Probabilistic Optical Flow Estimation for Large Pixel Displacements Utilizing Egomotion Flow Compensation

Volker Willert[1], Jens Schmüdderich[2], Julian Eggert[1],
Christian Goerick[1], Edgar Körner[1]
Honda Research Institute[1], Carl-Legien-Straße 30,
D-63073, Offenbach/Main, Germany
University of Bielefeld[2], PO Box 100131 ,
D-33501 Bielefeld, Germany

Abstract

The pixel movements in an image sequence grabbed by a camera that is mounted on a mobile platform comprise the superposition of several motion components. These motion components are caused by the egomotion of the camera and by the different movements of the objects seen by the camera. Utilizing sensory information from a calibrated stereo rig and egomotion measurements of the mobile platform we develop a probabilistic framework that estimates optical flow relative to the visual flow induced by the egomotion. Despite rapid egomotion changes and a large range of pixel movements the proposed Dynamic Bayesian Network allows to infer the optical flow induced by moving objects. This is used to segregate moving individuals from static background while the stereo rig is moving. We present optical flow and figure-background segmentation results by applying this general framework to image sequences captured by the humanoid robot ASIMO while he is walking and observing moving people.

1 Introduction

For humans the visual flow is a very useful source of information to describe the dynamics of the observed visual scene. It comprises different levels of motion complexity along the processing stream starting from low-level attentive mechanisms up to detailed motion analyses, e.g. to classify the movement of specific objects possessing characteristical motion patterns.

In this paper, we take advantage of two peculiarities of the visual flow. On the one hand, image motion is a dynamic feature of an image sequence and the longer this spatiotemporal information is observed the more precise and detailed we can estimate and predict the motion exploiting e.g. spatiotemporal constraints. Therefore, it is natural not to stick to the visual information within a certain time interval to get an estimation but applying a proper filtering technique that is able to generate, confirm and refine motion hypotheses over time. On the other hand, the visual flow is induced by several sources, like the movement of the observer and the movement of the objects that are observed.

BMVC 2008 doi:10.5244/C.22.69

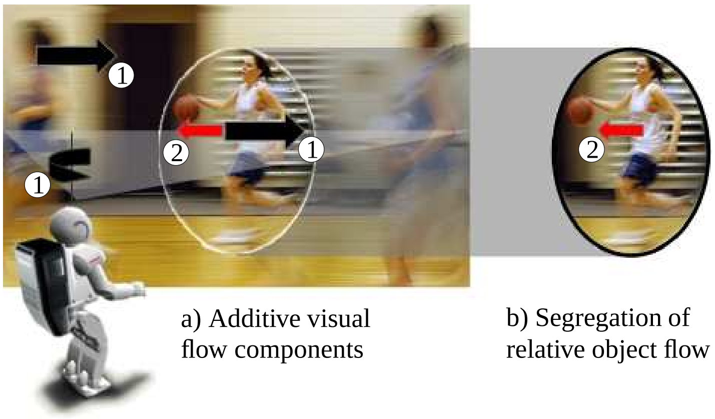

1
2
1
1
2
a) Additive visual flow components
b) Segregation of relative object flow

Figure 1: The basic idea of consecutive motion estimation. Flow components that are induced by the observer ① and the observed objects ② are split up to isolate moving objects.

Normally, this results in a large range of pixel displacements that have to be considered for motion measurement. Together with the inherent ambiguity in the measurement process, e.g. because of the lack of structural information, object deformations, or illumination changes, the uncertainty of the motion estimation increases with the velocity search range. This is because the generative model assumptions on how an image is generated dependent on temporal preceding images become more and more inadequate with the increase of pixel displacements between temporal consecutive frames.

We try to utilize both visual flow aspects 1) the *dynamics* and 2) the *superposition* of flow components to develop a motion estimation system for a moving platform that is capable of isolating object induced flow components from the egomotion flow. In case of rapid ego-movements of a robot while observing moving objects, like sketched in Fig. 1 a), a separation of the visual flow induced by movements of the objects ② from the egomotion induced flow field ①, like depicted in Fig. 1 b), has two advantages for describing the dynamics of the scene. First, it decreases the ambiguity in the estimation process because the velocity range that has to be covered to detect the visual flow components is split up into two independent measurements based on different sensory information. Second, we are able to treat temporal integration of the objects induced flow components independent from the temporal integration of the egomotion flow induced by the observer.

