# 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

5 Abstract-For the interaction of a mobile robot with a dynamic 6 environment, the estimation of object motion is desired while the 7 robot is walking and/or turning its head. In this paper, we describe 8 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 1 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

[^0]to undergo translatory and rotational movements in 3-D with 43 sudden velocity changes.

We propose a system for the computation of the OF induced 45 by independently moving objects, which we call object proper 46 motion (OPM). To achieve this goal under those demanding 47 conditions described earlier, we utilize a novel combination of 48 known algorithms to compute the OF, depth from binocular 49 disparity, and camera movement from kinematics.

An overview of the system is shown in Fig. 1. In the first step, 51 we use the forward kinematics to compute the movement of the 52 camera, occurring in the time interval from $m$ to $n=m+1.53$ Combining this with depth information from binocular dis- 54 parity, we can estimate where a static point in the image at 55 time $m$ moved due to EM and obtain the EMF. Hence, by 56 knowing where a point in the image at time $n$ originated, we 57 can compensate the EM effect by moving the point back to its 58 original position in the second step, resulting in an image freed 59 from EM effects. Afterward, this image is used for a calculation 60 of the OF measured relative to the EMF. This step-by-step 61 movement estimation allows the reduction of the OF search 62 range, which does not only reduce computation time but also 63 decreases the possibility for ambiguities in the OF measurement 64 and thereby leads to qualitatively better results. Ideally, the 65 OF measured in this way should be zero for unmoving objects 66 and otherwise describe their proper motion. However, because 67 binocular disparity and OF are particularly noisy signals, we 68 finally incorporate the measurement errors into a filter mecha- 69 nism to reject invalid velocity estimations.

The details of this system are described in Section II. Prior 71 to this, we give a review of related work in Section I-A. In 72 Section I-B, we shortly summarize neurobiological evidence 73 to get an idea of the OPM estimation in the human brain. 74 Section III will demonstrate the feasibility of our approach 75 by showing the achieved results in a real-world scenario. In 76 addition, a detailed parameter discussion concerning the tuning 77 and the interrelations is given.

## A. Related Work

The field, where the estimation of OPM during EM is 80 commonly addressed, is the car domain. Here, this estimation 81 is important to measure the speed of other cars relative to 82 the observing car, which allows to identify them as eventual 83 obstacles. Most approaches rely on the calculation of OF. This 84 flow is a superposition of the EMF and the OPM. A common 85 procedure to decompose the OF and extract the desired OPM 86 tries to estimate the movement of the camera from the visual 87


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

88 flow fields, where the underlying models for camera movement 89 highly differ in complexity. An overview of the different ap90 proaches and their characteristics is given in [4] and [5]. For 91 proper extraction of the camera movement, it is hence crucial 92 that the OF is primarily caused by EM effects and not by 93 independently moving objects-which we cannot assume for 94 a robot facing moving people.
95 The effect of translatory EM on points in the image is highly 96 influenced by their distances to the camera in a way that distant 97 points induce smaller flow vectors than closer ones. Some of 98 the approaches actually measure the distances of these points; 99 however, the majority of methods assume that all points lie on a 100 plane running through the position of the camera and the focus 101 of expansion at the horizon. A more detailed explanation is 102 given in [4].
103 Whereas this procedure can be very suited for the car do104 main, no such simplifying assumption about the environment is 105 appropriate for autonomously acting robots in dynamic scenes. 106 Thus, different ways to handle EM have to be found. The central 107 idea of using existing knowledge about the movement of the 108 camera was introduced by Lewis [6]. Letting a robot walk in a 109 circle, he computes the OF resulting from the robot's movement 110 over a textured ground. In conjunction with the robot's gait 111 phase and joint angles, a neural network is trained to learn the 112 EMF. Afterward, differences between this EMF and the mea113 sured OF can be used to detect obstacles in the path of the robot.
114 Whereas the movement of the robot in [6] was very con115 stricted, Fidelman et al. proved that EMF can be learned even 116 for more complex movements [7]. In their approach, the neural 117 network is provided with the recent OF calculation in addition 118 to the walk phase and joint angles of an AIBO robot. The neural 119 net predicts the flow for the next time step, allowing to compare 120 it with the actually measured OF in order to classify differences 121 as OPM.
122 Meanwhile, the idea of calculating the movement of the cam123 era and using it to compute the EMF has also been described for 124 the car. In [8], readings from gyroscopic sensors or GPS signals 125 are applied to estimate the camera movement. Likewise to other 126 approaches in cars, the depth is not measured but estimated 127 from the plane assumption, as described earlier.
128 Another approach using the AIBO platform actually com129 putes the EMF, depending on the measured robot joints instead 130 of using neural networks [9]. However, all the results of the 131 provided approaches are not very promising, which might be 132 due to a missing integration of depth information. In [7] and 133 [10], depth information is implicitly included by providing the 134 calculated OF field, which depends on the distances of points to 135 the camera. Nevertheless, this approach cannot cope properly
with the objects having a distance different from the objects in 136 the training phase: The difference in distances results in an OF 137 deviating from the learned one. Hence, even nonmoving points 138 can be classified as OPM.

The combination of knowledge about the robot movement 140 with distance information and OF measurements is described 141 in [11]. Having cameras fixed to the robot, Overett et al. try 142 to measure the odometry data and compute the resulting EMF, 143 considering the depth. This flow is afterward subtracted from 144 the estimated OF, and residual vectors are used to indicate the 145 OPM. Unfortunately, the noisy data from odometry force the 146 authors to manually measure the distance passed by the robot, 147 preventing the system from running in real time. 148
Aside from [12], none of the presented methods precomputes 149 the movement of the camera and searches for movement rela- 150 tive to it. In acquiring the effects of camera movement superim- 151 posed with those effects of independently moving objects, the 152 OF needs to analyze a higher range of movement. This does 153 not only increase the computation time but also corrupts OF 154 estimations by producing increased ambiguities. 155

Fardi et al. overcome this problem in their pedestrian detec- 156 tion approach by using a compensation similar to the one in our 157 system in order to cancel out the known EM of a car [12]. The 158 compensation and succeeding estimation of OF are computed 159 on a preselected region of interest, where the depth information 160 consists of a single measurement gained from a laser range 161 scanner.

Very recently, Rabe et al. proposed another solution by 163 computing flow and disparity from visual features tracked in the 164 image [13]. Fusing the results in a Kalman filter provides 6-D 165 information about location and motion for a set of points. To- 166 gether with the readings from inertial sensors, this information 167 is used to compensate the EM effects and obtain the actual 3-D 168 OPM. The results are very convincing, at least for the moderate 169 movement of a car.

However, to our knowledge, no procedure was presented 171 which is able to handle the noisy character of depth, OF, and the 172 fierce effects of camera movement induced by a legged robot, 173 combining them to make reliable estimations of OPM.

