Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Biological Motion Estimation with Spatiotemporal Integration

Julian Eggert, Volker Willert, Jens Schmüdderich

2007

Preprint:

This is an accepted article published in HRI Global Workshop. The final authenticated version is available online at: https://doi.org/[DOI not available]

Biological Motion Estimation with Spatiotemporal Integration

J. Eggert, V. Willert and J. Schmüdderich

Abstract— **We present a framework and study a model for biologically inspired motion estimation. It comprises several "areas" (corresponding roughly to V1-MT) which locally analyze motion input at different levels of granularity, coupled to functionally distinct areas which concentrate on the extraction of more global motion patterns (corresponding e.g. to MST). The areas are coupled reciprocally with each other, with "higher" areas providing prior knowledge or constraints that bias the processing at lower areas. In addition, within each area, spatial and temporal integration, mediated by lateral connections and internal dynamics, is used to resolve the inherent ambiguity of motion signals and to take advantage of coherence in naturally occurring stimuli. Our model leads to results that are quantitatively comparable to state-of-the-art techniques. We show how several psychophysical as well as physiological effects observed during motion processing can be explained by our model. We also show implementation results of our model in the application contexts of motion segmentation and egomotion compensation for ASIMO.**

I. Introduction

Despite many years of progress, motion processing continues to puzzle the mind of researchers involved in understanding vision. Basic aspects such as local motion filtering have been widely studied (see e.g. [10], [11]). What is most striking about motion processing is its observed temporal dynamics. Since we are dealing with motion, this seems like a trivial statement, but in fact, it is not. What we address here is the capability of the biological motion extraction system of primates to accumulate motion information over time to make sense of an otherwise highly ambiguous input. As an example, single neurons exhibit responses with a time-course that could be interpreted as: "Measure motions locally first, then look if they are consistent and can be combined with measurements of other neurons".

The selective integration of motion signals has to occur by means of reciprocal influences between neurons and areas. A single neuron receiving sensory feedforward input is tied up by its aperture problem - i.e., it only sees a very restricted part of the visual world and therefore is only able to give rough guesses about occurring motion direction and speed. Subsequent areas may collect the signals from several neurons of lower levels and may have less restrictions on the region of the visual world they analyse. Nevertheless, with pure feedforward integration they would simply *average*, loosing access to subtle details hidden in the ambiguity of the responses from lower level neurons. One way to alleviate this problem is to allow reciprocal information transfer in the sense that lower level neurons are able to interprete their measurements in the context of the knowledge of the higher level neurons. Similarly, in a single area, a neuron is influenced by the activities of its neighbours. Therefore, lower and higher areas of motion processing as well as the neurons within one area have to work in concert to make sense of a visual scene. In addition, other areas unrelated to motion processing (e.g., neurons providing information about spatial structures, borders, and border ownerships) could influence motion processing by the same means, providing biases that allow a better interpretation of the motion stimulus. This is what *mid-level motion processing* is largely about.

What are necessary ingredients for motion systems with reciprocal biasing and context integration? First of all, it would be necessary to represent separately the different types of information. The biasing (via top-down or lateral connections) is a sort of *prediction* that incorporates the network context. In addition, we need a separate observation resp. *measurement* that is provided mainly by feedforward connections, and finally the synthesizing *estimation* that improves the mesaurement by taking the prediction into account. This of course resembles very much the Bayesian foundations of prior, likelihood and posterior, and indeed, we will formulate part of the mathematics of our model using probabilistic arguments. The foundations for parts of this have been laid out already some years ago by Burgi, Yuille and Grzywacz [2], but astonishingly did not find entry into the mainstream models of biologically inspired visual motion processing [1]. In this paper, we will show that indeed we can build a sort of "canonical" model of mid-level visual motion processing, including interacting low and high level motion extracting areas as well as areas for the analysis of large motion patterns, based on a mixture of probabilistic and connectionist ideas. Nearly all assumptions needed to do this actually have their origin in the spatiotemporal structure of the motion inputs caused by the properties of the underlying physical world, which the brain probably discovered long ago. We will also show that the performance of such a system scales up to be used for real scenes and in real applications, and how our model can be used to explain well-known vision phenomena related to motion perception.

II. Motion processing in the brain

Visual motion processing is an important resource of informations that have a strong impact on behavioral decisions both in humans and monkeys. Similarly, for artificial systems, motion signals play a prominent role in interpreting a dynamic visual world. Motion estimations give rise to the

J. Eggert and V. Willert are with the Honda Research Institute Europe, emails: [Julian.Eggert, Volker.Willert]@honda-ri.de

J. Schmüdderich is PhD student at HRI-Europe in cooperation with the University of Bielefeld, email: jschmued@techfak.uni-bielefeld.de

[1] A reason might be the very probabilistic as well as theoretical focus of [2].

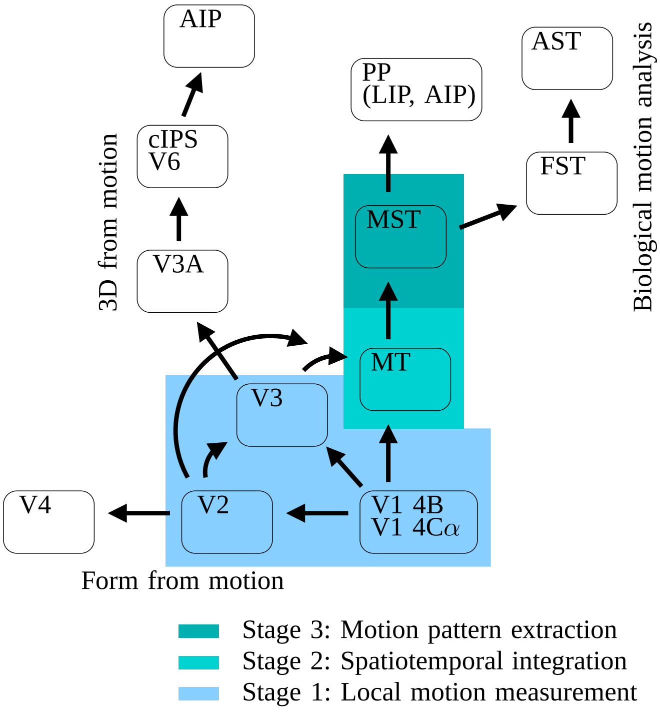
AIP
PP
(LIP, AIP)
AST
cIPS
V6
FST
MST
V3A
MT
V3
V4
V2
V1 4B
V1 4Cα
3D from motion
Biological motion analysis
Form from motion
Stage 3: Motion pattern extraction
Stage 2: Spatiotemporal integration
Stage 1: Local motion measurement

Fig. 1. Brain areas and feedforward information trasnferring connectivity involved in motion processing. Highlighted are 3 stages, which can be coarsely mapped to the model presented in this paper.

detection and perception of moving objects, as well as to the perception of self-motion. Motion perception also constitutes the basis for the control of eye movements, both for smooth object pursuit as well as for follow-up saccades.

