Optic Flow Integration at Multiple Spatial Frequencies - Neural Mechanism and Algorithm

Cornelia Beck, Pierre Bayerl, and Heiko Neumann

Dept. of Neural Information Processing, University of Ulm, Germany
{cornelia.beck,pierre.bayerl,heiko.neumann}@uni-ulm.de

Abstract. In this work we present an iterative multi-scale algorithm for motion estimation that follows mechanisms of motion processing in the human brain. Keeping the properties of a previously presented neural model of cortical motion integration we created a computationally fast algorithmic implementation of the model. The novel contribution is the extension of the algorithm to operate on multiple scales without the disadvantages of typical coarse-to-fine approaches. Compared to the implementation with one scale our multi-scale approach generates faster dense flow fields and reduces wrong motion estimations. In contrast to other approaches, motion estimation on the fine scale is biased by the coarser scales without being corrupted if erroneous motion cues are generated on coarser scales, e.g., when small objects are overlooked. This multi-scale approach is also consistent with biological observations: The function of fast feedforward projections to higher cortical areas with large receptive fields and feedback connections to earlier areas as suggested by our approach might contribute to human motion estimation.

1 Introduction

The detection of motion in our environment is part of our everyday life. Humans are capable to estimate motion very precisely and very fast. While ongoing research tries to resolve the detailed processing mechanism in the human brain, it is agreed on that areas V1, MT and MST play important roles in the motion processing pathway [1]. In contrast to the simple task of motion detection for a human, implementing motion detection in technical applications remains a difficult problem (see [2] for an overview of existing technical approaches). For instance, a moving robot provided with cameras should be capable to detect moving objects in its environment in real-time to avoid collision. Therefore, the motion estimation has to be fast, of high quality, and the motion estimates should be available for every position of the surrounding.
We developed a model for motion estimation that is based on the mechanisms observed in human motion processing (see Sect. 2) [3]. This neural model simulates areas V1 and MT, including feedforward as well as feedback connections [4]. To shorten the computing time, we reimplemented the model in an algorithmic version [5] and improved its results by adding multiple processing scales as described in Sect. 3. Furthermore, we will explain the biological motivation of this new approach.

2 Neural model

Approaches in computer vision for optic flow estimation use, e.g., regularization or Bayesian models to achieve globally consistent flow fields [6, 7]. Another possibility to approach this problem is to build a model corresponding to the neural processing in our visual system. We previously presented such a model for optic flow detection based on the first stages of motion processing in the brain [3]. Therein, areas V1 and MT of the dorsal pathway are simulated. In model area V1 a first detection of optic flow is realized, model area MT estimates the optic flow of larger regions and is thus, e.g., capable to solve the aperture problem [8]. The principle processing components of this neural model are feedforward, feedback and lateral connections.

Both modules V1 and MT comprise at each spatial location a certain number of neurons tuned to different velocities. For efficient computation, we need to discretize and limit the velocity space. Each neuron has a certain activity rate describing the likelihood of its represented velocity. The input net_{IN} for module V1 of the model represents the similarity of the image structure of two images at different time steps that is calculated using modified Reichardt detectors [9]. It is modulated with the feedback net_{FB} from module MT (1). This multiplicative feedback only enhances activated neurons, but will not create new activities. In the process of feedforward integration, signal $v^{(1)}$ is integrated with Gaussian isotropic filters in both the velocity and the spatial domain (2), "*" denotes the convolution operation. Finally, lateral shunting is effected at each location to strengthen the activity of unambiguous motion signals (3). The equations are identical for module MT, but the integration process (2) uses a larger spatial neighborhood and there is no feedback to MT.

$$\delta_t v^{(1)} = -v^{(1)} + net_{\mathrm{IN}} \cdot \left(1 + C \cdot net_{\mathrm{FB}}\right) \quad . \tag{1}$$

$$\delta_t v^{(2)} = \left(v^{(2)}\right)^2 + v^{(1)} * G_{(\mathrm{space})} * G_{(\mathrm{velocity})} \quad \cdot \tag{2}$$

$$\delta_t v^{(3)} = -0.01 \cdot v^{(3)} + v^{(2)} - \left(1/\left(2n\right) + v^{(3)}\right) \cdot \sum_{\Delta x} v^{(2)} \quad . \tag{3}$$

3 Algorithmic multi-scale model

A technical problem of the neural algorithm is the fact that the high number of neurons leads to large memory costs and to long simulation times. Thus, starting from the neural model for optic flow estimation, we have previously developed an efficient algorithmic version that behaves similar to the neural model [5].

