Action Recognition with a Bio–inspired Feedforward Motion Processing Model: The Richness of Center-Surround Interactions

Maria-Jose Escobar and Pierre Kornprobst

Odyssée project team, INRIA Sophia-Antipolis, France
{mjescoba,pkornp}@sophia.inria.fr

Abstract. Here we show that reproducing the functional properties of MT cells with various center–surround interactions enriches motion representation and improves the action recognition performance. To do so, we propose a simplified bio–inspired model of the motion pathway in primates: It is a feedforward model restricted to V1-MT cortical layers, cortical cells cover the visual space with a foveated structure and, more importantly, we reproduce some of the richness of center-surround interactions of MT cells. Interestingly, as observed in neurophysiology, our MT cells not only behave like simple velocity detectors, but also respond to several kinds of motion contrasts. Results show that this diversity of motion representation at the MT level is a major advantage for an action recognition task. Defining motion maps as our feature vectors, we used a standard classification method on the Weizmann database: We obtained an average recognition rate of 98.9%, which is superior to the recent results by Jhuang et al. (2007). These promising results encourage us to further develop bio–inspired models incorporating other brain mechanisms and cortical layers in order to deal with more complex videos.

1 Introduction

Action recognition in real scenes remains a challenging problem in computer vision. Until recently, most proposed approaches considered simplified sequence databases and relied on simplified assumptions or heuristics. Some examples of these kind of approaches are [1,2,3,4,5], where one could find therein other references and further information.

Motion is the key feature for a wide class of computer vision (CV) approaches: Existing methods consider different motion representations or characteristics, such as coarse motion estimation, global motion distribution, local motion feature detection or spatio-temporal structure learning [6,7,8,9,10,11,12]. Following this general idea which is to consider motion as an informative cue for action recognition (AR), we present a bio-inspired model for motion estimation and representation. Interestingly, it is confirmed that in the visual system the motion pathway is also very much involved in the AR task [10], but of course other

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part IV, LNCS 5305, pp. 186–199, 2008.

brain areas (e.g., the form pathway) and mechanisms (e.g., top-down attentional mechanisms) are also involved to analyze complex general scenes.

Among recent bio-inspired approaches for AR, [13] proposed a model for the visual processing in the dorsal (*motion*) and ventral (*form*) pathways. They validated their model in the AR task using stick figures constructed from real sequences. More recently, [14] proposed a feedforward architecture, which can be seen as an extension of [15]. In [14], the authors mapped their model to the cortical architecture, essentially V1 (with simple and complex cells). The only clear bio-inspired part is one of the models for S1 units and the pooling aspect. The use of spatio-temporal chunks seems to be supported also but the authors never claim any biological relevance for the corresponding subsequent processing stages (from S2 to C3). The max operator is also controversial and not supported in neurophysiology because it mainly does not allow feedbacks.

In this article, we follow the same objective as in [14], which is to propose a bio-inspired model of motion processing for AR in real sequences. Our model will be a connection-based network, in which a large number of neuron-like processing units operate in parallel. Each unit *neuron* will have an 'activation level' *membrane potential* that represents the strength of a particular feature in the environment. Here, our main contribution will be to better account for the visual system properties, and in particular, at MT layer level: We reproduce part of the variety of center-surround interactions [16,17]. Then, in order to prove the relevance of this extended motion description, we will show its benefits on the AR application, and compare our results with the ones obtained by [14].

This article presents the model described in Fig. 1 and it is organized as follows. Section 2 presents the core of the approach which is a biologically-inspired model of motion estimation, based on a feedforward architecture. As we previously mentioned, the aim of this article is to show how a bio-inspired model can be used in a real application such as AR. Note that we also studied some low-level properties of the model concerning motion processing [18] but those studies are out of the scope of this article. The first stage (Section 2.1) is the local motion extraction corresponding to the V1 layer, with a discrete foveated organization. The output of this layer is fed to the MT layer (Section 2.2), which is composed of a set of neurons whose dynamics are defined by a conductance-based neuron model. We define the connectivity between V1 and MT layers according to neurophysiology, which defines the center-surround interactions of a MT neuron. The output of the MT layer is a set of neuron membrane potentials, whose values indicate the presence of a certain velocity or contrasts of velocities. Then, in Section 3, we consider the problem of AR based on the MT layer activity. In this section we also present the experimental protocol, some validations and a comparison with the approach presented by [14]. Interestingly, we show how the variety of surround-interactions in MT cells found in physiology allows the improvement of the recognition performances. We conclude in Section 4.