## B. Neurobiological Evidence

From the previous section, we can conclude that the detection 176 of OPM during EM is a very challenging task. Thus, we have 177 to ask ourselves how humans and most mammals are able to 178 solve it with so much prosperity. For example, humans, running 179 at their top speed, still manage to perceive people standing in 180 some distance, waving their hands, despite the shaking of the 181

182 whole body. In this section, we will summarize some evidence 183 from neurobiological and behavioral studies, serving us as 184 inspiration for our proposed method.
185 An interesting review to the perception of motion is given by 186 Albright [3], where some important stages of motion processing 187 in the brain are explained. These stages seem to be covered to 188 a great extent by the so-called "M pathway," of which we are 189 particularly interested in the processing of two areas in superior 190 temporal sulcus: the middle temporal (MT) visual area and 191 medial superior temporal (MST) area.
192 There is some evidence that the directional selectivity of 193 neurons in MT can be interpreted as serving the creation of a 194 representation for the retinal flow [3], which is somehow similar 195 to the technical OF. Maunsell and Van Essen found neurons in 196 this area to be selective to binocular disparity [14].
197 The created retinal flow is analyzed in area MST, where the 198 dorsal subdivision of MST area (MSTd) seems to pay special 199 respect to OF patterns occurring during EM [15]-[18]. For the 200 estimation of EM effects, the distances of points to the observer 201 are again crucial for correct motion estimation: While points 202 further away than the point of fixation move in the same direc203 tion as the moving observer, nearer points move in the opposite 204 direction. Roy and Wurtz proofed that MST shows exactly this 205 motion selectivity dependent on binocular depth [19].
206 In addition, behavioral studies suggest that correct motion 207 estimation also needs to include vestibular information, encod208 ing changes in head direction [20]-[22]. On a neuronal level, 209 this was confirmed by studies with MSTd cells that are reactive 210 during EM in darkness, indicating usage of the vestibular 211 system [23]-[25].
212 Nevertheless, without one additional aspect, the perception 213 of moving objects would be difficult. During fixation of a 214 moving object, the object is kept in the center of the fovea, 215 resulting in no displacement on the retina. Because objects are 216 perceived as moving even during smooth eye pursuit, the per217 ceived movement must be a superposition of the object's motion 218 on the retina and eye movement. MST neurons were found to 219 perform this kind of summations during eye movement [26] 220 and also for static eye positions [27].
221 We do not try to model the neural activities in detail, but we 222 would rather argue in favor of understanding and transferring 223 the abstracted biological principals. We will show that follow224 ing this paradigm leads to a system which is advantageous for 225 dynamic behavior in dynamic scenes. For example, supposing 226 the system would not incorporate the knowledge of the head 227 movement, it would have to estimate it based on the flow and 228 depth information-with the described drawbacks in dynamic 229 scenes. Moreover, without the incorporation of depth, the re230 sulting system could not perform better than the ones using 231 model assumptions of the depth. Finally, without the separation 232 of EM estimation, the system could not perceive movement at 233 all while moving itself.

## II. Approach

235 In the following, we will describe the details of our approach, 236 starting with the computation of the EMF in Section II-A. The 237 compensation of EM and the estimation of the relative OF are
afterward explained in Section II-B. In Section II-C, we present 238 the filtering applied to the relative flow.

## A. Computation of EMF

240
The computation of EMF results in a flow field $E^{m}=\left\{\mathbf{e}_{i}^{m}\right\} 241$ indexed by $i$, which describes where each point in image $I^{m}$ at 242 time $m$ has shifted to in $I^{n}$ at time $n$, caused by the movement 243 of the camera. For this computation, it is assumed that the 244 environment is static and the points did not move themselves. 245 To account for the fact that the absolute values of the vectors $\mathbf{e}_{i}^{m} 246$ highly depend on the distances of the corresponding points to 247 the camera, the effect of camera movement is calculated in a 248 3-D camera-related space. Therefore, we first combine each 249 2-D point $\left(x_{i}^{m}, y_{i}^{m}\right)$ in the image with its binocular disparity 250 $d_{i}^{m}$ to define a 3-D point in homogenous image coordinates as 251

$$
\begin{equation*}
\mathbf{p}_{I, i}^{m}=\left(x_{i}^{m}, y_{i}^{m}, d_{i}^{m}, 1\right)^{\mathrm{T}} \tag{1}
\end{equation*}
$$

This point can then be reconstructed in camera coordinates 252 by computing the homogenous transformation matrix $\mathbf{T}_{C \leftarrow I} 253$ from image to camera coordinates analogous to [28] and by 254 multiplying $\mathbf{p}_{I, i}^{m}$ with it

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

To compute the position of each point relative to the camera 256 at the next time step, we have to know how the camera 257 moved. This knowledge is obtained from the robot kinematics. 258 Therefore, we use the robot state vector $\mathbf{s}^{m}$ which contains 259 the position of the upper body, as well as the angles of all 260 joints at time $m$ [29]. With this vector, we can compute the 261 forward kinematics, as presented in [30], returning us the 262 transformations from the initial world-coordinate system to 263 each robot segment, including the transformation from world 264 to camera coordinates $\mathbf{T}_{C \leftarrow W}^{m}$. Inverting this transformation 265 results in $\mathbf{T}_{W \leftarrow C}^{m}$, which describes the transformation from 266 camera to world coordinates. In addition, we can compute 267 the current transformation $\mathbf{T}_{C \leftarrow W}^{n}$ by applying $\mathbf{s}^{n}$. Because 268 this calculation is with respect to an initial world point, the 269 resulting transformations also encode camera displacements 270 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 272 calculated from transferring the point to world coordinates by 273 multiplying the last transformation $\mathbf{T}_{W \leftarrow C}^{m}$ and passing it back 274 to camera coordinates with the actual transformation $\mathbf{T}_{C \leftarrow W}^{n} 275$

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

This process is shown in Fig. 2.
Because we are not interested in the new 3-D-position of 277 the points but rather in the 2-D-flow in the image, we have 278 to project each point back into the image plane $I^{n}$. We do 279 this by computing a homogenous projection matrix $\mathbf{T}_{I \leftarrow C}$, as 280 described in [28], and multiply the new point $\mathbf{P}_{C, i}^{n}$ with it 281

$$
\begin{equation*}
\mathbf{p}_{I, i}^{n}=\mathbf{T}_{I \leftarrow C} \mathbf{p}_{C, i}^{n} \tag{4}
\end{equation*}
$$

Note that the spatial constance of the matrices $\mathbf{T}_{C \leftarrow I}, \mathbf{T}_{W \leftarrow C}^{m}, 282$ $\mathbf{T}_{C \leftarrow W}^{n}$, and $\mathbf{T}_{I \leftarrow C}$ allows us to precompute their product, 283


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

284 resulting in one single transformation matrix for each time step, 285 which is multiplied with each point $\mathbf{p}_{I, i}^{m}$.
286 Knowing $\mathbf{p}_{I, i}^{n}$ and $\mathbf{p}_{I, i}^{m}$, the shift of a point in the image plane 287 due to camera motion can now be expressed by

$$
\begin{equation*}
\mathbf{e}_{i}^{m}=\binom{x_{i}^{n}-x_{i}^{m}}{y_{i}^{n}-y_{i}^{m}} \tag{5}
\end{equation*}
$$

