Estimating Object Proper Motion Using Optical Flow, Kinematics, and Depth Information

Jens Schmüdderich, Volker Willert, Julian Eggert, Sven Rebhan, Christian Goerick, Gerhard Sagerer, Edgar Körner

2008

Preprint:

This is an accepted article published in IEEE Systems, Man, and Cybernetics Part B: Cybernetics. The final authenticated version is available online at: https://doi.org/[DOI not available]

Estimating Object Proper Motion Using Optical Flow, Kinematics, and Depth Information

Jens Schmüdderich, Volker Willert, Julian Eggert, Sven Rebhan, Christian Goerick, Gerhard Sagerer, *Member, IEEE*, and Edgar Körner

***Abstract*—For the interaction of a mobile robot with a dynamic environment, the estimation of object motion is desired while the robot is walking and/or turning its head. In this paper, we describe a system which manages this task by combining depth from a stereo camera and computation of the camera movement from robot kinematics in order to stabilize the camera images. Moving objects are detected by applying optical flow to the stabilized images followed by a filtering method, which incorporates both prior knowledge about the accuracy of the measurement and the uncertainties of the measurement process itself. The efficiency of this system is demonstrated in a dynamic real-world scenario with a walking humanoid robot.**

***Index Terms*—Disparity, egomotion (EM), kinematics, motion, optical flow (OF).**

I. Introduction

THE ABILITY to visually perceive motion is believed to be highly beneficial for surviving in a dynamic environment. Therefore, it is not surprising to see that movement is one of the most important cues to attract visual attention [1]. Interestingly, most mammals are able to perceive it while they are moving themselves—either by rotating the head and the eyes, moving the whole body, or even while they are running. The gathered information is then used for controlling their own movement [2] or tracking moving objects by keeping them centered in the fovea [3].

This motivates us to realize a neurobiologically inspired system for a humanoid robot, which is capable of measuring visual motion while the robot is moving itself. The effects of egomotion (EM) on the optical flow (OF), which is the so-called EM flow (EMF), produced by the robot makes this task quite challenging. Primarily, this is caused by the large number of degrees of freedom and the complex influence of each robot segment on the position of the camera: For example, bending the knee may, in some situations, cause the robot, and hence the camera, to shift heavily to the side. In other situations, the robot might be standing on the other leg, and thus, knee bending does not affect the camera position at all. In general, the autonomous movement of a walking humanoid robot causes the camera to undergo translatory and rotational movements in 3-D with sudden velocity changes.

We propose a system for the computation of the OF induced by independently moving objects, which we call object proper motion (OPM). To achieve this goal under those demanding conditions described earlier, we utilize a novel combination of known algorithms to compute the OF, depth from binocular disparity, and camera movement from kinematics.

An overview of the system is shown in Fig. 1. In the first step, we use the forward kinematics to compute the movement of the camera, occurring in the time interval from m to $n = m + 1$. Combining this with depth information from binocular disparity, we can estimate where a static point in the image at time m moved due to EM and obtain the EMF. Hence, by knowing where a point in the image at time n originated, we can compensate the EM effect by moving the point back to its original position in the second step, resulting in an image freed from EM effects. Afterward, this image is used for a calculation of the OF measured relative to the EMF. This step-by-step movement estimation allows the reduction of the OF search range, which does not only reduce computation time but also decreases the possibility for ambiguities in the OF measurement and thereby leads to qualitatively better results. Ideally, the OF measured in this way should be zero for unmoving objects and otherwise describe their proper motion. However, because binocular disparity and OF are particularly noisy signals, we finally incorporate the measurement errors into a filter mechanism to reject invalid velocity estimations.

The details of this system are described in Section II. Prior to this, we give a review of related work in Section I-A. In Section I-B, we shortly summarize neurobiological evidence to get an idea of the OPM estimation in the human brain. Section III will demonstrate the feasibility of our approach by showing the achieved results in a real-world scenario. In addition, a detailed parameter discussion concerning the tuning and the interrelations is given.

A. Related Work

The field, where the estimation of OPM during EM is commonly addressed, is the car domain. Here, this estimation is important to measure the speed of other cars relative to the observing car, which allows to identify them as eventual obstacles. Most approaches rely on the calculation of OF. This flow is a superposition of the EMF and the OPM. A common procedure to decompose the OF and extract the desired OPM tries to estimate the movement of the camera from the visual

Manuscript received October 16, 2007; revised January 25, 2008. This paper was recommended by Associate Editor Q. Ji.

J. Schmüdderich and G. Sagerer are with the Applied Computer Science Group, Bielefeld University, 33501 Bielefeld, Germany.

V. Willert, J. Eggert, S. Rebhan, C. Goerick, and E. Körner are with the Honda Research Institute Europe GmbH, 63073 Offenbach, Germany.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2008.925657

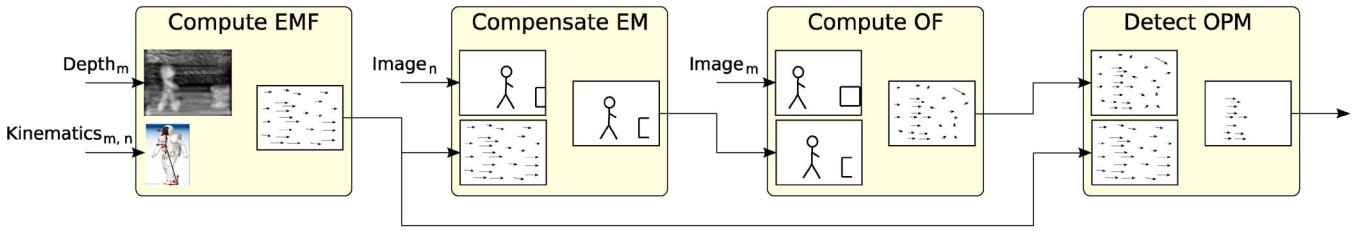


Fig. 1. System overview. Used modules for the computation of OPM.

flow fields, where the underlying models for camera movement highly differ in complexity. An overview of the different approaches and their characteristics is given in [4] and [5]. For proper extraction of the camera movement, it is hence crucial that the OF is primarily caused by EM effects and not by independently moving objects—which we cannot assume for a robot facing moving people.

The effect of translatory EM on points in the image is highly influenced by their distances to the camera in a way that distant points induce smaller flow vectors than closer ones. Some of the approaches actually measure the distances of these points; however, the majority of methods assume that all points lie on a plane running through the position of the camera and the focus of expansion at the horizon. A more detailed explanation is given in [4].

Whereas this procedure can be very suited for the car domain, no such simplifying assumption about the environment is appropriate for autonomously acting robots in dynamic scenes. Thus, different ways to handle EM have to be found. The central idea of using existing knowledge about the movement of the camera was introduced by Lewis [6]. Letting a robot walk in a circle, he computes the OF resulting from the robot's movement over a textured ground. In conjunction with the robot's gait phase and joint angles, a neural network is trained to learn the EMF. Afterward, differences between this EMF and the measured OF can be used to detect obstacles in the path of the robot.

Whereas the movement of the robot in [6] was very constricted, Fidelman *et al.* proved that EMF can be learned even for more complex movements [7]. In their approach, the neural network is provided with the recent OF calculation in addition to the walk phase and joint angles of an AIBO robot. The neural net predicts the flow for the next time step, allowing to compare it with the actually measured OF in order to classify differences as OPM.

Meanwhile, the idea of calculating the movement of the camera and using it to compute the EMF has also been described for the car. In [8], readings from gyroscopic sensors or GPS signals are applied to estimate the camera movement. Likewise to other approaches in cars, the depth is not measured but estimated from the plane assumption, as described earlier.

Another approach using the AIBO platform actually computes the EMF, depending on the measured robot joints instead of using neural networks [9]. However, all the results of the provided approaches are not very promising, which might be due to a missing integration of depth information. In [7] and [10], depth information is implicitly included by providing the calculated OF field, which depends on the distances of points to the camera. Nevertheless, this approach cannot cope properly with the objects having a distance different from the objects in the training phase: The difference in distances results in an OF deviating from the learned one. Hence, even nonmoving points can be classified as OPM.

The combination of knowledge about the robot movement with distance information and OF measurements is described in [11]. Having cameras fixed to the robot, Overett *et al.* try to measure the odometry data and compute the resulting EMF, considering the depth. This flow is afterward subtracted from the estimated OF, and residual vectors are used to indicate the OPM. Unfortunately, the noisy data from odometry force the authors to manually measure the distance passed by the robot, preventing the system from running in real time.

Aside from [12], none of the presented methods precomputes the movement of the camera and searches for movement relative to it. In acquiring the effects of camera movement superimposed with those effects of independently moving objects, the OF needs to analyze a higher range of movement. This does not only increase the computation time but also corrupts OF estimations by producing increased ambiguities.

Fardi *et al.* overcome this problem in their pedestrian detection approach by using a compensation similar to the one in our system in order to cancel out the known EM of a car [12]. The compensation and succeeding estimation of OF are computed on a preselected region of interest, where the depth information consists of a single measurement gained from a laser range scanner.

Very recently, Rabe *et al.* proposed another solution by computing flow and disparity from visual features tracked in the image [13]. Fusing the results in a Kalman filter provides 6-D information about location and motion for a set of points. Together with the readings from inertial sensors, this information is used to compensate the EM effects and obtain the actual 3-D OPM. The results are very convincing, at least for the moderate movement of a car.

