What can we expect from a V1-MT feedforward architecture for optical flow estimation?

Fabio Solari, Manuela Chessa, N. V. Kartheek Medathati, Pierre Kornprobst

What can we expect from a V1-MT feedforward architecture for optical flow estimation?

Fabio Solari[a], Manuela Chessa[a], N. V. Kartheek Medathati[b], Pierre Kornprobst[b]

[a] *University of Genoa, DIBRIS, Italy*
[b] *Inria, Neuromathcomp team, Sophia Antipolis, France*

Abstract

Motion estimation has been studied extensively in neuroscience in the last two decades. Even though there has been some early interaction between the biological and computer vision communities at a modelling level, comparatively little work has been done on the examination or extension of the biological models in terms of their engineering efficacy on modern optical flow estimation datasets. An essential contribution of this paper is to show how a neural model can be enriched to deal with real sequences. We start from a classical V1-MT feedforward architecture. We model V1 cells by motion energy (based on spatio-temporal filtering), and MT pattern cells (by pooling V1 cell responses). The efficacy of this architecture and its inherent limitations in the case of real videos are not known. To answer this question, we propose a velocity space sampling of MT neurones (using a decoding scheme to obtain the local velocity from their activity) coupled with a multi-scale approach. After this, we explore the performance of our model on the Middlebury dataset. To the best of our knowledge, this is the only neural model in this dataset. The results are promising and suggest several possible improvements, in particular to better deal with discontinuities. Overall, this work provides a baseline for future developments of bio-inspired scalable computer vision algorithms and the code is publicly available to encourage research in this direction.

Keywords: optical flow, spatio-temporal filters, motion energy, V1, MT, benchmarking

1. Introduction

Interpretation of visual motion information is a key competency for biological vision systems to survive in a dynamic world but also for artificial vision systems to process videos efficiently. As such, visual motion estimation has been studied extensively by both biological vision and computer vision communities. The question is to estimate optical flow, which is defined by 2-D vectors at sample locations of the visual image that describe temporal displacements of moving scene elements within the sensor's frame of reference. This displacement vector field constitutes the image flow representing apparent 2-D motions resultant from the 3-D velocities being projected onto the sensor. Such 2-D motions are observable only through intensity variations as a consequence of the relative change between an observer (eye or camera) and the surfaces or objects in a visual scene.

In the past two decades efforts by computer vision researchers have led to development of a large number of models for the computation of optical flow (see [1] for a review). In addition to modeling efforts to solve this task, a prominent achievement in computer vision has been to develop publicly available benchmarking datasets to evaluate and compare models in natural image scenarios. These benchmarking datasets have spurred a great deal of research resulting in new models, however, despite this large amount of work in this area, the problem still remains hard to solve as many of the models either lack consistent accuracy across video sequences or have a high computational cost.

On the other hand the neural mechanisms underlying motion analysis in the visual cortex have been extensively studied with a lot of emphasis on understanding the function of cortical areas V1 [2, 3] and MT [4], which play a crucial role in motion estimation (see [5, 6, 7] for reviews). Neurons in V1 are found to respond when motion direction is perpendicular to the contrast of the underlying pattern, while neurons in MT are found to respond best to a particular speed irrespective of the underlying contrast orientation and thus are believed to be solving the local motion estimation problem.

Several computational models have been proposed based on the available experimental data. Initially models focussed on motion sensitive cells in V1 (complex cells). Using the conceptual framework of receptive fields (RF) the responses were explained using Gabor functions [8], and spatio-temporal motion energy [9]. Then few attempts were made to recover the motion vectors directly from the motion energy representation [10, 11]. One could

call these models as being at the interface between computer vision and biological vision. These initial attempts were later on leveraged and extended to explain the properties of MT neurons by considering a feedforward pooling from V1 cells followed by divisive normalisation [12, 13, 14]. Apart from this class of linear-non linear feedforward models other attempts were made to simulate the information processing by V1-MT layers using lateral or feedback interactions for solving the aperture problem, by considering a pure velocity space representation and various kinds of local motion estimation [15, 16, 17, 18, 19].

Even though there was some early interaction among the biological and computer vision communities at a modeling level (see, e.g., [20, 21, 13]), comparatively little work has been done for examining or extending the models proposed in biology in terms of their engineering efficacy on modern optical flow estimation datasets. In this work, we take a step towards filling the critical gap between biological and computer vision communities (see [22] for a more general discussion), focusing on visual motion estimation leveraging and testing ideas proposed in biology in terms of building scalable algorithms. This is a challenging task as many of the models proposed in biology are confined to highly primed stimuli or often only examine a local decision making process such as a receptive field property, which demands non-trivial extensions to be made before the ideas could be tested on complex real world datasets.

In this paper, we focus on the V1-MT feedforward class of models, which can be seen as equivalent to the popular and well studied Lucas-Kanade approach [23] (see [24]). Our goal is to propose a bio-inspired model benchmarked on a state-of-the-art dataset, providing to the computer vision community a baseline model which can be extended by incorporating further findings from biology. The two key contributions of our work can be stated as follows: (i) Proposing a velocity space sampling of tuned MT neurones and a scheme to decode the local velocity from the activity of these neurones. Most of the experimental studies were focussed on single cell responses of V1 or MT neurones to a subset of stimuli, thus ignoring how does the overall population encode the true velocity vectors. We address this problem by our sampling and decoding scheme. (ii) Examining the efficacy of V1-MT feedforward processing in natural image scenarios. The stimuli used in various experiments are highly homogeneous and do not cover the spatio-temporal filtering plane as in the case of natural images [25]. Thus the efficacy of the system and inherent limitations in case of natural stimuli are not known.

This is explored by considering Middlebury dataset, which comprises complex natural stimuli.

The paper is organized as follows: In Sec. 2 we present our V1-MT feedforward architecture for optical flow estimation (called FFV1MT). Our model has three main steps: The two first steps model V1 cells and MT pattern cells following classical ideas from the literature. The third step is a decoding stage to extract the optical flow from MT population response. In Sec. 3 we present the algorithmic details of this model, which are an essential contribution here, since they allow this V1-MT architecture to be applied to real videos. In particular, we propose a multi-scale approach to deal with large ranges of speeds found in natural scenes. In Sec. 4 we evaluate our approach on several kinds of videos. We use test sequences to show the intrinsic properties of our approach and we benchmark our approach using the Middlebury dataset [26].

2. Feedforward V1-MT model for optical flow estimation

2.1. General overview

In general, the pattern selectivity of MT cells can be explained by following two different approaches [6]: the motion computation can be related to some kind of 2-D feature extraction mechanism, or based on intersection of constraints (IOC) mechanisms. For the former approach, the consequence is that the aperture problem does not affect the motion processing, though little evidence for a feature-tracking mechanism are reported [27, 28, 29]. The latter approach is based on geometric relationships among the local velocity estimates.

The model we study in this paper is based on a non-linear integration of the V1 afferents to obtain the MT pattern cells [7]. In particular, the IOC mechanism is indirectly considered through localized activations of V1 cells [12, 13, 14]. It is a three-step feedforward model: Step 1 corresponds to the V1 simple and complex cells, Step 2 corresponds to the MT pattern cells and Step 3 corresponds to a decoding stage to obtain the optical flow from the MT population response. In term of modeling, Steps 1 and 2 follow a classical view, while Step 3 has been introduced to solve the task of optical flow. An illustration of our model called FFV1MT is given in the figure next to Tab. 1 (see also Fig. 1 for a more detailed illustration of the computations involved).