A limitation of the neural as well as of the algorithmic model is that the optic flow is only evaluated on a single scale, i.e., the similarities in the input images are calculated on a single spatial resolution of the input image. This leads to problems when spatially low-frequency areas are moving in an image sequence. On a fine scale, this "coarse" motion may not be detected, as there exist many ambiguities when calculating the similarity measure for V1 between pairs of

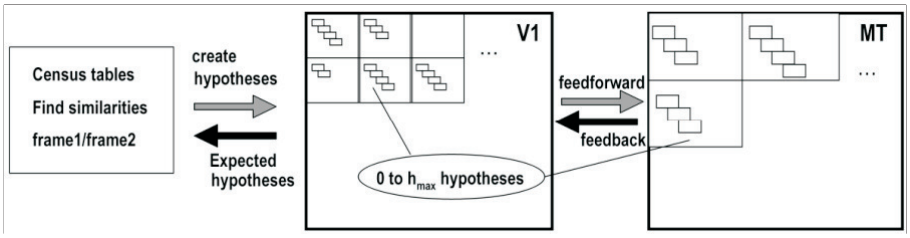

Fig. 1. Iterative model of optic flow estimation: First, the similarity of two input frames is calculated using the Census values. The hypotheses of V1 influence the creation of the initial optic flow hypotheses, if the number of Census values is bigger than $h_{\max}$. Locations in V1 are spatially integrated and subsampled for the optic flow estimation of MT. The *feedback* of MT modulates the likelihood of hypotheses in V1.

frames (see example in Fig. 2). To solve this problem we extend the algorithmic model with coarser scales of motion estimation. Such a "multi-scale processing" scheme was proposed, e.g., by Simoncelli [10]. In general, these algorithms are realized with "image pyramids" where motion estimation of coarser scales influences the estimation of finer scales as "initial guess" [10, 11]. The processing of the input image in resolutions of different spatial frequencies provides more information for the motion estimation.
Considering the biological basis of our model, coarser processing scales can be integrated in a plausible way. We believe that fast motion processing on a coarser scale can modulate the feedforward projection within the fine scale of motion estimation from V1 to MT [1]. This can be accomplished via feedback of the processing of a coarser representation of the input image from area V1 to MST. Our approach is in line with the "facilitation" of visual object recognition in human brain via expectations in the prefrontal cortex which act as "initial guess" as proposed by Bar [12]. In the following subsection we will explain the algorithmic version of the neural model with one processing scale. Thereafter, the extension of this model from one scale to two and more scales is presented, that combines fast optic flow estimation with improved qualitative performance due to the integration of multiple scales.

3.1 Algorithmic single-scale model

The algorithmic model consists of two different modules V1 and MT like the neural model (see Fig. 1). For the extraction of motion correspondences between two frames of an image sequence the algorithm uses a similarity measure of the class of rank-order approaches: The "Census transform" [13] provides an implicit local description of the world. Accordingly, possible motion correspondences between two frames of an image sequence can be extracted at locations with the same Census values in both frames. Initially, we extract motion correspondences (hypotheses) in V1 for identical Census values which show less than $h_{\max}$ possible correspondences in the second frame. Each hypothesis includes a likelihood

