Adaptive V1-MT model for motion perception

Shuai Li, Xiaoguang Fan, Yuelei Xu and Jinke Huang
Aeronautics and Astronautics Engineering College, Air Force Engineering University
Xi'an, Shaanxi 710038 - China
[e-mail: lishuailisuai@163.com]
*Corresponding author: Shuai Li

Received June 19, 2017; revised January 10, 2018; accepted April 11, 2018; published January 31, 2019

Abstract

Motion perception has been tremendously improved in neuroscience and computer vision. The baseline motion perception model is mediated by the dorsal visual pathway involving the cortex areas the primary visual cortex (V1) and the middle temporal (V5 or MT) visual area. However, few works have been done on the extension of neural models to improve the efficacy and robustness of motion perception of real sequences. To overcome shortcomings in situations, such as varying illumination and large displacement, an adaptive V1-MT motion perception (Ad-V1MTMP) algorithm enriched to deal with real sequences is proposed and analyzed. First, the total variation semi-norm model based on Gabor functions (TV-Gabor) for structure-texture decomposition is performed to manage the illumination and color changes. And then, we study the impact of image local context, which is processed in extra-striate visual areas II (V2), on spatial motion integration by MT neurons, and propose a V1-V2 method to extract the image contrast information at a given location. Furthermore, we take feedback inputs from V2 into account during the polling stage. To use the algorithm on natural scenes, finally, multi-scale approach has been used to handle the frequency range, and adaptive pyramidal decomposition and decomposed spatio-temporal filters have been used to diminish computational cost. Theoretical analysis and experimental results suggest the new Ad-V1MTMP algorithm which mimics human primary motion pathway has universal, effective and robust performance.

Keywords: Bio-inspired approach, V1, MT, TV-Gabor, motion perception, optical flow

http://doi.org/10.3837/tiis.2019.01.021 **ISSN : 1976-7277**

1. Introduction

Motion perception in visual scenes, the process of extracting and interpreting visual motion information of any vision system, has been studied extensively in last decades. Visual motion perception, in technical terms, could be described as optical flow estimation, a vector field indicating local velocity (both direction and speed) at each retinal sequence location. Given that the mammals' vision system has evolved highly efficient for complex motion information interpretation, understanding the neural mechanisms behind the visual motion analysis would be very beneficial and has spurred many researchers to investigate mechanisms for optical flow estimation. An accurate estimation of vision motion can be used effectively for many tasks, such as autonomous navigation, 3D scene (environment) interpretation, and target tracking [1-3].

Ever since the work of Horn and Schunck [4], many efforts have been made to increase the accuracy and reliability of optical flow estimation from sequence motion with a large number of methods, such as coarse-to-fine model to solve large displacement motion [2, 5], high-order filter constancy to overcome the influence of lighting varying [6], bilateral filtering to preserve motion boundaries [7], temporal averaging of image derivatives [8]. In addition, a prominent benchmarking dataset has been developed to evaluate and compare publicly estimation algorithms in natural image sequences [9]. However, the main remaining challenges, such as occlusions, motion discontinuities and large displacements, are still difficult to deal with because the existing models that either lack robustness or have very complex computational process.

On the other hand, psychophysical and neurophysiological results of the visual cortex have extensively inspired investigation on visual motion interpretation. To analyze visual motion, neural mechanisms compute oriented elements by filtering the real sequences in both space and time. Neurons in visual cortex area V1 and MT are found to play a vital role in motion perception [10-13]. Previous studies on highly primed stimuli have shown that V1 neurons are selective for spatiotemporal orientation, while MT neurons respond best to a velocity (speed and direction). However, visual motion interpretation is not only processed in areas V1 and MT, other areas also contribute greatly to motion information expression, such as neurons in V2 and V4 project the orientation or color of local edges of motion scenes to the motor cortex.

Even though many biological models of motion processing have been proposed to solve the local motion estimation problem, all these linear-nonlinear filtering models still mainly focus on spatially homogeneous motion inputs, such as random-dots, gratings and plaids, and largely ignore the temporal aspect. Thus, dynamical models based on experimental data have been proposed to explain the diffusion of non-ambiguous local motion cues [11, 14]. At first using Gabor functions or spatiotemporal oriented filters, models of motion sensitive V1 neurons (simple neurons and complex neurons) were made to explain the responses of receptive fields (RF) to visual motion. For instance, [11, 15, 16] mimicked V1 simple neurons with a linear model followed by tuned and unturned normalization. Rectified RFs were localized in space-time and tuned for spatiotemporal orientation. Through combining simple neuron afferents, the motion energy model can be used to explain many characteristics of V1 complex neuron responses. Then linear-nonlinear V1-MT feed-forward models were proposed to explain properties of MT neurons. However, it departs from the highly nonlinearity and adaptability of visual system. Moreover, they are barely simulated the lateral or feedback interactions of RFs and evaluated on complex real sequences.