However, to our knowledge, no procedure was presented which is able to handle the noisy character of depth, OF, and the fierce effects of camera movement induced by a legged robot, combining them to make reliable estimations of OPM.

B. Neurobiological Evidence

From the previous section, we can conclude that the detection of OPM during EM is a very challenging task. Thus, we have to ask ourselves how humans and most mammals are able to solve it with so much prosperity. For example, humans, running at their top speed, still manage to perceive people standing in some distance, waving their hands, despite the shaking of the

whole body. In this section, we will summarize some evidence from neurobiological and behavioral studies, serving us as inspiration for our proposed method.

An interesting review to the perception of motion is given by Albright [3], where some important stages of motion processing in the brain are explained. These stages seem to be covered to a great extent by the so-called "M pathway," of which we are particularly interested in the processing of two areas in superior temporal sulcus: the middle temporal (MT) visual area and medial superior temporal (MST) area.

There is some evidence that the directional selectivity of neurons in MT can be interpreted as serving the creation of a representation for the retinal flow [3], which is somehow similar to the technical OF. Maunsell and Van Essen found neurons in this area to be selective to binocular disparity [14].

The created retinal flow is analyzed in area MST, where the dorsal subdivision of MST area (MSTd) seems to pay special respect to OF patterns occurring during EM [15]–[18]. For the estimation of EM effects, the distances of points to the observer are again crucial for correct motion estimation: While points further away than the point of fixation move in the same direction as the moving observer, nearer points move in the opposite direction. Roy and Wurtz proofed that MST shows exactly this motion selectivity dependent on binocular depth [19].

In addition, behavioral studies suggest that correct motion estimation also needs to include vestibular information, encoding changes in head direction [20]–[22]. On a neuronal level, this was confirmed by studies with MSTd cells that are reactive during EM in darkness, indicating usage of the vestibular system [23]–[25].

Nevertheless, without one additional aspect, the perception of moving objects would be difficult. During fixation of a moving object, the object is kept in the center of the fovea, resulting in no displacement on the retina. Because objects are perceived as moving even during smooth eye pursuit, the perceived movement must be a superposition of the object's motion on the retina and eye movement. MST neurons were found to perform this kind of summations during eye movement [26] and also for static eye positions [27].

We do not try to model the neural activities in detail, but we would rather argue in favor of understanding and transferring the abstracted biological principals. We will show that following this paradigm leads to a system which is advantageous for dynamic behavior in dynamic scenes. For example, supposing the system would not incorporate the knowledge of the head movement, it would have to estimate it based on the flow and depth information—with the described drawbacks in dynamic scenes. Moreover, without the incorporation of depth, the resulting system could not perform better than the ones using model assumptions of the depth. Finally, without the separation of EM estimation, the system could not perceive movement at all while moving itself.

II. Approach

In the following, we will describe the details of our approach, starting with the computation of the EMF in Section II-A. The compensation of EM and the estimation of the relative OF are afterward explained in Section II-B. In Section II-C, we present the filtering applied to the relative flow.

A. Computation of EMF

The computation of EMF results in a flow field $E^m = \{\mathbf{e}_i^m\}$ indexed by i, which describes where each point in image I^m at time m has shifted to in I^n at time n, caused by the movement of the camera. For this computation, it is assumed that the environment is static and the points did not move themselves. To account for the fact that the absolute values of the vectors $\mathbf{e}_i^m$ highly depend on the distances of the corresponding points to the camera, the effect of camera movement is calculated in a 3-D camera-related space. Therefore, we first combine each 2-D point (x_i^m, y_i^m) in the image with its binocular disparity d_i^m to define a 3-D point in homogenous image coordinates as

$$\mathbf{p}_{I,i}^m = (x_i^m, y_i^m, d_i^m, 1)^{\mathrm{T}}. \tag{1}$$

This point can then be reconstructed in camera coordinates by computing the homogenous transformation matrix $\mathbf{T}_{C\leftarrow I}$ from image to camera coordinates analogous to [28] and by multiplying $\mathbf{p}_{I,i}^m$ with it

$$\mathbf{p}_{C,i}^m = \mathbf{T}_{C\leftarrow I}\mathbf{p}_{I,i}^m. \tag{2}$$

To compute the position of each point relative to the camera at the next time step, we have to know how the camera moved. This knowledge is obtained from the robot kinematics. Therefore, we use the robot state vector $\mathbf{s}^m$ which contains the position of the upper body, as well as the angles of all joints at time m [29]. With this vector, we can compute the forward kinematics, as presented in [30], returning us the transformations from the initial world-coordinate system to each robot segment, including the transformation from world to camera coordinates $\mathbf{T}_{C\leftarrow W}^m$. Inverting this transformation results in $\mathbf{T}_{W\leftarrow C}^m$, which describes the transformation from camera to world coordinates. In addition, we can compute the current transformation $\mathbf{T}_{C\leftarrow W}^n$ by applying $\mathbf{s}^n$. Because this calculation is with respect to an initial world point, the resulting transformations also encode camera displacements caused by the movement of the robot base.

Assuming a static point $\mathbf{p}_{C,i}^m$, the new position $\mathbf{p}_{C,i}^m$ can be calculated from transferring the point to world coordinates by multiplying the last transformation $\mathbf{T}_{W\leftarrow C}^m$ and passing it back to camera coordinates with the actual transformation $\mathbf{T}_{C\leftarrow W}^n$

$$\mathbf{p}_{C,i}^n = \mathbf{T}_{C\leftarrow W}^n\mathbf{T}_{W\leftarrow C}^m\mathbf{p}_{C,i}^m. \tag{3}$$

This process is shown in Fig. 2.

Because we are not interested in the new 3-D-position of the points but rather in the 2-D-flow in the image, we have to project each point back into the image plane I^n. We do this by computing a homogenous projection matrix $\mathbf{T}_{I\leftarrow C}$, as described in [28], and multiply the new point $\mathbf{P}_{C,i}^n$ with it

$$\mathbf{p}_{I,i}^n = \mathbf{T}_{I\leftarrow C}\mathbf{p}_{C,i}^n. \tag{4}$$

Note that the spatial constance of the matrices $\mathbf{T}_{C\leftarrow I}$, $\mathbf{T}_{W\leftarrow C}^m$, $\mathbf{T}_{C\leftarrow W}^n$, and $\mathbf{T}_{I\leftarrow C}$ allows us to precompute their product,

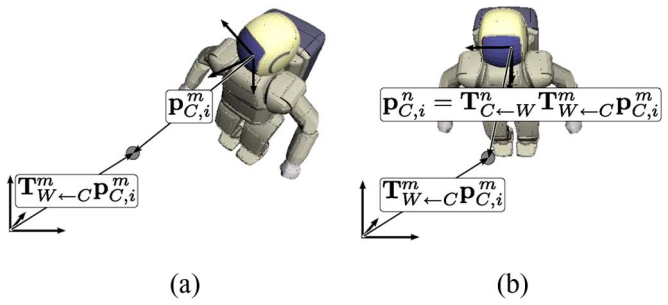


Fig. 2. (a) Point $\mathbf{p}^m_{C,i}$ in camera and world coordinates. (b) Description of the static point at time n, using the camera transformations.

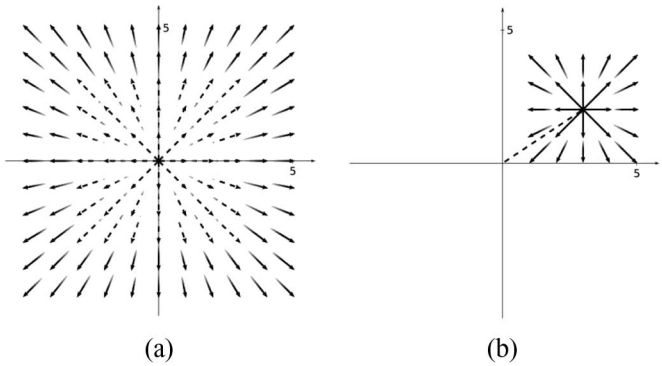


Fig. 3. (a) Spectrum of the OF motion if EM effects are not canceled out in advance. The solid line visualizes an assumed OPM of two-pixel displacement magnitude in each direction, and the dashed line denotes an assumed EM effect of three-pixel displacement. The interval which has to be acquired by the OF is a superposition of the two. (b) Reduced spectrum for searching relative to EMF.

resulting in one single transformation matrix for each time step, which is multiplied with each point $\mathbf{p}^m_{I,i}$.

Knowing $\mathbf{p}^n_{I,i}$ and $\mathbf{p}^m_{I,i}$, the shift of a point in the image plane due to camera motion can now be expressed by

$$\mathbf{e}^m_i = \begin{pmatrix} x^n_i - x^m_i \\ y^n_i - y^m_i \end{pmatrix}. \tag{5}$$

The reliability of this shift vector depends on the quality of the depth component z^m_i of $\mathbf{p}^m_{C,i}$ and the kinematics accuracy for the robot's movement estimation from time m to n.