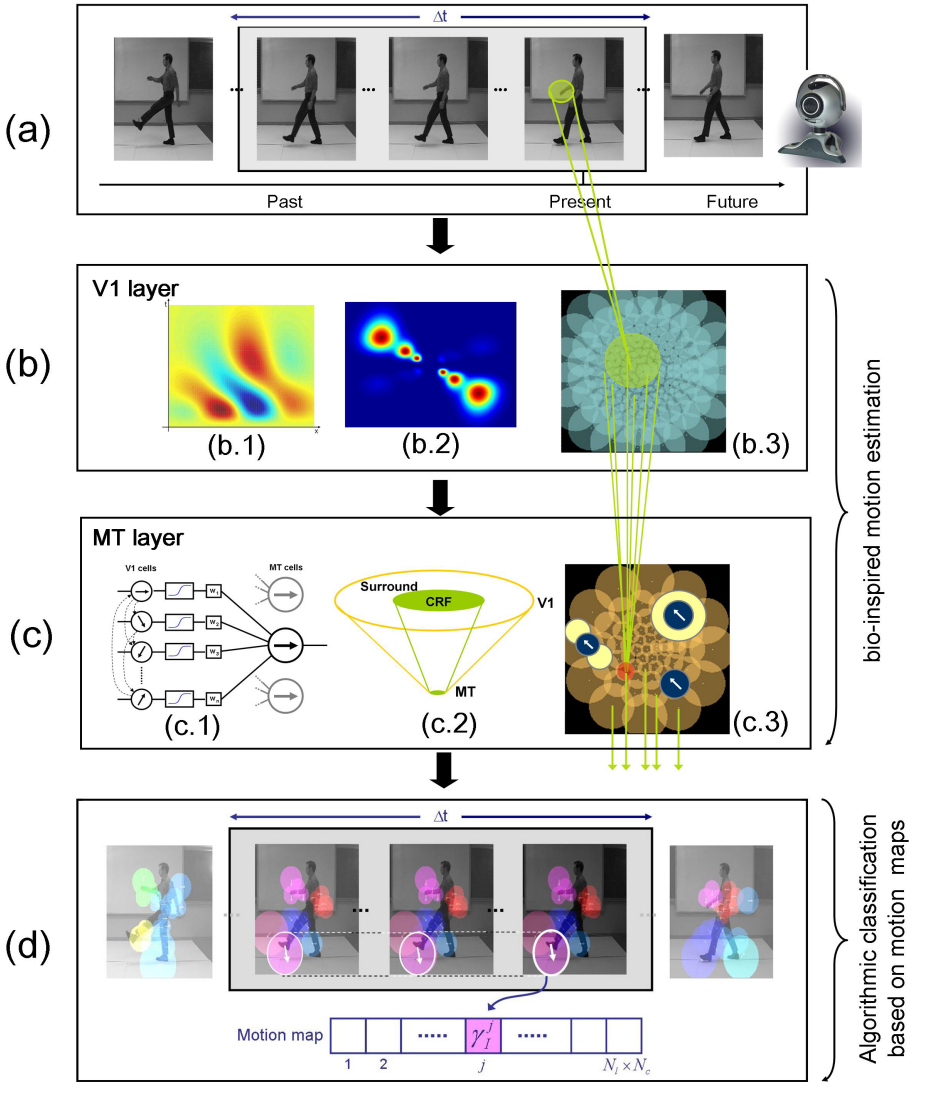

Fig. 1. Block diagram showing the different steps of our approach from the input image sequence as stimulus until the *motion map* encoding the motion pattern. (a) We use a real video sequence as input, the input sequences are preprocessed in order to have contrast normalization and centered moving stimuli. To compute the *motion map* 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 the activity of each V1 cell. (c) V1 outputs feed the MT cells which integrate the information in space and time. (d) The *motion map* is constructed calculating the mean activation of MT cells inside the sliding temporal window. The *motion map* has a length of $N_L \times N_c$ elements, where N_L is the number of MT layers of cells and N_c is the number of MT cells per layer. This *motion map* characterizes and codes the action stimulus.

2 Bio–inspired Motion Analysis Model

Several bio-inspired motion processing models have been proposed in the literature [19,20,21,22,23], 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. [24,25].

2.1 V1 Layer: Local Motion Detectors

Our V1 model is defined by a bank of energy motion detectors as a local motion estimation. The processing is done through energy filters which is a reliable and biologically plausible method for motion information analysis [26]. Each energy motion detector will emulate a complex cell, which is formed by a non-linear combination of V1 simple cells (see [27] for V1 cells classification). Note that the complex cells will be tuned for the direction of motion θ (and a range of speeds).

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 neurons.