Here, we draw inspiration from mammals' vision system to build our approach on results of neuroscience. Our key contributions are to (i) analyze and extend current neural models by adding feedback connectivity and controlling the pooling between V1, V2and MT, which is believed important to investigate ambiguity regions such as aperture or blank wall and adaptive pool image structure and contrast, (ii) modify motion perception methods to produce state-of-the-art velocity estimation through adaptive pyramidal approach and coarse-to-fine method, and focus on dealing with illumination-dependent problem through the total variation semi-norm model based on Gabor functions (TV-Gabor) model. Moreover, (iii) propose several strategies to decode the velocity of visual motion and adapt computationally efficient mechanisms to cope with problems encountered in real sequences.

In this paper, we briefly review biological vision solution of the motion processing and computational problems in motion perception, and the Heeger and Simoncelli framework of mimicking human visual processing stages, on which the new model is based in Section 2. Section 3 provides an overview of our Ad-V1MTMP model for motion interpretation and estimation, and focuses on exploiting several methods for extracting low-level features of the scene, and on the effect of image local contrast and texture in motion perception through considering neurons in area V2. Section 4 evaluates the performance of Ad-V1MTMP model on standard Middlebury datasets and real illumination changing sequences, while Section 5 is left to conclusion.

2. Related Work

2.1 Biological vision

In visual neuroscience, state-of-the-art models have been proposed to explain the properties of low-level motion processing of neural mechanisms [17-18]. **Fig. 1** depicts a schematic view of the hierarchical networks of motion analysis. Visual cortex, a complex hierarchical expression structure in the processing of visual information, is consist of two pathways: i) “What” ventral visual pathway, which is aimed at extracting the complex static features for hierarchical cognitive expression tasks, such as pattern recognition and image classification; (ii) "Where" dorsal visual pathway, which provides a distributed representation of visual features for the visual control of actions, such as motion perception and accurate positioning.

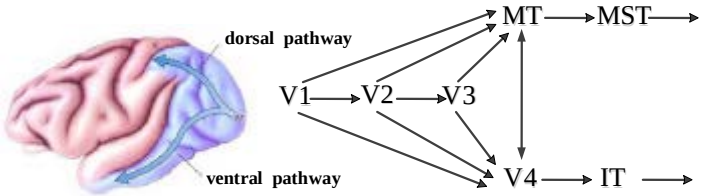


Fig. 1. The hierarchical networks in primate cerebral cortex

Although the two pathways proceed visual information independently, populations of hierarchical layers units may provide a deeper scene understanding and encode distributed representations with a more efficient computational framework. First layer composed of V1 simple and complex neurons extracts local motion information through a set of spatiotemporal filters. V1 simple neurons are always mimicked with a set of band-pass filters like Gabor functions, while complex neurons integrate information at different direction and frequency

by performing non-linear operations over feedback afferents. Gabor functions also provide us with both adaptive frequency and directional image decomposition, which can be used to deal with illumination changes. MT neurons pool and combine the responses of populations of V1 complex neurons over a wide range of spatio-temporal scales to percept the velocity vector. However, not only areas V1 and MT implement context modulations by center-surround intra- and inter-interactions, other areas like V2 and V4, also play very important role in visual motion perception. Above all, the processing of motion perception is based on a set of biological vision mechanisms.

2.2 The Heeger and Simoncelli model

In this paper, we explore the possible framework of mimicking human visual processing stages to provide motion information and features of visual inputs. The Heeger and Simoncelli (HS) model is a most canonical mechanism to detect local image velocity. HS model contains two layers through feedforward connections to deal with different visual problems, described by **Fig. 2.**

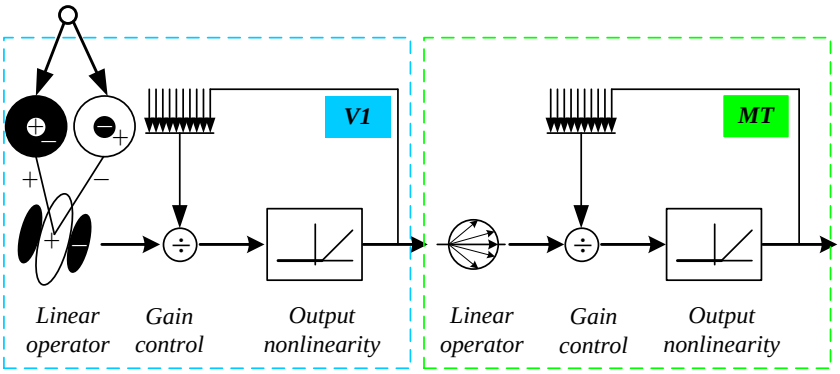


Fig. 2. The Heeger and Simoncelli model overview

The V1 layer is often divided into the simple neurons and the complex neurons, whose responses can be obtained with a series of nonlinear combination of filter responses, and the most critical work is the appropriate analysis and simulation of V1 RF. First, we characterize the visual stimulus, $I(p,t)$, by its local contrast,

$$A(p,t) = [I(p,t)/I_{avg} - 1] \; ; \; p=(x,y) \tag{1}$$

where I_{avg} is the spatial-temporal average of the input stimulus in space and time.

Then, the linear response of the *i*th simple neuron, $L_i(t)$, is given by convolving its local RF ($W_i(p,t)$ with the Retina output ($A(p,t)$),

$$L_i(t) = \iiint W_i(p,T)A(p,t-T)dxdydT + \alpha \tag{2}$$

where $W_i(p,t)$ denotes the RFs of V1 simple neurons with a set of directional third derivatives of Gaussian filters, and α is a very small constant.