For the latter, we assume an increasing error with increasing movement amplitude. By expressing the camera translation in this time interval with $\mathbf{t}^m_d$ and its rotation with $\mathbf{t}^m_r$, we can approximate the kinematics-based variance $(\sigma^m_k)^2$ by

$$(\sigma^m_k)^2 = \|\mathbf{t}^m_d\|^2 + \|\mathbf{t}^m_r\|^2. \tag{6}$$

Note that this variance is spatially independent and only depends on the robot's movement.

The credibility of the depth measurement z^m_i relates to the amount of correlation found by the disparity algorithm [31] between the left and the right camera image. This is expressed in terms of some confidence value $c^m_i \in [0,1]$. In addition, we have to account for the decreasing accuracy with increasing distance. In [32], it is shown that sensitivity for depth estimation z from disparity d decreases with the squared distance

$$\frac{\partial d}{\partial z} = -\frac{b \cdot f}{z^2 \cdot q} \tag{7}$$

where b denotes the baseline, f is the focal length, and q is the pixel size of the camera.

Because we want to describe the likelihood for $\mathbf{e}^m_i$ in terms of some Gaussian probability distribution, we express a decay of reliability in terms of an increasing covariance $\mathbf{\Sigma}^m_{e,i}$ by subsuming the different aspects of credibility

$$\mathbf{\Sigma}^m_{e,i} = \mathbf{1} \cdot w_e \left(w_k \left(\sigma^m_k\right)^2 + w_d \sigma_d \left(z^m_i\right)^2 + w_c \sigma_c \left(c^m_i\right)^2 \right) \tag{8}$$

with

$$\begin{aligned} \sigma_d(z)^2 &= \frac{z^2 \cdot q}{b \cdot f} \qquad (9) \\ \sigma_c(c)^2 &= (c + k_1)^2 + k_2, \\ k_1 &= \frac{1}{2}\left(\sigma_d(z_{\min})^2 - \sigma_d(z_{\max})^2 - 1\right) \\ k_2 &= \sigma_d(z_{\max})^2 - k_1^2. \qquad (10) \end{aligned}$$

The identity matrix is denoted by $\mathbf{1}$, and $z_{\min}$ and $z_{\max}$ are the minimum and maximum assumed distances, a behaviorally relevant point might have from the camera. The constants k_1 and k_2 in (10) use these distances to ensure that the variance $\sigma_c(c=0)^2$ equals $\sigma_d(z=z_{\max})$ and $\sigma_c(c=1)^2$ equals $\sigma_d(z=z_{\min})$, respectively. This is useful because it adjusts the scales of both variables to each other.

The factors w_c, w_d, and w_k are utilized to weigh the individual influence of ambiguities occurring during the measurement of confidence, distance, or kinematics. They should be based on the present scene, the used disparity algorithm, and kinematics precision. Finally, the scalar w_e is used to adjust the range of the whole covariance $\mathbf{\Sigma}^m_{e,i}$ to the covariance $\mathbf{\Sigma}^m_{r,i}$ computed for the OF (see hereafter).

B. Computation of Flow Relative to EMF

Because we have computed the effects of EM on the image, we are now able to compute the OF $R^m = \{\mathbf{r}^m_i\}$ between I^m and I^n relative to this EMF. In comparison to other approaches which cancel out the EM effects after the computation of the OF, this reduces the spectrum of the motion to be acquired. This effect, which is shown in Fig. 3, does not only reduce computation time but also improves the OF estimations by reducing ambiguities, a fact which is evaluated in Section III.

We can compensate the EM effects by warping the images. The two possible approaches of forward and inverse mapping are discussed in detail in [32].

In our case, forward mapping can be written as

$$\tilde{I}^m\left(\mathbf{p}^m_{I,i} + \mathbf{e}^m_i\right) = I^m\left(\mathbf{p}^m_{I,i}\right) \tag{11}$$

where $\tilde{I}^m$ is equal to the old image I^m but it is freed from the measured EM effects. This kind of forward mapping has two major drawbacks. Because $\mathbf{e}^m_i$ usually encodes real-valued shifts, the data points $\mathbf{p}^m_{I,i} + \mathbf{e}^m_i$ may not lie inside the grid and require complicated interpolations. In addition, it is not guaranteed that each point in the warped image is targeted by the sum of the original position and shift, leading to holes in the image.

(a)

(b)

Fig. 4. (a) Disrupted results of pixelwise warping under extreme body movements. (b) Effect of filling-in holes in the EMF with the averaged flow.

The more convenient solution lies in the usage of backward mapping to warp the actual image I^n back to $\tilde{I}^n$

$$\tilde{I}^n\left(\mathbf{p}_{I,i}^m\right) = I^n\left(\mathbf{p}_{I,i}^m + \mathbf{e}_i^m\right). \tag{12}$$

Because the data points are passed as arguments for the resulting image $\tilde{I}^n$, holes in the image cannot occur. The problem of real-valued shifts is tackled by using bilinear interpolation in the source image. Aside from the deviations caused by inaccurate depth measurements, $\tilde{I}^n$ and I^m should only differ in points with individual object motion.

Nevertheless, in some situations with extreme body movement, holes in the depth image D^m with $c_i^m = 0$ can cause artifacts in the pixelwise warped image, as shown in Fig. 4(a). These artifacts lead to errors in the computed OF, and hence, they can affect surrounding regions even if those regions have valid depth estimations. In a first step, these artifacts are reduced by filling invalid regions with suitable assumptions obtained from valid EMF estimations. We utilize a rather simple model by taking the average EMF $\hat{v}_E^m$ gained from valid depth estimations to fill the gaps [see Fig. 4(b)]. To account for the decreased reliability of the warped image, and hence of the OF, we create a penalty map P, which is high for points near- and inside invalid depth regions and zero otherwise

$$P\left(\mathbf{p}_{I,i}^m\right) = \begin{cases} \infty, & \text{if } \exists j \in \Omega_i \text{ with } c_j^m = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Ω_i denotes the indices of all points in an eight neighborhood around $\mathbf{p}_{I,i}^m$.

By passing I^m and $\tilde{I}^n$ to the OF algorithm described in [33], we get a velocity estimation $\mathbf{r}_i^m$ for each point, which is relative to the estimated EMF. The algorithm also computes a covariance $\mathbf{\Sigma}_{C,i}^m$, which gives a confidence measure for the OF vectors, assuming pixelwise independent Gaussian noise.

For the inclusion of the warping-based penalty, we compute the compound variance $\mathbf{\Sigma}_{r,i}^m$ at point $\mathbf{p}_{I,i}^m$ as

$$\mathbf{\Sigma}_{r,i}^m = \mathbf{1} \cdot P\left(\mathbf{p}_{I,i}^m\right) + \mathbf{\Sigma}_{C,i}^m. \tag{14}$$

For invalid points, the choice of $\mathbf{1} \cdot P(\mathbf{p}_{I,i}^m) \gg \Sigma_{C,i}^m$ ensures a negligible influence of the confidence-based variance. This is necessary because the artifacts in the warped image can create artificial edges and thereby decrease $\mathbf{\Sigma}_{C,i}^m$ locally.

The used OF algorithm [33] realizes probabilistic prediction over time, considering spatial relations for the transition. This enables the system to iteratively make reliable calculations of motion in unstructured image regions by taking the previous estimations into account. The ideal outcome of this algorithm would be a vector field, which is zero for unmoving objects and otherwise denotes their proper motion.

However, because the OF and the depth measurements from disparity are very noisy signals, we need some more filtering for the detection of OPM. This is described in the following section.

C. Detecting OPM

In the recent sections, we introduced two approaches for the calculation of image flow, which use very different methods and hence show different characteristics. Because the computation of the EMF is based on disparity and kinematics, it can acquire the effects of EM on image points, as long as the points are not moving. In contrast, the OF also works for moving points. By adding the EMF to the relative OF, we get an overall flow, which designates the compound retinal movement of OPM and EM effects. For unmoving points, this flow should equal the EMF, whereas it should be different for moving points. In this section, a measurement for the significance of this distance is introduced and used to extract OPM vectors from the OF.

This measurement is derived from a stochastic assumption about the estimated flows, the depth, and the images. Therefore, we have to define some stochastic variables for each point $\mathbf{p}_{I,i}$ in the image, describing a distribution of all measurements.[1] Each vector of the overall flow mentioned earlier is represented by the variable ϑ_i, whereas ϵ_i describes the EM vector for each point $\mathbf{p}_{I,i}$. The random variables $I = \{I^m, I^n\}$ and $D = D^m$ specify the observed source and depth images.

[1]This procedure is identical for each time step; thus, we drop the time indices for convenience.

The principal idea for the approach is to estimate the probability of measuring the same velocity ν_i from the EMF and the compound OF, assuming the corresponding point is static. That is, if some point did not move, the velocity described by ϵ_i should not differ too much from ϑ_i, and the likelihood to measure some identical velocity ν_i from both methods should be high. In contrast, a moving point results in different outcomes for ϵ_i and ϑ_i, and the likelihood to measure the same velocity ν_i from the two methods is very low. Concluding a high joint probability $\rho(\vartheta_i = \nu_i, \epsilon_i = \nu_i, I, D)$ indicates a static point, whereas a low probability marks a moving point.