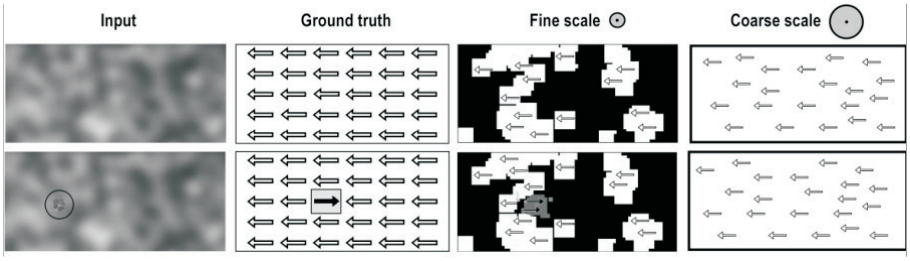

Fig. 2. Optic flow detection with the algorithmic model with one scale. First row: Results for an input image containing a spatially low frequency structure. Second row: Results for an input image with a small moving object in front of the moving background. The object is indicated by the black circle (not part of the input image). The second column represents the ground truth of the input. In the third and the fourth column the resulting flow fields of MT are shown after the first iteration using a fine and alternatively a coarse scale. Black positions indicate positions without motion estimation, white positions represent movement to the left, gray positions to the right. One iteration of the fine scale algorithm takes about 0.6 seconds, using the coarse scale, one iteration takes less than 0.1 seconds (Pentium IV, 3GHz)

initialized to 1.0 which indicates the likelihood of a particular velocity at a certain position. The restriction on $h_{\max}$ identical Census values results in at most $h_{\max}$ correspondences selected at each pixel (we use $h_{\max} = 4$). This means, in comparison to the neural model, that we only simulate the $h_{\max}$ neurons representing the velocities with the highest likelihood at each position. The extracted motion hypotheses include a lot of wrong hypotheses in addition to the correct ones caused, e.g., by the motion aperture problem. To improve the estimations, the hypotheses of the first module are spatially integrated to generate the hypotheses for the second module MT, which represents motion at a coarser spatial resolution. Hypotheses that are supported by adjacent positions have an advantage during the subsampling process in comparison to spatially isolated motion hypotheses. Again, only the $h_{\max}$ hypotheses with the strongest likelihood are kept at each position in the second module. These hypotheses are iteratively used as feedback for motion estimation in the first module. The recurrent signal modulates the likelihood of predicted hypotheses in the first module as in the neural model. In addition, the feedback influences the input to V1: Hypotheses are also created at positions with ambiguous motion hypotheses (Census value appears more than $h_{\max}$ times in the second image) if the velocity at this position corresponds to one of the velocities of the feedback (i.e., the velocity is expected). This procedure is necessary in the algorithmic model to compensate that only $h_{\max}$ hypotheses are represented at a position in contrast to the neural model where each possible velocity is computed at a position.

3.2 Integration of multiple scales

In the first row of Fig. 2 the results of the single-scale algorithm are presented for an input image (320x144 pixel) consisting of a spatially low-frequency texture like clouds that are moving to the left. Only few hypotheses are generated in V1 and MT when using the algorithm on the fine scale. In comparison to this, using the same algorithm but a coarse version of the input images, all the positions of MT show (the correct) motion hypotheses. This is due to the fact that the movement in the coarse scale is less ambiguous.
However, single-scale models operating only on a coarse scale have another disadvantage: Small moving objects with minor luminance contrast are overlooked by the model, as they are effaced during the subsampling and the motion integration process. An example is given in the second row of Fig. 2, where a small rectangle (17x17 pixel) is moving to the right in front of a background moving to the left. Whereas the model using the coarse scale completely overlooks the objects after the first iteration, the object is detected in the fine scale.
To combine the advantages of the fine and the coarse scale we need to integrate the feedback of at least one coarser scale (e.g., V1-MST) to our single-scale model. In doing so, the estimations of the fine scale need to be protected. For this reason, the coarser scales in this algorithm do only contribute to the estimation of motion in the next finer scale where the ambiguity is high. The calculation of motion estimation in a model with two scales will be calculated by the following way: In a first step, motion hypotheses of the coarse scale are created. Thereafter, the motion correspondences in V1 of the fine scale are calculated. Just if the motion at a position is ambiguous (i.e., more than $h_{\max} = 4$ hypotheses), the feedback of the coarse scale is used for the selection of possible motion hypotheses. Thereby it is combined with the feedback of MT of the fine scale (i.e., both modules contribute to the creation of new hypotheses at ambiguous positions). Adding more scales can be realized in an analog way. A technical detail to keep fast processing in the algorithm that has to be considered is that the resolution of the feedback from the coarse scale has to be in the same resolution as the motion hypotheses of the module receiving the feedback (fine scale). This can easily be realized if the motion hypotheses of the coarse scale are created of frames with greater temporal distance Δt (subsampling rate (fine scale to coarse scale) corresponds to Δt). In the neural model this would not be necessary.