288 The reliability of this shift vector depends on the quality of the 289 depth component $z_{i}^{m}$ of $\mathbf{p}_{C, i}^{m}$ and the kinematics accuracy for 290 the robot's movement estimation from time $m$ to $n$.
291 For the latter, we assume an increasing error with increasing 292 movement amplitude. By expressing the camera translation in 293 this time interval with $\mathbf{t}_{d}^{m}$ and its rotation with $\mathbf{t}_{r}^{m}$, we can 294 approximate the kinematics-based variance $\left(\sigma_{k}^{m}\right)^{2}$ by

$$
\begin{equation*}
\left(\sigma_{k}^{m}\right)^{2}=\left\|\mathbf{t}_{d}^{m}\right\|^{2}+\left\|\mathbf{t}_{r}^{m}\right\|^{2} \tag{6}
\end{equation*}
$$

295 Note that this variance is spatially independent and only de296 pends on the robot's movement.
297 The credibility of the depth measurement $z_{i}^{m}$ relates to the 298 amount of correlation found by the disparity algorithm [31] 299 between the left and the right camera image. This is expressed 300 in terms of some confidence value $c_{i}^{m} \in[0,1]$. In addition, we 301 have to account for the decreasing accuracy with increasing 302 distance. In [32], it is shown that sensitivity for depth estimation $303 z$ from disparity $d$ decreases with the squared distance

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

304 where $b$ denotes the baseline, $f$ is the focal length, and $q$ is the 305 pixel size of the camera.
306 Because we want to describe the likelihood for $\mathbf{e}_{i}^{m}$ in terms 307 of some Gaussian probability distribution, we express a decay 308 of reliability in terms of an increasing covariance $\boldsymbol{\Sigma}_{e, i}^{m}$ by 309 subsuming the different aspects of credibility

$$
\begin{equation*}
\boldsymbol{\Sigma}_{e, i}^{m}=\mathbf{1} \cdot w_{e}\left(w_{k}\left(\sigma_{k}^{m}\right)^{2}+w_{d} \sigma_{d}\left(z_{i}^{m}\right)^{2}+w_{c} \sigma_{c}\left(c_{i}^{m}\right)^{2}\right) \tag{8}
\end{equation*}
$$

310 with

$$
\begin{align*}
\sigma_{d}(z)^{2} & =\frac{z^{2} \cdot q}{b \cdot f}  \tag{9}\\
\sigma_{c}(c)^{2} & =\left(c+k_{1}\right)^{2}+k_{2}, \\
k_{1} & =\frac{1}{2}\left(\sigma_{d}\left(z_{\min }\right)^{2}-\sigma_{d}\left(z_{\max }\right)^{2}-1\right) \\
k_{2} & =\sigma_{d}\left(z_{\max }\right)^{2}-k_{1}^{2} . \tag{10}
\end{align*}
$$



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.

The identity matrix is denoted by $\mathbf{1}$, and $z_{\min }$ and $z_{\max }$ are 311 the minimum and maximum assumed distances, a behaviorally 312 relevant point might have from the camera. The constants $k_{1} 313$ and $k_{2}$ in (10) use these distances to ensure that the vari- 314 ance $\sigma_{c}(c=0)^{2}$ equals $\sigma_{d}\left(z=z_{\max }\right)$ and $\sigma_{c}(c=1)^{2}$ equals 315 $\sigma_{d}\left(z=z_{\text {min }}\right)$, respectively. This is useful because it adjusts the 316 scales of both variables to each other.

The factors $w_{c}, w_{d}$, and $w_{k}$ are utilized to weigh the individ- 318 ual influence of ambiguities occurring during the measurement 319 of confidence, distance, or kinematics. They should be based on 320 the present scene, the used disparity algorithm, and kinematics 321 precision. Finally, the scalar $w_{e}$ is used to adjust the range of 322 the whole covariance $\boldsymbol{\Sigma}_{e, i}^{m}$ to the covariance $\boldsymbol{\Sigma}_{r, i}^{m}$ computed for 323 the OF (see hereafter).

## B. Computation of Flow Relative to EMF

Because we have computed the effects of EM on the image, 326 we are now able to compute the OF $R^{m}=\left\{\mathbf{r}_{i}^{m}\right\}$ between $I^{m} 327$ and $I^{n}$ relative to this EMF. In comparison to other approaches 328 which cancel out the EM effects after the computation of the 329 OF, this reduces the spectrum of the motion to be acquired. 330 This effect, which is shown in Fig. 3, does not only reduce 331 computation time but also improves the OF estimations by 332 reducing ambiguities, a fact which is evaluated in Section III. 333

We can compensate the EM effects by warping the images. 334 The two possible approaches of forward and inverse mapping 335 are discussed in detail in [32].

In our case, forward mapping can be written as

$$
\begin{equation*}
\tilde{I}^{m}\left(\mathbf{p}_{I, i}^{m}+\mathbf{e}_{i}^{m}\right)=I^{m}\left(\mathbf{p}_{I, i}^{m}\right) \tag{11}
\end{equation*}
$$

where $\tilde{I}^{m}$ is equal to the old image $I^{m}$ but it is freed from 338 the measured EM effects. This kind of forward mapping has 339 two major drawbacks. Because $\mathbf{e}_{i}^{m}$ usually encodes real-valued 340 shifts, the data points $\mathbf{p}_{I, i}^{m}+\mathbf{e}_{i}^{m}$ may not lie inside the grid 341 and require complicated interpolations. In addition, it is not 342 guaranteed that each point in the warped image is targeted by 343 the sum of the original position and shift, leading to holes in the 344 image.


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.

346 The more convenient solution lies in the usage of backward 347 mapping to warp the actual image $I^{n}$ back to $\tilde{I}^{n}$

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

348 Because the data points are passed as arguments for the result349 ing image $\tilde{I}^{n}$, holes in the image cannot occur. The problem 350 of real-valued shifts is tackled by using bilinear interpolation 351 in the source image. Aside from the deviations caused by 352 inaccurate depth measurements, $\tilde{I}^{n}$ and $I^{m}$ should only differ 353 in points with individual object motion.
354 Nevertheless, in some situations with extreme body move355 ment, holes in the depth image $D^{m}$ with $c_{i}^{m}=0$ can cause 356 artifacts in the pixelwise warped image, as shown in Fig. 4(a). 357 These artifacts lead to errors in the computed OF, and hence, 358 they can affect surrounding regions even if those regions have 359 valid depth estimations. In a first step, these artifacts are 360 reduced by filling invalid regions with suitable assumptions 361 obtained from valid EMF estimations. We utilize a rather 362 simple model by taking the average EMF $\hat{v}_{E}^{m}$ gained from 363 valid depth estimations to fill the gaps [see Fig. 4(b)]. To 364 account for the decreased reliability of the warped image, and 365 hence of the OF, we create a penalty map $P$, which is high
for points near- and inside invalid depth regions and zero 366 otherwise

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

$\Omega_{i}$ denotes the indices of all points in an eight neighborhood 368 around $\mathbf{p}_{I, i}^{m}$.