Reflecting our assumptions about the dependencies of ϵ_i, ϑ_i, D, and I, this joint distribution can be decomposed to the following:

$$\rho(\vartheta_i = \nu_i, \epsilon_i = \nu_i, I, D) = \rho(\vartheta_i = \nu_i|\epsilon_i, I) \, \rho(\epsilon_i = \nu_i|D) \, \rho(I) \, \rho(D). \quad (15)$$

Because we make no prior assumptions about the source images and the depth, the corresponding variables are uniformly distributed, and hence, they have a negligible influence on the distribution. Using the precomputed results for the EMF and OF, we can approximate the conditional distributions with the following:

$$\rho(\vartheta_i = \nu_i|\epsilon_i, I) \propto \mathcal{N}_{\nu_i}(\mathbf{r}_i + \mathbf{e}_i, \mathbf{\Sigma}_{r,i}) \quad (16)$$

$$\rho(\epsilon_i = \nu_i|D) \propto \mathcal{N}_{\nu_i}(\mathbf{e}_i, \mathbf{\Sigma}_{e,i}). \quad (17)$$

Adding $\mathbf{e}_i$ in (16) accounts for the warping of the image—whereas $\mathbf{r}_i$ encodes a flow relative to the EMF, the sum of $\mathbf{e}_i$ and $\mathbf{r}_i$ makes it an absolute flow and allows the comparison with the EMF. With this approximation, the joint distribution is proportional to the product of two Gaussians, being defined as

$$\mathcal{N}_{\nu_i}(\mathbf{r}_i + \mathbf{e}_i, \mathbf{\Sigma}_{r,i}) \mathcal{N}_{\nu_i}(\mathbf{e}_i + \mathbf{\Sigma}_{e,i}) = L_i \cdot \mathcal{N}_{\nu_i}(\mathbf{c}_i, \mathbf{C}_i) \quad (18)$$

with

$$\begin{aligned} \mathbf{c}_i &= \mathbf{\Sigma}_{r,i}(\mathbf{\Sigma}_{r,i}\mathbf{\Sigma}_{e,i})^{-1}\mathbf{e}_i + \mathbf{\Sigma}_{e,i}(\mathbf{\Sigma}_{r,i}\mathbf{\Sigma}_{e,i})^{-1}(\mathbf{r}_i + \mathbf{e}_i) \\ \mathbf{C}_i &= \mathbf{\Sigma}_{r,i}(\mathbf{\Sigma}_{r,i}\mathbf{\Sigma}_{e,i})^{-1}\mathbf{\Sigma}_{e,i} \\ L_i &= \mathcal{N}_{\mathbf{e}_i}(\mathbf{r}_i + \mathbf{e}_i, \mathbf{\Sigma}_{r,i} + \mathbf{\Sigma}_{e,i}). \end{aligned}$$

A visualization of this product is shown in Fig. 5. The mean value $\mathbf{c}_i$ of the resulting distribution can be interpreted as that identical velocity, which is most likely to be measured by both algorithms—the EMF and the OF, while finding a value, which fits the hypothesis of a common velocity best, is always possible, the factor L_i is a measure to describe how well $\mathbf{c}_i$ actually fits in the light of the calculated displacements and variances. For the evaluation of OPM, we are not interested in the value of the vector $\mathbf{c}_i$ but whether such vector is likely to occur. Thus, the rejection of OF estimations is based on this value L_i and can be further simplified by applying the logarithm to L_i

$$\begin{aligned} L_i &= \mathcal{N}_{\mathbf{e}_i}(\mathbf{r}_i + \mathbf{e}_i, \mathbf{\Sigma}_{r,i} + \mathbf{\Sigma}_{e,i}) \\ &= z \cdot e^{-\frac{1}{2}(\mathbf{e}_i - (\mathbf{r}_i + \mathbf{e}_i))^{\mathrm{T}}(\mathbf{\Sigma}_{r,i} + \mathbf{\Sigma}_{e,i})^{-1}(\mathbf{e}_i - (\mathbf{r}_i + \mathbf{e}_i))} \\ &\propto -\mathbf{r}_i{}^{\mathrm{T}}(\mathbf{\Sigma}_{r,i} + \mathbf{\Sigma}_{e,i})^{-1}\mathbf{r}_i. \end{aligned} \quad (19)$$

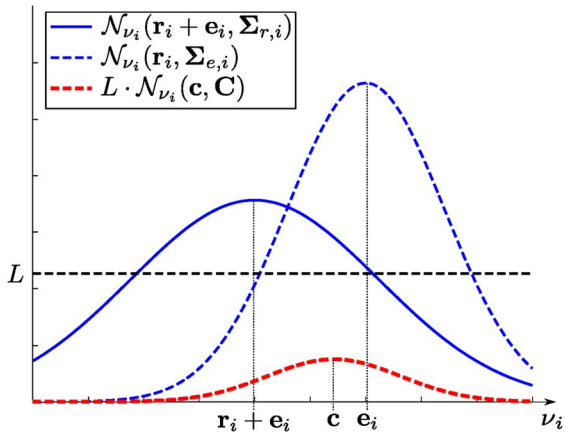


Fig. 5. One-dimensional plot of the resulting distribution $L_i \cdot \mathcal{N}_{\nu_i}(\mathbf{c}_i, C_i)$ from $\mathcal{N}_{\nu_i}(\mathbf{r}_i + \mathbf{e}_i, \sum_{r,i})$ and $\mathcal{N}_{\nu_i}(\mathbf{e}_i, \sum_{e,i})$, as well as the likelihood L.

That is, the decision whether some optically measured velocity is classified as OPM is based on the absolute value of that velocity scaled by the variances of the EMF and the OF. Defining $\mathbf{r}_i^{\mathrm{T}}(\Sigma_{r,i} + \Sigma_{e,i})^{-1}\mathbf{r}_i$ as squared Mahalanobis norm $\|\mathbf{r}_i\|_M^2$, we can conclude that a big Mahalanobis norm indicates a moving point; thus, we rely on the computation of the OF. Formally, this can be expressed as

$$\mathbf{o}_i = \begin{cases} \mathbf{0}, & \text{if } \|\mathbf{r}_i\|_M < \theta_{\mathrm{M}} \\ \mathbf{r}_i, & \text{otherwise.} \end{cases} \quad (20)$$

III. Experiments and Evaluation

In this section, we would like to show the feasibility of our approach by demonstrating its abilities in a scenario where the combination of our cues is highly recommendable. This scenario shows typical interaction with ASIMO and makes the detection of OPM a difficult task. Nevertheless, we evaluate our results quantitatively and qualitatively, showing robust OPM measurements unseen for legged robots.

We will also evaluate the system design by analyzing the results of the different steps and their integration for the detection of OPM. Further on, the benefits of our step-by-step movement estimation will be demonstrated by comparing computation time and results with the outcome of an overall estimated OF.

A. Experiment Description

The experiments are carried out with a Honda ASIMO robot, as presented in [34]. The computation is performed on a Pentium 4 single core with 3.4 GHz, and the images are captured with a constant frame rate of 12 Hz.

In our scene, ASIMO is initially located in front of one person P1 (black shirt), standing at a distance of approximately 2.8 m. A second person P2 (white shirt) is approaching the robot from behind, and both people walk to the right, passing ASIMO's view field (see Fig. 6). Meanwhile, ASIMO is walking forward on a path that can roughly be described as an inverted S-shape, which is superimposed with a rotation of the body at about 45° in the second half. From the endpoint, he walks backward toward its starting position, turning his body

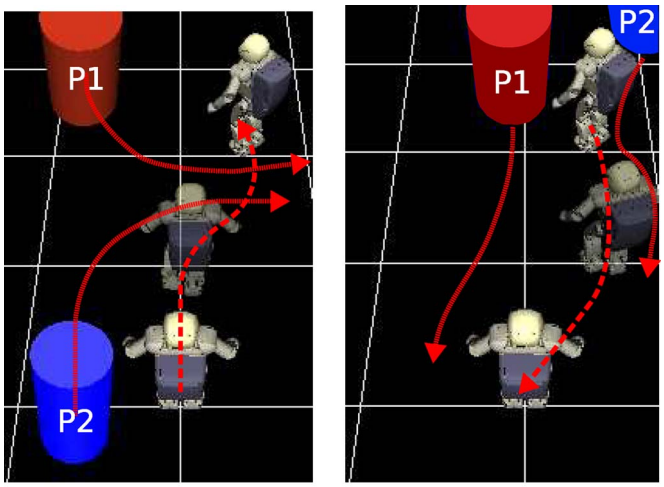


Fig. 6. Visualization of ASIMO's walking path by the dashed line and those of persons P1 and P2 by the solid lines. (a) Scenario during ASIMO's forward movement. (b) During its backward movement.

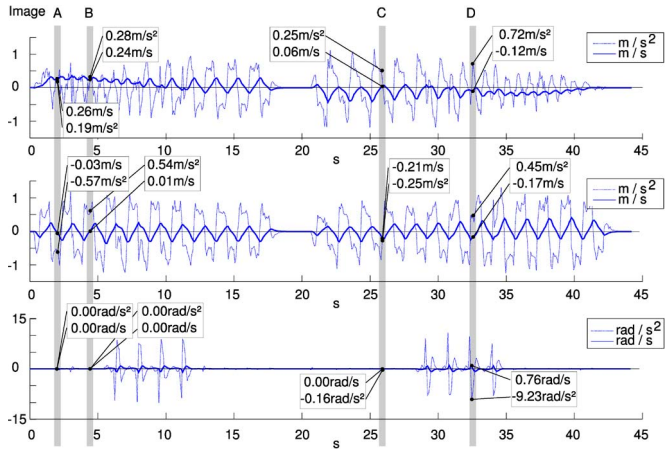


Fig. 7. Plot of velocities and acceleration for translatory (top row) x- and (center row) y-movement, as well as rotation around the (bottom row) z-axis. The movement in these plots depicts the movement of the left camera. The time steps where the evaluation images were taken are highlighted by gray vertical bars.

straightforward again while P2 is approaching and passing him to the right.