4 Results

The following results are obtained with a multi-scale model of the presented algorithm containing two scales of motion detection. In the coarse scale the input images processed are four times smaller than the original ones. For this subsampling process, the image is blurred with a Gaussian filter ($sigma = 4$). The integration of motion hypotheses of the second module of the coarse scale (module MST) reduces its size to a fourth of its subsampled input image from V1. The images shown here always represent the hypotheses of MT (the second

module of the fine scale), as these are the final results of the motion estimation. First, we tested the developed multi-scale model with spatially low-frequency image sequences where the motion is not detected in the fine scale (see Fig. 2/first row). The detected motion hypotheses of the multi-scale model are presented in Fig. 3/first row. In contrast to the results of the fine scale model nearly all positions represent a motion hypothesis. The direction of the hypotheses is also correct. The result is close to the optimal result of the coarse scale model. Second, the image sequence with a small moving object not detected within the coarse scale model was used as input to the multi-scale model (see Fig. 2/second row). Whereas the coarse scale model ignores the movement of the small object opposite to the background, the multi-scale model clearly indicates a rectangle in the center of the image moving to the right as depicted in Fig. 3/second row. The proportion of detected movement at the positions of the object and the background in the input image is shown in Fig. 4. Only the multi-scale model is close to 100 percent detection for both the background and the object after only one iteration.

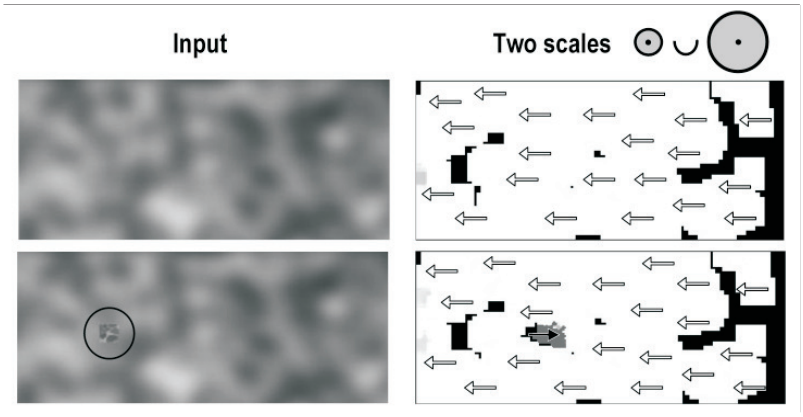

Fig. 3. Motion hypotheses of MT of the two scale model. The input images (first column) are the same as in Fig. 2. In the second column the motion hypotheses of module MT are shown. Black positions are positions where no movement was detected, white positions indicate movement to the left, gray positions represent movement to the right. One iteration of the multi-scale algorithm takes about 0.7 seconds

We further compared our multi-scale approach to the fine scale model using the Yosemite Sequence (316x256 pixel, version with clouds) as input images [2]. An exemplary image of this sequence is presented in Fig. 5(a). For this artificial sequence the ground truth of the optic flow is provided which enables us to evaluate the quality of extracted motion hypotheses (see Fig. 5(b)). Gray positions represent movement to the right, white positions movement to the left, at black positions no motion hypothesis is created. The results of module MT of the fine scale model are presented in Fig. 5(d)-(f) for three iterations of the