Here, we define the final output of the *i*th simple neuron as

$$RS_i(t) = K_1 \lfloor L_i(t) \rfloor^2 \Big/ (\sum_j \lfloor L_j(t) \rfloor^2 + \sigma_1^2) \tag{3}$$

where K_1 reflects the maximum outputs of V1 simple neurons, σ_1 is the semi-saturation

constant.

Using the motion energy model, V1 complex neuron responses combine a quadrature pair of simple neuron responses:

$$C_i(t) = \sum_j c_{ij} RS_j(t) \tag{4}$$

here, the weights C_{ij} are all positive.

$$Q_i(t) = \sum_j p_{ij} C_j(t) + \beta \tag{5}$$

$$P_i(t) = K_2 \lfloor Q_i(t) \rfloor^2 \Big/ (\sum_j \lfloor Q_j(t) \rfloor^2 + \sigma_2^2) \tag{6}$$

MT layer contains populations of pattern neurons whose velocity selectivity is described via the combination of V1 complex neuron afferents, described as equation 5 and 6.

After above processes, HS model can compute motion energies for perceiving the local velocity at each point. But this linear-nonlinear filtering model only focuses on spatially homogeneous motion inputs, such as random-dots, gratings and plaids. The local analyzer also cannot perceive the velocity along the gradient for the case of local luminance changes only at one orientation in aperture problem, observe to any kind of motion in blank wall problem for the absence of illumination contrast, and arrive at an accurate estimation in multiple motions or multiple objects case. Therefore, it can be extended as a way of motion perception in natural scenes.

3. Adaptive V1-MT motion perception (Ad-V1MTMP)

Above neural mechanisms of the HS model involving a feedforward processing are proposed to account for biological tests on motion perception of homogeneously stimuli. However, it is not sufficient to handle visual challenges, like blank wall problem, strong textured regions, and occlusion boundaries. In this paper, Gabor filters are adopted to simulate the RFs of V1 neuron, and further image texture features could be obtained through structure-texture decomposition technique based on TV-Gabor, which is useful to manage the illumination and color changes. Then, this paper takes the feedback inputs from area V2 and lateral connectivity in area MT into account by considering the image local contrast at a given location. Moreover, we use series of methods to deal with different visual problems, such as coarse-to-fine refinement, multi-scale approach, and adaptive pyramidal decomposition.

3.1 Illumination adaptive processing

In this paper, we use the structure-texture decomposition to handle illumination-dependent problems under shadow or shading reflection conditions. More formally, the texture image is a kind of residuals that reflects the homogeneity property of surface structure with slow change or periodic change. Different from image intensity and structure, image texture is a visual feature represented by the grayscale distribution of pixels and their neighborhoods. By employing total variation (TV) regularization, [18] had provided implementation details to preserve high frequencies or discontinuities of images. TV-Gabor model defined in [18] is

$$\inf_u \ (J(u) + \frac{\eta}{2} \left\| \sqrt{K}(f - u) \right\|_{L^2}^2) \ ; \ \sqrt{K}^2 = K \tag{7}$$

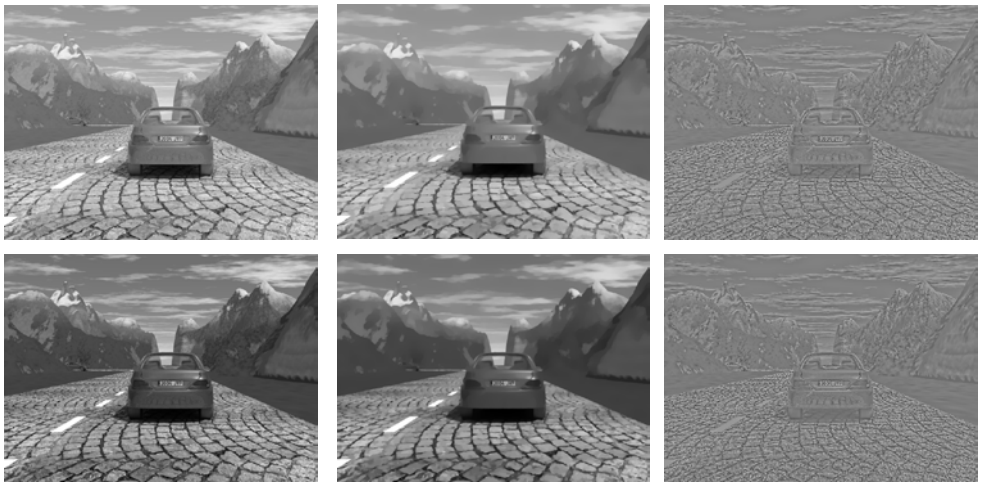
where $J(\cdot)$ denoting the variation (TV) regularization is the indicator function of some closed

convex set. H is Hilbert spaces defined thanks to the operator K.