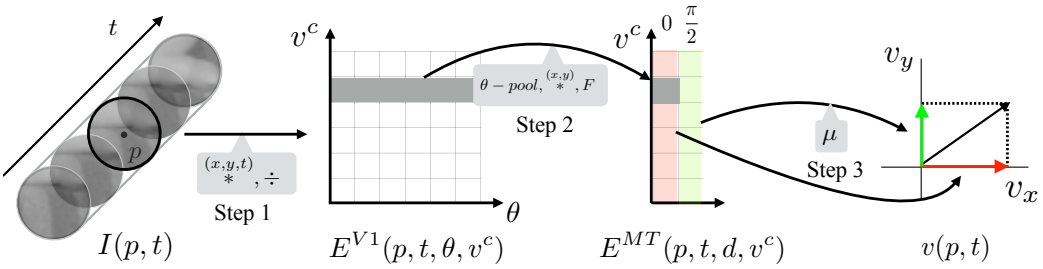


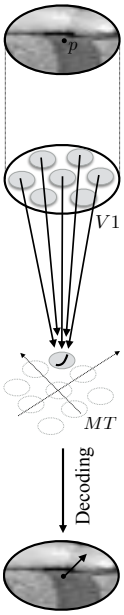
Figure 1: FFV1MT Model overview: It is a three-step feedforward model, where Step 1 corresponds to the V1 layer (obtained by a non-separable spatio-temporal filtering and a normalisation), Step 2 corresponds to MT layer (obtained by pooling V1 responses first with respect to θ, then in a local spatial neighbourhood, and applying a static nonlinearity) and Step 3 is velocity estimation (obtained by a weighted average of MT responses).

This model is inspired from previous works from visual neuroscience [10, 13, 14] and in Tab. 1, we summarise what are the main differences. In the seminal paper of Heeger [10] a first motion estimation model is introduced to compute the optical flow. Steps 1 and 2 of our model are similar to the ones presented in [13], but in the latter the optical flow is not estimated. It is worth to note that the model proposed in [14] is described in the parameter space, whereas we present a model in the (p,t) space that is able to estimate the optical flow of real-world sequences. All the model, but [14], introduce a processing stage to avoid responses to ambiguous low frequency textures. Finally, we propose an empirical sampling scheme of the two-dimensional velocity space, which provides competitive estimates while reducing the computational cost significantly when compared to [13].

2.2. Description of the FFV1MT model

Let us consider a grayscale image sequence $I(p,t)$, for all positions $p = (x,y)$ inside a domain Ω and for all time $t > 0$. Our goal is to find the optical flow $v(p,t) = (v_x, v_y)(p,t)$ defined as the apparent motion at each position p and time t.

Step 1 : V1 (Motion energy estimation and normalization). In the V1-layer two sub-populations of neurons are involved in the information processing, namely V1-direction selective simple cells and complex cells. Simple cells are characterised by the preferred direction θ of their contrast sensitivity in the spatial domain and their preferred velocity v^c in the direction orthogonal to their contrast orientation often referred to as component speed. The RFs



Model characteristics	Heeger [10]	Simoncelli and Heeger [13]	Rust et al. [14]	FFV1MT
V1 cell model	Gabor filters	Third derivative of a Gaussian	Direction space only	Gabor filters as in [10]
MT pooling	N.A.	Yes	Yes	Yes
MT nonlinearity	N.A.	Yes	Yes	Yes
MT population sampling	N.A.	Dense	Direction space only	Principal axes only
Decoding	Least-square on motion energy	No	No	Linear
Multi scale	Yes	No	No	Yes
Coarse-to-fine	No	No	No	Yes

Table 1: Comparison of our model FFV1MT with respect to other most related work.

of the V1 simple cells are classically modelled using band-pass filters in the spatio-temporal domain. In order to achieve low computational complexity, the spatio-temporal filters are decomposed into separable filters in space and time. Spatial component of the filter is described by Gabor filters $\mathcal{H}$ and temporal component by an exponential decay function $\mathcal{P}$. Given the peak spatial and temporal frequencies f_s and f_t of a receptive field, we define the following complex filters by:

$$\mathcal{H}(p, \theta, f_s) = Be^{\left(\frac{-(x^2+y^2)}{2\sigma^2}\right)} e^{j2\pi(f_s cos(\theta)x + f_s sin(\theta)y)}, \tag{1}$$

$$\mathcal{P}(t, f_t) = e^{\left(-\frac{t}{\tau}\right)} e^{j2\pi(f_t t)}, \tag{2}$$

where σ and τ define the spatial and temporal scales, respectively. Denoting the real and imaginary components of the complex filters $\mathcal{H}$ and $\mathcal{P}$ as $\mathcal{H}_e, \mathcal{P}_e$ and $\mathcal{H}_o, \mathcal{P}_o$ respectively, and a preferred velocity v_c related to the frequencies by the relation

$$v^c = \frac{f_t}{f_s}, \tag{3}$$

we introduce the odd and even spatio-temporal filters defined as follows,

$$\begin{aligned}\mathcal{G}_o(p,t,\theta,v^c) &= \mathcal{H}_o(p,\theta,f_s)\mathcal{P}_e(t,f_t) + \mathcal{H}_e(p,\theta,f_s)\mathcal{P}_o(t,f_t),\\ \mathcal{G}_e(p,t,\theta,v^c) &= \mathcal{H}_e(p,\theta,f_s)\mathcal{P}_e(t,f_t) - \mathcal{H}_o(p,\theta,f_s)\mathcal{P}_o(t,f_t).\end{aligned} \tag{4}$$

These odd and even symmetric and tilted (in space-time domain) filters characterize V1 simple cells. Using these expressions, we define the response of simple cells, either odd or even, with a preferred direction of contrast sensitivity θ in the spatial domain, with a preferred velocity v^c and with a spatial scale σ by

$$R_{o/e}(p,t,\theta,v^c) = (\mathcal{G}_{o/e}(\cdot,\cdot,\theta,v^c) \overset{(x,y,t)}{*} I)(p,t). \tag{5}$$

Fig. 2(a) shows the amplitude power spectra of the spatio-temporal filters $\mathcal{G}_o(p,t,\theta,v^c)$ (the same is for $\mathcal{G}_e(p,t,\theta,v^c)$) in the frequency domain. The shape of the amplitude power spectra of the filters' bank is due to the combination of the odd and even functions ($\mathcal{H}_o$, $\mathcal{H}_e$, $\mathcal{P}_o$, and $\mathcal{P}_e$) given in (4).

The complex cells are described as a combination of the quadrature pair of simple cells (5) by using the motion energy formulation

$$E(p,t,\theta,v^c) = R_o(p,t,\theta,v^c)^2 + R_e(p,t,\theta,v^c)^2,$$

followed by a normalisation: Considering a finite set of orientations $\theta = \theta_1 \dots \theta_N$, the final V1 response is defined by

$$E^{V1}(p,t,\theta,v^c) = \frac{E(p,t,\theta,v^c)}{\sum_{i=1}^{N} E(p,t,\theta_i,v^c) + \varepsilon}, \tag{6}$$

where $0 < \varepsilon \ll 1$ is a small constant to avoid divisions by zero in regions with no energy (when no spatio-temporal texture is present). The main property of V1 is its tuning to the spatial orientation of the visual stimulus, since the preferred velocity of each cell is related to the direction orthogonal to its spatial orientation.