To show the high dynamic movement occurring during the experiments, we plot the camera movement and acceleration over time in Fig. 7. The plot shows the translatory movement in the x- and y-direction, as well as the rotation around the z-axis of the left camera.[2]

The important aspect in this scenario is that the view field is dominated by moving people or objects. This makes a modeling of depth or the environment, as described in Section I-A, inappropriate because each person would violate the expectation. Due to the dominance of OPM in the view field, the estimation of camera movement from OF would also not be possible without segmentation information.

For the evaluation, we extract four images from the captured stream, which are shown in column 1 in Fig. 11. They are chosen to represent the different aspects of interaction scenarios. The first image (1, a) is used to evaluate the system's ability to detect movement at a high distance (about 2.8 m) and taken while the robot is slowly moving forward. Image (1, b) shows the two people walking from the left side of the image to the right, with P2 walking closer to the robot than P1. This scene is considered as key scene, because both people are within the interaction range and separable based on their different velocities. In addition, ASIMO is walking forward with moderate speed. The pictures (1, c) and (1, d) are captured while ASIMO is walking backward and stepping from one foot to the other. Image (1, c) is chosen to examine the ability of detecting motion for considerable small body parts. During image (1, d), the robot abruptly performs an additional rotation of the upper body. This scene will demonstrate the system's ability to handle jerky camera rotation (see Fig. 7), and it will also be used to evaluate the benefits of a step-by-step movement estimation.

B. Parameter Evaluation

The overall aim of the system is the detection and exact estimation of OPM relative to the EM for all points in the image. To evaluate the influence of different parameters on the system's performance, we did the following investigations. The choice for the motion range to be captured by the system is based on the magnitude of movement in our scene. To cope with people passing very close to the robot, the acquired range of displacements for the OF is chosen in a range from $[-10, 10]$ in the x-direction to $[-2, 2]$ in the y-direction, using an image resolution of 200×150 pixels. For the calculation of the overall OF, i.e., the EM-uncompensated flow, the displacements are chosen in the range from $[-22, 22]$ in the x-direction to $[-4, 4]$ in the y-direction, accounting for the fact that this flow has to acquire the maximum range of movement occurring in the stream, which is composed of EM effects superimposed with OPM. The classification of OPM as in (20) depends on the relative flow $\mathbf{r}_i$, the variances $\mathbf{\Sigma}_{r,i}$, and $\mathbf{\Sigma}_{e,i}$, as well as the threshold θ_M. Because $\mathbf{r}_i$ and $\mathbf{\Sigma}_{r,i}$ are measured, and hence do not depend on any parameter, we will focus on the evaluation of $\mathbf{\Sigma}_{e,i}$ and θ_M.

$\mathbf{\Sigma}_{e,i}$ is based on the choice for w_e, w_k, w_d, and w_c. Setting $w_e = 0.07$ adjusts the scale of $\mathbf{\Sigma}_{e,i}$ to match the measured $\mathbf{\Sigma}_{r,i}$. In our evaluation, we neglected the influence of the kinematics-based variance by assigning $w_k = 0$. This accounts for the fact that the worst case error from kinematics is lower than 2 mm, owing to ASIMO's high-precision encoders and its stiffness. In comparison to this, the resolution in distance estimation from disparity drops to 69 mm for a point that is 2.5 m away from the camera,[3] assuming that the disparity algorithm found the right displacement. The error increases by 69 mm for each one pixel of wrong displacement.

The remaining parameters w_d, w_c, and θ_M are chosen based on the receiver operating characteristic (ROC) of the system in order to find the optimal parameter settings for OPM detection.

[2]The movement of the remaining components is negligible and hence not shown.

[3]From (7): $b = 74$, $q = 0.004$, $f = 4.902$, and $\partial d = 1$ result in $\partial z = 68.9$ (neglecting dimensions).

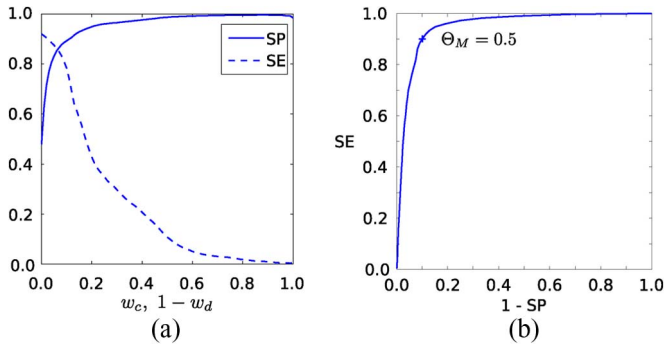


Fig. 8. (a) Systems SP and SE dependent on the weights w_c and w_d. (b) Plot of SE and SP dependent on θ_M.

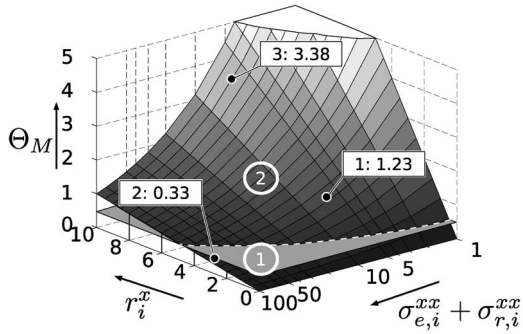


Fig. 9. Classification border dependent on the relative velocities, accumulated variances, and θ_M.

Therefore, we create some ground-truth data $G(i) \in \{0, 1\}$ from the four images in Fig. 11, indicating for each pixel i whether it moved [$G(i) = 1$] or not [$G(i) = 0$]. We also define our system's output as

$$\Phi(i) = \begin{cases} 1, & \text{if } \mathbf{o}_i = \mathbf{r}_i \\ 0, & \text{otherwise.} \end{cases}$$

Further on, the sensitivity SE and specificity SP are defined as

$$SE = \frac{TP}{TP + FN} \tag{21}$$

$$SP = \frac{TN}{TN + FP} \tag{22}$$

with

$$\begin{aligned} TP &= |\{i | G(i) = 1 \cap \Phi(i) = 1\}| && \text{(true}-\text{positive)} \\ FP &= |\{i | G(i) = 0 \cap \Phi(i) = 1\}| && \text{(false}-\text{positive)} \\ TN &= |\{i | G(i) = 0 \cap \Phi(i) = 0\}| && \text{(true}-\text{negative)} \\ FN &= |\{i | G(i) = 1 \cap \Phi(i) = 0\}| && \text{(false}-\text{negative).} \end{aligned}$$

For the evaluation of w_c, w_d, and θ_M, we use an iterative procedure. Because the threshold θ_M depends on the choice of w_c and w_d, whereas these weights are determined based on the system's output which depends again on θ_M, there is circular dependence between the three evaluated parameters. For this reason, we start with a fixed $\theta_M = 0.1$, rejecting almost no velocities, and vary $w_c \in [0, 1]$ and w_d accordingly by choosing $w_d = 1 - w_c$. The resulting SE and SP for these values, as shown in Fig. 8(a), show a rapid decline of SE with increasing w_c and decreasing w_d. Hence, $\sigma_d(z)^2$ appears to be a more suitable approximation of the EMF variance than $\sigma_c(c)^2$. Those w_c and w_d which elicit equal values for SE and SP are considered as optimum, because they represent a tradeoff between a high true-positive rate and a high true-negative rate. As shown by the intersection of SP and SE in the plot, this results in a choice of $w_c = 0.1$ and $w_d = 0.9$, respectively.

For θ_M, we use these determined weights and vary $\theta_M \in [0, 5]$. The resulting ROC curve in Fig. 8(b) shows a plot of $1 - SP$ against SE. Again, the best choice for θ_M is determined from equal SE and SP values, resulting in $\theta_M = 0.5$ for our scenario. The described process is repeated iteratively until convergence, which is achieved after two iterations in our experiments. In addition, the obtained parameters have also been validated on a large-range image sequence by means of visual inspection.

The described evaluation shows that the choice for θ_M is crucial for the system's performance; thus, we also perform some qualitative analysis of this threshold, which is useful for cases in which no ground truth is available. Fig. 9 shows the relation of the x-component[4] of relative velocities r_i^x, accumulated variance $\sigma_{e,i}^{xx} + \sigma_{r,i}^{xx}$, and threshold θ_M. The surface labeled by a white "2" shows the maximal θ_M value for each combination of relative velocity and accumulated variance, which would classify this specific combination as OPM. For example, the point 1 with $r_i^x \approx 3$ and $\sigma_{e,i}^{xx} + \sigma_{r,i}^{xx} \approx 5.6$ is rejected from OPM with $\theta_M > 1.23$. The points 1, 2, and 3 correspond to those points shown in Fig. 10.

The plane labeled by the white "1" visualizes the choice $\theta_M = 0.5$: Those points on the mesh lying higher than this plane are classified as OPM, those below are rejected.