If mobile robots move around in a *static* environment, the projection of the environment onto the robot cameras induces a flow field that is *exclusively* caused by the egomotion of the robot and varies with the 3D profile of the scene. Visual SLAM [2] or egomotion computation approaches [8] utilize these dependencies to estimate the pose of a moving camera and the scene structure usually assuming that a sparse and temporal stable set of point-to-point correspondences of static image features can be extracted. Additional sensing of the body movement via proprioception combined with the information of the visual flow allows for dense depth estimation which is called *Structure from Motion* [7]. As a reverse operation to egomotion-based depth estimation, the expected visual flow generated by egomotion can be inferred by combining body movement and scene depth

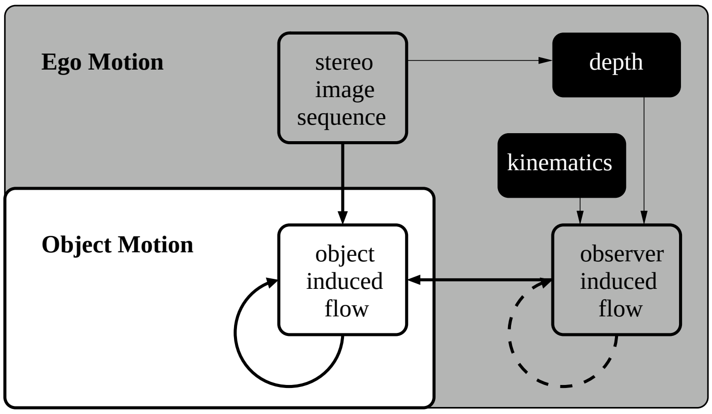


Figure 2: System overview for sequential refinement of motion estimation. An egomotion compensated image sequence is processed by a probabilistic recurrent filter to obtain the movements of objects.

information using depth cues like e.g. extracted from binocular disparity [5]. Unfortunately, in most cases the environment is *not static* but contains moving objects. These induce flow field components onto the robots cameras which deviate from the flow field as it is predicted from egomotion for static scenes. Therefore, we focus on the estimation of visual flow caused by a combination of egomotion and object flow components assuming uncertain probabilistic binocular disparity and optical flow information and confident deterministic body movement information.

There are already some approaches, e. g. [3, 5, 6], trying to estimate the image flow of a moving observer including the motion of objects moving relative to the observer. Basically, they differ 1) in the accuracy of depth information which can be directly measured or modelled indirectly, e.g. via planar surface assumptions 2) in whether they apply temporal filtering or not and 3) in whether the estimate is only done for sparse feature points or all pixels in the image. We are not aware of methods that split the process into precomputing the egomotion flow and search for the object flow relative to it in combination with a spatiotemporal filter for dense object flow fields.

To tackle the problem of extracting moving objects and estimate their optical flow fields despite egomotion of the observer, we set up a structure as depicted in Fig. 2, allowing the system to compensate for egomotion effects. We estimate the image flow induced by egomotion as described in Sec. 2.1 assuming a static scene by utilizing the robots kinematics and depth information from binocular disparity. According to this predicted flow each image is warped so that we get an egomotion compensated image. The sequential motion estimation described in Sec. 2.2 then occurs on the basis of compensated images, so that only the relative visual flow is extracted. With the continuous image streams and the compensated images as input data to a recurrent motion estimation system we are able to extract, integrate and predict the optical flow induced by moving objects (separated from the ego-flow) with all the advantages of probabilistic spatiotemporal filtering. Section 3 provides results of the estimation capabilities of the proposed motion estimation framework which are shortly discussed in Sec. 4.