The direction-selectivity (DS) refers to the property of a neuron to respond 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 (Fig. 1 (b.1)). Let us consider two spatio-temporal oriented simple cells, $F^a_{\theta,f}$ and $F^b_{\theta,f}$, spatially oriented in the direction θ, and spatio-temporal frequency oriented to $f = (\bar{\xi}, \bar{\omega})$, where $\bar{\xi}$ and $\bar{\omega}$ are the spatial and temporal maximal responses, respectively:

$$\begin{aligned} F^a_{\theta,f}(x,y,t) &= F^{odd}_{\theta}(x,y)H_{fast}(t) - F^{even}_{\theta}(x,y)H_{slow}(t), \\ F^b_{\theta,f}(x,y,t) &= F^{odd}_{\theta}(x,y)H_{slow}(t) + F^{even}_{\theta}(x,y)H_{fast}(t). \end{aligned} \tag{1}$$

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 $H_{slow}(t)$ are defined by:

$$H_{fast}(t) = T_{3,\tau}(t) - T_{5,\tau}(t), \quad \text{and} \quad H_{slow}(t) = T_{5,\tau}(t) - T_{7,\tau}(t), \tag{2}$$

where $T_{\eta,\tau}(t)$ is a Gamma function defined by $T_{\eta,\tau}(t) = \frac{t^\eta}{\tau^{\eta+1}\eta!} \exp\left(-\frac{t}{\tau}\right)$, which models the series of synaptic and cellular delays in signal transmission, from retinal photoreceptors to V1 afferents serving as a plausible approximation of biological findings [28].

Remark that the causality of $H_{fast}(t)$ and $H_{slow}(t)$ generates a more realistic model than the one proposed by [22] (see also [14]), where the Gaussian proposed as temporal profile is non-causal and inconsistent with V1 physiology.

The frequency analysis is required to a right design of our filter bank. For a given speed, the filter covers a specified region of the spatio-temporal frequency

domain. The quotient between the highest temporal frequency activation ($\bar{\omega}$) and the highest spatial frequency ($\bar{\xi}$) is the speed of the filter. So, the filter will be able to detect the motion for a stimulus whose spatial frequency lies 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 spatio-temporal frequency orientation (Fig. 1 (b.2)).

Complex Cells are also direction-selective neurons, however they include other characteristics that cannot be explained by a linear combination of the input stimulus. The complex cell property that we want to keep in this model is the invariance to contrast polarity.

Based on [26], we define the ith V1 complex cell, 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

$$C_{\mathbf{x}_i,\theta_i,f_i}(t) = \left[\left(F^a_{\theta_i,f_i} * I\right)(\mathbf{x}_i, t)\right]^2 + \left[\left(F^b_{\theta_i,f_i} * I\right)(\mathbf{x}_i, t)\right]^2, \tag{3}$$

where the symbol $*$ represents the spatio-temporal convolution between the simple cells defined in (1) and the input sequence $I(\mathbf{x}, t)$. With this definition, the cell response is independent of stimulus contrast sign and constant in time for a drifting grating as input stimulus.

Finally, it is well known in biology that the V1 output shows several nonlinearities due to: response saturation, response rectification, or contrast gain control [29]. In order to obtain a nonlinear saturation in the V1 response, the V1 output is passed through a sigmoid function $S(\cdot)$, where the respective parameters were tuned to have a suitable response in the case of drifting gratings as inputs. So, finally the V1 output will be given by:

$$r_i^{V1} = S\left(C_{\mathbf{x}_i,\theta_i,f_i}(t)\right). \tag{4}$$

2.2 MT Layer: Higher Order Motion Analysis

Modeling Dynamics of MT Neurons. In this article, the dynamics of the MT neurons are modeled by a simplified conductance-based neuron (without input currents) [30]. Considering a MT neuron i, its membrane potential $u_i^{MT}(t)$ evolves in time according to the conductance-driven equation:

$$\begin{aligned}\tau \frac{du_i^{MT}(t)}{dt} = \; & G_i^{exc}(t)\left(E^{exc} - u_i^{MT}(t)\right) + G_i^{inh}(t-\delta)\left(E^{inh} - u_i^{MT}(t)\right) \\ & + g^L\left(E^L - u_i^{MT}(t)\right),\end{aligned} \tag{5}$$

where E^{exc}, E^{inh} and $E^L = 0$ are constant which typical values of 70mV, -10mV and 0mV, respectively. According to (5), $u_i^{MT}(t)$ will belong to the interval $\left[E^{inh}, E^{exc}\right]$ and it will be driven by several influences. The first term refers to input pre-synaptic neurons and it will push the membrane potential $u_i^{MT}(t)$ towards E^{exc}, with a strength defined by $G_i^{exc}(t)$. Similarly, the second term also coming from pre-synaptic neurons will drive $u_i^{MT}(t)$ towards E^{inh} with a

strength $G_i^{inh}(t)$. Finally, the last term will drive $u_i^{MT}(t)$ towards the resting potential E^L with a constant strength given by g^L. The constant δ, typically 30ms, is the delay associated to the inhibitory effect.