Here, K denotes the convolution operator, so K^{-1} is equal to $1/K$ in the Fourier domain. During structure-texture decomposition, frequencies not included in image texture should be penalized with convolution K, so as K^{-1}. Thus, Gabor functions could be used to account for properties of the inverse kernel.

$$h_k = \cos(2\pi vk)\frac{1}{\sqrt{2\pi\sigma^2}}exp(-\frac{k^2}{2\sigma^2});\ v \in (0, 0.5] \tag{8}$$

where v is the texture frequencies.



(a) original image (b) the structure part (c) the texture part

Fig. 3. Decomposition of synthetic input images by TV-Gabor model

After the above processes, original images are decomposed into two parts. One is the structural part, which depicts the smooth and structure features of the input. The other one is the textural part, corresponding to its fine and detail features. TV-Gabor model designs a set of Hilbert spaces based on Gabor functions, which ensures us making the most of a-priori knowledge of the spatio-temporal texture information. Fig. 3 shows the decomposition results of synthetic input images. It clearly demonstrated that this model provides us with adaptive image decomposition. Through the comparison of the decomposition of both two original images under different illumination conditions, experiment results prove the assumption that motion perception with image textural part is not perturbed by shadow and shading reflection artifacts, which cover large image regions.

3.2 Area V1: Divisive Normalization and Motion Energy

The biological role of orientation-selective neurons is believed to be the extraction of contrast changes and local contour information. The input stimulus is also described as $A(p, t)$, characterized by its local contrast and defined as the equation 1. Because the RFs of V1 simple neurons are basically simulated by band-pass filters at present, we also define the response of simple neurons with a set of tilted three-dimensional Gabor filters. First, we define the RF spatio-temporal filters with a given spatial frequencies f_s and temporal frequencies f_t as

$$\mathcal{H}(p, \theta, f_s) = Be^{(-(x^2+y^2)/(2\sigma^2))}e^{j2\pi(f_s \cos(\theta)x + f_s \sin(\theta)y))} \tag{9}$$

$$\mathcal{P}(t, f_t) = e^{(-t/\tau)}e^{j2\pi(f_t t)} \tag{10}$$

where σ is the spatial scale parameter of the RF filter, and it represents the bandwidth of the spatial tuning function. τ defines the temporal scales parameter of the filter.

Here, we use the odd and even symmetric 3-D Gabor filters, $\mathcal{G}_o$ and $\mathcal{G}_e$, to replace the third derivative of Gaussian filters. And the tilted Gabor filters is defined through the real and imaginary filters, $\mathcal{H}_e, \mathcal{H}_o$ and $\mathcal{P}_e, \mathcal{P}_o$,

$$\mathcal{G}_o(p,t,\theta,v^c) = \mathcal{H}_o(p,\theta,f_s)\mathcal{P}_e(t,f_t) + \mathcal{H}_e(p,\theta,f_s)\mathcal{P}_o(t,f_t) \tag{11}$$

$$\mathcal{G}_e(p,t,\theta,v^c) = \mathcal{H}_e(p,\theta,f_s)\mathcal{P}_e(t,f_t) - \mathcal{H}_e(p,\theta,f_s)\mathcal{P}_o(t,f_t) \tag{12}$$

where the preferred velocity v^c is expressed as $v^c = f_t / f_s$.

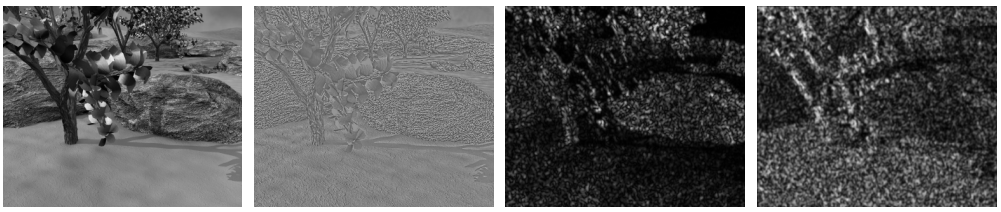
Moreover, to better preserve image texture edges, the responses of V1 complex neurons are obtained according to the sum of absolute value of the odd and even filters,

$$R_{o/e}(p,t,\theta,v^c) = (\mathcal{G}_{o/e}(\cdot,\cdot,\theta,v^c) \overset{(x,y,t)}{*} A)(p,t) \tag{13}$$

$$E(p,t,\theta,v^c) = \left|R_o(p,t,\theta,v^c)\right| + \left|R_e(p,t,\theta,v^c)\right| \tag{14}$$

Through a normalization to avoid the special case where the denominator is zero (regions without energy), the output of V1 layer is defined by

$$E^{\mathrm{V1}}(p,t,\theta,v^c) = \frac{E(p,t,\theta,v^c)}{\sum_{i=1}^{N} E(p,t,\theta,v^c) + \varepsilon} \tag{15}$$



(a) Grove2 (b) texture images (c) V1 simple cell (d) V1 complex cell

Fig. 4. Computation results of Grove2 in area V1

3.3 Area MT: V2-Modulated Pooling

According to experimental results, MT region belongs to the intermediate region in the visual cortex for processing motion information. MT neurons can process the local motion information input by effectively pooling the outputs of the V1 complex neurons. However, the method of pooling V1-afferents linearly does not, which leads to poor velocity estimation results by blurring edges boundaries. Thus, it is advantageous to make the V1 to MT pooling adaptive as a function of texture edges. To define a measure of local context (contrast and image texture), we take the role of responses from area V2 into account. First, using the complex filters $\mathcal{H}(p,\theta,f_s)$ in equation 9, we define

$$R(p) = (R_{\theta_1}(p), \cdots, R_{\theta_N}(p)) \; ; \; R_{\theta_i}(p) = |\mathcal{H} * I|(p) \tag{16}$$

$R_{\theta_i}(p)$ is maximal at a given edge orientation θ_i when crossing the prefered direction at the position $p(x, y)$.