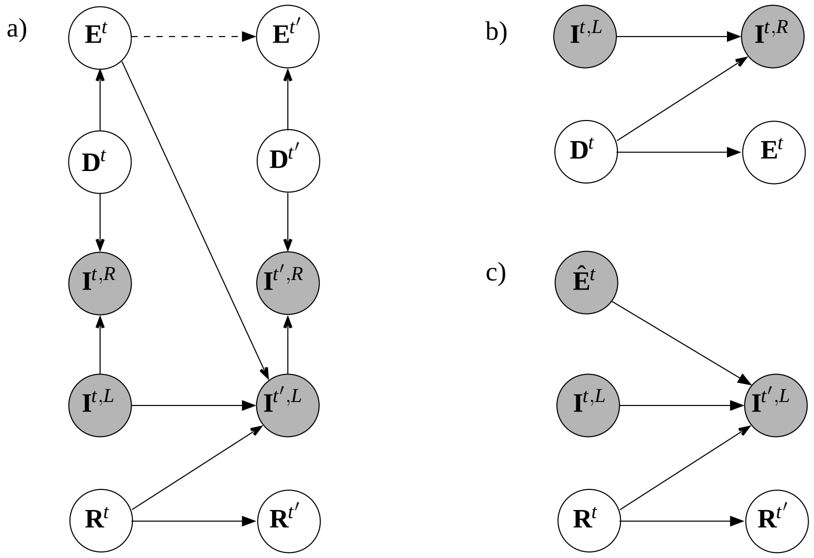


Figure 3: Dynamic Bayesian Network for sequential refinement of motion estimation. The nodes of observed variables are shaded grey, while the latent variables are denoted unshaded. Here, t' is an abbreviation for $t+1$.

2 Dynamic Bayesian Network Model

First of all, the variables and their dependencies given by the probabilistic grapical model depicted in Fig. 3, which in our case is a Dynamic Bayesian Network (DBN), are defined. We assume a generative model for the image sequences of a stereo rig in canonical configuration $\mathbf{I}^{1:T,L}$ and $\mathbf{I}^{1:T,R}$ of T images of the left L and right R camera both with image range X at equidistant points in time t as illustrated by the graphical model in Fig. 3 a). The observed variables are the grey value images $\mathbf{I}^{t,L}, \mathbf{I}^{t,R} \in \mathbb{R}^X$ at every time slice t. The hidden variables are 1) the disparity field $\mathbf{D}^t$ with entries $d^t_{\mathbf{x}} \in \mathbb{N}_0$, 2) the egomotion flow field $\mathbf{E}^t$ with entries $\mathbf{e}^t_{\mathbf{x}} \in \mathbb{R}^2$ and 3) the object flow $\mathbf{R}^t$ with entries $\mathbf{r}^t_{\mathbf{x}} \in \mathbb{Z}^2$, defined at all pixel locations $\mathbf{x} \in \mathbb{N}^2$ of the image. All hidden variables refer to the left camera but for the sake of clarity we neglect the index L. The egomotion of the left camera is introduced as parameters to the network and given by the deterministic state vector $\mathbf{s}^t = (T^t_x, T^t_y, T^t_z, \Omega^t_x, \Omega^t_y, \Omega^t_z)^T$ with the camera translation vector $\mathbf{T}^t = (T^t_x, T^t_y, T^t_z)^T$ and the camera rotation angles $\mathbf{\Omega}^t = (\Omega^t_x, \Omega^t_y, \Omega^t_z)^T$.

Since the observable $\mathbf{I}^{t+1,L}$ is a *head-to-head* node with respect to the path from $\mathbf{E}^t$ to $\mathbf{R}^t$ it follows from d-separation [1] that $\mathbf{E}^t$ and $\mathbf{R}^t$ are not independent. To reduce complexity of the model, we approximate the DBN in Fig. 3 a) by two separate networks shown in Fig. 3 b) and c) that split the computation in one Bayesian Network for the estimation of $\mathbf{E}^t$ and one DBN for the estimation of $\mathbf{R}^t$ assuming that the maximum aposteriori (MAP) estimate of the egomotion flow $\hat{\mathbf{E}}^t$ is observable. Further on, we neglect the temporal transition of the egomotion flow $P(E^{t+1}|E^t)$ (dashed arrow in Fig. 3 a)) because the specific egomotion of the humanoid robot ASIMO which we use to test our algorithm on is rapidly changing while he is walking and therefore a prediction of the movement is not straight forward.