Step 2: MT pattern cells response. MT neurones exhibit velocity tuning irrespective of the contrast orientation. This is believed to be achieved by pooling afferent responses in both spatial and orientation domains followed by a non-linearity [13]. The responses of an MT pattern cell tuned to the speed v^c and to direction of speed d can be expressed as follows:

$$E^{MT}(p,t,d,v^c) = F\left(\sum_{i=1}^{N} w_d(\theta_i) G_{\sigma_{pool}} \overset{x,y}{*} E^{V1}(p,t,\theta_i,v^c)\right), \tag{7}$$

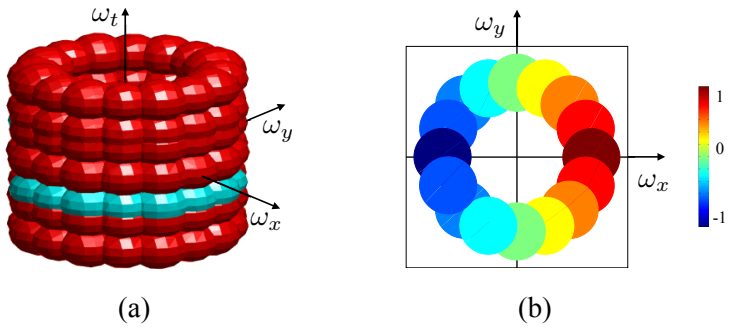


Figure 2: Representation of the V1 RFs in the frequency domain. (a) The iso-surface of the power spectra of the considered spatio-temporal filter bank that models the V1 cells. The spatial radial peak frequency of the filters is constant and the temporal frequency changes, thus the frequency bands have a cylinder-like shape. The V1 cells afferent to a population of MT cells for a specific v^c are highlighted in cyan. (b) The weights $w_d(\theta)$ used to pool the afferent V1 cells. In particular, the weights refer to a cosine weighting function, with values from -1 to 1 as in the colormap.

where $G_{\sigma_{pool}}$ denotes a Gaussian kernel of standard deviation σ_{pool} for the spatial pooling, $F(s) = exp(s)$ is a static nonlinearity chosen as an exponential function [30, 14], and w_d represents the MT linear weights that give origin to the MT tuning. In Fig. 2(a) the power spectra of the filters corresponding to the V1 cells afferent to a population of MT cells tuned to a specific v^c are represented in cyan. Such afferent cells are weighted through the $w_d(\theta)$, as shown in Fig. 2(b).

Physiological evidence suggests that w_d is a smooth function with central excitation and lateral inhibition. Cosine function shifted over various orientations is a potential function that could satisfy this requirement to produce the responses for a population of MT neurones [31]. Considering the MT linear weights shown in [14], $w_d(\theta)$ is defined by

$$w_d(\theta) = \cos(d - \theta) \quad d \in [0, 2\pi[. \tag{8}$$

This choice allows to obtain direction tuning curves of pattern cells that behave as in [14]. However, considering MT neurones that span over the 2-D velocity space with a preferred set of tuning speed directions in $[0, 2\pi[$ and also a multiplicity of tuning speeds is not necessary to encode velocity. A sampling along the cardinal axes is sufficient to recover the full velocity vector: since cosine functions shifted over various orientations (see Eq. (8)) can be described by the linear combination of an orthonormal basis (i.e.,

sine and cosine functions), all the V1 afferent information is encoded by two populations of MT neurons (see Eq. (7)). For this reason, in this paper, we sample the velocity space using two MT populations tuned to the directions $d = 0$ and $d = \pi/2$ with varying tuning speeds.

Step 3: Decoding. In this step we wonder how optical flow can be estimated by decoding the population responses of the MT neurones. Indeed, a unique velocity vector cannot be recovered by activity of a single velocity tuned MT neurone as multiple scenarios could evoke the same activity, but unique vector can be recovered based on the activity of a population. In this paper, we present a decoding step which was not present in [13, 14] to decode the MT population. We adopt a linear combination approach to decode the MT population response as in [32, 33]:

$$\begin{cases} v_x(p,t) = \sum_{i=1}^{M} v_i^c E^{MT}(p,t,0,v_i^c), \\ v_y(p,t) = \sum_{i=1}^{M} v_i^c E^{MT}(p,t,\pi/2,v_i^c). \end{cases} \tag{9}$$

2.3. An extension to deal with discontinuities: The FFV1MT–TF model

The FFV1MT approach described in this section relies on isotropic spatial smoothing at V1 level and isotropic pooling from V1 to MT. There is no mechanism to deal with motion discontinuities. In this section, we propose a simple extension of the FFV1MT model to show how discontinuities could be preserved. The idea is to introduce an iterative diffusion process between MT cells, which could be interpreted as the effect of lateral connections inside the MT population. The way nearby cells exchange information depends on their respective tuning speeds and directions, but it can also depend on the local context of the image. For example, local contrast and luminance information can modulate neurones characteristics and connections.

To model this idea, we propose a solution based on the trilateral filter (TF) which is an extension of the linear Gaussian filtering. Bilateral and trilateral filter have been extensively used in the context of nonlinear image smoothing leading to many applications (see [34] for a review). They provide a simple way to take discontinuities into account. Considering each population of MT cells tuned to a specific value of d and v^c as a spatial map, the goal is to apply TF in space to each map $E^{MT}(\cdot, t, d, v^c)$. This model is called FFV1MT–TF.

Denoting $E^{MT}(p,t,d,v^c)$ by $E^{MT}(p)$ for sake of simplicity, one iteration

of TF on $E^{MT}(p)$ is defined by:

$$TF_{\alpha,\beta,\gamma}[E^{MT}](p) = \frac{1}{N(p)} \int_{p' \in \Omega} f_\alpha(\|p - p'\|) f_\beta(E^{MT}(p') - E^{MT}(p)) \\ f_\gamma(I(p',t) - I(p,t)) E^{MT}(p') dp', \tag{10}$$

where

$$f_\mu(s) = \exp(s^2/\mu^2) \quad s \in \mathbb{R}, \tag{11}$$

α, β and γ are parameters defining the smoothing properties of TF and $N(p)$ is the normalising term

$$N(p) = \int_{p' \in \Omega} f_\alpha(\|p - p'\|) f_\beta(E^{MT}(p') - E^{MT}(p)) f_\gamma(I(p',t) - I(p,t)) dp'.$$

The interpretation of (10) is that, to estimate the new activity of an MT cell located at position p after one pass of TF, we average MT cell activities which are close in space, which have a similar activity, and which correspond to positions having similar luminance. The resulting filtered energy $TF_{\alpha,\beta,\gamma}[E^{MT}](p)$ is smoothed while main discontinuities are preserved and enhanced according to energy and luminance discontinuities. Several iterations of this filter can be made depending on the degree of smoothing desired.

3. Making the approach applicable to real videos

This kind of V1-MT feedforward architecture presented in Sec. 2 was initially proposed to explain recorded neural activities and mainly applied on synthetic homogeneous images such as moving gratings and plaids. They were not designed to be a systematic alternative to computer vision algorithms to work on real videos. In this section, we propose algorithmic solutions to make this V1-MT feedforward architecture applicable to real videos so that it could be benchmarked using state-of-the-art dataset.

3.1. Multiscale approach

One critical point in dealing with real videos is to be able to deal with a large range of speeds. As detailed in Sec. 2, the V1-like RFs are modelled through spatio-temporal filters. In order to keep as low as possible the computational load of the model, only one spatial radial peak frequency f_s has

been considered. This is in contrast with the physiological findings, since information in natural images is spread over a wide range of frequencies, it is necessary to use a mechanism that allows to get information from the whole range of frequency.

In this paper, we propose a multi-scale approach as illustrated in Fig. 3. This is a classical approach used in computer vision. It consists in (i) a pyramidal decomposition with L levels [35] and (ii) a coarse-to-fine refinement [36], which is a computationally efficient way to take into account the presence of different spatial frequency channels in the visual cortex and their interaction.

Using this approach, the spatial distance between corresponding points is reduced, thus yielding to a more precise estimate, since the residual values of the velocities lie in the filters' range. This also allows large displacements to be estimated which is a crucial aspect when dealing with real sequences. Interestingly, at a functional level, there is an experimental evidence that MT neurons seems to follow a coarse-to-fine strategy [37] suggesting that motion signals become more refined over time.

The equivalence between a multi-scale approach and the corresponding multi-resolution approach is shown in Fig. 4. The multi-scale analysis is performed by using three banks of Gabor filters with different spatial peak radial frequencies, each separated by an octave scale. The multi-resolution approach is obtained by iteratively low-pass filtering and subsampling the input image, then only the outermost bank of filter (i.e., the highest frequency one) is applied.