By passing $I^{m}$ and $\tilde{I}^{n}$ to the OF algorithm described in 370 [33], we get a velocity estimation $\mathbf{r}_{i}^{m}$ for each point, which is 371 relative to the estimated EMF. The algorithm also computes a 372 covariance $\Sigma_{C, i}^{m}$, which gives a confidence measure for the OF 373 vectors, assuming pixelwise independent Gaussian noise. 374

For the inclusion of the warping-based penalty, we compute 375 the compound variance $\boldsymbol{\Sigma}_{r, i}^{m}$ at point $\mathbf{p}_{I, i}^{m}$ as

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

For invalid points, the choice of $1 \cdot P\left(\mathbf{p}_{I, i}^{m}\right) \gg \Sigma_{C, i}^{m}$ ensures a 377 negligible influence of the confidence-based variance. This is 378 necessary because the artifacts in the warped image can create 379 artificial edges and thereby decrease $\Sigma_{C, i}^{m}$ locally.

The used OF algorithm [33] realizes probabilistic prediction 381 over time, considering spatial relations for the transition. This 382 enables the system to iteratively make reliable calculations of 383 motion in unstructured image regions by taking the previous 384 estimations into account. The ideal outcome of this algorithm 385 would be a vector field, which is zero for unmoving objects and 386 otherwise denotes their proper motion.

However, because the OF and the depth measurements from 388 disparity are very noisy signals, we need some more filtering 389 for the detection of OPM. This is described in the following 390 section.

## C. Detecting OPM

In the recent sections, we introduced two approaches for the 393 calculation of image flow, which use very different methods and 394 hence show different characteristics. Because the computation 395 of the EMF is based on disparity and kinematics, it can acquire 396 the effects of EM on image points, as long as the points are 397 not moving. In contrast, the OF also works for moving points. 398 By adding the EMF to the relative OF, we get an overall flow, 399 which designates the compound retinal movement of OPM and 400 EM effects. For unmoving points, this flow should equal the 401 EMF, whereas it should be different for moving points. In this 402 section, a measurement for the significance of this distance is 403 introduced and used to extract OPM vectors from the OF. 404

This measurement is derived from a stochastic assumption 405 about the estimated flows, the depth, and the images. Therefore, 406 we have to define some stochastic variables for each point $\mathbf{p}_{I, i} 407$ in the image, describing a distribution of all measurements. ${ }^{1} 408$ Each vector of the overall flow mentioned earlier is represented 409 by the variable $\vartheta_{i}$, whereas $\epsilon_{i}$ describes the EM vector for each 410 point $\mathbf{p}_{I, i}$. The random variables $I=\left\{I^{m}, I^{n}\right\}$ and $D=D^{m} 411$ specify the observed source and depth images.

412

[^1]413 The principal idea for the approach is to estimate the prob414 ability of measuring the same velocity $\nu_{i}$ from the EMF and 415 the compound OF , assuming the corresponding point is static. 416 That is, if some point did not move, the velocity described by $\epsilon_{i}$ 417 should not differ too much from $\vartheta_{i}$, and the likelihood to mea418 sure some identical velocity $\nu_{i}$ from both methods should be 419 high. In contrast, a moving point results in different outcomes 420 for $\epsilon_{i}$ and $\vartheta_{i}$, and the likelihood to measure the same velocity $421 \nu_{i}$ from the two methods is very low. Concluding a high joint 422 probability $\rho\left(\vartheta_{i}=\nu_{i}, \epsilon_{i}=\nu_{i}, I, D\right)$ indicates a static point, 423 whereas a low probability marks a moving point.
424 Reflecting our assumptions about the dependencies of $\epsilon_{i}$, $425 \vartheta_{i}, D$, and $I$, this joint distribution can be decomposed to the 426 following:

$$
\begin{align*}
\rho\left(\vartheta_{i}=\nu_{i}, \epsilon_{i}=\nu_{i}, I, D\right)= & \rho\left(\vartheta_{i}=\nu_{i} \mid \epsilon_{i}, I\right) \\
& \rho\left(\epsilon_{i}=\nu_{i} \mid D\right) \rho(I) \rho(D) . \tag{15}
\end{align*}
$$

427 Because we make no prior assumptions about the source im428 ages and the depth, the corresponding variables are uniformly 429 distributed, and hence, they have a negligible influence on the 430 distribution. Using the precomputed results for the EMF and 431 OF , we can approximate the conditional distributions with the 432 following:

$$
\begin{align*}
\rho\left(\vartheta_{i}=\nu_{i} \mid \epsilon_{i}, I\right) & \propto \mathcal{N}_{\nu_{i}}\left(\mathbf{r}_{i}+\mathbf{e}_{i}, \boldsymbol{\Sigma}_{r, i}\right)  \tag{16}\\
\rho\left(\epsilon_{i}=\nu_{i} \mid D\right) & \propto \mathcal{N}_{\nu_{i}}\left(\mathbf{e}_{i}, \boldsymbol{\Sigma}_{e, i}\right) . \tag{17}
\end{align*}
$$

433 Adding $\mathbf{e}_{i}$ in (16) accounts for the warping of the image434 whereas $\mathbf{r}_{i}$ encodes a flow relative to the EMF, the sum of $\mathbf{e}_{i}$ 435 and $\mathbf{r}_{i}$ makes it an absolute flow and allows the comparison 436 with the EMF. With this approximation, the joint distribution is 437 proportional to the product of two Gaussians, being defined as

$$
\begin{equation*}
\mathcal{N}_{\nu_{i}}\left(\mathbf{r}_{i}+\mathbf{e}_{i}, \boldsymbol{\Sigma}_{r, i}\right) \mathcal{N}_{\nu_{i}}\left(\mathbf{e}_{i}+\boldsymbol{\Sigma}_{e, i}\right)=L_{i} \cdot \mathcal{N}_{\nu_{i}}\left(\mathbf{c}_{i}, \mathbf{C}_{i}\right) \tag{18}
\end{equation*}
$$

438 with

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

439 A visualization of this product is shown in Fig. 5. The mean 440 value $\mathbf{c}_{i}$ of the resulting distribution can be interpreted as that 441 identical velocity, which is most likely to be measured by 442 both algorithms-the EMF and the OF, while finding a value, 443 which fits the hypothesis of a common velocity best, is always 444 possible, the factor $L_{i}$ is a measure to describe how well $\mathbf{c}_{i}$ 445 actually fits in the light of the calculated displacements and 446 variances. For the evaluation of OPM, we are not interested 447 in the value of the vector $\mathbf{c}_{i}$ but whether such vector is likely 448 to occur. Thus, the rejection of OF estimations is based on this 449 value $L_{i}$ and can be further simplified by applying the logarithm 450 to $L_{i}$

$$
\begin{align*}
L_{i} & =\mathcal{N}_{\mathbf{e}_{i}}\left(\mathbf{r}_{i}+\mathbf{e}_{i}, \boldsymbol{\Sigma}_{r, i}+\boldsymbol{\Sigma}_{e, i}\right) \\
& =z \cdot e^{-\frac{1}{2}\left(\mathbf{e}_{i}-\left(\mathbf{r}_{i}+\mathbf{e}_{i}\right)\right)^{\mathrm{T}}\left(\boldsymbol{\Sigma}_{r, i}+\boldsymbol{\Sigma}_{e, i}\right)^{-1}\left(\mathbf{e}_{i}-\left(\mathbf{r}_{i}+\mathbf{e}_{i}\right)\right)} \\
& \propto-\mathbf{r}_{i}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_{r, i}+\boldsymbol{\Sigma}_{e, i}\right)^{-1} \mathbf{r}_{i} . \tag{19}
\end{align*}
$$



Fig. 5. One-dimensional plot of the resulting distribution $L_{i} \cdot \mathcal{N}_{\nu_{i}}\left(\mathbf{c}_{i}, C_{i}\right)$ from $\mathcal{N}_{\nu_{i}}\left(\mathbf{r}_{i}+\mathbf{e}_{i}, \sum_{r, i}\right)$ and $\mathcal{N}_{\nu_{i}}\left(\mathbf{e}_{i}, \sum_{e, i}\right)$, as well as the likelihood $L$.