The MT neuron i is a part of a neural network where the input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ are obtained by pooling the activity of all the pre-synaptic neurons connected to it (Fig. 1). Each MT cell has a receptive field built from the convergence of pre-synaptic afferent V1 complex cells (Fig. 1 (c.1)). The excitatory inputs forming $G_i^{exc}(t)$ are related with the activation of the classical receptive field (CRF) of the MT cell; whereas $G_i^{inh}(t)$ afferents are the cells forming the surround interactions that could modulate or not the response of the CRF [16,17] (Fig. 1(c.2)). The surround does not elicit responses by itself, it needs the CRF activation to be considered. According to this, the total input conductances $G_i^{exc}(t)$ and $G_i^{inh}(t)$ of the post-synaptic neuron i are defined by

$$G_i^{exc}(t) = \max\left(0, \sum_{j\in\Omega_i} w_{ij} r_j^{V1} - \sum_{j\in\Omega_i'} w_{ij} r_j^{V1}\right), \quad G_i^{inh}(t) = \sum_{j\in\Phi_i} w_{ij} r_j^{V1}, \quad (6)$$

where $\Omega_i = \{j \in \text{CRF} \mid \varphi_{ij} < \pi/2\}$, $\Omega_i' = \{j \in \text{CRF} \mid \varphi_{ij} > \pi/2\}$ and $\Phi_i = \{j \in \text{Surround} \mid \varphi_{ij} < \pi/2\}$, and where the connection weight w_{ij} is the efficacy of the synapse from neuron j to neuron i, which is proportional to the angle φ_{ij} between the two preferred motion direction-selectivity of the V1 and MT cell. It is important to remark that the values of the conductances will be always greater or equal to zero, and their positive or negative contribution to $u_i^{MT}(t)$ is due to the values of E^{exc} and E^{inh}.

The connection weights w_{ij} will be given by

$$w_{ij} = k_c w_{cs}(\mathbf{x}_i - \mathbf{x}_j)\cos(\varphi_{ij}), \qquad 0 \le \varphi_{ij} \le \pi, \quad (7)$$

where k_c is an amplification factor, 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. The weight $w_{cs}(\cdot)$ is associated to the distance between the MT cell positioned at $\mathbf{x}_i = (x_i, y_i)$ and the V1 cell positioned at $\mathbf{x}_j = (x_j, y_j)$, but also depends on the CRF or surround associated to the MT cell.

Remark. Many studies on MT focused on motion direction selectivity (DS), but very few on speed selectivity (see, e.g., [31,32,33]), 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 (6): Our MT cells pool V1 cells just considering their motion DS, and not their spatio-temporal tuning. However, note that it is also possible to pool differently V1 cells in order to extract some speed information, as proposed for example in [22,23,34]. As a result, one could obtain a velocity field qualitatively similar to an optical flow (i.e., one velocity per position).

Modeling the Richness of Surround Modulations. The activation of a MT neuron inside its CRF can be modulated by the activation of a surround area [16],

which is usually ignored in most MT-like models. In most cases this modulation is inhibitory, but Huang et al. [35] showed that this interaction, depending on the input stimulus, can be also integrative. The direction tuning of the surround compared with the center tends to be either the same or opposite, but rarely orthogonal.

Half of MT neurons have asymmetric receptive fields introducing anisotropies in the processing of the spatial information [16]. The neurons with asymmetric receptive fields seem to be involved in the encoding of important surfaces features, such as slant and tilt or curvature. Their geometry is the main responsible of the direction tuning of the MT cell and it changes along time.

Considering this, we included four types of MT cells (Fig. 2): One basic type of cell just only activated by its CRF, and three other types with inhibitory surrounds. We claim that inhibitory surrounds contain key information about the motion characterization (such as motion contrasts), as we will illustrate in Section 3. The tuning direction of the surround is always the same as the CRFs, but their spatial geometry changes, from symmetric to asymmetric-unilateral and asymmetric-bilateral surround interactions. It is important to mention that this approach is a coarse approximation of the real receptive field shapes.

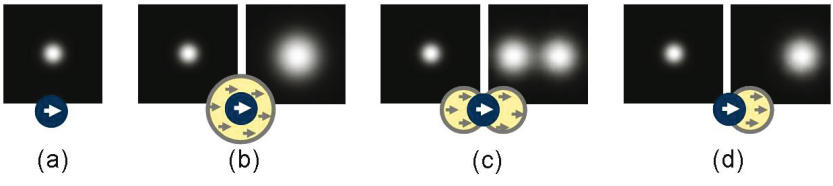

Fig. 2. MT center-surround interactions modeled in our approach. The classical receptive field CRF (a) is modeled with a Gaussian. All the surrounds from (b) to (d) are also modeled by Gaussians. In (b) the surround is symmetric. The two groups of cells with asymmetric surrounds are represented in (c) and (d). (c) has a bilateral asymmetric surround and (d) is a unilateral asymmetric surround. There is an important presence of anisotropic surround interactions in MT cells: In [16,17], the authors showed that within the MT cells with surround suppression, the configuration (b) is present only in the 25% of the cells, while (c) and (d) cover the resting percentage with a presence of 50% and 25%, respectively.

