action recognition, Image and Vision Computing 28 (2010) 976–990.
[47] N. Priebe, C. Cassanello, S. Lisberger, The neural representation of speed in macaque area MT/V5, Journal of Neuroscience 23 (13) (2003) 5650–5661.
[48] Rodriguez, M., Shah, M., 2007. Detecting and segmenting humans in crowded scenes. In: ACM MM.
[49] N. Rust, V. Mante, E. Simoncelli, J. Movshon, How mt cells analyze the motion of visual patterns, Nature Neuroscience 9 (2006) 1421–1431.
[50] A. Safford, E. Hussey, R. Parasuraman, J. Thompson, Object-based attentional modulation of biological motion processing: Spatiotemporal dynamics using functional magnetic resonance imaging and electroencephalography, The Journal of Neuroscience 30 (27) (2010) 9064–9073.
[51] Serre, T., apr 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.
[52] E.P. Simoncelli, D. Heeger, A model of neuronal responses in visual area MT, Vision Research 38 (1998) 743–761.
[53] M. Smith, N. Majaj, A. Movshon, Dynamics of motion signaling by neurons in macaque area mt, Nature Neuroscience 8 (2) (2005) 220–228.
[54] É. Tlapale, G.S. Masson, P. Kornprobst, Modelling the dynamics of motion integration with a new luminance-gated diffusion mechanism, Vision Research 50 (17) (2010) 1676–1692.
[55] J. Tsotsos, Y. Liu, J. Martinez-Trujillo, M. Pomplun, E. Simine, K. Zhou, Attending to visual motion, Computer Vision and Image Understanding 100 (2005) 3–40.
[56] P. Turaga, R. Chellappa, V. Subrahmanian, O. Udrea, Machine recognition of human activities: a survey, IEEE Transactions on Circuits and Systems for Video Technology 18 (11) (2008) 1473–1488.
[57] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2010.
[58] A. Watson, A. Ahumada, Model of human visual-motion sensing, J Opt Soc Am A 2 (2) (1985) 322–342.
[59] M. Weliky, J. Fiser, R. Hunt, D. Wagner, Coding of natural scenes in primary visual cortex, Neuron 37 (2003) 703–718.
[60] D. Xiao, S. Raiguel, V. Marcar, J. Koenderink, G.A. Orban, Spatial heterogeneity of inhibitory surrounds in the middle temporal visual area, Proceedings of the National Academy of Sciences 92 (24) (1995) 11303–11306.
[61] D.K. Xiao, S. Raiguel, V. Marcar, G.A. Orban, The spatial distribution of the antagonistic surround of MT/V5 neurons, Cereb Cortex 7 (7) (1997) 662–677.
[62] Xing, J., Ai, H., Lao, S., 2009. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In: cvp (2009), p. 1200+1207.
[63] Yeffet, L., Wolf, L., sept 2009. Local trinary patterns for human action recognition. In: Proceedings of the 12th International Conference on Computer Vision. pp. 492–497.
[64] Zelnik-Manor, L., Irani, M., 2001. Event-based analysis of video. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition. Vol. 2. pp. 123–128.
[65] T. Zhao, R. Nevatia, B. Wu, Segmentation and tracking of multiple humans in crowded environments, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (7) (2007) 1198–1211.

María-José Escobar received the Ms. in Electronic Engineer and Electronic Engineer diploma from UTFSM, Chile, in 2003 and her PhD in Computer Science from Nice Sophia Antipolis University, France, in 2009. Her PhD concerns bio-inspired models for motion estimation and analysis applied to human action recognition and motion integration, and it was supervised by Pierre Kornprobst and Thierry Vieville, in close collaboration with neurophysiologists. Since January 2010, she has been a researcher at the Electronics Engineering Department, Universidad Técnica Federico Santa María (UTFSM), Chile. Her research interests cover computational neuroscience, biological vision, motion perception, spiking neural networks, natural image analysis.

Pierre Kornprobst received his PhD in Mathematics from Nice Sophia Antipolis University, France, in 1998. He is a research scientist at INRIA. His research interests include calculus of variations, nonlinear PDEs and numerical analysis as applied to image processing (e.g., image restoration, super resolution, inpainting, motion estimation and motion recognition). He is the co-author of the 2002 monograph Mathematical Problems in Image Processing (Springer). Since 2002, he focuses on the exploration of the brain from the mathematical and computational perspectives. His goal is to bridge the gap between biological and computational vision by proposing novel bio-inspired models.