3.2. Boundary conditions

The problem of boundary conditions arises as soon as we need to consider values outside the domain of definition Ω. Even with simple Gaussian smoothing, when estimating results close to the boundaries, one needs to access values outside Ω. This is solved generally by choosing some boundary conditions like Neumann or Dirichlet. However, in our case, using such assumptions might introduce some strong errors at the boundaries. For this reason, we proposed instead to work inside an inner region denoted by Ω_{in} in which only available values are taken into account (so that no approximation or assumption has to be made), and then to interpolate values in the remaining outer region denoted by Ω_{out}. Note that this is an important issue to consider, especially because we use a multi-scale approach since errors

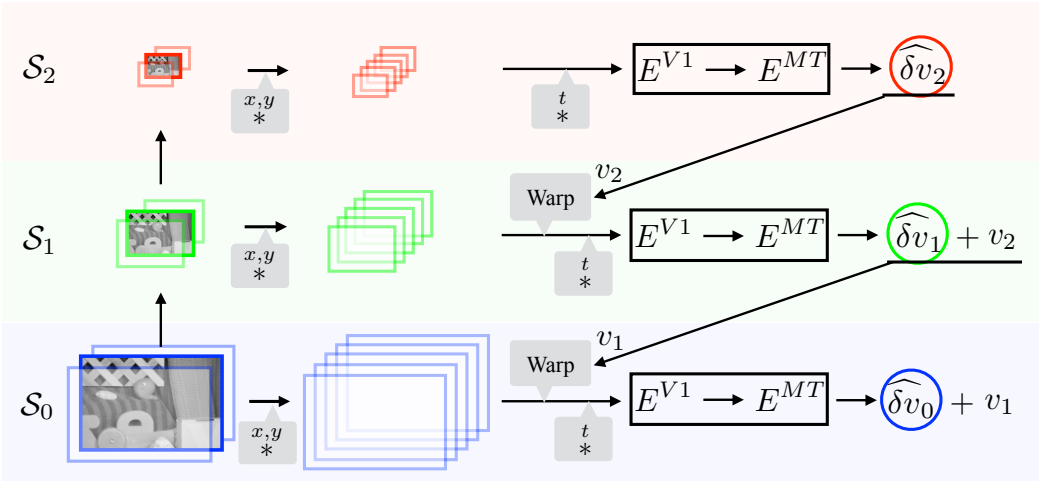


Figure 3: Multi-scale approach: In this example, three scales are represented ($L = 3$). Pyramidal decomposition is denoted by $\mathcal{S}_l$ with $(l = 0 \ldots L-1)$ ($l = 0$ is the finer scale). At a scale l, the estimated residual optical flow $(\widehat{\delta v_l})$ plus the optical flow coming from the coarser scale (v_{l+1}) is used to warp the sequence of the spatially filtered images at scale $l-1$.

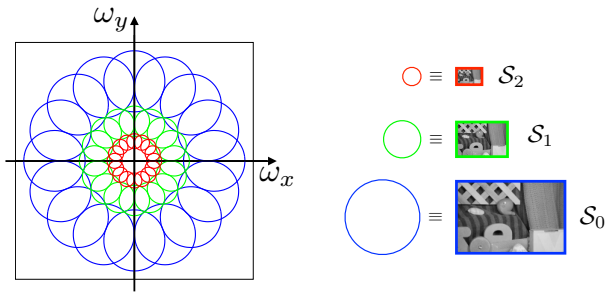


Figure 4: Equivalence between a multi-scale approach and the corresponding multi-resolution approach. This figure shows the amplitude spectra of three banks of Gabor filters with three spatial peak radial frequencies and eight spatial orientation: this frequency representation is a slice obtained for a fixed ω_t, the $(\omega_x, \omega_y, \omega_t)$ amplitude spectra of the bank of filters is shown in Fig. 2. Processing the image at full resolution by using the three banks of filters is equivalent to apply the outermost bank of filters to the three subsampled images.

done at the boundaries at low scales can spread a lot as scales are getting finer.

The way to defined the outer region Ω_{out} is illustrated in Fig. 5(a). It is constructed by first taking into account the region $\mathcal{B}_1$ in which V1 cells would

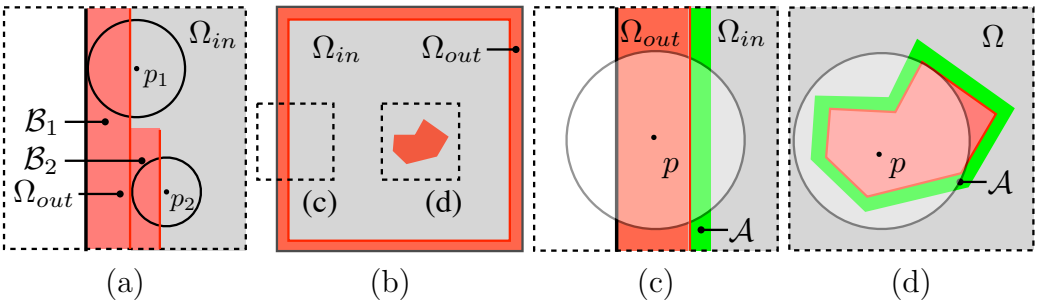


Figure 5: Illustration of the filling-in approach used to deal with boundary conditions and the unreliable regions. (a) How inner domain Ω_{in} (in grey) is defined taking into account V1 filter spatial size and V1 to MT pooling. Ω_{out} (in red) corresponds to $\mathcal{B}_1 \cup \mathcal{B}_2$ (see text). (b) Image domain showing the inner region Ω_{in} where exact computations can be done (i.e., without any approximation), the outer region Ω_{out} where an interpolation scheme is applied, and an example of unreliable region explained in (d). (c) Illustration of the interpolation scheme for a pixel $p \in \Omega_{out}$, showing the spatial neighbourhood associated with the spatial support of the integration and in green the region $\mathcal{A}$ which is used to estimate the interpolated values. (d) Same as (c) but in the case of an unreliable region.

need values outside Ω, and then the regions $\mathcal{B}_2$ corresponding to MT cells that would pool information from V1 cells in $\mathcal{B}_1$. So we have $\Omega_{out} = \mathcal{B}_1 \cup \mathcal{B}_2$ and $\Omega_{in} = \Omega \backslash \Omega_{out}$. Given this definition of inner and outer regions (Fig. 5(b)), the idea is to make all the estimations in Ω_{in} and to interpolate values in the outer region Ω_{out} (Fig. 5(c)). Given E^{MT} estimated in Ω_{in}, we propose that

$$E^{MT}(p) = \frac{1}{N(p)} \int_{p' \in \mathcal{A}} f_\alpha(\|p - p'\|) f_\gamma(I(p) - I(p')) E^{MT}(p') dp' \quad \forall p \in \Omega_{out}, \tag{12}$$

where $\mathcal{A}$ contains pixels at the inner boundary of Ω_{in} (green region) where E^{MT} is well estimated, function f_μ is defined as in (11), α and γ are parameters and $N(p)$ is a normalizing term

$$N(p) = \int_{p' \in \mathcal{A}} f_\alpha(\|p - p'\|) f_\gamma(I(p) - I(p')) dp'.$$

This method is based on luminance similarities using the same idea as developed in Sec. 2.3. Note that other interpolation methods could be used instead.

3.3. Unreliable regions

A problem is found with regions having a null spatio-temporal content, which happens for example in the blank wall problem. In that case, locally,

it is not possible to find a velocity. Given a threshold T, a pixel p will be categorised as unreliable if and only if $E^{MT}(p,t,d,v^c) < T$ for all d and v^c. For these pixels, the same interpolation as (12) is proposed (Fig. 5(d)).