Moving objects can be better segregated from a background (motion-based pop-out effects) than non-moving ones. The perception of a 2D shape is enhanced, both at its borders (kinetic boundaries and boundary continuation) as well as at its surface (by grouping and filling-in based on motion coherence effects), making its perception more vivid than in the stationary case. The 3D perception of shape is also enhanced by motion signals, and depth-from-motion directly contributes to estimations about 3D aspects of a scene.

It is commonly agreed that visual areas V1, MT and MST contribute to human motion processing, establishing a sort of "motion pathway". This involves a number of anatomically interconnected visual areas and their subdivisions along the dorsal processing streams in the brain [1]. Mainly fed by neurons of the magnocellular type of the LGN (which exhibit larger spatial receptive fields, higher temporal resolution and higher contrast sensitivity than their parvocellular counterparts), motion processing starts in neurons from layers 4 (4cα) and 6 in V1 [4]. This is the earliest stage (if processing starts at the eyes) that some neurons exhibit a selectivity both for orientation and direction of a stimulus; i.e., they have a response characteristic that can be described by a *spatiotemporal receptive field*. Motion processing then continues preferentially in the thick cytochrome oxidase stripes in V2, and areas V3, MT, MST, and possibly lateral and ventral intraparietal areas LIP and VIP [4]. From these areas, motion information then influences distinct portions of the prefrontal and premotor cortices contributing to higher cognitive activity and behaviour generation.

Areas V1, V2 and V3 constitute complementary areas for low-level motion measurement. The neurons are largely arranged in a retinotopic order, with varying receptive field sizes and already some degree of featural specialization in the different areas. Motion-selective neurons from these areas exhibit sensitivity for motion direction, but little sensitivity for motion speed [8].

Neurons from MT are also arranged retinotopically and exhibit direction selectivity. They have receptive field sizes that are about tenfold those of V1 neurons. In addition, they are more selective to different motion speeds [7]. What makes MT neurons special is that they do not only respond to their locally measurable motion, but take information of the context into account. So rather than responding to locally measurable motion, they integrate motion information from lower areas and MT itself to arrive at a consistent motion perception. In addition, it is assumed that they also incorporate information from other (e.g., non-motion related) pathways and higher processing areas.

Area MT projects onto area MST, which has even larger receptive fields. Among other things, MST neurons respond selectively to large-scale motion patterns, like expansion and rotation, which can cover the entire visual field. If multiple different local motions are present, the MST neurons for the different motions respond stronger than their MT counterparts, with MT exhibiting a more pronounced winner-takes-all characteristic [9]. Grouping processes seem to play a more prominent role in MST than in MT, but the details of the MST functionality are still to be investigated. It is further assumed that MST plays a role in the control of eye movements and egomotion-related estimations.

III. A 3-STAGE CANONICAL MOTION PROCESSING MODEL

Inspired from the biological findings, we built a model for human motion processing that consists of 3 main stages. In a first stage, local motion measurements are extracted from the visual signal. In a second stage, the local motion measurements are combined over time and space to take context effects into account. Finally, in a third stage, characteristic motion patterns are analyzed and a labeling of spatial positions with respect to the motion patterns occurs. Figure 2 shows the 3 stages of the network together with the information flow between the stages.

The task of the first stage is the detection of local changes in the visual scene and a measurement of the local displacements that may have led to these changes. Here, a system encounters the full range of ambiguities that are inherently present during motion estimation, caused by the aperture problem, physical overlapping, transparency of moving plaids and lack of measurable texture. What has to be solved is the correspondence problem: Which part of the scene is moving where? Of course it is not possible to solve this problem without taking into account further information.

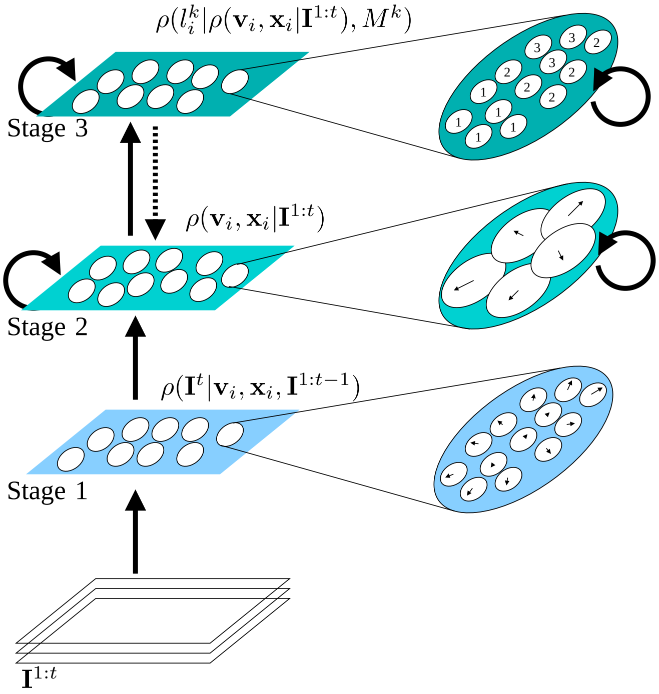
$\rho(l_i^k|\rho(\mathbf{v}_i, \mathbf{x}_i|\mathbf{I}^{1:t}), M^k)$
Stage 3
$\rho(\mathbf{v}_i, \mathbf{x}_i|\mathbf{I}^{1:t})$
Stage 2
$\rho(\mathbf{I}^t|\mathbf{v}_i, \mathbf{x}_i, \mathbf{I}^{1:t-1})$
Stage 1
$\mathbf{I}^{1:t}$
3 3 2
2 3 2
1 2 2
1 1
1 1

Fig. 2. The 3 stages of the motion estimation framework. In a first stage, local motion measurements are extracted from the visual signal. In a second stage, the local motion measurements are integrated over time and space. In a third stage, spatial positions are labeled (here: labels $1, 2, 3$) according to their participation in distinct motion patterns. The notations on the stages are used in the subsequent explanatory sections IV-A, IV-B and IV-C.

Therefore, in the first stage, we do not intend to solve this problem but instead represent the results in such a way that later stages may be able to handle it: We use “motion probabilities” to explicitly incorporate the ambiguities into the measurement process.

The first stage would correspond to the motion-sensitive entry neurons in areas V1, V2 and V3. How exactly the local measurement process occurs has previously been thoroughly investigated in model studies of biological experiments, but this is not considered to be of crucial importance here, as long as we can start with a “sufficiently good” first motion measurement. Moreover, it is also of little importance which signal we take to start with, like e.g. the intensity of the image, aspects of its color, or visual data preprocessed in any other way.

Due to the ambiguities, it is difficult to directly estimate the motion of larger components of the scene, such as moving objects, from the local motion measurements of stage 1. Higher level motion estimations therefore have to rely on grouping the local signals. This occurs in a sort of “binding” process, by integrating measurements that are consistent with object motion hypotheses and segregating them from inconsistent ones. In our model, this occurs by means of recurrent connectivity and a spatiotemporal integration of the local measurements in an area that we consider as functionally homologous to visual processing area MT. The spatiotemporal integration incorporates prior knowledge about motion measurement statistics and can be regarded as an implementation of the Gestalt law of combined common fate and good continuation (coupling space and time in a consistent way). Consequences are that dynamic aspects are introduced, leading to motion estimations that build up and improve over time.