In conclusion, the choice for θ_M can simply be done by choosing one combination of relative flow and variance, which should serve as a classification border.

C. Quantitative Results

We use the ROC of the θ_M evaluation as quantitative quality measure for the system's ability to detect OPM. The accuracy of the OPM is not evaluated for two reasons. Assuming an accurately estimated EMF, the accuracy of OPM is dependent on the OF precision, which is investigated in [33]. In case of an inaccurate EMF, the resulting OPM would not designate pure object motion but a mixture with effects from EM. As can be seen hereafter, this does not seem to be the case in our scenario, otherwise the OF would contain systematic errors.

To summarize the overall system performance in one single value, we compute the area below the ROC curve in Fig. 8(b). Our system achieves a value of 0.92, where 1.0 would indicate an optimal classifier. This value is quite high despite the fact that the patchwise computation of disparity and flow cannot determine exact object borders but always tends to surround objects.

By setting $\theta_M = 0.5$, an SE of 0.89 and an SP of 0.90 are achieved, i.e., 89% of all moving pixels are detected, and 90% of all rejected points are actually not moving.

[4]The results are analogous for the y-component.

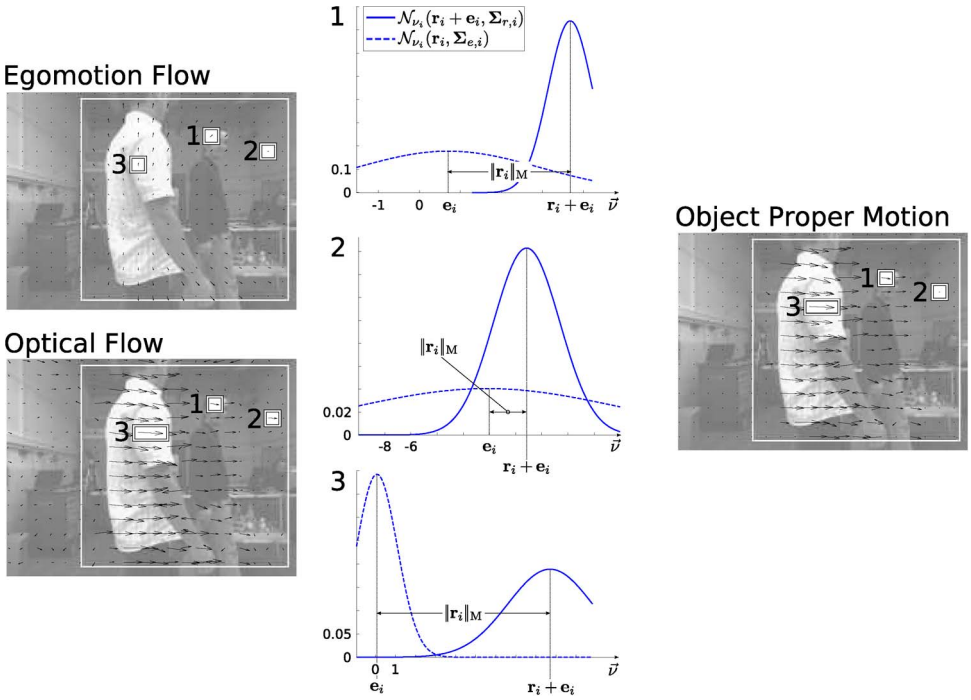


Fig. 10. This figure shows the computed EMF, relative OF, and OPM. The direction of movement is visualized by arrows, whereas arrow length and gray value visualize the speed. Exemplarily, the velocities of three points are highlighted, and their representing normal distributions are shown. The Mahalanobis norm used for filtering is visualized qualitatively as the distance between the normal distributions.

One iteration of the system lasts about 400 ms. The vast majority is spend on the computation of the OF estimation, which takes in an average of 387 ms whereas the computation of EMF and warping and detection of OPM take less than 5 ms each.

D. Qualitative Results

Fig. 10 shows the computed velocities and the rejection mechanism exemplarily for the three points 1, 2, and 3 of our key scene. An overall inspection of the results and their quality is given later on.

The EM vector for point 1 in the image reflects the forward movement of the robot by showing a displacement pointing away from the center of expansion located on person P2. Because the point is relatively far away, the EM vector has a small absolute value and also shows a high variance in the plotted normal distribution. The corresponding OF vector represents a slow movement to the right and exhibits a small variance due to the high amount of image structure in the surroundings. This small variance leads to a high Mahalanobis norm in (20), so that the OF vector is maintained as OPM.

The second point lies on the wall, which is more than 9 m away from the camera. Hence, $\mathbf{e}_i$ is almost zero, and it is accompanied by a high variance. Because the lack of image structure in this region causes also a high variance of the OF, the Mahalanobis norm is small even so the Euclidean distance between $\mathbf{e}_i$ and $\mathbf{r}_i$ is similar to the one for point 1. Accordingly, this flow vector is rejected.

Point 3 is clearly indicated as OPM, resulting from a large distance between $\mathbf{r}_i$ and $\mathbf{e}_i$ in combination with a very low variance $\Sigma_{e,i}$ which accounts for the small distance to the camera.

To visualize the results for the entire view field, Fig. 11 shows the four left camera images taken from the described stream of 571 images.

The depth dependence of the EMF prevents the system from operating on areas with a lack of depth information. At the left side of the image, this originates from the displacement shift used for disparity computation, whereas the right, top, and bottom borders are the effect of image rectification. The remaining area is marked by a white rectangle in the gray images.

Columns 2, 3, and 4 in Fig. 11 visualize the different flow fields for these marked areas. The gray value in these images represents the flow magnitude for each pixel, whereas the exact flow vector is shown for every 17th pixel in the x- and y-direction.

The radial expanding EMF typical for pure forward translatory movement is clearly visible in image (2, b), whereas the slow robot movement during scene A causes this flow to be visible only for very close points at the bottom. The flow field in (2, c) is characterized by the robot's translation to the right. The depth dependence of the EMF is visible by showing larger velocities for points closer to the camera than for more distant points. Despite the robot's translation to the right during image (1, d), the associated flow field in (2, d) is dominated by the rotation of the robot's upper body to the left. Because the EMF for purely rotational movement does not depend on depth, this flow contains almost equal velocity vectors to the right. However, it is observable in all scenes that the EMF caused by close objects occupies areas which are bigger than the objects themselves. This is the already mentioned effect of the patchwise disparity estimation. The holes in the EMF, which are particularly visible in image (2, d), are the results of insufficient texture.

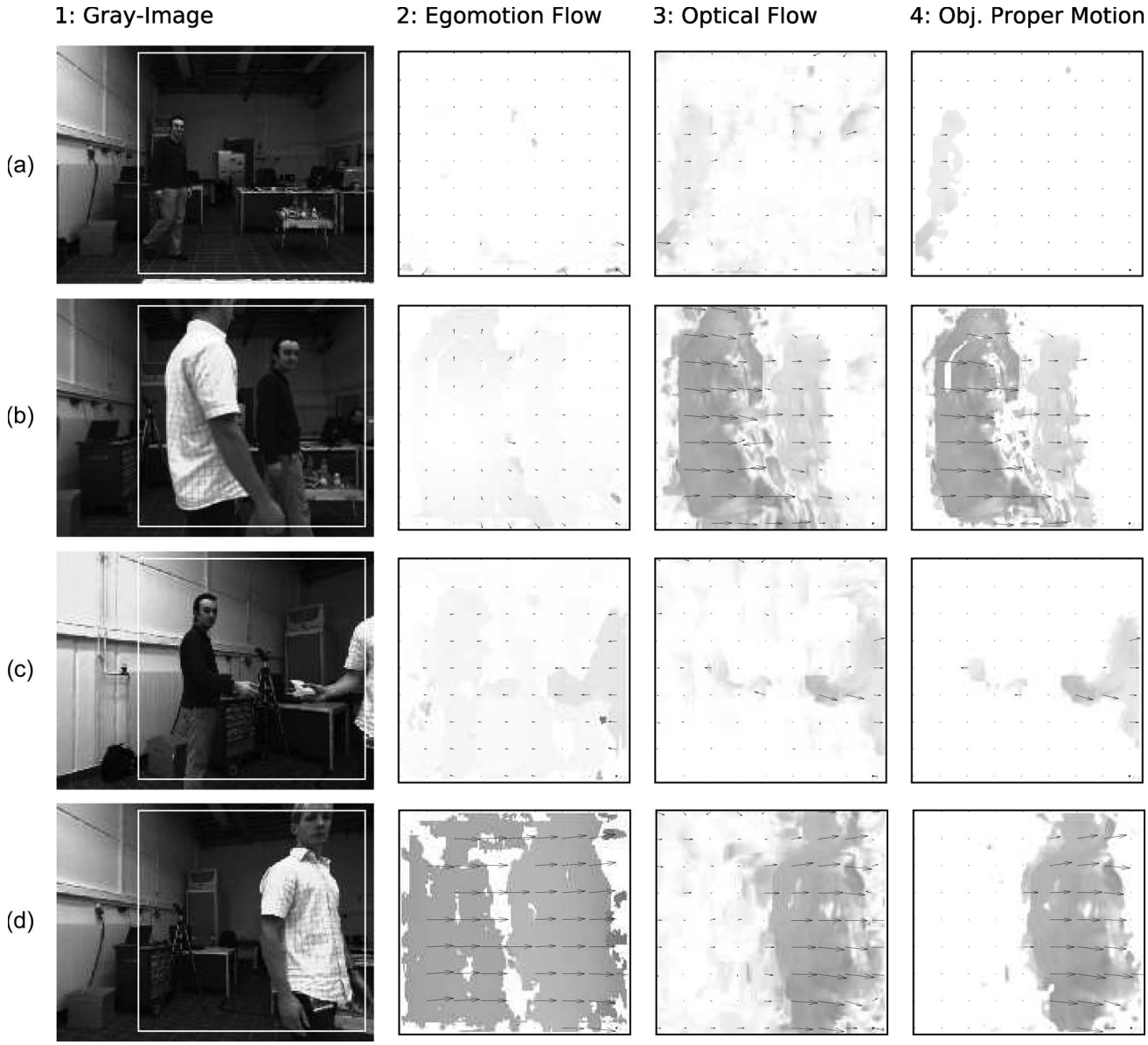


Fig. 11. Column 1 shows the different images captured from the stream. The EMF, OF, and OPM for the area surrounded by the white rectangle are presented in columns 2, 3, and 4 at the right of their corresponding gray images.