3 Action Recognition Based on MT Activity

3.1 Describing Motion Activity by a *Motion Map*

In this section, we use a standard supervised classification method which has no biological inspiration. To do this, one needs to define the correspondence between the input space (here the space of sequences) and a feature space, but also a notion of distance between feature vectors. We considered the simpler case of *supervised* classification which means that for some inputs, the class is known (training set). Then, considering a new sequence to be analyzed, we will estimate the corresponding feature vector and find the best class with a classifier.

Concerning our problem, we define below feature vectors as motion maps, which represent averaged MT cells activity in a temporal window.

Motion Map as a Feature Vector. At time t, given a video stream $I(\mathbf{x}, t)$ between $[t-\Delta t, t]$, we define the feature vector (from now on called *motion map*, see Fig. 1(c)) as the vector which represents the average membrane potential of the MT neurons in a temporal window $[t - \triangle t, t]$:

$$H_I(t, \triangle t) = \left\{\gamma_j^I(t, \triangle t)\right\}_{j=1,\ldots,N_l \times N_c}, \tag{8}$$

with $\gamma_j^I(t, \triangle t) = \frac{1}{\Delta t} \int_{t-\triangle t}^{t} u_j^{MT}(s)ds$, and where N_l is the number of MT layers and N_c is the number of MT cells per layer.

The *motion map* defined in (8) 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.

Definition of a Distance Measure. We propose a measure discrimination to evaluate the similarities between two motion maps $H_I(t, \triangle t)$ and $H_J(t', \triangle t')$, defined by

$$\mathcal{D}(H_I(t, \triangle t), H_J(t', \triangle t')) = \frac{1}{N_l\, N_c} \sum_{l=1}^{N_l\, N_c} \frac{(\gamma_l^I(t, \triangle t) - \gamma_l^J(t', \triangle t'))^2}{\gamma_l^J(t, \triangle t) + \gamma_l^J(t', \triangle t')}. \tag{9}$$

This measure refers to the *triangular discrimination* introduced by [36]. Other measures derived from statistics, such as *Kullback-Leiber* (KL) could also be used. The experiments done using, e.g., the KL measure showed no significant improvements. Note that (9) and the motion representation (8) can be seen as an extension of [37].

3.2 Experiments

Implementation Details. We considered luminosity and contrast normalized videos of size 210×210 pixels, centered on the action to recognize. Given V1 cells modeled by (3), we consider 9 layers of V1 cells. Each layer is built with V1 cells tuned with the same spatio-temporal frequency and 8 different orientations. The 9 layers of V1 cells are distributed in the frequency space in order to tile the whole space of interest (maximal spatial frequency of 0.5 pixels/sec and a maximal temporal frequency of 12 cycles/sec). The centers of the receptive fields are distributed according to a radial log-polar scheme with a foveal uniform zone. The limit between the two regions is given by the radius of the V1 fovea R_0 (80 pixels). The cells with an eccentricity less than R_0 have an homogeneous density and receptive fields size. The cells with an eccentricity greater than R_0 have a density and a receptive field size depending on its eccentricity, giving a total of 4473 cells per layer.

The MT cells are also distributed in a log-polar architecture, but in this case R_0 is 40 pixels giving a total of 144 cells per layer. Different layers of MT cells conform our model. Four different surround interactions were used in the MT construction (see Fig. 2). Each layer, with a certain surround interaction, has 8 different directions.