Finally, in a third stage homologous to MST, we use the motion estimations from the area MT and analyze different global motion patterns. This includes the job of sorting out which motion estimations and which spatial positions are recruited by any motion pattern, i.e., we are dealing with a binding process between the motion features and the motion patterns. Such a binding process is usually divided conceptually into the processes of segregation (parsing) and integration (grouping) of the local measurements, with the aim to arrive at a figure-based scene description where elementary visual “objects” can be separated from each other. The integration process can by its own be divided into two subprocesses, one dealing with (I) propagation of information between local motion measurements to overcome ambiguities inherent in the sensory process, and the other one dealing with (II) the segmentation of regions of movements that are interpretable as object motions from the rest of the scene. In our model, the segregation occurs by using a labeling process, attributing labels to different positions according to their motion measurements and their consistency with the motion pattern. Recurrent interactions are very important here, since they mediate the propagation of information and the assignment of the labels via competition that constitutes the basis for the segmentation. Here again, prior knowledge is incorporated in terms of Gestalt laws of good continuation, reflecting that points that are retinotopically close to each other and have similar labels usually belong to the same object and move according to a certain motion pattern.

IV. Model Description

A. Local motion estimation

Disregarding the exact biological mechanisms of local motion estimation, we can assume that at time t, for every stream of incoming image data $\mathbf{I}^1, ..., \mathbf{I}^t$ resp. incoming image patch of small size around a retinotopic location $\mathbf{x}$, typically a number of local motion detectors is used to measure the translational motions for different velocities and directions $\mathbf{v}$. This means that for each $\mathbf{x}$, velocity measurements are extracted using a local neighbourhood of $\mathbf{x}$ with a predefined aperture size to get estimations

$$\rho(\mathbf{v}, \mathbf{x}|\mathbf{I}^1, ..., \mathbf{I}^t) \qquad (1)$$

for the different velocities, in contrast to models that extract a single velocity estimate. The advantage of such an approach is that the system is provided with information about the uncertainties inherent in the local measurements, such as occurring in cases of ambiguous and multiple motions. Subsequent grouping processes can then combine the local measurements of different positions $\mathbf{x}$ in a way that takes their uncertainties into account.

If we are working with discrete timesteps and two-frame estimations, eq. 1 reads as the probability that a velocity $\mathbf{v}$ is locally present around location $\mathbf{x}$ at time t under the assumption of image data $\mathbf{I}^t$ *and* previous timesteps image data $\mathbf{I}^{1:t-1}$, with $\mathbf{I}^{1:t-1} := \mathbf{I}^1, ..., \mathbf{I}^{t-1}$ being the set of all past measurements up to time $t-1$.

Eq. 1 constitutes a way of *distributed*, or population coding of the local motion estimates. This stands in contrast to direct measurements of velocity and direction of motion like in gradient optical flow approaches [3], [5]. The whole idea is to use the full information available in the population code to improve subsequent motion estimation processes that rely on this information. The population code can be seen in analogy to the sets of motion-selective neurons forming cortical maps and hypercolumns, with all $\rho(\mathbf{v}|\mathbf{I}^{1:t}, \mathbf{x})$ for fixed $\mathbf{x}$ and t (resp. $\mathbf{I}^{1:t}$) being the output of a hypercolumn of motion-sensitive neurons looking at a common retinotopic location $\mathbf{x}$ with the same aperture.

In a Bayesian interpretation, the estimation $\rho(\mathbf{v}, \mathbf{x}|\mathbf{I}^{1:t})$ is gained by combining a likelihood with a prior, in a way that

$$\rho(\mathbf{v}, \mathbf{x}|\mathbf{I}^{1:t}) \propto \rho(\mathbf{I}^t|\mathbf{v}, \mathbf{x}, \mathbf{I}^{1:t-1})\, \rho(\mathbf{v}, \mathbf{x}) \ , \qquad (2)$$

with the corresponding normalization of $\rho(\mathbf{v}, \mathbf{x}|\mathbf{I}^{1:t})$. The likelihood constitutes the actual measuring process and indicates how certain/probable a measurement of local image data $\mathbf{I}^t$ is around the retinotopic position $\mathbf{x}$ for a given assumed physical velocity $\mathbf{v}$. In the biological view of fig. 1, the likelihood corresponds to the activity of the motion selective cells from areas V1, V2 and V3. This is combined with a velocity prior $\rho(\mathbf{v}, \mathbf{x})$ (something like a "top-down expectation" in the neural sense) to get the Bayes' posterior in the usual way, expressed in the activity of the cells of area MT.

In addition, local motion signals are propagated non-locally within MT from one cell to another to be able to resolve the ambiguities inherently contained in the measurement process. This propagation sometimes has to occur over extended retinotopic regions, if there is e.g. a large object with a rigid-body motion pattern. The propagation can be seen to occur over space (spatial integration) and time (temporal prediction), in a way that it builds up iteratively by spreading from the areas where the motion measurement is not ambiguous to areas with larger uncertainties, taking as relay points the already disambiguated regions.

Both spatial integration and temporal prediction influence the Bayesian posterior calculation by modification of the prior. This means that instead of 2, we will use

$$\rho(\mathbf{v}, \mathbf{x}|\mathbf{I}^{1:t}) \propto \rho(\mathbf{I}^t|\mathbf{v}, \mathbf{x}, \mathbf{I}^{1:t-1})\, \rho^t(\mathbf{v}, \mathbf{x}, \mathbf{I}^{1:t-1}) \qquad (3)$$

with a prior $\rho^t(\mathbf{v}, \mathbf{x}, \mathbf{I}^{1:t-1})$ that depends on the inputs at previous timesteps $\mathbf{I}^{1:t-1}$. The prior is calculated anew for each timestep depending on the spatial configuration of the locally measured motions and the temporal coherence assumptions.

The spatiotemporal integration enhances coherence and causality as inherent properties of motion-related visual signals. Spatial coherence is one important cue for the visual system, which can be seen in analogy to contour integration processes. In motion estimation, it seems reasonable for the visual system to assume that sets of local motion detectors with receptive fields that are close to each other tend to arrive at similar measurement results, both because their apertures are overlapping and because the motion itself may be spatially extended. Temporal prediction is the second source of information which can be used to refine the local motion measurements. Ambiguous motion information can sometimes be resolved by temporal coherence. The underlying assumption is that motion is usually temporally continuous, that is, to a first approximation the motion will continue with the same velocity and direction at the next timestep. Combining the two ideas of spatial integration and temporal prediction, we get

$$\rho^t(\mathbf{v}, \mathbf{x}, \mathbf{I}^{1:t-1}) := \int_{\mathbf{x}'} \mathbf{W}^{\mathbf{x}}_{\mathbf{x}'}\, \rho(\mathbf{v}, \mathbf{x}' - \mathbf{v}\Delta t|\mathbf{I}^{1:t-1})\, \mathrm{d}\mathbf{x}' \quad (4)$$

as a spatiotemporally integrating prior to eq. 3. It expresses that the cells from MT expect velocities that are gained by selectively averaging ($\mathbf{W}^{\mathbf{x}}_{\mathbf{x}'}$) the results from previous timesteps that are compatible with the assumption of temporal coherence ($\mathbf{x}' - \mathbf{v}\Delta t|\mathbf{I}^{1:t-1}$).