4. Results

4.1. Parameters settings

Table 2 gives parameters used in our simulations. The size of the spatial support of the V1 RF was chosen so that fine details in real-world sequences at high image resolution could be processed. V1 and MT RFs process the visual signal within an average time of 200 ms [38, 37], which corresponds to five frames for a standard video acquisition device, thus we have chosen the temporal support of the filters in order to match this constraint. With this choice, we can not have tuning to velocities higher than one pixel per frame (ppf), i.e., one ppf corresponds to the maximum temporal frequency (see (3)) that can be sampled for the Nyquist theorem. This limitation has been addressed here by considering a multi-scale approach, as explained in Sec. 3.1. The number of scales depends on the size of the input images and on the speed range (a priori unknown). For the Middlebury videos we chose six spatial scales. It is worth noting that to avoid the introduction of a loss of balance between the convolutions with the even and odd Gabor filters, the contribution of the DC component is removed [39]. Finally, we set the support of the spatial pooling $G_{\sigma_{pool}}$ to five which is in accordance with findings reported in literature [40, 41].

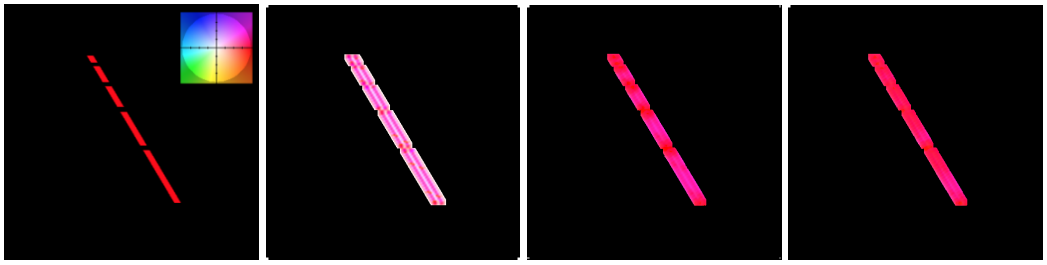
4.2. Analysis of proposed approaches

In this section, we evaluate the proposed FFV1MT model using synthetic and real sequences to show the intrinsic properties of our approach. When ground truth optical flow is available, average angular error (AAE) and endpoint error (EPE) will be estimated (with associated standard deviations) [26].

The influence of the number of spatial scales is shown in Fig. 6. In this sequence a dashed bar moves rightward with velocity (2,0) ppf. Results show that increasing the number of scales improves the results. It is worth noting that the aperture problem is correctly solved by considering three spatial scales in the small segments, whereas five spatial scales are needed to handle longer segments, though a residual optical flow at the finest scale is not

Description	Parameter	Value	Equation
V1			
RF spatial scale	σ	2.27 pixels	(1)
... and spatial support	SS	11×11 pixels,	(1)
Time constant of the exp. decay	τ	2.5 frames	(2)
... and temporal support	TS	5 frames	(2)
Spatial radial peak frequency	f_s	0.25 cycles/pixel	(1)
Temporal radial peak frequencies	f_t	$\{0, 0.10, 0.15, 0.23\}$ cycles/frame	(2)
Number of spatial contrast orientations	N	8 (from 0 to π)	(6)
... and sampling	θ_i	$\theta = k\pi/N, k = 0..N-1$	(6)
Number of component speeds	M	7	(3)
... and sampling	v^c	$\{-0.9, -0.6, -0.4, 0, 0.4, 0.6, 0.9\}$	(3)
Semi-saturation constant	ε	10^{-9}	(6)
MT			
Std dev of the Gaussian spatial pooling	σ_{pool}	0.9 pixels	(7)
... and spatial support		5×5 pixels	(7)
Decoding step			
Number of MT direction tuning directions		2	(9)
... and sampling	d	$\{0, \pi/2\}$	(9)
Algorithm			
Number of scales	L	6	
Spatial parameter of the interpolation	α	2.5 pixels	(12)
Luminance parameter of interpolation	γ	1/6 of luminance range	(12)
Other parameters for FFV1MT-TF model			
Spatial parameter	α	$\{0.50, 0.83, 1.16, 1.50, 1.83\}$ as a function of spatial scale	(10)
Range parameter	β	1/6 of energy range	(10)
Luminance parameter	γ	1/6 of luminance range	(10)

Table 2: Parameter values used in our simulations for the FFV1MT model and its extension FFV1MT-TF. Equation number refers to the equation where it has been first introduced.



Ground truth	FFV1MT ($L = 1$)	FFV1MT ($L = 3$)	FFV1MT ($L = 5$)
	AAE=52.21±4.43	AAE=10.32±5.78	AAE=6.20±3.82
	EPE=1.79±0.08	EPE=0.36±0.22	EPE=0.24±0.14

Figure 6: Influence of the number of spatial scales. The FFV1MT model is tested with L=1, 3 and 5 scales. The color code used to show optical flow is in the inset on the first image. This color code will be used in all figures to represent optical flow. Note that the aperture problem is partially solved by considering a scale-space approach, where the effective receptive field size of MT increases and thus takes into consideration 2-D cues that are present at a distance. This can be readily observed by the results on bars with different lengths.

correctly recovered in the middle of the longest segment, since the spatial support of the RFs is too small with respect to the visual feature.

The next example in Fig. 7 is on another synthetic video that represents a textured shape moving on top of a translating background. Optical flow result show a good estimation of the optical flow except in the neighbourhood of objects boundaries (which are also here motion boundaries). The FFV1MT–TF approach looks qualitatively better, however it does not improve the quantitative performance. It might be due to the noisy texture of this synthetic sequence.

In order to analyze the roles of the different stages of the model, Fig. 8 shows the V1 and MT activities. The first row shows $\|E^{V1}\|_\theta(p, v^c) = \left(\sum_{i=1}^{N} E^{V1}(p, \theta_i, v^c)^2\right)^{1/2}$: the activities do not identify specific tuning speeds, since all the spatial orientations are pooled in the norm and the tuning speeds are component speeds, i.e., they are orthogonal to the spatial orientation of the cell. The second row shows $\|E^{V1}\|_{v^c}(p, \theta) = \left(\sum_{i=1}^{M} E^{V1}(p, \theta, v_i^c)^2\right)^{1/2}$: the cells are elicited by the spatial orientation of the shape, the V1 layer shows a tuning on the spatial orientation. The third and fourth rows show $E^{MT}(p, 0, v^c)$ and $E^{MT}(p, \pi/2, v^c)$ maps, respectively. At MT layer, a speed tuning emerges: on the left, the energies are higher for the region related

to the shape, this means that there is a negative speed for the horizontal and vertical velocities related to the shape. On the right, the energies are higher for the background (for the third row, only), since the background moves rightwards. These results confirm that the V1 layer has a tuning on the spatial orientation (cells respond to the spatial orientation of the shape), whereas at MT layer, a speed tuning no more related to spatial orientation emerges (i.e., the aperture problem is solved).

In Fig. 9 we show the distribution of E^{MT} at different positions to understand its relation to velocities. By observing the distribution of MT energies in four different positions on the original image (indicated as (a), (b), (c) and (d) in Fig. 7), we see how the MT layer encodes the velocities. In particular: the behaviours in (a) and (c) are affected by the values of the neighboring borders, thus there are no prominent activities; in (b), which corresponds to a point on the foreground shape sufficiently far from borders given the actual spatial support of the filters, cells tuned to negative speeds (v_1^c) on both horizontal and vertical direction (E^{MT} with $d = 0$ and $d = \pi/2$, respectively) have the maximum response; in (d), which corresponds to a point on the background, only the response of the horizontal direction has a maximum for positive horizontal speed (v_7^c).

Fig. 10 shows the results of the FFV1MT model on the classical realistic Yosemite sequence with clouds. We obtain AAE=5.57 which is better than former biologically-inspired models such as the original Heeger approach (AAE=11.74, with 44.8% of reliable pixels,[42]) and the neural model from Bayerl and Neumann (AAE=6.20, [43]). One can also make comparisons with standard computer vision approaches such as Pyramidal Lucas and Kanade (AAE=6.41), modified Horn and Schunk (AAE=5.48 with 32.9% of reliable pixels, [42]) and 3DCLG (AAE=6.18, [44]), showing a better performance of the FFV1MT. The FFV1MT–TF approach shows a slightly better performance in particular close to motion discontinuities.