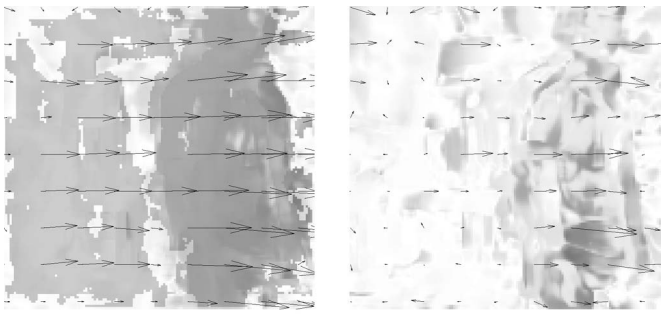
The relative OF in column 3 clearly acquires the people's movement to the right in images a, b, and d. It also captures the divergent movement of the arms of P1 and P2 in image c, which is caused by handing over the stamper from the left to the right person.

All four flow fields include false-positive velocities. For fast movement as in scenes b and d, this originates from the flow spatiotemporal prediction which makes person P2 drag a "trail" of movement behind it (for further explanations of this effect, see [33]). The remaining errors can be classified as correspondence problems occurring at straight borders and textureless regions.

The OPM shown in the last column contains almost none of these false positives, except for some small patches in image (4, d). In particular, the described "trail" of movement is removed. Moreover, the amount of false negatives is considerably low and only visible in image (4, b) for some areas on the arm of person P2.

The qualitative effect of the preceding EM compensation and subsequent OF estimation in contrast to the computation of the absolute OF without EM compensation becomes visible in a comparison of the overall flows shown in Fig. 12. They should acquire a superposition of OPM and EMF. Whereas the flow in Fig. 12(a) derives from a summation of the EMF (2, d) and relative flow (3, d), the one in Fig. 12(b) is computed on the uncompensated input images. Because the person P2 is moving contrarily to the camera's rotation, it should exhibit large velocities to the right, which is actually the case in Fig. 12(a). In contrast, the uncompensated flow captures neither P2's movement nor the movement of the camera. Evidently, the abrupt change in camera rotation forces the OF algorithm to cope with a measurement conflicting with the assumption from temporal integration and hence results in an inhomogeneous distorted flow field.

The influence of a displacement vector set that is more than four times as large as the one of the relative flow is also reflected



(a) (b)

Fig. 12. This figure shows a comparison between the overall flow (a) with preceding EM compensation and (b) without.

by the computation time. Whereas the estimation of the relative flow takes nearly 400 ms, the computation of the overall flow lasts almost 3000 ms.

E. Discussion

The evaluation shows that the presented integration of EMF and OF is suitable for real-world scenarios, which is also reflected by the high SE and SP computed on ground truth images created for the key scenarios of robot interaction in indoor scenes. The rejection based on the Mahalanobis norm of the relative OF includes the reliability of depth and flow estimations and allows the compensation of qualitatively weak measurements by stronger ones. For example, distant and, hence, unreliable points can show OPM if the corresponding OF estimation is credible and vice versa.

Running with 2.5 Hz, the system is at the border to being real-time ready. Nevertheless, a higher frame rate is desired and necessary for the interaction with the robot. Most of the computation time is spend for the estimation of the OF, and it was high in our scenario due to the large ranges required for a proper acquisition of people walking close to the camera. In more sophisticated interaction scenarios, where people kept a distance of approximately 1 m, we reduced the displacement range and ran the system with 7 Hz.

By comparing an OF without preceding EM compensation with our relative one, we could prove that the proposed step-by-step movement estimation is a key feature for the reduction of both computation time and ambiguities in the OF measurement.

The EMF used in our system proved to be quite accurate. As we pointed out, inaccuracy in the EMF would still enable the system to detect OPM, at least if the inaccuracy is modeled accordingly by the flow variance. However, the estimated OPM would not acquire pure-object-caused motion but a superposition with the EMF and hence be inaccurate.

Unfortunately, our system cannot cope with errors that occur likewise in OF and disparity estimation. Both these methods search for correlations between images in a patchwise manner. This has two implications for our system: It fails to detect movement for large homogenous regions and blurs object boundaries due to the patchwise computation. The used temporal integration of the OF described in [33] helps to overcome the first aspect; hence, it does not occur frequently for the continuous movement shown in our evaluation. Nevertheless, to cope with the second aspect, the system would need to incorporate more accurate information about motion and depth discontinuities.

IV. Conclusion

In this paper, a system being capable of perceiving OPM from a moving platform has been presented. The effects of a step-by-step movement estimation, including the compensation of EM prior to the OF computation, are central for the robustness of the system against the firm EM of the robot. Robustness against noise in the depth and flow estimation appears to result from the probabilistic rejection mechanism, which neglects velocities based on their amplitude and reliability.

From a macroscopic system perspective, we plan to reduce the system's unidirectional dependence on the accuracy of the EMF. In biology, this is achieved by using not only the proprioception and depth information for motion estimation, as described in Section I-B, but also vice versa: The motion estimation is combined with proprioception to determine the depth, as well as the depth and motion are used to figure the proprioception. Because our system provides segmentation information about moving objects, we can separate the non-moving parts of a scene and enhance the depth estimation, using OF and kinematics. For robots with less precise kinematics, a fusion of OF and depth could be used to improve the kinematics accuracy.

The movement computed by this system will be used in the future for the attraction of visual attention, as well as real-time object interaction.

Acknowledgment

The authors would like to thank M. Gienger for the support with the kinematics computation and reviewing this paper, M. Toussaint for the fruitful discussions about probabilistic fusion, and B. Bolder, C. Karaoguz, and H. Janssen for making the experiments possible.

References

[1] J. Wolfe and T. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nat. Rev. Neurosci.*, vol. 5, no. 6, pp. 495–501, Jun. 2004.

[2] W. H. Warren, B. A. Kay, W. D. Zosh, A. P. Duchon, and S. Sahuc, "Optic flow is used to control human walking," *Nat. Neurosci.*, vol. 4, no. 2, pp. 213–216, Feb. 2001.

[3] T. Albright, "Cortical processing of visual motion," in *Visual Motion and its Use in the Stabilization of Gaze*, J. Wallman and F. Miles, Eds. New York: Elsevier, 1993, ch. 9, pp. 177–201.

[4] T. Tian, C. Tomasi, and D. Heeger, "Comparison of approaches to egomotion computation," in *Proc. IEEE CS Conf. CVPR*, Jun. 18–20, 1996, pp. 315–320.

[5] T. Zhang and C. Tomasi, "On the consistency of instantaneous rigid motion estimation," *Int. J. Comput. Vision*, vol. 46, no. 1, pp. 51–79, Jan. 2002.

[6] M. A. Lewis, "Detecting surface features during locomotion using optic flow," in *Proc. IEEE Int. Conf. Robotics Automation*, 2002, vol. 1, pp. 305–310.

[7] P. Fidelman, T. Coffman, and R. Miikkulainen, "Detecting motion in the environment with a moving quadruped robot," in *RoboCup-2006: Robot Soccer World Cup X*, vol. 4434, G. Lakemeyer, E. Sklar, D. Sorenti, and T. Takahashi, Eds. Berlin, Germany: Springer Verlag, 2007, pp. 219–231.