B. Full derivation of the spatiotemporal integration in MT

The spatiotemporal integration according to eq. 4 is *one* (and additionally, the simplest) possible choice of incorporating predictions via recurrent connectivity into motion processing. In this section, we present the full derivation of how such an integration can be theoretically justified. We start from an overall state vector

$$\mathbf{S} := \{\mathbf{V}, \mathbf{X}\} \qquad (5)$$

that comprises the *vector field*, i.e., the set of local velocities $\mathbf{V} := \{\mathbf{v}_i\}_i$ at all spatial locations $\mathbf{X} := \{\mathbf{x}_i\}_i$ (with the index $i = 1 \dots I$ running over I "particles" with attached positions and velocities, like e.g. in a retinotopic map, where the particles i represent a number of motion selective cells with fixed retinotopic receptive field locations $\mathbf{x}_i$).

The problem of probabilistic motion estimation can be seen as a particular case of the estimation of the state of a system that changes over time using a series of (noisy) measurements. The dynamic state estimation can be achieved by constructing the posterior probability density function of the state based on all available information. The process of state estimation usually involves two stages: *prediction* and *update*, and occurs at every timestep when a new measurement is received.

The predictive prior can then be calculated (for a discrete state space and timesteps numbered $1, \dots, t$) according to [2]

$$\rho^t(\mathbf{V}, \mathbf{X}|\mathbf{I}^{1:t-1}) = \qquad (6)$$
$$\int_{\mathbf{V}'} \int_{\mathbf{X}'} \rho(\mathbf{V}, \mathbf{X}|\mathbf{V}', \mathbf{X}')\ \rho(\mathbf{V}', \mathbf{X}'|\mathbf{I}^{1:t-1})\, \mathrm{d}\mathbf{X}'\mathrm{d}\mathbf{V}'$$

[2] Usually known as the Chapman-Kolmogorov equation

This equation expresses that the new prediction for the velocity probability density is given by the last timestep $t-1$ estimate weighted with the transition probability $\rho(\mathbf{V},\mathbf{X}|\mathbf{V}',\mathbf{X}')$ from the last timestep velocity $\mathbf{V}'$ to the new velocity $\mathbf{V}$ for all combinations of positions. The probabilistic model of the state evolution $\rho(\mathbf{V},\mathbf{X}|\mathbf{V}',\mathbf{X}')$ is assumed to be known and describes the knowledge about the state transitions from one timestep to the next.

At timestep t, a new measurement $\mathbf{I}^t$ becomes available, and this can be used to update the state estimation from the predictive prior by combining it with the velocity likelihood via Bayes' rule. We therefore arrive at the update equation (this is the analogous equation to eq. 3)

$$\rho(\mathbf{V},\mathbf{X}|\mathbf{I}^{1:t}) \propto \rho(\mathbf{I}^t|\mathbf{V},\mathbf{X},\mathbf{I}^{1:t-1})\,\rho^t(\mathbf{V},\mathbf{X}|\mathbf{I}^{1:t-1}) \ . \quad (7)$$

This involves the measurement model $\rho(\mathbf{I}^t|\mathbf{V},\mathbf{X},\mathbf{I}^{1:t-1})$ indicating the likelihood that image data $\mathbf{I}^t$ is measured if local velocities $\mathbf{V}$ at $\mathbf{X}$ and previous timestep measurements $\mathbf{I}^{1:t-1}$ are assumed. Spatial correlations between velocities at positions $\mathbf{x}$ and $\mathbf{x}'$ are in this case hidden in the state evolution model $\rho(\mathbf{V},\mathbf{X}|\mathbf{V}',\mathbf{X}')$ and the measurement model $\rho(\mathbf{I}^t|\mathbf{V},\mathbf{X},\mathbf{I}^{1:t-1})$.

Although it is beneficial to have the analytical description of the time course of the full motion model for the entire vector field with eqs. 6 and 7, it is unfeasible to use this equation directly to calculate the local velocities. Therefore, we now care how we can use the full motion model for $\mathbf{V},\mathbf{X}$ to get expressions for the single $\mathbf{v}_i$, $\mathbf{x}_i$. For each vector of the vector field (consisting of velocity $\mathbf{v}_i$ and position $\mathbf{x}_i$), marginalizing out (i.e., integrating over) all the $\mathbf{v}_j$, $\mathbf{x}_j$ with $j \neq i$ we can set up a local predictive prior in analogy to the spatiotemporally integrating prior from eq. 4

$$\begin{aligned}&\rho^t(\mathbf{v}_i,\mathbf{x}_i|\mathbf{I}^{1:t-1}) = &(8)\\ &\int_{\mathbf{V}'}\int_{\mathbf{X}'} \rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{V}',\mathbf{X}')\ \rho(\mathbf{V}',\mathbf{X}'|\mathbf{I}^{1:t-1})\,\mathrm{d}\mathbf{X}'\mathrm{d}\mathbf{V}' \ ,\end{aligned}$$

and in analogy to the Bayes estimation eq. 3 we get the posterior for the local vectors according to

$$\rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{I}^{1:t}) \propto\ \rho(\mathbf{I}^t|\mathbf{v}_i,\mathbf{x}_i,\mathbf{I}^{1:t-1})\,\rho^t(\mathbf{v}_i,\mathbf{x}_i|\mathbf{I}^{1:t-1}) \ . \quad (9)$$

What we see from eq. 8 is that all other past velocity estimations (i.e., the entire vector field) very annoyingly influence the local velocity estimation via the $\mathbf{V}'$, $\mathbf{X}'$, making purely local expressions for the local velocity estimations impossible. We now make 2 assumptions. First, we assume that the posterior probabilities of the vectors of the vector field given the past inputs factorize, implying that they can be estimated independently from each other,

$$\begin{aligned}&\rho(\mathbf{V},\mathbf{X}|\mathbf{I}^{1:t}) &(10)\\ &\quad = \rho(\mathbf{v}_1,\mathbf{x}_1,\ldots,\mathbf{v}_I,\mathbf{x}_I|\mathbf{I}^{1:t})\\ &\quad = \prod_j \rho(\mathbf{v}_j,\mathbf{x}_j|\mathbf{I}^{1:t}) \ .\end{aligned}$$

This is reasonable if one e.g. thinks of the vectors as being attached to different spatial positions, so that basically we are saying here that estimations at one position can be made independently of estimations at another position [3].