That is, the decision whether some optically measured velocity 451 is classified as OPM is based on the absolute value of that 452 velocity scaled by the variances of the EMF and the OF. 453 Defining $\mathbf{r}_{i}^{\mathrm{T}}\left(\Sigma_{r, i}+\Sigma_{e, i}\right)^{-1} \mathbf{r}_{i}$ as squared Mahalanobis norm 454 $\left\|\mathbf{r}_{i}\right\|_{M}^{2}$, we can conclude that a big Mahalanobis norm indicates 455 a moving point; thus, we rely on the computation of the OF. 456 Formally, this can be expressed as 457

$$
\mathbf{o}_{i}= \begin{cases}\mathbf{0}, & \text { if }\left\|\mathbf{r}_{i}\right\|_{M}<\theta_{\mathrm{M}}  \tag{20}\\ \mathbf{r}_{i}, & \text { otherwise }\end{cases}
$$

## III. Experiments and Evaluation

In this section, we would like to show the feasibility of our 459 approach by demonstrating its abilities in a scenario where 460 the combination of our cues is highly recommendable. This 461 scenario shows typical interaction with ASIMO and makes the 462 detection of OPM a difficult task. Nevertheless, we evaluate our 463 results quantitatively and qualitatively, showing robust OPM 464 measurements unseen for legged robots. 465
We will also evaluate the system design by analyzing the re- 466 sults of the different steps and their integration for the detection 467 of OPM. Further on, the benefits of our step-by-step movement 468 estimation will be demonstrated by comparing computation 469 time and results with the outcome of an overall estimated OF. 470

## A. Experiment Description

The experiments are carried out with a Honda ASIMO robot, 472 as presented in [34]. The computation is performed on a Pen- 473 tium 4 single core with 3.4 GHz , and the images are captured 474 with a constant frame rate of 12 Hz . 475
In our scene, ASIMO is initially located in front of one 476 person P1 (black shirt), standing at a distance of approximately 477 2.8 m . A second person P2 (white shirt) is approaching the 478 robot from behind, and both people walk to the right, pass- 479 ing ASIMO's view field (see Fig. 6). Meanwhile, ASIMO is 480 walking forward on a path that can roughly be described as an 481 inverted S-shape, which is superimposed with a rotation of the 482 body at about $45^{\circ}$ in the second half. From the endpoint, he 483 walks backward toward its starting position, turning his body 484


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.

485 straightforward again while P2 is approaching and passing him 486 to the right.
487 To show the high dynamic movement occurring during the 488 experiments, we plot the camera movement and acceleration 489 over time in Fig. 7. The plot shows the translatory movement in 490 the $x$ - and $y$-direction, as well as the rotation around the $z$-axis 491 of the left camera. ${ }^{2}$
492 The important aspect in this scenario is that the view field is 493 dominated by moving people or objects. This makes a modeling 494 of depth or the environment, as described in Section I-A, inap495 propriate because each person would violate the expectation. 496 Due to the dominance of OPM in the view field, the estimation 497 of camera movement from OF would also not be possible 498 without segmentation information.
499 For the evaluation, we extract four images from the captured 500 stream, which are shown in column 1 in Fig. 11. They are

[^2]chosen to represent the different aspects of interaction sce- 501 narios. The first image ( 1, a) is used to evaluate the system's 502 ability to detect movement at a high distance (about 2.8 m ) 503 and taken while the robot is slowly moving forward. Image 504 $(1, b)$ shows the two people walking from the left side of the 505 image to the right, with P2 walking closer to the robot than 506 P1. This scene is considered as key scene, because both people 507 are within the interaction range and separable based on their 508 different velocities. In addition, ASIMO is walking forward 509 with moderate speed. The pictures ( $1, \mathrm{c}$ ) and ( $1, \mathrm{~d}$ ) are captured 510 while ASIMO is walking backward and stepping from one foot 511 to the other. Image $(1, \mathrm{c})$ is chosen to examine the ability of 512 detecting motion for considerable small body parts. During 513 image ( $1, \mathrm{~d}$ ), the robot abruptly performs an additional rotation 514 of the upper body. This scene will demonstrate the system's 515 ability to handle jerky camera rotation (see Fig. 7), and it will 516 also be used to evaluate the benefits of a step-by-step movement 517 estimation.

## B. Parameter Evaluation

The overall aim of the system is the detection and exact esti- 520 mation of OPM relative to the EM for all points in the image. To 521 evaluate the influence of different parameters on the system's 522 performance, we did the following investigations. The choice 523 for the motion range to be captured by the system is based 524 on the magnitude of movement in our scene. To cope with 525 people passing very close to the robot, the acquired range of 526 displacements for the OF is chosen in a range from $[-10,10] 527$ in the $x$-direction to $[-2,2]$ in the $y$-direction, using an image 528 resolution of $200 \times 150$ pixels. For the calculation of the overall 529 OF, i.e., the EM-uncompensated flow, the displacements are 530 chosen in the range from $[-22,22]$ in the $x$-direction to $[-4,4] 531$ in the $y$-direction, accounting for the fact that this flow has 532 to acquire the maximum range of movement occurring in the 533 stream, which is composed of EM effects superimposed with 534 OPM. The classification of OPM as in (20) depends on the 535 relative flow $\mathbf{r}_{i}$, the variances $\boldsymbol{\Sigma}_{r, i}$, and $\boldsymbol{\Sigma}_{e, i}$, as well as the 536 threshold $\theta_{M}$. Because $\mathbf{r}_{i}$ and $\boldsymbol{\Sigma}_{r, i}$ are measured, and hence do 537 not depend on any parameter, we will focus on the evaluation 538 of $\boldsymbol{\Sigma}_{e, i}$ and $\theta_{M}$.
$\boldsymbol{\Sigma}_{e, i}$ is based on the choice for $w_{e}, w_{k}, w_{d}$, and $w_{c}$. Setting 540 $w_{e}=0.07$ adjusts the scale of $\boldsymbol{\Sigma}_{e, i}$ to match the measured $\boldsymbol{\Sigma}_{r, i} .541$ In our evaluation, we neglected the influence of the kinematics- 542 based variance by assigning $w_{k}=0$. This accounts for the fact 543 that the worst case error from kinematics is lower than $2 \mathrm{~mm}, 544$ owing to ASIMO's high-precision encoders and its stiffness. In 545 comparison to this, the resolution in distance estimation from 546 disparity drops to 69 mm for a point that is 2.5 m away from 547 the camera, ${ }^{3}$ assuming that the disparity algorithm found the 548 right displacement. The error increases by 69 mm for each one 549 pixel of wrong displacement.