2.1 Egomotion Estimation

The network in Fig. 3 b) is precisely defined by the specification of 1) the observation likelihood $P(\mathbf{I}^{t,R}|\mathbf{I}^{t,L},\mathbf{D}^t)$ of a pair of stereo images $(\mathbf{I}^{t,L},\mathbf{I}^{t,R})$ with their corresponding disparity field $\mathbf{D}^t$ and 2) the transition probability $P(\mathbf{E}^t|\mathbf{D}^t;\mathbf{s}^t)$ from the disparity field $\mathbf{D}^t$ to the egomotion flow $\mathbf{E}^t$ given the egomotion parameters $\mathbf{s}^t$. We assume the likelihood $P(\mathbf{I}^{t,L})$ to be a uniform distribution which can be neglected. For both the observation likelihood $P(\mathbf{I}^{t,R}|\mathbf{I}^{t,L},\mathbf{D}^t)$ and the transition probability $P(\mathbf{E}^t|\mathbf{D}^t;\mathbf{s}^t)$ we assume that they factorize over the image w.r.t. $\mathbf{E}^t$ and $\mathbf{D}^t$, i.e.,

$$P(\mathbf{I}^{t,R}|\mathbf{I}^{t,L},\mathbf{D}^t) := \prod_{\mathbf{x}} \ell(\mathbf{I}^{t,R}|\mathbf{I}^{t,L},\mathbf{e}^t_{\mathbf{x}}) \tag{1}$$

$$P(\mathbf{E}^t|\mathbf{D}^t;\mathbf{s}^t) := \prod_{\mathbf{x}} P(\mathbf{e}^t_{\mathbf{x}}|d^t_{\mathbf{x}};\mathbf{s}^t)\ . \tag{2}$$

This allows us to maintain only factored beliefs over $\mathbf{E}^t$ during inference making the approach computationally practicable. The likelihood measure is defined as

$$\ell(\mathbf{I}^{t,R}|\mathbf{I}^{t,L},d^t_{\mathbf{x}}) := \mathcal{N}(\mathbf{I}^{t,R}_{\mathbf{x}}|\lambda\mathbf{I}^{t,L}_{\mathbf{x}-d^t_{\mathbf{x}}}+\kappa,\mathbf{\Sigma}_{\ell x}) \propto e^{-\frac{1}{2}(1-C^{2,t}_{\mathbf{x}})/(\alpha(1+C^{2,t}_{\mathbf{x}})+\varepsilon)}\ . \tag{3}$$

For details on the notation and the derivation we refer to the next subsection 2.2 since the likelihood measurement in (3) is analogous to the likelihood measurement in (12) with the only difference that it is based on a correspondence measure between stereo images $(\mathbf{I}^{t,L},\mathbf{I}^{t,R})$ instead of a correspondence measure between temporal consecutive images $(\mathbf{I}^{t,L},\mathbf{I}^{t+1,L})$. Using the general mapping from disparity and egomotion to egomotion flow as given in [4] we define the transition probability for the egomotion flow given the disparity as

$$P(\mathbf{e}^t_{\mathbf{x}}|d^t_{\mathbf{x}};\mathbf{s}^t) := \mathcal{N}(\mathbf{e}^t_{\mathbf{x}}|\mu^t_{\mathbf{x}}(d^t_{\mathbf{x}},\mathbf{s}^t),\mathbf{\Sigma}_e)\ , \tag{4}$$

$$\mu^t_{\mathbf{x}}(d^t_{\mathbf{x}},\mathbf{s}^t) = \begin{pmatrix} \frac{T_zq}{bf}d^t_{\mathbf{x}}x + y\Omega_z + xy\Omega_x - x^2\Omega_y - (\frac{T_xq}{bf}d^t_{\mathbf{x}} + \Omega_y) \\ \frac{T_zq}{bf}d^t_{\mathbf{x}}y - x\Omega_z - xy\Omega_y + y^2\Omega_x - (\frac{T_yq}{bf}d^t_{\mathbf{x}} + \Omega_y) \end{pmatrix}\ . \tag{5}$$

Here, q denotes the pixel size, b the baseline and f the focal length. For the disparity prior $P(d^t_{\mathbf{x}}) := \mathcal{N}(d^t_{\mathbf{x}}|0,\sigma_d)$ we prefer small disparities to force unreliable measurements being far away in depth. Marginalizing $d^t_{\mathbf{x}}$ we are able to infer the egomotion flow as follows

$$P(\mathbf{e}^t_{\mathbf{x}}|\mathbf{I}^{t,R},\mathbf{I}^{t,L}) \propto \sum_{d^t_{\mathbf{x}}} \ell(\mathbf{I}^{t,R}|\mathbf{I}^{t,L},d^t_{\mathbf{x}})P(d^t_{\mathbf{x}})P(\mathbf{e}^t_{\mathbf{x}}|d^t_{\mathbf{x}};\mathbf{s}^t)\ . \tag{6}$$