Second, we start with a particle-to-particle state evolution (i.e., a description of how the state $\mathbf{v}_j'$, $\mathbf{x}_j'$ of a particle j at the last timestep influences the state $\mathbf{v}_i$, $\mathbf{x}_i$ of a particle i at the current timestep)

$$\rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{v}_j',\mathbf{x}_j') \ , \quad (11)$$

gained from marginalizing $\rho(\mathbf{V},\mathbf{X}|\mathbf{V}',\mathbf{X}')$, and assume that the state evolution model for a single vector i can be factorized according to

$$\begin{aligned}&\rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{v}_1',\mathbf{x}_1',\ldots,\mathbf{v}_I',\mathbf{x}_I') &(12)\\ &\quad = 1-\prod_k[1-\rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{v}_k',\mathbf{x}_k')]\end{aligned}$$

which, by multiplying out, can be approximated by

$$\begin{aligned}&\rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{v}_1',\mathbf{x}_1',\ldots,\mathbf{v}_I',\mathbf{x}_I') &(13)\\ &\quad \approx \sum_k \rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{v}_k',\mathbf{x}_k') + \text{terms } O(\rho^2)\end{aligned}$$

since higher combinations of ρ always include some very low probabilities between inconsistent pairs of velocities and positions and can therefore be neglected.

This is reasonable if one thinks that the local state $\mathbf{v}_i$, $\mathbf{x}_i$ is gained from the previous local state $\mathbf{v}_1'$, $\mathbf{x}_1'$ *or* $\mathbf{v}_2'$, $\mathbf{x}_2'$ *or* ... .

Starting now from eq. 8 (predictive prior for a single velocity vector) we make use of the factorization property eq. 10 for the $\rho(\mathbf{V}',\mathbf{X}'|\mathbf{I}^{1:t-1})$ so as to get

$$\begin{aligned}&\rho^t(\mathbf{v}_i,\mathbf{x}_i|\mathbf{I}^{1:t-1}) = &(14)\\ &\int_{\mathbf{v}_1'}\int_{\mathbf{x}_1'}\cdots\int_{\mathbf{v}_I'}\int_{\mathbf{x}_I'} \rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{v}_1',\mathbf{x}_1',\ldots,\mathbf{v}_I',\mathbf{x}_I') \times\\ &\quad \times \prod_j \rho(\mathbf{v}_j',\mathbf{x}_j'|\mathbf{I}^{1:t-1})\,\mathrm{d}\mathbf{x}_I'\mathrm{d}\mathbf{v}_I'\ldots\mathrm{d}\mathbf{x}_1'\mathrm{d}\mathbf{v}_1'\end{aligned}$$

In a second step, we use the factorization property eq. 12 resp. its approximation eq. 13 with eq. 14 to arrive at

$$\begin{aligned}&\rho^t(\mathbf{v}_i,\mathbf{x}_i|\mathbf{I}^{1:t-1}) \propto &(15)\\ &\int_{\mathbf{v}_1'}\int_{\mathbf{x}_1'}\cdots\int_{\mathbf{v}_I'}\int_{\mathbf{x}_I'} \sum_k \rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{v}_k',\mathbf{x}_k') \times\\ &\quad \times \prod_j \rho(\mathbf{v}_j',\mathbf{x}_j'|\mathbf{I}^{1:t-1})\,\mathrm{d}\mathbf{x}_I'\mathrm{d}\mathbf{v}_I'\ldots\mathrm{d}\mathbf{x}_1'\mathrm{d}\mathbf{v}_1' \ .\end{aligned}$$

For a fixed k, $\mathbf{v}_k'$, $\mathbf{x}_k'$, and for all $\mathbf{v}_j'$, $\mathbf{x}_j'$ with $j \neq k$, we can then move the integral right up to the product sign and integrate, which leads to factors 1. The result is

$$\begin{aligned}&\rho^t(\mathbf{v}_i,\mathbf{x}_i|\mathbf{I}^{1:t-1}) \propto &(16)\\ &\sum_k \int_{\mathbf{v}_k'}\int_{\mathbf{x}_k'} \rho(\mathbf{v}_i,\mathbf{x}_i|\mathbf{v}_k',\mathbf{x}_k')\rho(\mathbf{v}_k',\mathbf{x}_k'|\mathbf{I}^{1:t-1})\,\mathrm{d}\mathbf{x}_k'\mathrm{d}\mathbf{v}_k' \ .\end{aligned}$$

Using 2 reasonable assumptions eqs. 10 and 12, we have therefore arrived at a closed form eq. 16 for the local state

[3] Which is true for most pairs of positions, since only selected other positions and velocities are able to influence a local $\mathbf{v}_i$, $\mathbf{x}_i$.

prediction which together with the local update equation 9 can now be used to calculate the local motion estimates over time using a framework as motivated and introduced in section IV-A.

One remarks remain to be made at this point. Whereas the Bayesian update eq. 9 works locally (only $\mathbf{v}_i$, $\mathbf{x}_i$ appear), the predictive prior eq. 16 is not, since nonlocal influences are integrated using the two-point state evolution $\rho(\mathbf{v}_i, \mathbf{x}_i|\mathbf{v}'_k, \mathbf{x}'_k)$. Nevertheless, this state evolution is much easier to handle than the full $\rho(\mathbf{V}, \mathbf{X}|\mathbf{V}', \mathbf{X}')$. Indeed, for an implementation we will reduce $\rho(\mathbf{v}_i, \mathbf{x}_i|\mathbf{v}'_k, \mathbf{x}'_k)$ further, which can in many cases be done without causing harm to the velocity estimations. On the other hand, eq. 16 should not be simplified to much, since in the brain it may implicitly contain assumptions about the world that cause the velocity sensations, i.e., it can be used to comprise knowledge about scene statistics and spatiotemporal correlations between velocity estimations at different positions in subsequent timesteps. We will come back to this issue in the application examples of sections V-B and V-C.

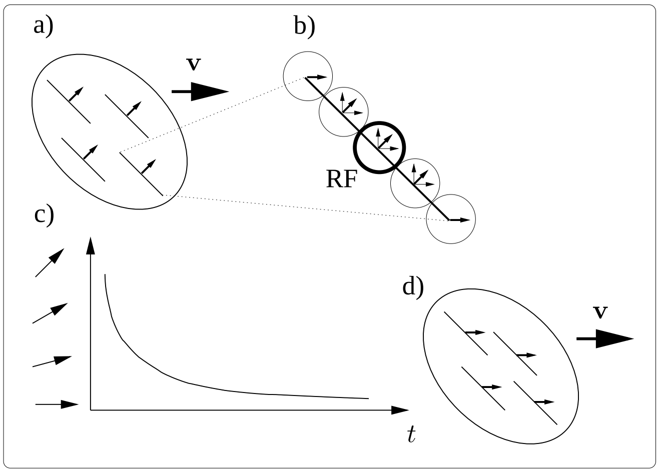

Fig. 3. Time course of the motion selectivity of neurons in MT. The pattern a) moves to the right, but due to the aperture problem b) the neurons are only able to see movement orthogonal to the line orientations. c) At first, the main activity is at neurons with orthogonal motion selectivities, but after a time period of about 70 ms the activity starts to shift towards neurons that reflect the "true" object motion, as shown in d).