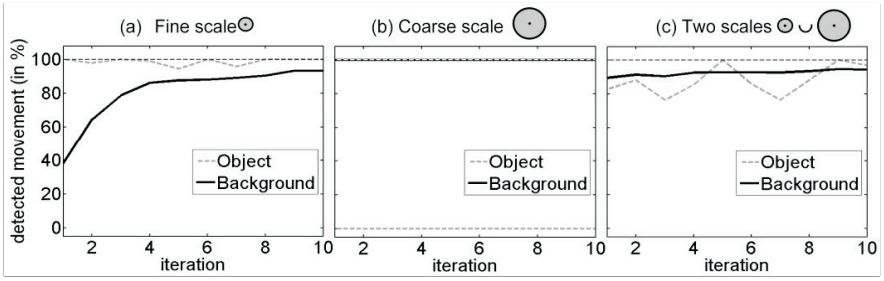

Fig. 4. Detection of background and object for input image sequence with moving object. The percent of the positions of the background and the object with the correct detected direction are shown. The dashed line at 100 percent represents the reference. **(a)** The fine scale model detects the movement of the complete object after the first iteration, but it needs 4 iterations to cover the main parts of the background. **(b)** shows the results for the coarse scale model that detects the background movement immediately, but completely misses the object, independently of the number of iterations. **(c)** In the two scale model about 85 percent of the background movement is detected after one iteration as well as the main parts of the object.

algorithm. After the first iteration the positions in MT representing at least one motion hypothesis add up to 85 percent. Similar to the texture in input image of Fig. 2 the spatially low-frequency texture of the sky causes problems to the fine scale model. Thus, mainly positions in the sky do not have motion hypotheses (black positions). At the same time, in comparison to the smooth original flow field (see Fig. 5(b)), there are some wrong hypotheses especially in the lower left area of module MT caused by the aperture problem when using only a small scale.

The motion hypotheses of MT for three iterations of the multi-scale algorithm are shown in Fig. 5(g)-(i). Just after one iteration, 98 percent of the MT positions represent motion hypotheses. The flow field contains only few positions which represent motion hypotheses that differ from the smooth flow field of the image. The aperture problem is solved due to the coarse scale added. Thus, only one iteration of the multi-scale algorithm provides a flow field comparable to the ground truth for nearly every position. A comparison of the quality of the motion hypotheses of the two algorithms is presented in Fig. 5(c). The multi-scale model achieves better results in the median angular error of the motion hypotheses than the fine scale model in MT. The better results of the multi-scale model after the first iteration is even more significant, if we take into account that 98 percent of all positions in MT of the multi-scale model cause a lower error than the 85 percent of positions with hypotheses of the fine scale model.