The remaining parameters $w_{d}, w_{c}$, and $\theta_{M}$ are chosen based 551 on the receiver operating characteristic (ROC) of the system in 552 order to find the optimal parameter settings for OPM detection. 553

[^3]

Fig. 8. (a) Systems $S P$ and $S E$ dependent on the weights $w_{c}$ and $w_{d}$. (b) Plot of $S E$ and $S P$ dependent on $\theta_{M}$.

554 Therefore, we create some ground-truth data $G(i) \in\{0,1\}$ 555 from the four images in Fig. 11, indicating for each pixel $i$ 556 whether it moved $[G(i)=1]$ or not $[G(i)=0]$. We also define 557 our system's output as

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

558 Further on, the sensitivity $S E$ and specificity $S P$ are defined as

$$
\begin{align*}
& S E=\frac{T P}{T P+F N}  \tag{21}\\
& S P=\frac{T N}{T N+F P} \tag{22}
\end{align*}
$$

559 with

$$
\begin{array}{ll}
T P=|\{i \mid G(i)=1 \cap \Phi(i)=1\}| & \\
F P=|\{i \mid G(i)=0 \cap \Phi(i)=1\}| & \\
\text { (frue-positive) } \\
T N=|\{i \mid G(i)=0 \cap \Phi(i)=0\}| & \\
\text { (true-negative) } \\
F N=|\{i \mid G(i)=1 \cap \Phi(i)=0\}| & \\
\text { (false-negative). }
\end{array}
$$

560 For the evaluation of $w_{c}, w_{d}$, and $\theta_{M}$, we use an iterative 561 procedure. Because the threshold $\theta_{M}$ depends on the choice 562 of $w_{c}$ and $w_{d}$, whereas these weights are determined based 563 on the system's output which depends again on $\theta_{M}$, there is 564 circular dependence between the three evaluated parameters. 565 For this reason, we start with a fixed $\theta_{M}=0.1$, rejecting 566 almost no velocities, and vary $w_{c} \in[0,1]$ and $w_{d}$ accordingly 567 by choosing $w_{d}=1-w_{c}$. The resulting $S E$ and $S P$ for these 568 values, as shown in Fig. 8(a), show a rapid decline of $S E$ with 569 increasing $w_{c}$ and decreasing $w_{d}$. Hence, $\sigma_{d}(z)^{2}$ appears to be a 570 more suitable approximation of the EMF variance than $\sigma_{c}(c)^{2}$. 571 Those $w_{c}$ and $w_{d}$ which elicit equal values for $S E$ and $S P$ 572 are considered as optimum, because they represent a tradeoff 573 between a high true-positive rate and a high true-negative rate. 574 As shown by the intersection of $S P$ and $S E$ in the plot, this 575 results in a choice of $w_{c}=0.1$ and $w_{d}=0.9$, respectively.
576 For $\theta_{M}$, we use these determined weights and vary $577 \theta_{M} \in[0,5]$. The resulting ROC curve in Fig. 8(b) shows a plot 578 of $1-S P$ against $S E$. Again, the best choice for $\theta_{M}$ is deter579 mined from equal $S E$ and $S P$ values, resulting in $\theta_{M}=0.5$ 580 for our scenario. The described process is repeated iteratively 581 until convergence, which is achieved after two iterations in our 582 experiments. In addition, the obtained parameters have also


Fig. 9. Classification border dependent on the relative velocities, accumulated variances, and $\theta_{M}$.
been validated on a large-range image sequence by means of 583 visual inspection.

The described evaluation shows that the choice for $\theta_{M}$ is cru- 585 cial for the system's performance; thus, we also perform some 586 qualitative analysis of this threshold, which is useful for cases 587 in which no ground truth is available. Fig. 9 shows the relation 588 of the $x$-component ${ }^{4}$ of relative velocities $r_{i}^{x}$, accumulated 589 variance $\sigma_{e, i}^{x x}+\sigma_{r, i}^{x x}$, and threshold $\theta_{M}$. The surface labeled by 590 a white " 2 " shows the maximal $\theta_{M}$ value for each combination 591 of relative velocity and accumulated variance, which would 592 classify this specific combination as OPM. For example, the 593 point 1 with $r_{i}^{x} \approx 3$ and $\sigma_{e, i}^{x x}+\sigma_{r, i}^{x x} \approx 5.6$ is rejected from 594 OPM with $\theta_{M}>1.23$. The points 1,2 , and 3 correspond to 595 those points shown in Fig. 10.

The plane labeled by the white " 1 " visualizes the choice 597 $\theta_{M}=0.5$ : Those points on the mesh lying higher than this 598 plane are classified as OPM, those below are rejected. 599

In conclusion, the choice for $\theta_{M}$ can simply be done by 600 choosing one combination of relative flow and variance, which 601 should serve as a classification border.

## C. Quantitative Results

We use the ROC of the $\theta_{M}$ evaluation as quantitative quality 604 measure for the system's ability to detect OPM. The accuracy 605 of the OPM is not evaluated for two reasons. Assuming an 606 accurately estimated EMF, the accuracy of OPM is dependent 607 on the OF precision, which is investigated in [33]. In case of an 608 inaccurate EMF, the resulting OPM would not designate pure 609 object motion but a mixture with effects from EM. As can be 610 seen hereafter, this does not seem to be the case in our scenario, 611 otherwise the OF would contain systematic errors.

612
To summarize the overall system performance in one single 613 value, we compute the area below the ROC curve in Fig. 8(b). 614 Our system achieves a value of 0.92 , where 1.0 would indicate 615 an optimal classifier. This value is quite high despite the fact 616 that the patchwise computation of disparity and flow cannot 617 determine exact object borders but always tends to surround 618 objects.

By setting $\theta_{M}=0.5$, an $S E$ of 0.89 and an $S P$ of 0.90 are 620 achieved, i.e., $89 \%$ of all moving pixels are detected, and $90 \% 621$ of all rejected points are actually not moving.

[^4]

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.

623 One iteration of the system lasts about 400 ms . The vast ma624 jority is spend on the computation of the OF estimation, which 625 takes in an average of 387 ms whereas the computation of EMF 626 and warping and detection of OPM take less than 5 ms each.

## 627 D. Qualitative Results

628 Fig. 10 shows the computed velocities and the rejection 629 mechanism exemplarily for the three points 1,2 , and 3 of our 630 key scene. An overall inspection of the results and their quality 631 is given later on.
632 The EM vector for point 1 in the image reflects the forward 633 movement of the robot by showing a displacement pointing 634 away from the center of expansion located on person P2. Be635 cause the point is relatively far away, the EM vector has a small 636 absolute value and also shows a high variance in the plotted 637 normal distribution. The corresponding OF vector represents a 638 slow movement to the right and exhibits a small variance due 639 to the high amount of image structure in the surroundings. This 640 small variance leads to a high Mahalanobis norm in (20), so that 641 the OF vector is maintained as OPM.
642 The second point lies on the wall, which is more than 9 m 643 away from the camera. Hence, $\mathbf{e}_{i}$ is almost zero, and it is 644 accompanied by a high variance. Because the lack of image 645 structure in this region causes also a high variance of the OF, 646 the Mahalanobis norm is small even so the Euclidean distance 647 between $\mathbf{e}_{i}$ and $\mathbf{r}_{i}$ is similar to the one for point 1. Accordingly, 648 this flow vector is rejected.
649 Point 3 is clearly indicated as OPM, resulting from a large 650 distance between $\mathbf{r}_{i}$ and $\mathbf{e}_{i}$ in combination with a very low 651 variance $\Sigma_{e, i}$ which accounts for the small distance to the 652 camera.