C. Extraction of motion patterns

The result of the motion measurement and the spatiotemporal integration is the motion estimation $\rho(\mathbf{v}_i, \mathbf{x}_i|\mathbf{I}^{1:t})$ (eq. 9), which we coarsely localize in stage 2 of our motion processing model, eventually corresponding to area MT.

Sitting on top of this is our stage 3 (corresponding to area MST), used for motion pattern extraction. Here, the system builds up a small number of "motion pattern models" M^k and the target is to estimate the *labeling probability*

$$\rho(l_i^k|\rho(\mathbf{v}_i, \mathbf{x}_i|\mathbf{I}^{1:t}), M^k) \tag{17}$$

that a particle i (corresponding to a location in space with a velocity vector attached to it) contributes to a motion pattern k, given the current motion estimation.

The pattern assigment (i.e., the labeling) occurs in 3 steps:

1) Label measurement. Here, the match between the current motion estimation and the expected motion pattern generated from M^k is computed.
2) Spatiotemporal constraints. The labels compete with each other for a unique assignment to a model M^k, forcing a winner-take-all behaviour like for MST neurons (see section II). In addition, recurrent interactions tend to assign similar labels for spatially neighbouring particles, implementing Gestalt laws for label assignment.
3) Model adaptation and estimation. This occurs by minimization of the difference between the motion pattern generated from the M^k's and the motion estimations $\rho(\mathbf{v}_i, \mathbf{x}_i|\mathbf{I}^{1:t})$ evaluated using the pattern assignment probabilities $\rho(l_i^k|\rho(\mathbf{v}_i, \mathbf{x}_i|\mathbf{I}^{1:t}), M^k)$.

The type of expected motion patterns depends from the application area. Motion patterns can be very simple, like e.g. for motion-based object segmentation, where it is assumed that an entire object has the same motion at all its constituting locations. They can also be more complex, like e.g. when searching for expanding patterns as are measured for flow fields during egomotion, or rotating patterns as occur when we tilt our heads. Furthermore, motion patterns can be imposed from domains that are unrelated with motion processing itself, expressing that the system is e.g. expecting a particular dynamics in its retinal input because the observer is moving and the system knows about this from other sources than visual motion estimation.

In sections V-B and V-C, we will show two examples of motion pattern extraction, one for layer separation and another one for egomotion compensation, based on the model presented in the previous sections.

V. Results and Applications

In the following subsections we show how our model can be used to simulate and explain experimental phenomena found in physiological and psychophysical measurements. We have found that a broad range of motion-related phenomena can be at least reproduced with our model, like special motion illusions. Here we explain 2 selected effects, and afterwards we present 2 more application oriented extensions of our motion estimation framework.

A. Psychophysical and experimental data

If a stimulus with oriented edges moves horizontally as shown in fig. 3 a), local motion detectors with small receptive fields respond strongest if their selectivity is tuned to movement orthogonal to the line orientations, since they are only able to "see" a very limited portion of the stimulus. Only receptive fields at the edge endpoints (fig. 3 b) can detect the true direction of motion. Interestingly, after about 70 ms, the neuronal activities in area MT begin to shift from the orthogonal towards the true direction, as shown in fig. 3 c. After about 150 ms, only those neurons remain active whose selectivity is tuned to the true, horizontal motion, regardless of the orientation of the single edges.

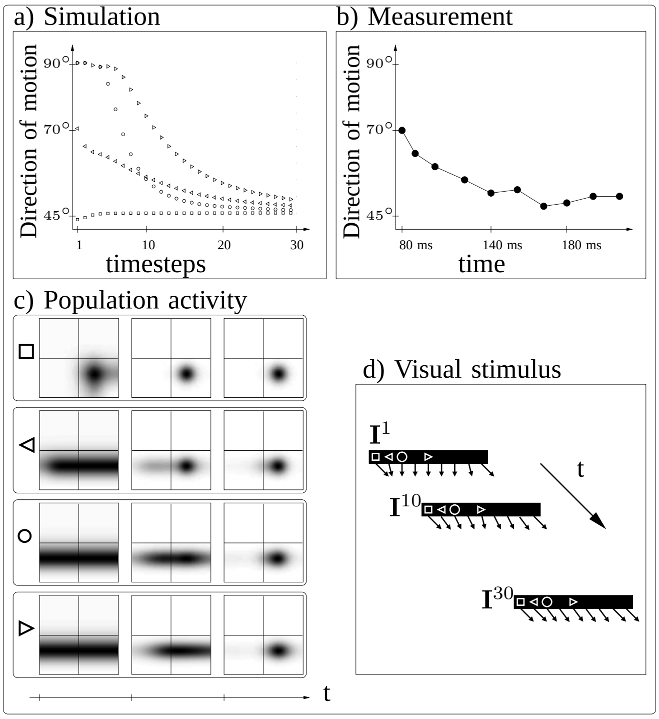

Fig. 4. Resolution of ambiguous motion signals by spatiotemporal integration in our model. a) Time course of motion estimations at 4 different positions of a moving bar stimulus. b) Comparison with an experimentally measured curve. c) Activity of the velocity hypercolumn for all 4 points. d) The estimated motion plotted together with the stimulus.

This reflects the fact that a spatiotemporal integration occurred that propagated information from the edge endpoints towards the inner, more ambiguous parts of the edges.

In fig. 4, we show the results of the neuronal activity from our model for a bar stimulus moving diagonally to the bottom right. In a), we show the simulated time course of motion estimations at 4 different positions of the stimulus. The curve for 70° can be compared with the experimentally measured curve b) from [6]. Below, in c), we show the activities at 3 selected times of the neuronal population encoding all velocities at 4 retinal positions, as marked on the bar in d) (the zero-velocity is at the center of each small diagram, at the lower right quadrant we find the neurons responsive for displacements to the lower right, etc.). It can be seen how the activities refine and shrink with time so that after 30 timesteps they encode the true diagonal velocity of the stimulus.

A psychophysical effect that can be nicely explained with our model is the motion-based hysteresis effect. In this experiment, a display is presented with dots moving homogeneously to one direction. With increasing time, single dots are selected and their direction of motion inverted, until all points move homogenously into the opposite direction [12]. The setup is shown in the bottom row of fig. 5.

The perception of such a stimulus set exhibits a marked hysteresis effect. Subjects report a sensated "homogeneous motion", meaning a sensated motion of *all* points into the same direction, even if a considerable percentage of points already moves into an opposite direction. It seems that outliers are suppressed until a certain threshold is reached.

In figure 5 a) and b), we plotted the perceived vs. the real proportion $\mathbf{v}_{r/rl}$ of points moving homogeneously into one direction. We define $\mathbf{v}_{r/rl} = \sum \mathbf{v}_{\text{right}} / (\sum \mathbf{v}_{\text{right}} + \sum \mathbf{v}_{\text{left}})$. The perceived $\mathbf{v}_{r/rl}$ is gained from our model, by evaluating the number of motion detectors "voting" for the right or the left direction. In a), we see the hysteresis effect caused by the recurrent connectivity and the spatiotemporal integration in the system. In b), we have switched off the recurrence and the hysteresis effect vanishes.