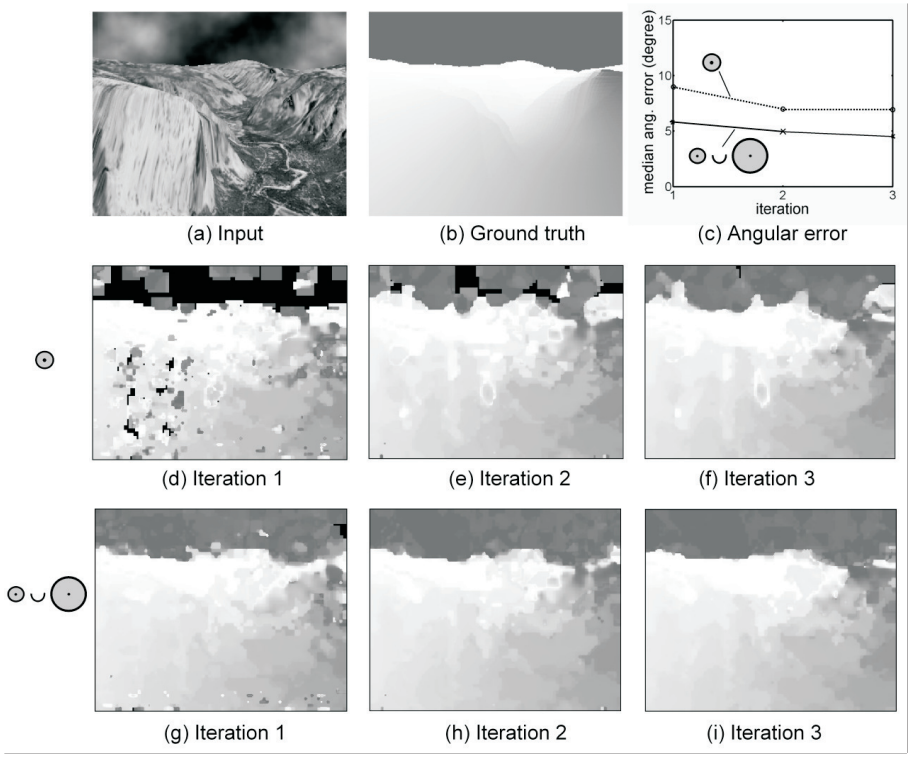

Fig. 5. (a) One input image of the Yosemite Sequence ($t = 3$), **(b)** ground truth for the optic flow of frame 3 and 4. For the motion in the sky we assume horizontal motion to the right as proposed in [2]. **(c)** shows the median angular error in degree of module MT. The motion estimations in MT for the Yosemite sequence using the fine scale algorithm are shown in the second row. **(d)** After the first iteration a lot of positions in the sky do not contain a motion hypothesis (black positions). Furthermore, there are some wrong motion hypotheses in the lower left. **(e)** After the second iteration the positions representing motion hypotheses in the sky is higher, the flow field contains less errors. **(f)** After three iterations more than 98 percent of the positions represent motion hypotheses, the flow field is similar to the ground truth. In the third row the results for the multi-scale algorithm are presented. **(g)** After the first iteration almost every position holds a motion hypothesis, even in the coarse structure of the sky the movement to the right is correctly indicated. The flow field contains only few errors. **(h)(i)** the positions disturbing the smoothness of the flow field are corrected in the second and third iteration

.

5 Discussion and Conclusion

We presented a multi-scale algorithm for optic flow estimation based on a neural model. In contrast to other multi-scale approaches [10], this algorithm does not propagate the error of coarser scales to the fine scale. This is ensured by the way coarse scales influences the motion estimation in the fine scale. Only if the estimation in the fine scale is highly ambiguous and if the motion estimation of the coarse scale is compatible with the motion correspondences in the fine scale, then the additional information of the large scale will be used. This avoids that small objects are overlooked on the fine scale [10]. Moreover, the extraction of the motion hypotheses is implemented in a way that the search space for corresponding positions in each scale comprises the entire image. This is not realized in many other multi-scale approaches [2].

In the examples presented for the multi-scale model we restricted the model to two scales. This is due to the fact that for the employed input sequences a combination of one coarse and one fine scale was sufficient to get estimations for all positions after one iteration. Thus, adding coarser scales would not further improve the estimations. Nevertheless, the algorithm can be extended to more scales in a straightforward way. Concerning the computing time of the algorithm, more scales do not increase the time for one iteration considerably. This follows from the faster processing of coarser scales where estimations for less positions have to be calculated. Furthermore, the time for the calculation could be further reduced by limiting the positions of the images to be processed. This could be done by a preclassification (e.g., corners [13]) or a limitation to positions with a certain minimum contrast.