To visualize the results for the entire view field, Fig. 11 shows 653 the four left camera images taken from the described stream of 654 571 images.

The depth dependence of the EMF prevents the system from 656 operating on areas with a lack of depth information. At the left 657 side of the image, this originates from the displacement shift 658 used for disparity computation, whereas the right, top, and bot- 659 tom borders are the effect of image rectification. The remaining 660 area is marked by a white rectangle in the gray images. 661

Columns 2, 3, and 4 in Fig. 11 visualize the different flow 662 fields for these marked areas. The gray value in these images 663 represents the flow magnitude for each pixel, whereas the exact 664 flow vector is shown for every 17th pixel in the $x$ - and $y-665$ direction.

666
The radial expanding EMF typical for pure forward trans- 667 latory movement is clearly visible in image (2, b), whereas 668 the slow robot movement during scene A causes this flow to 669 be visible only for very close points at the bottom. The flow 670 field in (2, c) is characterized by the robot's translation to the 671 right. The depth dependence of the EMF is visible by showing 672 larger velocities for points closer to the camera than for more 673 distant points. Despite the robot's translation to the right during 674 image (1, d), the associated flow field in (2, d) is dominated 675 by the rotation of the robot's upper body to the left. Because 676 the EMF for purely rotational movement does not depend on 677 depth, this flow contains almost equal velocity vectors to the 678 right. However, it is observable in all scenes that the EMF 679 caused by close objects occupies areas which are bigger than 680 the objects themselves. This is the already mentioned effect 681 of the patchwise disparity estimation. The holes in the EMF, 682 which are particularly visible in image ( $2, \mathrm{~d}$ ), are the results of 683 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.

685 The relative OF in column 3 clearly acquires the people's 686 movement to the right in images $a, b$, and d. It also captures the 687 divergent movement of the arms of P1 and P2 in image c , which 688 is caused by handing over the stamper from the left to the right 689 person.
690 All four flow fields include false-positive velocities. For 691 fast movement as in scenes $b$ and d, this originates from the 692 flow spatiotemporal prediction which makes person P2 drag 693 a "trail" of movement behind it (for further explanations of 694 this effect, see [33]). The remaining errors can be classified 695 as correspondence problems occurring at straight borders and 696 textureless regions.
697 The OPM shown in the last column contains almost none of 698 these false positives, except for some small patches in image $699(4, d)$. In particular, the described "trail" of movement is re700 moved. Moreover, the amount of false negatives is considerably 701 low and only visible in image $(4, b)$ for some areas on the arm 702 of person P2.

The qualitative effect of the preceding EM compensation 703 and subsequent OF estimation in contrast to the computation 704 of the absolute OF without EM compensation becomes visible 705 in a comparison of the overall flows shown in Fig. 12. They 706 should acquire a superposition of OPM and EMF. Whereas the 707 flow in Fig. 12(a) derives from a summation of the EMF (2, d) 708 and relative flow (3, d), the one in Fig. 12(b) is computed on 709 the uncompensated input images. Because the person P2 is 710 moving contrarily to the camera's rotation, it should exhibit 711 large velocities to the right, which is actually the case in 712 Fig. 12(a). In contrast, the uncompensated flow captures neither 713 P2's movement nor the movement of the camera. Evidently, the 714 abrupt change in camera rotation forces the OF algorithm to 715 cope with a measurement conflicting with the assumption from 716 temporal integration and hence results in an inhomogeneous 717 distorted flow field.

The influence of a displacement vector set that is more than 719 four times as large as the one of the relative flow is also reflected 720


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

721 by the computation time. Whereas the estimation of the relative 722 flow takes nearly 400 ms , the computation of the overall flow 723 lasts almost 3000 ms .

## 724 E. Discussion

725 The evaluation shows that the presented integration of EMF 726 and OF is suitable for real-world scenarios, which is also 727 reflected by the high $S E$ and $S P$ computed on ground truth 728 images created for the key scenarios of robot interaction in 729 indoor scenes. The rejection based on the Mahalanobis norm 730 of the relative OF includes the reliability of depth and flow 731 estimations and allows the compensation of qualitatively weak 732 measurements by stronger ones. For example, distant and, 733 hence, unreliable points can show OPM if the corresponding 734 OF estimation is credible and vice versa.
735 Running with 2.5 Hz , the system is at the border to being 736 real-time ready. Nevertheless, a higher frame rate is desired 737 and necessary for the interaction with the robot. Most of the 738 computation time is spend for the estimation of the OF, and it 739 was high in our scenario due to the large ranges required for 740 a proper acquisition of people walking close to the camera. In 741 more sophisticated interaction scenarios, where people kept a 742 distance of approximately 1 m , we reduced the displacement 743 range and ran the system with 7 Hz .
744 By comparing an OF without preceding EM compensation 745 with our relative one, we could prove that the proposed step-by746 step movement estimation is a key feature for the reduction of 747 both computation time and ambiguities in the OF measurement. 748 The EMF used in our system proved to be quite accurate. 749 As we pointed out, inaccuracy in the EMF would still enable 750 the system to detect OPM, at least if the inaccuracy is modeled 751 accordingly by the flow variance. However, the estimated OPM 752 would not acquire pure-object-caused motion but a superposi753 tion with the EMF and hence be inaccurate.
754 Unfortunately, our system cannot cope with errors that occur 755 likewise in OF and disparity estimation. Both these methods 756 search for correlations between images in a patchwise manner. 757 This has two implications for our system: It fails to detect 758 movement for large homogenous regions and blurs object 759 boundaries due to the patchwise computation. The used tempo760 ral integration of the OF described in [33] helps to overcome 761 the first aspect; hence, it does not occur frequently for the 762 continuous movement shown in our evaluation. Nevertheless,
to cope with the second aspect, the system would need to 763 incorporate more accurate information about motion and depth 764 discontinuities.

## IV. Conclusion

In this paper, a system being capable of perceiving OPM 767 from a moving platform has been presented. The effects of a 768 step-by-step movement estimation, including the compensation 769 of EM prior to the OF computation, are central for the robust- 770 ness of the system against the firm EM of the robot. Robustness 771 against noise in the depth and flow estimation appears to result 772 from the probabilistic rejection mechanism, which neglects 773 velocities based on their amplitude and reliability.

From a macroscopic system perspective, we plan to reduce 775 the system's unidirectional dependence on the accuracy of 776 the EMF. In biology, this is achieved by using not only the 777 proprioception and depth information for motion estimation, 778 as described in Section I-B, but also vice versa: The motion 779 estimation is combined with proprioception to determine the 780 depth, as well as the depth and motion are used to figure 781 the proprioception. Because our system provides segmentation 782 information about moving objects, we can separate the non- 783 moving parts of a scene and enhance the depth estimation, using 784 OF and kinematics. For robots with less precise kinematics, a 785 fusion of OF and depth could be used to improve the kinematics 786 accuracy.