Then assuming μ and σ^2 denotes the average and variance of *R*, respectively. Combined with the contrast measurement at a given position, and at the same time ensured that the contrast was not limited to a single direction, the local contrast is defined as

$$C(p) = f_{\xi}(\mu(R(p)))(1-\sigma^2(R(p))/\sigma_{max}^2) \tag{17}$$

where the term $f_{\xi}(\cdot)$ is the Rectified linear unit (ReLu), which is biological plausible and popular activation function, and $\xi \geq 0$ is the threshold. In regions with stronger texture this term is larger, whereas in regions with average value less than ξ this term equals to zero. The latter term denotes the strength of contrast in a single direction: the value of contrast in a single direction is larger, and the value of contrast in multiple directions is small.

The scale and anisotropy of spatial pooling function in MT motion integration stage are modulated by image local texture information, so as to achieve better edge preservation. MT pattern neurons response can be revised as

$$E^{\mathrm{MT}}(p,t,d,v^c) = F(\sum_{i=1}^{N} w_d(\theta_i)\mathbb{P}(E^{\mathrm{V1}}(p,t,\theta_i,v^c))) \tag{18}$$

where $F(x) = exp(x)$. MT weights $w_d(\theta)$ is a nonlinear function with central excitation and lateral inhibition, defined as $w_d(\theta) = \boldsymbol{cos}(d-\theta)$.

We define the spatial pooling strategy adapting itself to the image texture and motion discontinuities as follows,

$$\mathbb{P}(E^{\mathrm{V1}})(p,t,\theta_i,v^c) = \frac{1}{N}(\sum_{p'} g_{\alpha}(\|p-p'\|) f_{\varepsilon}(-\frac{\nabla R_{\theta_i}(p)}{\|\nabla R_{\theta_i}(p)\|+\varepsilon}(p'-p))E^{\mathrm{V1}}(p,t,\theta_i,v^c) \tag{19}$$

where $\nabla(\cdot)$ denotes a gradient function. As to term $g_i(\cdot)$, we use the L2-norm function. $f_{\xi}(\cdot)$ is an anisotropic weight enabling discontinuities be better preserved.

Through decoding the population responses of the MT neurons, the optical flow is finally estimated with a linear combination approach,

$$v_x(p,t) = \sum_{i=1}^{M} v_i^c E^{\mathrm{MT}}(p,t,0,v_i^c)\ ;\ v_y(p,t) = \sum_{i=1}^{M} v_i^c E^{\mathrm{MT}}(p,t,\pi/2,v_i^c) \tag{20}$$

4. Results and Discussion

During verification of our proposed algorithm, we conducted experiments on two data sets: the standard Middlebury benchmark [9] and the real scenes. The first one contains several challenges that optical flow estimation needs to solve, such as fast-moving objects, sharp edges, and target occlusions. In order to better visualize the experimental results, this paper calculates the results with the color-coding scheme proposed in [5]. The other one is used to evaluates the results of our algorithm under difficult conditions, such as illumination-changing and large displacement.

4.1 Performance evaluation on Middlebury dataset

In this paper, average angular error (AAE) and endpoint error (EPE) are estimated to evaluate the performance of models (See **Table 1**).

Table 1. A comparison of error measurements between following models

		Grove2	Grove3	RubberWhale	Urban3	Yosemite
HS	AAE±STD	17.61±31.86	22.24±34.65	21.62±41.70	32.39±64.87	11.74±12.93
	EPE±STD	1.52±2.61	6.48±7.04	1.92±2.39	7.47±8.53	1.36±1.28
FFV1MT	AAE±STD	4.28±10.25	9.72±19.34	10.20±17.67	15.11±35.28	3.41±5.44
	EPE±STD	0.29±0.62	1.13±1.85	0.34±0.54	1.88±3.27	0.16±0.18
Ad-V1MTMP	AAE±STD	3.85±9.72	9.87±18.33	8.78±14.74	12.54±32.43	3.15±3.10
	EPE±STD	0.23±0.46	1.14±1.56	0.37±0.48	1.33±2.21	0.12±0.13

The results of **Table 1** indicate that our Ad-V1MTMP model has a state-of-the-art performance of motion perception. Take Yosemite sequence (without clouds) for instance, we have AAE=3.15±3.10 for Ad-V1MTMP. However, compared with some existing biological inspired models such as the HS model (AAE=11.74 [19]) and the FFV1MT model (AAE=3.41±5.44 [15]), our model shows a great improvement.