[8] D. Comaniciu and B. Xie, "Real-time obstacle detection with a calibrated camera and known ego-motion," U.S. Patent 20 040 183 905, Sep. 23, 2004. to Siemens Corp, Tech. Rep. February, 2004. www.freepatentsonline.com/20040183905.html
[9] J. R. del Solar and P. A. Vallejos, "Motion detection and tracking for an AIBO robot using camera motion compensation and Kalman filtering," in *RoboCup 2004: Robot Soccer World Cup VIII*, vol. 3276, D. Nardi, M. Riedmiller, C. Sammut, and J. Santos-Victor, Eds. Berlin, Germany: Springer-Verlag, Mar. 2005, pp. 619–627.
[10] P. Fidelman, T. Coffman, R. Miikkulainen, and P. Stone, "Detecting motion in the world with a moving quadruped robot," Dept. Comput. Sci., Univ. Texas, Austin, TX, Tech. Rep. TR-05-37, 2005.
[11] G. Overett and D. Austin, "Stereo vision motion detection from a moving platform," in *Proc. Australas. Conf. Robot. Autom.*, Dec. 2004, pp. 1–11.
[12] B. Fardi, I. Seifert, G. Wanielik, and J. Gayko, "Motion-based pedestrian recognition from a moving vehicle," in *Proc. IEEE Symp. Intell. Veh.*, Jun. 2006, pp. 219–224.
[13] C. Rabe, U. Franke, and S. Gehrig, "Fast detection of moving objects in complex scenarios," in *Proc. IEEE Symp. Intell. Veh.*, Jun. 2007, pp. 398–403.
[14] J. H. Maunsell and D. C. Van Essen, "Functional properties of neurons in middle temporal visual area of the macaque monkey. II. Binocular interactions and sensitivity to binocular disparity," *J. Neurophysiol.*, vol. 49, no. 5, pp. 1148–1167, May 1983.
[15] K. Tanaka, K. Hikosaka, H. Saito, M. Yukie, Y. Fukada, and E. Iwai, "Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey," *J. Neurosci.*, vol. 6, no. 1, pp. 134–144, Jan. 1986.
[16] K. Tanaka, Y. Fukada, and H. A. Saito, "Underlying mechanisms of the response specificity of expansion/contraction and rotation cells in the dorsal part of the medial superior temporal area of the macaque monkey," *J. Neurophysiol.*, vol. 62, no. 3, pp. 642–656, Sep. 1989.
[17] C. J. Duffy and R. H. Wurtz, "Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli," *J. Neurophysiol.*, vol. 65, no. 6, pp. 1329–1345, Jun. 1991.
[18] C. J. Duffy and R. H. Wurtz, "Response of monkey MST neurons to optic flow stimuli with shifted centers of motion," *J. Neurosci.*, vol. 15, no. 7, pp. 5192–5208, Jul. 1995.
[19] J. P. Roy and R. H. Wurtz, "The role of disparity-sensitive cortical neurons in signalling the direction of self-motion," *J. Nature*, vol. 348, no. 6297, pp. 160–162.
[20] L. R. Harris, M. Jenkin, and D. C. Zikovitz, "Visual and non-visual cues in the perception of linear self-motion," *Exp. Brain Res.*, vol. 135, no. 1, pp. 12–21, Nov. 2000.
[21] Y. Gu, P. V. Watkins, D. E. Angelaki, and G. C. DeAngelis, "Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area," *J. Neurosci.*, vol. 26, no. 1, pp. 73–85, Jan. 2006.
[22] C. R. Fetsch, S. Wang, Y. Gu, G. C. Deangelis, and D. E. Angelaki, "Spatial reference frames of visual, vestibular, and multimodal heading signals in the dorsal subdivision of the medial superior temporal area," *J. Neurosci.*, vol. 27, no. 3, pp. 700–712, Jan. 2007.
[23] C. J. Duffy, "MST neurons respond to optic flow and translational movement," *J. Neurophysiol.*, vol. 80, no. 4, pp. 1816–1827, Oct. 1998.
[24] F. Bremmer, M. Kubischik, M. Pekel, M. Lappe, and K. P. Hoffmann, "Linear vestibular self-motion signals in monkey medial superior temporal area," *Ann. N.Y Acad. Sci.*, vol. 871, pp. 272–281, May 1999.
[25] W. K. Page and C. J. Duffy, "MST neuronal responses to heading direction during pursuit eye movements," *J. Neurophysiol.*, vol. 81, no. 2, pp. 596–610, Feb. 1999.
[26] H. Komatsu and R. H. Wurtz, "Relation of cortical areas MT and MST to pursuit eye movements. I. Localization and visual properties of neurons," *J. Neurophysiol.*, vol. 60, no. 2, pp. 580–603, Aug. 1988.
[27] F. Bremmer, A. Pouget, and K. P. Hoffmann, "Eye position encoding in the macaque posterior parietal cortex," *Eur. J. Neurosci.*, vol. 10, no. 1, pp. 153–160, Jan. 1998.
[28] O. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA: MIT Press, 1999.
[29] M. U. Gienger, H. Janssen, and C. Goerick, "Task-oriented whole body motion for humanoid robots," in *Proc. 5th IEEE-RAS Int. Conf. Humanoid Robots*. Tsukuba, Japan: IEEE Press, 2005, pp. 238–244.
[30] J. J. Craig, *Introduction to Robotics: Mechanics and Control*. Boston, M.A.: Addison-Wesley, 1989.
[31] K. Konolige, "Small vision system: Hardware and implementation," in *Proc. 8th. Int. Symp. Robot. Res.*, Oct. 1997, pp. 111–116.
[32] B. Jähne, *Digital Image Processing*, 6th ed. New York: Springer-Verlag, 2005.
[33] V. Willert, J. Eggert, J. Adamy, and E. Koerner, "Non-Gaussian velocity distributions integrated over space, time and scales," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 3, pp. 482–493, Jun. 2006.
[34] Honda, *The Honda Humanoid Robots*, 2001. [Online]. Available: http://www.honda-robots.com

Jens Schmüdderich received the Dipl.-Inform. degree in applied computer science in the natural sciences from Bielefeld University, Bielefeld, Germany, in 2006, where he is currently working toward the Ph.D. degree in the Applied Computer Science Group in cooperation with the Honda Research Institute Europe GmbH, Offenbach, Germany.

His fields of research are system design and representations of visual and auditory stimuli for behavior generation.

Volker Willert received the Dipl.Ing. degree in electrical engineering and the Ph.D. degree from the Darmstadt University of Technology, Darmstadt, Germany, in 2002 and 2006, respectively.

Since 2005, he has been with the Honda Research Institute Europe GmbH, Offenbach, Germany. His interests include visual scene dynamics, probabilistic machine learning, and modeling of cognitive systems.

Julian Eggert received the Ph.D. degree in physics from the Technical University of Munich, Munich, Germany, where he was working in the Theoretical Biophysics Department of Prof. J. L. van Hemmen.

He is currently with the Honda Research Institute Europe GmbH, Offenbach, Germany. His interests are the dynamics of spiking neurons and neuronal assemblies, large-scale models for the vision system, and gating in hierarchical neural networks via feedback and attention.

Sven Rebhan received the Dipl.Ing. degree in computational engineering from the Technical University of Ilmenau, Ilmenau, Germany, in 2005. He is currently working toward the Ph.D. degree from the Honda Research Institute Europe GmbH, Offenbach, Germany.

His interests are attention-driven visual processing, dynamic scene representation, and large-scale modeling of the vision system.

Christian Goerick received the Diploma degree in electrical engineering and the Ph.D. degree in electrical engineering and information processing from Ruhr-Universität Bochum, Bochum, Germany.

During his time in Bochum, he was Research Assistant, Doctoral Worker, Project Leader, and Lecturer in the Institute for Neural Computation and Chair for Theoretrical Biology. The research was concerned with biologically motivated computer vision for autonomous systems and learning theory of neural networks. He is currently a Chief Scientist with the Honda Research Institute Europe GmbH, Offenbach, Germany. His research interests are behavior-based vision, audition, behavior generation, cognitive robotics, advanced driver assistance systems, system architecture, and hard- and software environments.

Gerhard Sagerer (M'88) received the Diploma and Ph.D. (Dr. Ing.) degree in computer science from the University of Erlangen–Nurnberg, Erlangen, Germany, in 1980 and 1985, respectively, where he received the *venia legendi* (Habilitation) in computer science from the Technical Faculty in 1990.

In 1980–1990, he was with the Research Group for Pattern Recognition (Institut fur Informatik, Mustererkennung), University of Erlangen–Nurnberg. Since 1990, he has been a Professor of computer science with the University of Bielefeld, Beilefeld, Germany, where he is the Head of the research group for Applied Computer Science (Angewandte Informatik), a member of the academic senate in 1991–1993, and the Dean of the Technical Faculty in 1993–1995. He is the author, coauthor, or Editor of several books and technical articles. His fields of research are image and speech understanding including artificial intelligence techniques and the application of pattern understanding methods to natural science domains.

Dr. Sagerer is a member of the German Computer Society and the European Society for Signal Processing. In 1995, he was the Chairman of the Annual Conference of the German Society for Pattern Recognition. He is on the Scientific Board of the German Section of Computer Scientists for Peace and Social Responsibility (Forum InformatikerInnen fur Frieden und gesellschaftliche Verantwortung).

Edgar Körner received the Dr.Ing. degree in biomedical engineering and the Dr.Sci. degree in biocybernetics from the Technical University of Ilmenau, Ilmenau, Germany, in 1977 and 1984, respectively, where he became a Full Professor and Head of the Department of Neurocomputing and Cognitive Science in 1988.

In 1992–1997, he was a Chief Scientist with the Honda R&D Co., Ltd., Wako, Japan. In 1997, he was with the Honda R&D Europe, Offenbach, Germany, to establish the Future Technology Research Division, and since 2003, he has been the President of the Honda Research Institute Europe GmbH, Offenbach. His research interests focus on brainlike artificial neural systems for image understanding, smooth transition between signal–symbol processing, and self-organization of knowledge representation.