4.3. Performance evaluation on Middlebury dataset

In this section, we benchmark our approach on the computer vision dataset Middlebury [26][1]. The sequences in this dataset bring several challenges, such as sharp edges, high velocities and occlusions. Figure 11 show results obtained on the training dataset, which has public available ground

[1] http://vision.middlebury.edu/flow/data/

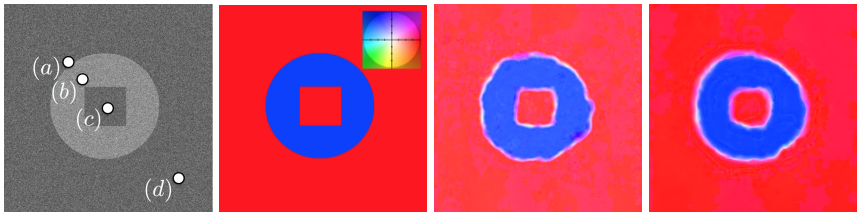


Figure 7: Results on a synthetic video: A translating shape is moving with velocity $v = (-3, -3)$ ppf on top of a translating background moving with velocity $v = (4, 0)$ ppf. Results are AAE=3.56±14.40, EPE=0.26±0.86. for FFV1MT and AAE=3.70±14.78, EPE=0.27±0.86 for FFV1MT–TF.

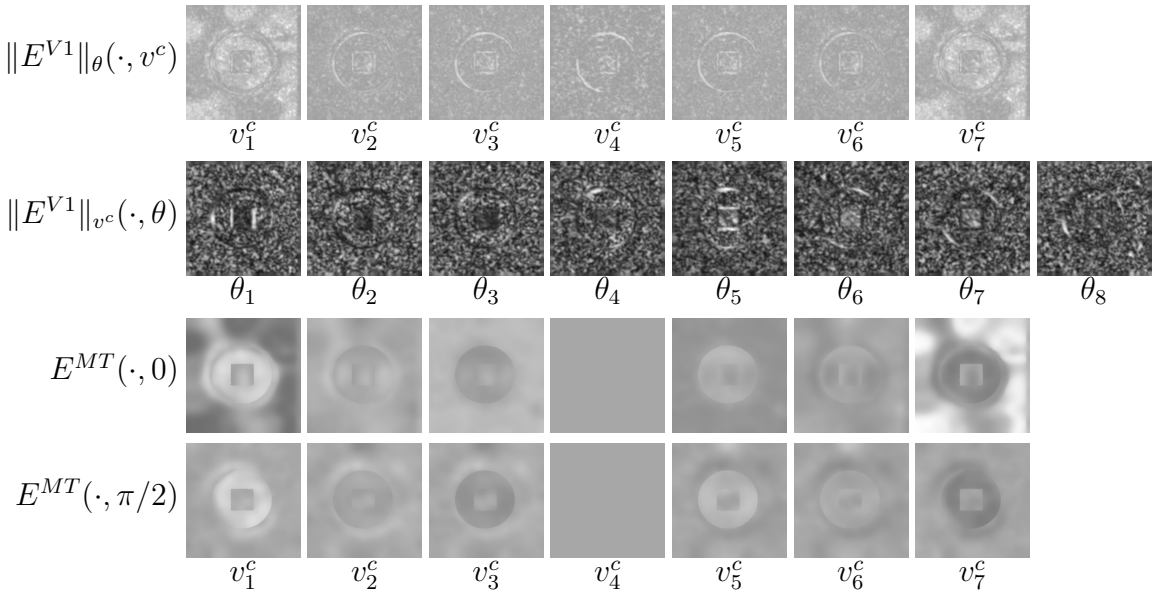


Figure 8: V1 and MT activities on the synthetic video shown in Fig. 7 (see text).

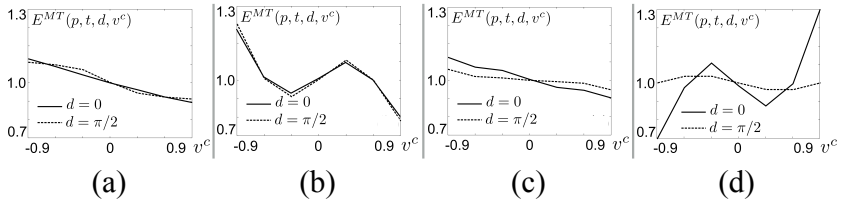
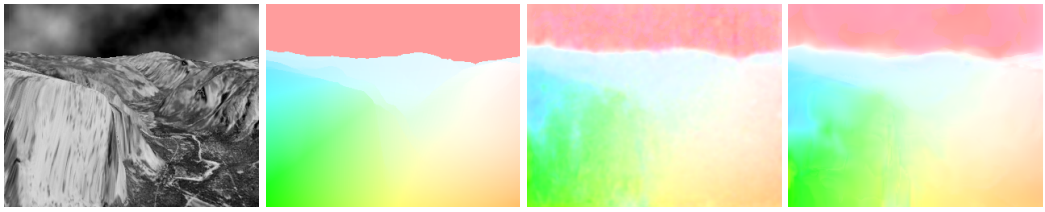


Figure 9: Distribution of MT energy at positions indicated in Fig. 7.



Sample input	Ground truth	FFV1MT	FFV1MT–TF
		AAE=5.57 ± 12.27	AAE=4.58 ± 9.35
		EPE=0.23 ± 0.39	EPE=0.20 ± 0.33

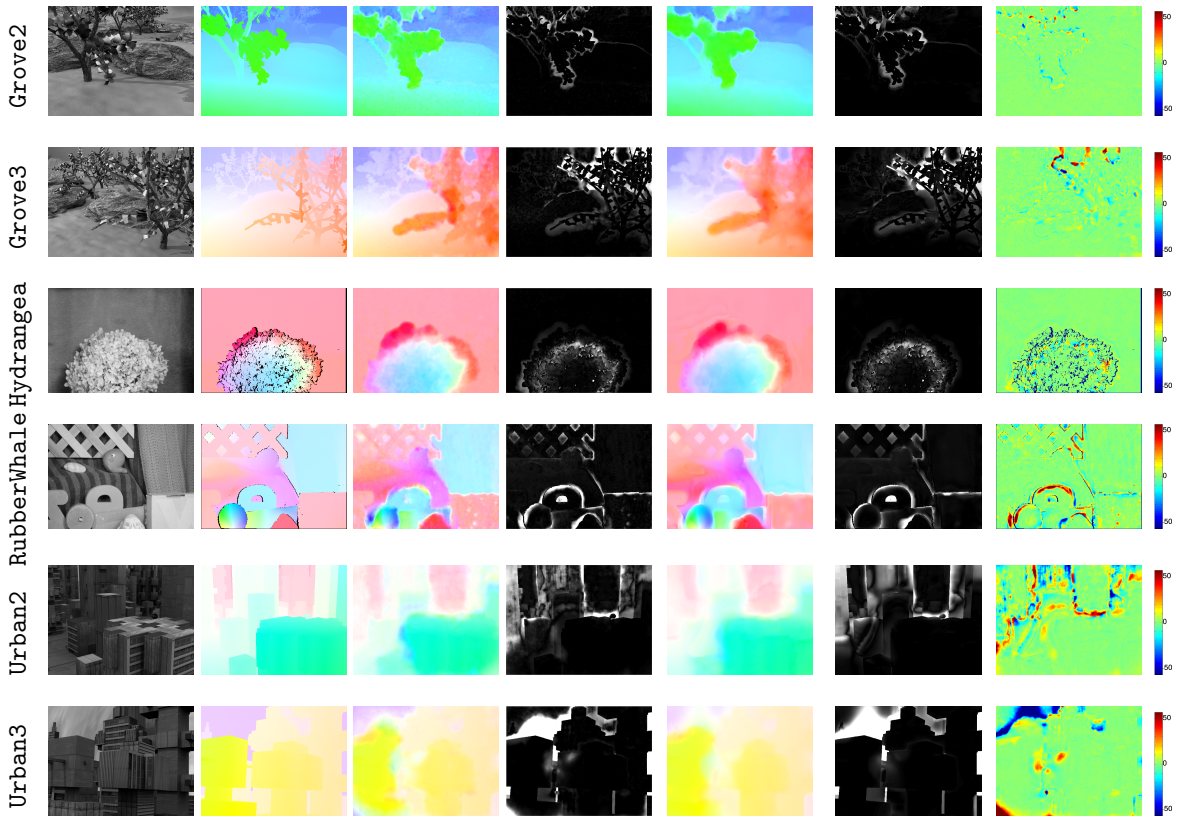
Figure 10: Performance of the FFV1MT and FFV1MT–TF models on the classical `Yosemite` sequence with clouds. The color code is the same as in Fig. 6.