Fig. 5 shows results of different approaches obtained on training dataset. The relative performance of our model compared with other methods can be shown subjectively by observing the real results of δAAE (δAAE=AAE_{FFV1MT}-$AAE_{Ad\text{-}V1MTMP}$), which are presented in the last column. Because we incorporate function principles in human visual system, namely contrast adaptation based on V2-modulated pooling, image structure based on TV-Gabor model and ambiguity based on lateral interaction, the δAAE maps highlight a better performance of our Ad-V1MTMP method (**Fig. 5(e)**) than FFV1MT (**Fig. 5(d)**) on image edges and details as expected. The baseline FFV1MT model involving a hierarchical structure from area V1 to area MT is initially proposed to process the homogeneously textured regions, hence smooth effects on edges and fine details still exist compared with the HS method (see Grove2 and Grove3 in **Fig. 5(c)** and **Fig. 5(d)**). However, the proposed method partially solves this issue by considering inputs from area V2, the improvements are prominent, as shown in RubberWhale and Urban3.

Fig. 6 shows the experiment performances of different approaches on several test sequences. Lower errors and better estimation results at the occlusions and sharp edges, which shows a better motion boundary preservation of our model (see, e.g., Urban sequences). In conditions such as high-speed motion and large occlusions, on which our method has good performance, because the multi-scale approach allows us to estimate motion on different scales with coarse-to-fine refinement and pyramidal decomposition, (see, e.g., Urban sequence). In the presence of sharp edges, our method could recover velocity vectors by taking the lateral connections between neurons and scale space issues into account (see, e.g., Wooden sequences).