The biological motivation of the multi-scale model is based on the observations that V1, MT and MST are main components of the motion processing pathway that includes feedforward and feedback connections [4]. While in V1 motion is detected only in a very small neighborhood, its projections to MT lead to an integration of the detected motion within a larger region [14]. Motion estimation in an even coarser spatial resolution is accomplished in MST, its neurons respond to planar, circular, and radial motion as well as to complex patterns of motion [15]. Low latencies of the first responses in V1, MT *and* MST [16] indicate a possible computation of a quick initial guess in higher areas, such as MST, which may in turn influence information processing in earlier areas such as V1 or MT via feedback connections. Because area MST receives its primary afferent inputs from area MT, such a computation may probably be realized in a feedforward manner via MT, but also direct connections from V1 to MST have been observed [1]. The prediction through a fast and spatially coarse "initial guess" is compatible to theories predicting that the context (here a large spatial context of motion information) may influence initial feature extraction [17, 12].

In conclusion, we presented a biologically motivated algorithm for optic flow integration on multiple processing scales that generates fast and reliable motion estimations.

6 Acknowledgements

This research has been supported in part by a grant from the European Union (EU FP6 IST Cognitive Systems Integrated project: Neural Decision-Making in Motion; project number 027198).

References

1. L.G. Ungerleider, J.V. Haxby, 'What' and 'where' in the human brain, Current Opinion in Neurobiology, 4, pp. 157-165, 1994
2. S.S. Beauchemin, J.L. Barron, The Computation of Optical Flow, ACM Computing Surveys, Vol. 27, no. 3, pp. 433-467, 1995
3. P. Bayerl, H. Neumann, Disambiguating Visual Motion through Contextual Feedback Modulation, Neural Computation, 16(10), pp. 2041-2066, 2004
4. J.M. Hupé, A.C. James, P. Girard, S.G. Lomber, B.R. Payne, J. Bullier, Feedback Connections Act on the Early Part of the Responses in Monkey Visual Cortex, J. Neurophys., 85, pp. 134-145, 2001
5. P. Bayerl, H. Neumann, Towards real-time: A neuromorphic algorithm for recurrent motion segmentation, Ninth International Conference on Cognitive and Neural Systems (ICCNS '05), Boston, USA, 2005
6. B.K.P. Horn, B.G. Schunk, Determining optical flow, Artificial Intelligence, 17, pp. 185-203, 1981
7. Y. Weiss, D.J. Fleet, Velocity likelihoods in biological and machine vision, Probabilistic models of the brain: Perception and neural function, pp. 81-100, Cambridge, MA:MIT Press, 2001
8. C.C. Pack, R.T. Born, Temporal dynamics of a neural solution to the aperture problem in cortical area MT, Nature, 409, pp. 1040-1042, 2001
9. E. Adelson, J. Bergen, Spatiotemporal energy models for the perception of motion, Optical Society of America A, 2/2, pp. 284-299, 1985
10. E. Simoncelli, Course-to-fine Estimation of Visual Motion, IEEE Eighth Workshop on Image and Multidimensional Signal Processing, Cannes France, Sept. 1993
11. P.J. Burt, E.H. Adelson, The Laplacian Pyramid as a Compact Image Code, IEEE Transactions On Communications, Vol. 31, No. 4, 1983
12. M. Bar, A Cortical Mechanism for Triggering Top-Down Facilitation in Visual Object Recognition, Journal of Cognitive Neuroscience, 15:4, pp. 600-609, 2003
13. F. Stein, Efficient Computation of Optical Flow Using the Census Transform, DAGM-Symposium 2004, pp. 79-86, 2004
14. T.D. Albright, Direction and orientation selectivity of neurons in visual area MT of the macaque, J. Neurophys., 52, pp. 1106-1130, 1984
15. C.J. Duffy, R.H. Wurtz, Sensitivy of MST Neurons to Optic Flow Stimuli. I. A Continuum of Response Selectivity to Large-Field Stimuli, J. Neurophys., 65, pp. 1329-1345, 1991
16. V.A.F. Lamme, P.R. Roelfsema, The distinct modes of vision offered by feedforward and recurrent processing, Trends Neurosci., 23, pp. 571-579, 2000
17. A. Torralba, A. Oliva, Statistics of natural image categories, Network: Comput. Neural Syst., 14, pp. 391-412, 2003