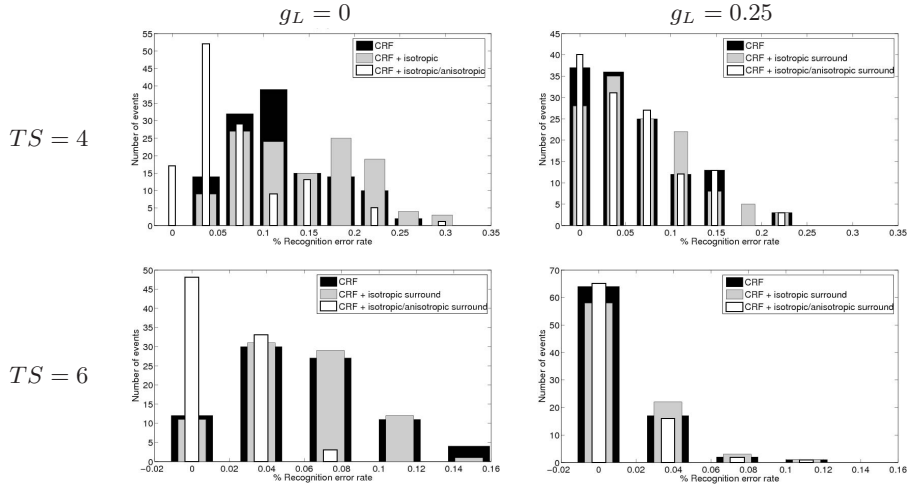

Fig. 3. Recognition error rate obtained for Weizmann database using the four different cells described in Fig. 2. We took all the combinations possibles considering 4 or 6 subjects in the training set (TS). For both cases, we ran the experiments with $g^L = 0$ and $g^L = 0.25$, and three surround-interactions: just CRF (black bars), CRF plus isotropic surround suppression (gray bars) and CRF plus isotropic and anisotropic surround suppression (red bars).

Experimental Protocol. In order to evaluate the performance of our algorithm, we used the Weizmann Database[1]: This database contains 9 different samples of different people doing 9 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 is variable and depends on the action.

We selected the actions of 4 or 6 (as in [14]) random subjects as training set (total of 36 or 64 sequences, respectively) and use the remaining 5 or 3 subjects for the test set (45 or 27 sequences, respectively). All the motion maps of the training set were obtained and stored in a data container. We used a RAW classifier[2]: When a new input sequence belonging to the test set is presented to the system, the motion map is calculated (with Δt covering here all the

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

[2] Note that we repeated the experiments with a standard SVM classifier but we did not get significant improvements in the recognition performance.

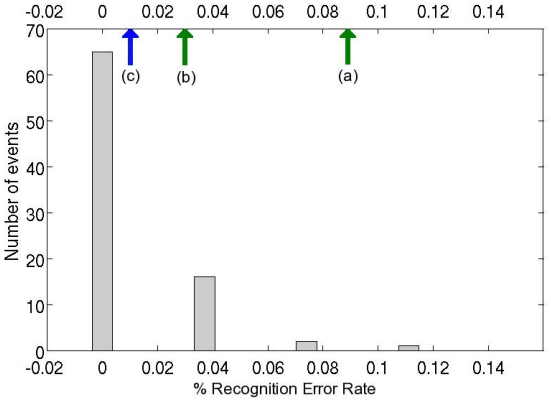

Fig. 4. Histograms obtained from the recognition error rates of our approach using all the cells defined in Fig. 2 for Weizmann database and the same experiment protocol used in [14]. The gray bars are our histogram obtained for $g^L = 0.25$. (a) Mean recognition error rate obtained by [14] (GrC_2, dense C_2 features): 8.9%/ ± 5.9. (b) Mean recognition error rate obtained by [14] (GrC_2, sparse C_2 features): 3.0%/ ± 3.0. (c) Mean recognition error rate obtained with our approach: 1.1%/ ± 2.1.

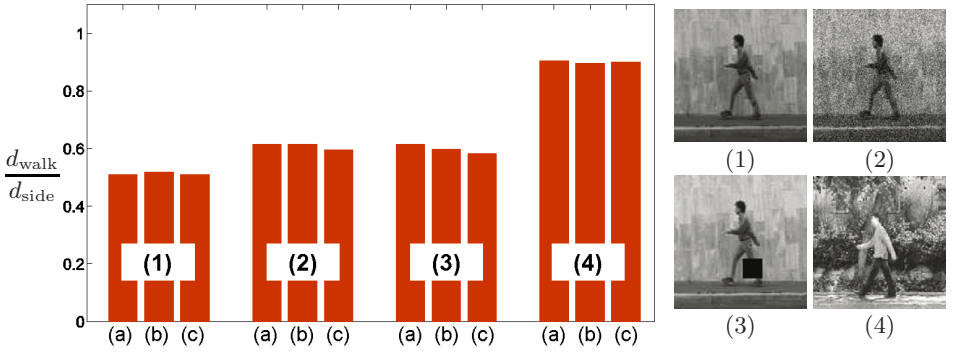

Fig. 5. Results obtained for the robustness experiments carried out for the three input sequences represented by the snapshots shown for *normal-walker* (1), *noisy* sequence (2), *legs-occluded* sequence (3) and *moving-background* sequence (4). In all the cases the recognition was correctly performed as *walk* and the second closest distance was to the class *side*. The red bars indicate the ratio between the distance to *walk* class and the distance to *side* class (d_{walk}/d_{side}). The experiments were done for the three configurations of surround-suppression: (a) just CRF, (b) CRF with isotropic surround and (c) CRF with isotropic/anisotropic surround ($g^L = 0.25$).

sequence) and it is compared using (9) to all motion maps stored in the training set. The class of the sequence with the shortest distance is assigned as the match class. The experiments were done considering every possible selection of 4 or 6 subjects, giving a total of 126 or 84 experiments. As output we obtained histograms showing the frequency of the recognition error rates.