truth. The AAEs and EPEs show that FFV1MT is able to recover reliable optical flows, though some issues remain open. Smooth effects are present on edges and fine details (see `Grove2` and `Grove3`), FFV1MT–TF partially solves this issue, as shown in `RubberWhale` and `Urban2`. The δAAE maps highlight the differences in the AAEs between FFV1MT and FFV1MT–TF, showing that the latter is better on edges as expected (red tones). In presence of high image velocity large occlusions occur, on which both approaches fail (see left-hand side of `Urban3`). In this case, the worst performance of FFV1MT–TF method is due to the fast movements of edges that undermines the luminance similarity principle on which it is based.

Figure 12 show results obtained on the test dataset. Higher errors coincide with occlusions (see, e.g., `Urban` sequence) and sharp edges (see, e.g., `Urban` and `Wooden` sequences), similarly to what was observed on the training set. Results can be further analysed through the Middleburg website and compared to a variety of state-of-the-art algorithms. It is worth noting that our FFV1MT model is the only neural model for motion estimation shown in the table so far.

5. Conclusion

In this paper, we have presented an approach that is based on models primarily developed to account for various physiological findings related to motion processing in primates. Starting from the classical hierarchical feedforward processing model involving V1 and MT cortical areas, which is usually limited to a single spatial scale, we have extended it to consider the whole range of frequency by adapting a multi scale approach and analysed

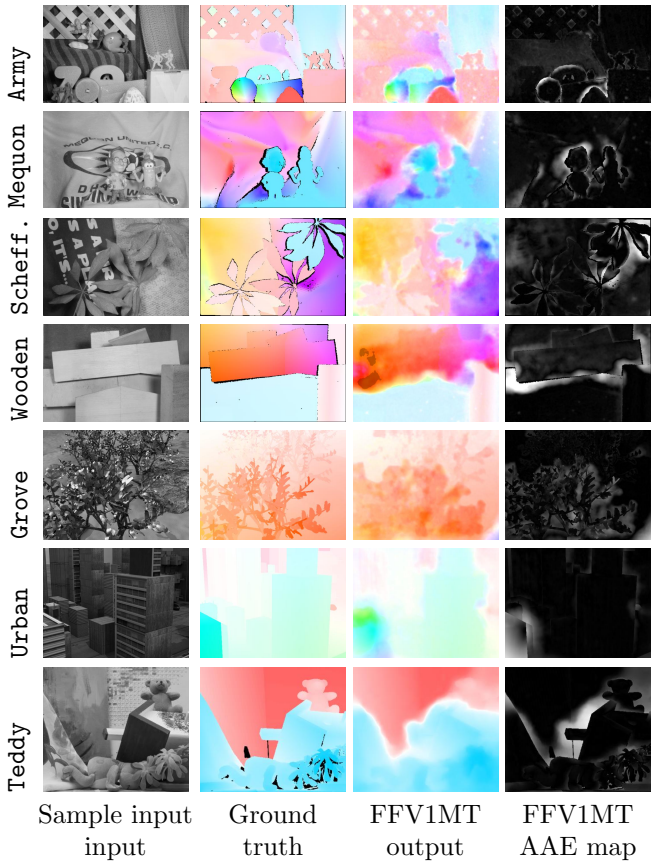


Sequence	FFV1MT		FFV1MT–TF	
	AAE ± STD	EPE ± STD	AAE ± STD	EPE ± STD
grove2	4.28 ± 10.25	0.29 ± 0.62	3.96 ± 10.00	0.26 ± 0.60
grove3	9.72 ± 19.34	1.13 ± 1.85	9.40 ± 18.39	1.10 ± 1.74
Hydrangea	5.96 ± 11.17	0.62 ± 0.96	5.74 ± 11.03	0.59 ± 0.86
RubberWhale	10.20 ± 17.67	0.34 ± 0.54	9.38 ± 15.21	0.31 ±0.46
urban2	14.51 ± 21.02	1.46 ± 2.13	13.62 ± 10.23	1.19 ±1.58
urban3	15.11 ± 35.28	1.88 ± 3.27	16.22 ± 2.21	1.69 ±2.85

Figure 11: Sample results and error measurements on Middlebury training set. δAAE=$AAE_{FFV1MT} - AAE_{FFV1MT--TF}$ is represented with a color code, where red and blue tones are for positive and negative values, respectively.

the efficacy of the approach in estimating the dense optical flow in real world scenarios by considering an efficient velocity decoding step.

Here, we show that a V1-MT feedforward model can be successfully used



Sequence	AAE All (Rank) - Disc. (Rank)	EPE All (Rank) - Disc. (Rank)
Army	12.02(102) - 23.3(102)	0.33(100) - 0.64(100)
Mequon	10.7(94) - 26.6(103)	0.79(94) - 1.90(103)
Schefflera	15.6(96) - 29.0(101)	1.33(104) - 1.90(101)
Wooden	16.6(102) - 36.3(105)	1.38(103) - 2.98(104)
Grove	6.51(105) - 6.40(103)	1.76(105) - 1.99(105)
Urban	16.2(104) - 30.7(105)	2.33(105) - 3.64(106)
Yosemite	3.41(74) - 5.44(88)	0.16(66) - 0.18(83)
Teddy	12.3(101) - 18.8(102)	1.81(100) - 2.64(100)

Figure 12: Sample results and error measurements of FFV1MT model on Middlebury test set. By the time of evaluation 107 algorithms are benchmarked by the website, and Rank indicates the relative performance of the method with respect to others for both the entire sequence (All) and for discontinuities (Disc.). The results are public at `http://vision.middlebury.edu/flow/eval`

to compute optical flow in real videos. We have tested the performance of our model using synthetic stimuli as well as the standard Middlebury dataset. A qualitative evaluation shows that model could recover velocity vectors in regions with coarse textures quite well, but typically fails to achieve robust estimates in regions with very fine texture or regions with sharp edges. This was expected, since the V1-MT feedforward model does not take into account the details of lateral interactions and scale space issues that need to be tackled in order to solve the blank wall problem. In order to address these problems, we proposed a simple extension of our baseline model using trilateral filtering at MT level as a way to simulate lateral interactions between MT cells. Results were slightly improved suggesting that one should further focus on lateral interactions and possibly feedback into the models to better deal with real videos.

Moreover, this work has opened up several interesting question, which could be of relevance to biologists as well, for example what could be afferent pooling strategy of MT when there are multiple surfaces or occlusion boundaries within the MT receptive field? Can a better dense optical flow map be recovered by considering different multi-scale strategies? These questions are currently under consideration.

We think that this work could act as a good starting point for building scalable computer vision algorithms for motion processing that are rooted in biology. For that reason we propose to share the code in order to encourage research in this direction. Our Matlab code for the FFV1MT model has been made available on ModelDB [45]: `http://senselab.med.yale.edu/modeldb/`.