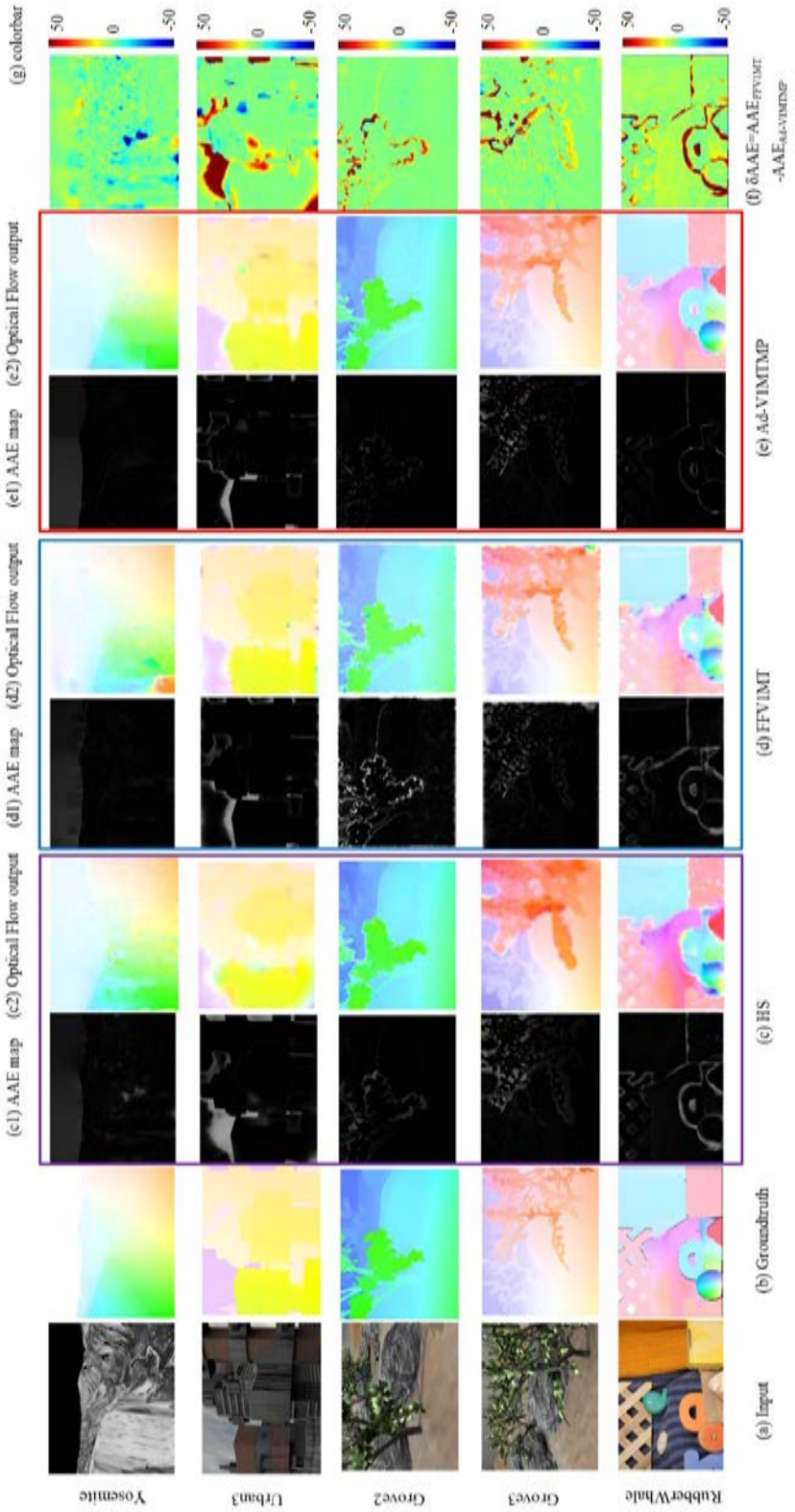


Fig. 5. Evaluation results on Middlebury training set

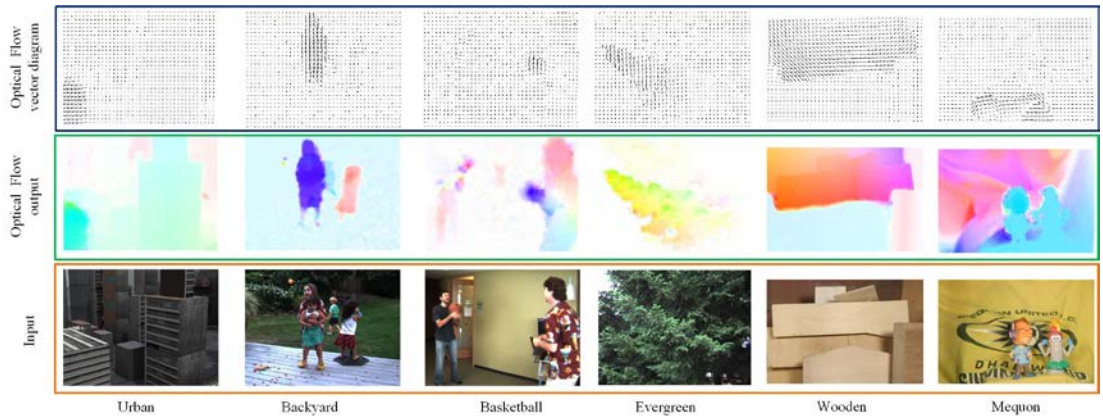


Fig. 6. Sample results on Middlebury test set

4.2 Performance evaluation on real scenes

Obtaining accurate optical flow is very important for sensing scene information and interpreting sequence images. Here, what interests us is only motion perception approach on real-world sequences under illumination-changing conditions and high-speed motion.

Fig. 7 shows the results of optical flow estimation with a car approaching sequences, where the upper one is the computation result of FFV1MT approach, and the other one is the proposed approach result. Because of high-speed motion in the input sequences and illumination changes (left images in **Fig. 7**), the computation of FFV1MT approach fails (upper-right image in **Fig. 7**). Due to defining adaptive frequency and directional image decomposition, artifacts are not visible in the structure-texture decomposed images (middle images in **Fig. 7**). The lower-right image in **Fig. 7** shows a better performance on the approaching car with flow discontinuities at image edges by using adaptive spatial pooling strategy. This demonstrates that our proposed approach is very robust to illumination changes.

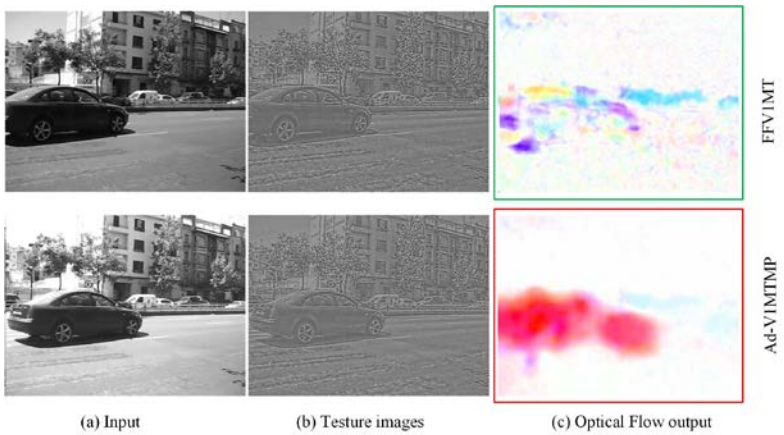


Fig. 7. The results of optical flow estimation for car approaching sequences.

5. Conclusion

Based on the analysis of primate visual system and existing biological approaches, we have proposed a new FFV1MT algorithm, which can fill the gap between studies in biological and computer vision for motion perception. In this paper, we analyze and extend current neural models by adding feedback connectivity and investigating the role of area V2 in ambiguity regions. We additionally propose several strategies to decode the velocity of visual motion and adapt computationally efficient mechanisms to cope with problems encountered in real sequences, such as mapping image sequences into illumination-independent sequences through the TV-Gabor model. We experimentally evaluated the proposed algorithm in two datasets: Firstly, we quantify the improvement in performance on Middlebury benchmark sequences. Secondly, we demonstrate that our new model has more accurate and robust performance on real-world sequences illumination-changing conditions and high-speed motion. In future work, we intend to focus on addressing other mechanisms for motion perception, for example recurrent interactions among MT neurons, or feedback connections from higher-order areas to area MT.

References

[1] Gulde T, Kärcher S, Curio C. "Vision-Based SLAM Navigation for Vibro-Tactile Human-Centered Indoor Guidance," in *Proc. of 14th European Conference on Computer Vision*, pp. 343-359, October 8-16, 2016. Article (CrossRef Link).