Applying the MAP estimate results in the expected egomotion flow

$$\hat{\mathbf{E}}^t = \{\hat{\mathbf{e}}^t_{\mathbf{x}}\}_{\mathbf{x}} = \{\mathrm{argmax}_{\mathbf{e}^t_{\mathbf{x}}} P(\mathbf{e}^t_{\mathbf{x}}|\mathbf{I}^{t,R},\mathbf{I}^{t,L})\}\ . \tag{7}$$

As long as only the MAP estimate $\hat{\mathbf{E}}^t$ and not the whole probability $P(\mathbf{e}^t_{\mathbf{x}}|\mathbf{I}^{t,R},\mathbf{I}^{t,L})$ is used for further processing the choice of $\mathbf{\Sigma}_e$ does not influence the result and can be neglected. This saves the marginalization in (6) and results in a direct mapping from the MAP estimate of the disparity $\hat{\mathbf{D}}^t$ to $\hat{\mathbf{E}}^t$ applying (5).

2.2 Object Motion Filtering

Now we define the network in Fig. 3 c) by the specification of 1) the observation likelihood $P(\mathbf{I}^{t+1}|\mathbf{I}^t,\hat{\mathbf{E}}^t,\mathbf{R}^t)$ of a pair of consecutive images $(\mathbf{I}^t,\mathbf{I}^{t+1})$ with their corresponding egomotion flow $\hat{\mathbf{E}}^t$ and relative object flow $\mathbf{R}^t$ and 2) the transition probability $P(\mathbf{R}^{t+1}|\mathbf{R}^t)$ of the relative object flow. Note, that from now on we neglect the index L also for the consecutive images. For both the observation likelihood $P(\mathbf{I}^{t+1}|\mathbf{I}^t,\hat{\mathbf{E}}^t,\mathbf{R}^t)$ and the $\mathbf{R}$-transition probability $P(\mathbf{R}^{t+1}|\mathbf{R}^t)$ we again assume that they factorize over the image but w.r.t. $\mathbf{R}^t$ and $\mathbf{R}^{t+1}$, i.e.,

$$P(\mathbf{I}^{t+1}|\mathbf{I}^t,\hat{\mathbf{E}}^t,\mathbf{R}^t) := \prod_{\mathbf{x}} \ell(\mathbf{I}^{t+1}|\mathbf{I}^t,\hat{\mathbf{E}}^t,\mathbf{r}^t_{\mathbf{x}}) \tag{8}$$

$$P(\mathbf{R}^{t+1}|\mathbf{R}^t) := \prod_{\mathbf{x}} P(\mathbf{r}^{t+1}_{\mathbf{x}}|\mathbf{R}^t)\ . \tag{9}$$

The likelihood measure is based on a generative model for probabilistic flow field computation as proposed in [9]. We assume that the likelihood factor $\ell(\mathbf{I}^{t+1}|\mathbf{I}^t,\hat{\mathbf{E}}^t,\mathbf{r}^t_{\mathbf{x}})$ of a local image velocity $\hat{\mathbf{e}}^t_{\mathbf{x}}+\mathbf{r}^t_{\mathbf{x}}$ (which is in our case a superposition of egomotion and object flow) should be related to finding a scaled λ and biased κ image patch $\lambda\mathbf{I}^t_{\mathbf{x}-\mathbf{r}^t_{\mathbf{x}}-\hat{\mathbf{e}}^t_{\mathbf{x}}}+\kappa$ centered around $\mathbf{x}-(\mathbf{r}^t_{\mathbf{x}}+\hat{\mathbf{e}}^t_{\mathbf{x}})$ at time t in the image $\mathbf{I}^{t+1}$ but centered around $\mathbf{x}$, denoted $\mathbf{I}^{t+1}_{\mathbf{x}}$. This leads to