Acknowledgments

KM and PK acknowledge funding from the EC IP project FP7-ICT-2011-8 no. 318723 (MatheMACS). We are thankful to Tom Morse and ModelDB [45] for distributing our code.

Bibliography

[1] D. Fortun, P. Bouthemy, C. Kervrann, Optical flow modeling and computation: a survey, Computer Vision and Image Understanding 134 (2015) 1–21.

[2] L. C. Sincich, J. C. Horton, The circuitry of V1 and V2: Integration of color, form, and motion, Annual Review of Neuroscience 28 (1) (2005) 303–326, pMID: 16022598.

[3] M. J. Rasch, M. Chen, S. Wu, H. D. Lu, A. W. Roe, Quantitative inference of population response properties across eccentricity from motion-induced maps in macaque V1, Journal of Neurophysiology 109 (5) (2013) 1233–1249.

[4] N. Rust, V. Mante, E. Simoncelli, J. Movshon, How MT cells analyze the motion of visual patterns, Nature Neuroscience 9 (2006) 1421–1431.

[5] 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.

[6] D. Bradley, M. Goyal, Velocity computation in the primate visual system, Nature Reviews Neuroscience 9 (9) (2008) 686–695.

[7] C. Pack, R. Born, Cortical mechanisms for the integration of visual motion, in: R. H. Masland, T. D. Albright, T. D. Albright, R. H. Masland, P. Dallos, D. Oertel, S. Firestein, G. K. Beauchamp, M. C. Bushnell, A. I. Basbaum, J. H. Kaas, E. P. Gardner (Eds.), The Senses: A Comprehensive Reference, Academic Press, New York, 2008, pp. 189 – 218.

[8] J. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, Journal of the Optical Society of America A 2 (1985) 1160–1169.

[9] E. Adelson, J. Bergen, Spatiotemporal energy models for the perception of motion, Journal of the Optical Society of America 2 (1985) 284–321.

[10] D. Heeger, Model for the extraction of image flow, Journal of the Optical Society of America 4 (8) (1987) 1455–1471.

[11] N. Grzywacz, A. Yuille, A model for the estimate of local image velocity by cells in the visual cortex, Proceeding of the Royal Society of London B 239 (1990) 129–161.

[12] G. C. Deangelis, I. Ohzawa, R. D. Freeman, Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. II. Linearity of

temporal and spatial summation, Journal of Neurophysiology 69 (4) (1993) 1118–1135.

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

[14] N. C. Rust, V. Mante, E. P. Simoncelli, J. A. Movshon, How MT cells analyze the motion of visual patterns, Nature Neuroscience 9 (11) (2006) 1421–1431.

[15] P. Bayerl, H. Neumann, Disambiguating visual motion through contextual feedback modulation, Neural Computation 16 (10) (2004) 2041–2066.

[16] P. Bayerl, H. Neumann, A fast biologically inspired algorithm for recurrent motion estimation, Pattern Analysis and Machine Intelligence, IEEE Transactions on 29 (2) (2007) 246–260.

[17] E. 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.

[18] U. Ilg, G. Masson, Dynamics of Visual Motion Processing: Neuronal, Behavioral, and Computational Approaches, SpringerLink: Springer e-Books, Springer Verlag, 2010.

[19] J. Bouecke, E. Tlapale, P. Kornprobst, H. Neumann, Neural mechanisms of motion detection, integration, and segregation: From biology to artificial image processing systems, EURASIP Journal on Advances in Signal Processing 2011, special issue on Biologically inspired signal processing: Analysis, algorithms, and applications.

[20] D. Heeger, Optical flow using spatiotemporal filters, The International Journal of Computer Vision 1 (4) (1988) 279–302.

[21] S. Nowlan, T. Sejnowski, Filter selection model for motion segmentation and velocity integration, J. Opt. Soc. Am. A 11 (12) (1994) 3177–3199.

[22] N. V. K. Medathati, H. Neumann, G. S. Masson, P. Kornprobst, Bio-inspired computer vision: Setting the basis for a new departure, Tech. Rep. 8698, INRIA (Mar. 2015).

[23] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: International Joint Conference on Artificial Intelligence, 1981, pp. 674–679.

[24] E. Simoncelli, E. H. Adelson, Computing optical flow distributions using spatio-temporal filters, Tech. rep., MIT Media Lab Vision and Modeling, Tech. Rep (1991).

[25] S. Nishimoto, J. L. Gallant, A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies, The Journal of Neuroscience 31 (41) (2011) 14551–14564.

[26] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, R. Szeliski, A database and evaluation methodology for optical flow, International Journal of Computer Vision 92 (1) (2011) 1–31.

[27] G. R. Stoner, T. D. Albright, V. S. Ramachandran, Transparency and coherence in human motion perception., Nature 344 (6262) (1990) 153–155.

[28] A. Noest, A. Van Den Berg, The role of early mechanisms in motion transparency and coherence, Spatial Vision 7 (2) (1993) 125–147.

[29] B. C. Skottun, Neuronal responses to plaids, Vision Research 39 (12) (1999) 2151 – 2156.

[30] L. Paninski, Maximum likelihood estimation of cascade point-process neural encoding models, Network: Computation in Neural Systems 15 (4) (2004) 243–262.

[31] J. H. Maunsell, D. C. Van Essen, Functional properties of neurons in middle temporal visual area of the macaque monkey. I. selectivity for stimulus direction, speed, and orientation, Journal of Neurophysiology 49 (5) (1983) 1127–1147.

[32] A. Pouget, K. Zhang, S. Deneve, P. E. Latham, Statistically efficient estimation using population coding, Neural Computation 10 (2) (1998) 373–401.

[33] K. R. Rad, L. Paninski, Information rates and optimal decoding in large neural populations., in: J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, K. Q. Weinberger (Eds.), NIPS, 2011, pp. 846–854.

[34] S. Paris, P. Kornprobst, J. Tumblin, F. Durand, Bilateral filtering: Theory and applications, Foundations and Trends in Computer Graphics and Vision 4 (1).

[35] C. A. J.R. Bergen, E.H. Adelson, P. Burt, J. Ogden, Pyramid methods in image processing, RCA Engineer 29 (1984) 33–41.

[36] E. P. Simoncelli, Course-to-fine estimation of visual motion, in: IEEE Eighth Workshop on Image and Multidimensional Signal Processing, 1993.

[37] C. C. Pack, R. T. Born, Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain, Nature 409 (2001) 1040–1042.

[38] G. C. DeAngelis, I. Ohzawa, R. D. Freeman, Receptive-field dynamics in the central visual pathways, Trends in Neurosciences 18 (10) (1995) 451 – 458.

[39] D. A. Clausi, M. E. Jernigan, Designing Gabor filters for optimal texture separability, Pattern Recognition 33 (11) (2000) 1835 – 1849.

[40] T. D. Albright, R. Desimone, Local precision of visuotopic organization in the middle temporal area (MT) of the macaque, Experimental Brain Research 65 (3) (1987) 582–592.

[41] P. Bayerl, H. Neumann, Disambiguating visual motion through contextual feedback modulation., Neural Computation 16 (10) (2004) 2041–2066.

[42] J. Barron, D. Fleet, S. Beauchemin, Performance of optical flow techniques, The International Journal of Computer Vision 12 (1) (1994) 43–77.

[43] P. Bayerl, H. Neumann, Disambiguating visual motion through contextual feedback modulation, Neural Computation 16 (10) (2004) 2041–2066.

[44] A. Bruhn, J. Weickert, C. Schnrr, Lucas/kanade meets horn/schunck: Combining local and global optic flow methods, International Journal of Computer Vision 61 (3) (2005) 211–231.

[45] M. L. Hines, T. Morse, M. Migliore, N. T. Carnevale, G. M. Shepherd, ModelDB: A database to support computational neuroscience, J. Comput. Neurosci. 17 (1) (2004) 7–11.