[2] Chessa M, Noceti N, Odone F, et al. "An integrated artificial vision framework for assisting visually impaired users," Computer *Vision and Image Understanding*, vol. 149, pp. 209-228, 2016. Article (CrossRef Link).

[3] McLaughlin N, Del Rincon J M, Miller P. "Enhancing linear programming with motion modeling for multi-target tracking," *IEEE Winter Conference on Applications of Computer Vision*, pp. 71-77, 2015. Article (CrossRef Link).

[4] B. Horn and B. Schunck. "Determining Optical Flow," *Artificial Intelligence*, vol. 17, pp. 185-203, 1981. Article (CrossRef Link).

[5] D. Fortun, P. Bouthemy, C. Kervrann, "Optical flow modeling and computation: a survey," Computer *Vision and Image Understanding*, vol. 134, pp. 1-21, 2015. Article (CrossRef Link).

[6] D. J. Butler, J. Wulff, G. B. Stanley, et al. "A naturalistic open source movie for optical flow evaluation," in *Proceedings of the 12th European Conference on Computer Vision-Volume Part VI*, pp. 611-625, 2012. Article (CrossRef Link).

[7] J. F. Aujol, G. Gilboa, T. Chan, et al. "Structure-texture image decomposition-modeling, algorithms, and parameter selection," *International Journal of Computer Vision*, , vol. 67, no. 1, pp. 111-136, 2006. Article (CrossRef Link).

[8] A. Wedel, T. Pock, C. Zach, et al. "An improved algorithm for TV-L1 optical flow," in *Proceedings of the Dagstuhl Seminar on Statistical and Geometrical Approaches to Visual Motion Analysis*, pp. 23-45, 2009. Article (CrossRef Link).

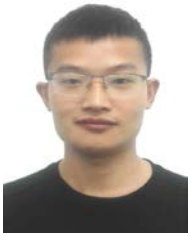
[9] S. Baker, D. Scharstein, J. Lewis, et al. "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol 92, no. 1, pp. 1-31, 2007. Article (CrossRef Link).

[10] D. Heeger, "Optical flow using spatiotemporal filters," *International Journal of Comput Vision*, vol 1, no. 4, pp. 279-302, 1988. Article (CrossRef Link).

[11] N.C. Rust, V. Mante, E.P. Simoncelli, J.A. Movshon, M.T. "How MT cells analyze the motion of visual patterns," *Nature Neuroscience*, vol 9, no. 11, pp. 1421-1431, 2006. Article (CrossRef Link).

[12] 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*, vol. 31, no. 41, pp. 14551-14564, 2011. Article (CrossRef Link).

[13] Born RT, and Bradley DC. "Structure and function of visual area MT," *Annual Review of Neuroscience*, vol. 28, no. 1, pp. 157-189, 2005. Article (CrossRef Link).
[14] Jianbo Xiao, Yu-Qiong Niu, Steven Wiesner, et al. "Normalization of neuronal responses in cortical area MT across signal strengths and motion directions," *Journal of Neurophysiology*, vol. 112, no. 6, pp. 1291-1306, 2014. Article (CrossRef Link).
[15] F. Solari, M. Chessa, N. V. K. Medathati, et al. "What can we expect from a V1-MT feedforward architecture for optical flow estimation?" *Signal Processing: Image Communication*, vol. 39, pp. 342-354, 2015. Article (CrossRef Link).
[16] B. Krekelberg, and van Wezel RJ. "Neural mechanisms of speed perception: transparent motion," *Journal of Neurophysiology*, vol. 110, no.9, pp. 2007-2018, 2013. Article (CrossRef Link).
[17] E. Simoncelli, D. Heeger. "A model of neuronal responses in visual area MT," *Vision Research*, vol. 38, no. 5, pp. 743-761, 1998. Article (CrossRef Link).
[18] H. Nagel, W. Enkelmann. "An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 565-593, 1986. Article (CrossRef Link)
[19] N. V. Kartheek Medathati, HeikoNeumann, G. Masson, et al. "Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision," *Computer Vision and Image Understanding*, vol. 150, pp. 1-30, 2016. Article (CrossRef Link).



Shuai Li received the M.S. degree in information and communication engineering from the Aeronautics and Astronautics Engineering College of Air Force Engineering University, China, in the year 2014. He is currently as a PHD student at the Aeronautics and Astronautics Engineering College of Air Force Engineering University, China. His research interests include computer vision and intelligent computation and signal processing.

Xiaoguang Fan is a professor and Ph.D. supervisor at the Aeronautics and Astronautics Engineering College of Air Force Engineering University. His main research interests are in the areas of airborne computer technology and intelligent information processing.

Yuelei Xu is a professor and M. S. supervisor at the Aeronautics and Astronautics Engineering College of Air Force Engineering University. His main research interests are in the areas of computer vision and intelligent computation and signal processing.

Jinke Huang received the M.S. degree in information and communication engineering from the Aeronautics and Astronautics Engineering College of Air Force Engineering University in the year 2013. He is currently as a PHD student at the Aeronautics and Astronautics Engineering College of Air Force Engineering University, China. His research interests include intelligent computation and signal processing.