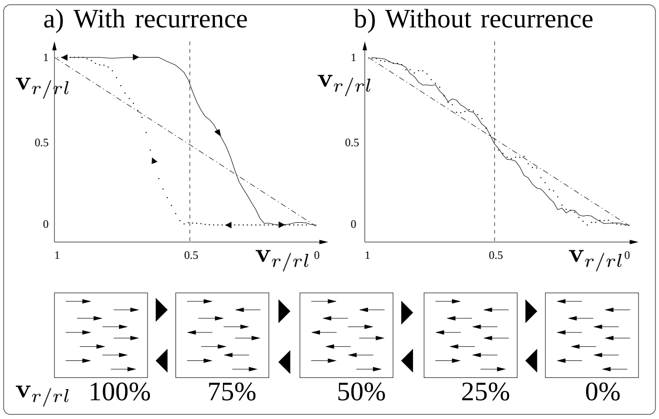

Fig. 5. Simulation of the hysteresis effect for a gradual change of the global direction of motion by direction inversion of single points.

B. *Layering using affine motion patterns*

For many application domains, planar continuous motion patterns suffice. In this example, we restrict to 3 affine motion patterns and 3 corresponding labels. Fig. 6 shows the time course of the label assignment and motion pattern estimation process. The original sequence is shown at the top row, the lower 3 rows show the label assignment process. At the beginning (left column), motion patterns are unspecific and the locations of the input are distributed evenly among the labels.

As explained in section IV-C, the motion estimation from section IV-B provides the basis for the label assignment. As soon as the motion patterns differentiate, the labels start to compete with each other. Those locations and velocities that best match with a pattern then drive the label dynamics. It can be seen that the assignment to the tree occurs quite rapidly (second row from top), followed by the flower-bed (third row) and the rest of the background (bottom row)

C. *Egomotion compensation*

When mobile robots move around in a *static* environment, the projection of the environment onto the robot's cameras induces an optical flow that is exclusively caused by the egomotion of the robot. Additional sensing of the body movement via proprioception allows for depth estimation of the scene because of motion parallax. As a reverse operation to egomotion-based depth estimation, the expected optical flow generated by egomotion can be inferred by combining body

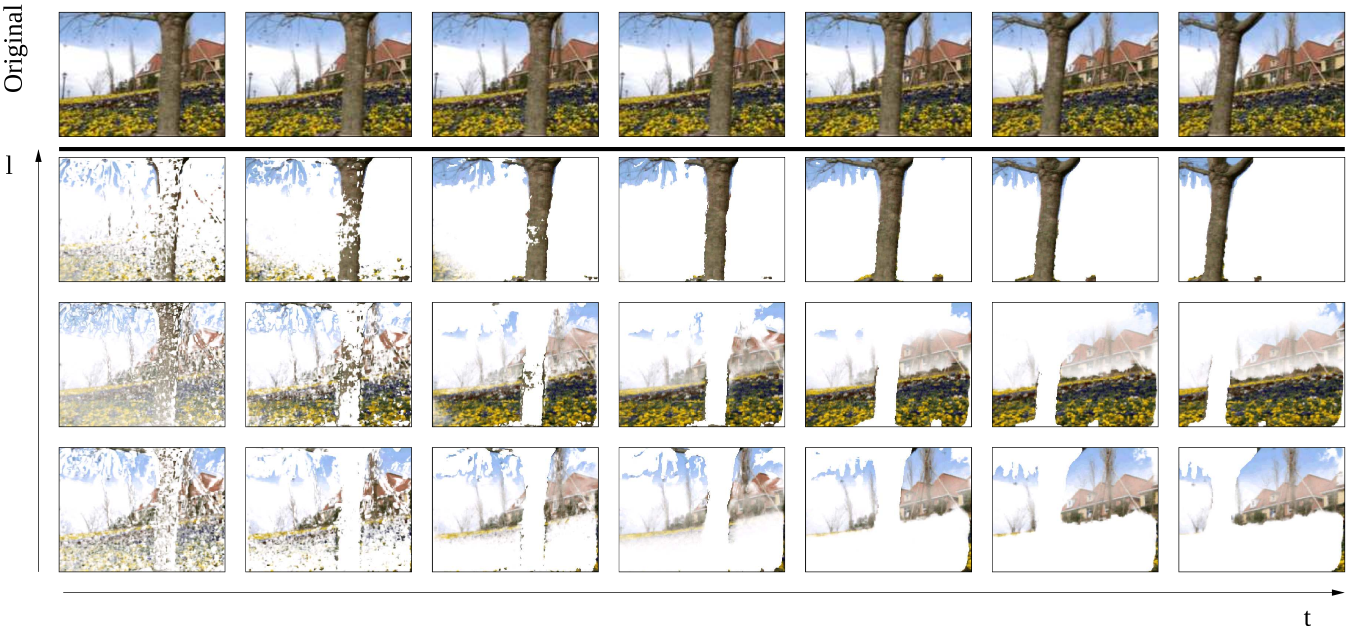

Fig. 6. "MPEG Flowergarden" sequence taken from a moving observer (top row). The different planes of the scene shift horizontally with different speeds depending on their depth: The tree moves fastest, the flower-bed at intermediate speed and the house and the sky background move very slowly. The system separates the overall measured motion into 3 layers of affine motion patterns, clearly segregating the different parts of the scene (bottom 3 rows).

movement and scene depth information using depth cues like e.g. extracted from binocular disparity. Unfortunately, in most cases the environment is not static but contains moving objects. These then induce optical flow components onto the robot's cameras which deviates from the optical flow as it is predicted from egomotion for static scenes. The overall optical flow is therefore always caused by a combination of ego- and object motion that cannot be separated without depth and body movement information.

To tackle the problem of extracting moving objects in a visual scene despite egomotion of the observer, we set up a structure as depicted in fig. 7, allowing the system to compensate for egomotion effects. We estimate the image flow induced by egomotion assuming a static scene by utilizing the robot's kinematics and depth information d from stereo vision with input data streams $\{\mathbf{I}^{r,1:t}, \mathbf{I}^{l,1:t}\}$ [4] (see *Egomotion flow* in fig. 7). According to this predicted flow each image $\mathbf{I}^{l,t+1}$ is warped so that we get an egoflow-compensated image $\hat{\mathbf{I}}^{l,t+1}$. All motion estimations then occur on the basis of the compensated image, so that only the *relative* optical flow is extracted. With the continuous image streams $\{\mathbf{I}^{l,1:t}, \hat{\mathbf{I}}^{l,1:t+1}\}$ as input data to our motion estimation system we are able to extract, integrate and predict the optical flow induced by moving objects (separated from the egoflow) with all the advantages of probabilistic spatiotemporal filtering mentioned beforehand (see *Relative object flow* in fig. 7 where for comparison we also show the *Overall flow* measured without egomotion compensation).