787
The movement computed by this system will be used in the 788 future for the attraction of visual attention, as well as real-time 789 object interaction.

## Acknowledgment

The authors would like to thank M. Gienger for the support 792 with the kinematics computation and reviewing this paper, 793 M. Toussaint for the fruitful discussions about probabilistic 794 fusion, and B. Bolder, C. Karaoguz, and H. Janssen for making 795 the experiments possible.

## REFERENCES

[1] J. Wolfe and T. Horowitz, "What attributes guide the deployment of visual 798 attention and how do they do it?" Nat. Rev. Neurosci., vol. 5, no. 6, 799 pp. 495-501, Jun. 2004.

800
[2] W. H. Warren, B. A. Kay, W. D. Zosh, A. P. Duchon, and S. Sahuc, "Optic 801 flow is used to control human walking," Nat. Neurosci., vol. 4, no. 2, 802 pp. 213-216, Feb. 2001.

803
[3] T. Albright, "Cortical processing of visual motion," in Visual Motion 804 and its Use in the Stabilization of Gaze, J. Wallman and F. Miles, Eds. 805 New York: Elsevier, 1993, ch. 9, pp. 177-201.

806
[4] T. Tian, C. Tomasi, and D. Heeger, "Comparison of approaches to ego- 807 motion computation," in Proc. IEEE CS Conf. CVPR, Jun. 18-20, 1996, 808 pp. 315-320.

809
[5] T. Zhang and C. Tomasi, "On the consistency of instantaneous rigid 810 motion estimation," Int. J. Comput. Vision, vol. 46, no. 1, pp. 51-79, 811 Jan. 2002.

812
[6] M. A. Lewis, "Detecting surface features during locomotion using op- 813 tic flow," in Proc. IEEE Int. Conf. Robotics Automation, 2002, vol. 1, 814 pp. 305-310.

815
[7] P. Fidelman, T. Coffman, and R. Miikkulainen, "Detecting motion in 816 the environment with a moving quadruped robot," in RoboCup-2006: 817 Robot Soccer World Cup X, vol. 4434, G. Lakemeyer, E. Sklar, 818 D. Sorenti, and T. Takahashi, Eds. Berlin, Germany: Springer Verlag, 819 2007, pp. 219-231.

821 [8] D. Comaniciu and B. Xie, "Real-time obstacle detection with a cal-

822 823
824 ibrated camera and known ego-motion," U.S. Patent 20040183 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. 134144, 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. 596610, 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 898 distributions integrated over space, time and scales," IEEE Trans. Syst., 899 Man, Cybern. B, Cybern., vol. 36, no. 3, pp. 482-493, Jun. 2006.
[34] Honda, The Honda Humanoid Robots, 2001. [Online]. Available: http:// 901 www.honda-robots.com


Jens Schmüdderich received the Dipl.-Inform. 903 degree in applied computer science in the nat- 904 ural sciences from Bielefeld University, Bielefeld, 905 Germany, in 2006, where he is currently working 906 toward the Ph.D. degree in the Applied Computer 907 Science Group in cooperation with the Honda Re- 908 search Institute Europe GmbH, Offenbach, Germany. 909
His fields of research are system design and repre- 910 sentations of visual and auditory stimuli for behavior 911 generation.


Volker Willert received the Dipl.Ing. degree in 913 electrical engineering and the Ph.D. degree from 914 the Darmstadt University of Technology, Darmstadt, 915 Germany, in 2002 and 2006, respectively.

916
Since 2005, he has been with the Honda Research 917 Institute Europe GmbH, Offenbach, Germany. His 918 interests include visual scene dynamics, probabilis- 919 tic machine learning, and modeling of cognitive 920 systems.


Julian Eggert received the Ph.D. degree in physics 922 from the Technical University of Munich, Munich, 923 Germany, where he was working in the Theoretical 924 Biophysics Department of Prof. J. L. van Hemmen. 925
He is currently with the Honda Research Institute 926 Europe GmbH, Offenbach, Germany. His interests 927 are the dynamics of spiking neurons and neuronal 928 assemblies, large-scale models for the vision system, 929 and gating in hierarchical neural networks via feed- 930 back and attention.


Sven Rebhan received the Dipl.Ing. degree in com- 932 putational engineering from the Technical Univer- 933 sity of Ilmenau, Ilmenau, Germany, in 2005. He is 934 currently working toward the Ph.D. degree from the 935 Honda Research Institute Europe GmbH, Offenbach, 936 Germany.

937
His interests are attention-driven visual process- 938 ing, dynamic scene representation, and large-scale 939 modeling of the vision system.


Christian Goerick received the Diploma degree in 941 electrical engineering and the Ph.D. degree in elec- 942 trical engineering and information processing from 943 Ruhr-Universität Bochum, Bochum, Germany. 944
During his time in Bochum, he was Research 945 Assistant, Doctoral Worker, Project Leader, and Lec- 946 turer in the Institute for Neural Computation and 947 Chair for Theoretrical Biology. The research was 948 concerned with biologically motivated computer vi- 949 sion for autonomous systems and learning theory of 950 neural networks. He is currently a Chief Scientist 951 with the Honda Research Institute Europe GmbH, Offenbach, Germany. His 952 research interests are behavior-based vision, audition, behavior generation, 953 cognitive robotics, advanced driver assistance systems, system architecture, and 954 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 ErlangenNurnberg. Since 1990, he has been a Professor 967 of computer science with the University of Bielefeld, Beilefeld, Germany, 968 where he is the Head of the research group for Applied Computer Science 969 (Angewandte Informatik), a member of the academic senate in 1991-1993, and 970 the Dean of the Technical Faculty in 1993-1995. He is the author, coauthor, or 971 Editor of several books and technical articles. His fields of research are image 972 and speech understanding including artificial intelligence techniques and the 973 application of pattern understanding methods to natural science domains.
974 Dr. Sagerer is a member of the German Computer Society and the European 975 Society for Signal Processing. In 1995, he was the Chairman of the Annual 976 Conference of the German Society for Pattern Recognition. He is on the Scien977 tific Board of the German Section of Computer Scientists for Peace and Social 978 Responsibility (Forum InformatikerInnen fur Frieden und gesellschaftliche 979 Verantwortung).


Edgar Körner received the Dr.Ing. degree in bio- 980 medical engineering and the Dr.Sci. degree in biocy- 981 bernetics from the Technical University of Ilmenau, 982 Ilmenau, Germany, in 1977 and 1984, respectively, 983 where he became a Full Professor and Head of 984 the Department of Neurocomputing and Cognitive 985 Science in 1988.

986
In 1992-1997, he was a Chief Scientist with 987 the Honda R\&D Co., Ltd., Wako, Japan. In 1997, 988 he was with the Honda R\&D Europe, Offenbach, 989 Germany, to establish the Future Technology Re- 990 search Division, and since 2003, he has been the President of the Honda 991 Research Institute Europe GmbH, Offenbach. His research interests focus on 992 brainlike artificial neural systems for image understanding, smooth transi- 993 tion between signal-symbol processing, and self-organization of knowledge 994 representation.

995


[^0]:    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

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

[^2]:    ${ }^{2}$ The movement of the remaining components is negligible and hence not shown.

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

[^4]:    ${ }^{4}$ The results are analogous for the $y$-component.