Results. In order to quantify the influence of the information coded by center-surround interactions, we did the experiments with the different configurations shown in Fig. 2. The cells were combined in order to create three different motion maps: just considering the CRF, CRF plus the isotropic surround interaction, and finally considering all the cells described in Fig. 2, i.e., with isotropic and anisotropic surround interactions. Results are summarized in the histograms shown in Fig. 3. Results show that $g^L > 0$ significantly improves the performance of our system, mainly because the constant leak term attracts the membrane potential of the cell to its resting value ($E^L = 0$), avoiding possible saturation. It is also important to remark that in the case $g^L = 0$, the effect of inhibitory surrounds (either isotropic or anisotropic) is stronger than the case of $g^L = 0.25$. The explanation is that the inhibitory surround is the only mechanism to reduce the activation of the cell. Maybe this effect can be compensated in the case of $g^L = 0.25$ by adding more relevance to the response of the cells with inhibitory surround. Remark that the results have a strong variability and so that the recognition performance highly depends on the sequences used to define the training set.

In the case where 6 random subjects were taken to construct the training set, we compared our results with [14]. As previously mentioned, we estimated the performance of our approach based on all the possible combinations (84), and not only on 5 random trials (as in [14]). In Fig. 4, we show the histogram with the different recognition error rates obtained with our approach using the motion maps generated for the CRF and isotropic/anisotropic surround interactions cells. We obtained an average recognition rate of 98.9% (i.e., mean error rate of 1.1%), which exceeds the results obtained by [14].

To test the robustness of our approach, we considered input sequences with different kinds of perturbations (Fig. 5): noise (case (2)), legs-occlusion (case (3)) and moving textured background (case (4)). Both *noisy* and *legs-occluded* sequences were created starting from the sequence shown in Fig. 5(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 Gaussian noise. The *moving-background* sequence was taken from [38]. For the original sequence and the three modified input sequences the recognition was correctly performed as *walk*. A graph with the ratio between the shortest distance to *walk* class and the distance to the second closest class (*side* for the all the cases) is shown in Fig. 5: the inclusion of the anisotropic surround interaction makes the model less sensitive to occlusions or noise.

4 Conclusion

We proposed a feedforward bio-inspired model of V1-MT cortical layers that can be used for solving several aspects of motion integration [18], but also high-level tasks such as AR for natural scenes stimuli. Our model offers an efficient platform to unveil the contribution of different components involved in visual

processing within a single experimental framework. One clear advantage of our model is that it is generic: Unlike [13], there is no need to tune the properties of local motion given the specific application of AR. 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. Interestingly, we showed that taking into account this diversity of MT cells improves the recognition performance. Our interpretation is that cells with inhibitory surrounds bring information related to velocity opponency or singularities in the velocity field of the input stimulus.

Future work will be focused on better exploiting the dynamical state of the MT layer. Here, we defined the feature vector as the motion map, which represents the average membrane potential of MT neurons in a temporal window. Since it is averaged, this representation obviously misses the information about the fine dynamical properties and the evolution of MT cells. For example, our goal will be to detect and take into account synchronizations and temporal correlations between cells.

Another perspective is about enriching the model with other brain functions or cortical layers. Of course, the motion pathway is not the only actor for AR in the visual system. Like every motion-based approach for AR, our approach is likely to be limited. 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. [19]).

Acknowledgements

This work was partially supported by the EC IP project FP6-015879, FACETS and CONICYT Chile. We also would like to thank John Tsotsos and Guillaume S. Masson for their valuable comments.

References

1. Gavrila, D.: The visual analysis of human movement: A survey. Computer Vision and Image Understanding 73(1), 82–98 (1999)
2. Goncalves, L., DiBernardo, E., Ursella, E., Perona, P.: Monocular tracking of the human arm in 3D. In: Proceedings of the 5th International Conference on Computer Vision, June 1995, pp. 764–770 (1995)
3. Mokhber, A., Achard, C., Milgram, M.: Recognition of human behavior by space-time silhouette characterization. Pattern Recognition Letters 29(1), 81–89 (2008)
4. Seitz, S., Dyer, C.: View-invariant analysis of cyclic motion. The International Journal of Computer Vision 25(3), 231–251 (1997)
5. Collins, R., Gross, R., Shi, J.: Silhouette-based human identification from body shape and gait. In: 5th Intl. Conf. on Automatic Face and Gesture Recognition, p. 366 (2002)

6. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: Proceedings of CVPR 2001, vol. 2, pp. 123–128 (2001)
7. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: Proceedings of the 9th International Conference on Computer Vision, vol. 2, pp. 726–734 (October 2003)
8. Laptev, I., Capuo, B., Schultz, C., Lindeberg, T.: Local velocity-adapted motion events for spatio-temporal recognition. Computer Vision and Image Understanding 108(3), 207–229 (2007)
9. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS, pp. 65–72 (2005)
10. Michels, L., Lappe, M., Vaina, L.: Visual areas involved in the perception of human movement from dynamic analysis. Brain Imaging 16(10), 1037–1041 (2005)
11. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial–temporal words. Internation Journal of Computer Vision 79(3), 299–318 (2008)
12. Wong, S.F., Kim, T.K., Cipolla, R.: Learning motion categories using both semantic and structural information. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 1–6 (June 2007)
13. Giese, M., Poggio, T.: Neural mechanisms for the recognition of biological movements and actions. Nature Reviews Neuroscience 4, 179–192 (2003)
14. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: Proceedings of the 11th International Conference on Computer Vision, pp. 1–8 (2007)
15. Serre, T., Wolf, L., Poggio, T.: Object recognition with features inspired by visual cortex. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 994–1000 (June 2005)
16. Xiao, D.K., Raiguel, S., Marcar, V., Orban, G.A.: The spatial distribution of the antagonistic surround of MT/V5 neurons. Cereb Cortex 7(7), 662–677 (1997)
17. Xiao, D., Raiguel, S., Marcar, V., Koenderink, J., Orban, G.A.: Spatial heterogeneity of inhibitory surrounds in the middle temporal visual area. Proceedings of the National Academy of Sciences 92(24), 11303–11306 (1995)
18. Escobar, M., Masson, G., Kornprobst, P.: A simple mechanism to reproduce the neural solution of the aperture problem in monkey area MT. Research Report 6579, INRIA (2008)
19. Tsotsos, J., Liu, Y., Martinez-Trujillo, J., Pomplun, M., Simine, E., Zhou, K.: Attending to visual motion. Computer Vision and Image Understanding 100, 3–40 (2005)
20. Nowlan, S., Sejnowski, T.: A selection model for motion processing in area MT of primates. J. Neuroscience 15, 1195–1214 (1995)
21. Rust, N., Mante, V., Simoncelli, E., Movshon, J.: How MT cells analyze the motion of visual patterns. Nature Neuroscience (11), 1421–1431 (2006)
22. Simoncelli, E.P., Heeger, D.: A model of neuronal responses in visual area MT. Vision Research 38, 743–761 (1998)
23. Grzywacz, N., Yuille, A.: A model for the estimate of local image velocity by cells on the visual cortex. Proc. R. Soc. Lond. B. Biol. Sci. 239(1295), 129–161 (1990)
24. Berzhanskaya, J., Grossberg, S., Mingolla, E.: Laminar cortical dynamics of visual form and motion interactions during coherent object motion perception. Spatial Vision 20(4), 337–395 (2007)
25. Bayerl, P., Neumann, H.: Disambiguating visual motion by form–motion interaction – a computational model. International Journal of Computer Vision 72(1), 27–45 (2007)

26. Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America A 2, 284–299 (1985)
27. Carandini, M., Demb, J.B., Mante, V., Tollhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., Rust, N.C.: Do we know what the early visual system does? Journal of Neuroscience 25(46), 10577–10597 (2005)
28. Robson, J.: Spatial and temporal contrast-sensitivity functions of the visual system. J. Opt. Soc. Am. 69, 1141–1142 (1966)
29. Albrecht, D., Geisler, W., Crane, A.: Nonlinear properties of visual cortex neurons: Temporal dynamics, stimulus selectivity, neural performance, pp. 747–764. MIT Press, Cambridge (2003)
30. Destexhe, A., Rudolph, M., Paré, D.: The high-conductance state of neocortical neurons in vivo. Nature Reviews Neuroscience 4, 739–751 (2003)
31. Priebe, N., Cassanello, C., Lisberger, S.: The neural representation of speed in macaque area MT/V5. Journal of Neuroscience 23(13), 5650–5661 (2003)
32. Perrone, J., Thiele, A.: Speed skills: measuring the visual speed analyzing properties of primate mt neurons. Nature Neuroscience 4(5), 526–532 (2001)
33. Liu, J., Newsome, W.T.: Functional organization of speed tuned neurons in visual area MT. Journal of Neurophysiology 89, 246–256 (2003)
34. Perrone, J.: A visual motion sensor based on the properties of V1 and MT neurons. Vision Research 44, 1733–1755 (2004)
35. Huang, X., Albright, T.D., Stoner, G.R.: Adaptive surround modulation in cortical area MT. Neuron. 53, 761–770 (2007)
36. Topsoe, F.: Some inequalities for information divergence and related measures of discrimination. IEEE Transactions on information theory 46(4), 1602–1609 (2000)
37. Zelnik-Manor, L., Irani, M.: Statistical analysis of dynamic actions. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(9), 1530–1535 (2006)
38. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Proceedings of the 10th International Conference on Computer Vision 2, 1395–1402 (2005)