Additionally, we take the reliability of the depth and motion estimates based on the (un)certainty of the probability density functions $\rho(\mathbf{v}, \mathbf{x}|\mathbf{I}^{l,1:t}, \hat{\mathbf{I}}^{l,1:t+1})$ and $\rho(d, \mathbf{x}|\mathbf{I}^{r,t}, \mathbf{I}^{l,t})$ into account, so that relative object flow vectors that are based on unreliable depth and motion information are neglected. The result is shown at the rightmost picture of figure 7 (*Relative object flow*). In this case, the robot (ASIMO) was moving backwards, resulting in a concentric egomotion-based optical flow and a rightwards oriented estimated egomotion flow at the arm of the person, while the arm itself also moved rightwards. Therefore, in the overall flow, the arm movement is hard to distinguish, whereas the egomotion compensation removes the egomotion and extracts the arm as the only moving part of an otherwise static scene.

VI. Conclusions

In this paper, we have presented a framework for motion estimation based on ideas originated from biological findings and psychophysical experiments, which indicate that the brain uses a spatiotemporal integration mechanism to overcome ambiguities inherent in the sensory process. We have shown how such a model can indeed be used for motion integration to overcome the limitations imposed by the aperture problem. We consider the model to be sufficiently rich to account for a large variety of psychophysical and physiological findings on motion processing in the brain, yet sufficiently simple so that it can be implemented efficiently which makes it applicable to practical applications with real-world images.

Acknowledgments

We would like to thank U. Koerner for support and advice with figure 1 and researching the biological basis of motion processing.

[4] r/l: right/left image

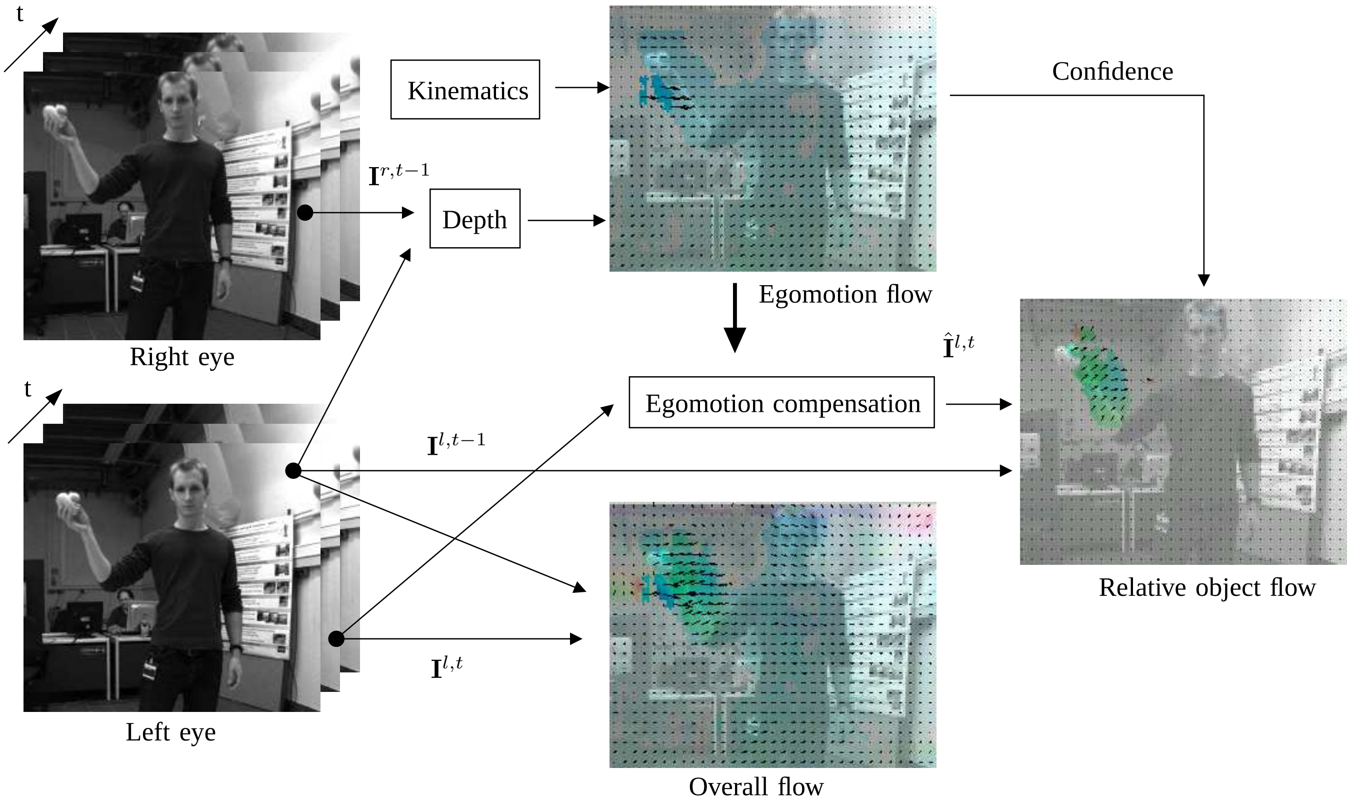

Fig. 7. Egomotion-compensated motion estimation system as tested with ASIMO. We extract depth information from disparity and use this together with robot's kinematics to calculate the predicted retinal flow for static scenes. Objects which move in the scene can then be extracted by their deviation from the predicted flow.

References

[1] D. Boussaoud, L. Ungerleider, and R. Desimone. Pathways for motion analysis: cortical connections of the medial superior temporal and fundus of the superior temporal visual areas in the macaque. *J. Comp. Neurol.*, 1990.

[2] Pierre-Yves Burgi, Alan L. Yuille, and Norberto M. Grzywacz. Probabilistic motion estimation based on temporal coherence. *Neural Computation*, 12(8):1839–1867, 2000.

[3] B. K. P. Horn and B. G. Schunk. Determining Optic Flow. *Artificial Intelligence*, 17:185–204, 1981.

[4] E. R. Kandel, J.H. Schwartz, and T.M. Jessell. *Principles of Neural Science*. McGraw-Hill, 1995.

[5] B. D. Lukas and T. Kanade. An Iterative Image-Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, Vancouver, Kanada, 1981.

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

[7] J. N. Priebe, R. C. Casanello, S. G. Lisberger, and G. Stephen. The neural representation of speed in macaque area mt/v5. *J. Neurosci.*, 2003.

[8] J. N. Priebe, S. G. Lisberger, and J. A. Movshon. Tuning for spatiotemporal frequency and speed in directionally selectiven neurons of macaque striate cortex. *J. Neurosci.*, 2006.

[9] G.H. Recanzone, R.H. Wurtz, and U. Schwarz. Responses of MT and MST Neurons to One and Two Moving Objects in the Receptive Field. *Journal of Neurophysiology*, 78:2904–2915, 1997.

[10] W. Reichardt. Autokorrelationsauswertung als Funktionsprinzip des Zentralnervensysems. *Z. Naturforsch. B*, 12:447–457, 1957.

[11] E.P. Simoncelli and D.J. Heeger. A model of neural responses in visual area MT. *Vision Research*, 38:743–761, 1998.

[12] D. Williams and G. Phillips. Cooperative phenomena in the perception of motion direction. *Optical Society of America*, 4:878–885, 1987.