$$\begin{aligned}\ell(\mathbf{I}^{t+1}|\mathbf{I}^t,\hat{\mathbf{E}}^t,\mathbf{r}^t_{\mathbf{x}}) &:= \mathcal{N}(\mathbf{I}^{t+1}_{\mathbf{x}}|\lambda\mathbf{I}^t_{\mathbf{x}-\mathbf{r}^t_{\mathbf{x}}-\hat{\mathbf{e}}^t_{\mathbf{x}}}+\kappa,\mathbf{\Sigma}_{\ell x}) \\ &\approx \mathcal{N}(\tilde{\mathbf{I}}^{t+1}_{\mathbf{x}}|\lambda\mathbf{I}^t_{\mathbf{x}-\mathbf{r}^t_{\mathbf{x}}}+\kappa,\mathbf{\Sigma}_{\ell x})\ ,\end{aligned} \tag{10}$$

$$\mathbf{\Sigma}_{\ell x} := \begin{pmatrix} \ddots & \cdots & \mathbf{0} \\ \vdots & \frac{\sigma^2_{\ell x}}{\mathcal{N}(\mathbf{x}'|\mathbf{x},\rho_I)} & \vdots \\ \mathbf{0} & \cdots & \ddots \end{pmatrix}\ . \tag{11}$$

For reasons of computational efficiency, we first warp the image $\mathbf{I}^{t+1}$ backward applying the estimated egomotion flow $\hat{\mathbf{E}}^t$ and using bilinear interpolation which results in the egomotion compensated image $\tilde{\mathbf{I}}^{t+1}$. The function $\mathcal{N}(\mathbf{x}'|\mathbf{x},\rho_I)$ implements an isotropic homogeneous Gaussian weighting of the neighborhood x' centered around x. The parameter ρ_I defines the spatial range of the image patches and $\sigma^2_{\ell x}$ the grey value variance which is assumed to be dependent on position x. Following the same reasoning as given in [9], λ and κ are chosen to always maximize the likelihood with respect to these parameters. Additionally, the grey value variance for a grey value at position x is chosen to be a function $\sigma^2_{\ell x} := \alpha(s^2(\tilde{\mathbf{I}}^{t+1}_{\mathbf{x}})+\lambda^2 s^2(\mathbf{I}^t_{\mathbf{x}-\mathbf{r}^t_{\mathbf{x}}}))+\varepsilon s^2(\tilde{\mathbf{I}}^{t+1}_{\mathbf{x}})$ of the variances $s^2(\tilde{\mathbf{I}}^{t+1}_{\mathbf{x}})$ and $s^2(\mathbf{I}^t_{\mathbf{x}-\mathbf{r}^t_{\mathbf{x}}})$ of the two grey value patches $\tilde{\mathbf{I}}^{t+1}_{\mathbf{x}}$ and $\mathbf{I}^t_{\mathbf{x}-\mathbf{r}^t_{\mathbf{x}}}$ that are compared. This leads to the final likelihood measurement

$$\ell(\mathbf{I}^{t+1}|\mathbf{I}^t,\hat{\mathbf{E}}^t,\mathbf{r}^t_{\mathbf{x}}) \propto e^{-\frac{1}{2}(1-C^t_{\mathbf{x}})/(\alpha(1+C^t_{\mathbf{x}})+\varepsilon)}\ , \tag{12}$$

incorporating the squared weighted empirical correlation coefficient $C^t_{\mathbf{x}}$ between the egomotion compensated grey value patch $\tilde{\mathbf{I}}^{t+1}_{\mathbf{x}}$ and $\mathbf{I}^t_{\mathbf{x}-\mathbf{r}^t_{\mathbf{x}}}$ (similar to (3)). It ensures minimal influence on the likelihood accuracy by local changes in illumination. Here, the parameter α defines the noise proportion caused by the projection onto the camera chip and ε the noise proportion that considers the incompleteness of the generative model.

For the definition of the transition probability $P(\mathbf{R}^{t+1}|\mathbf{R}^t)$ of the relative object flow we follow the ideas given in [10] by assuming that the relative flow field component $\mathbf{R}^t$ transforms according to itself. This means, that a flow vector $\mathbf{r}^{t+1}_{\mathbf{x}}$ at position $\mathbf{x}$ equals the previous flow vector $\mathbf{r}^t_{\mathbf{x}'}$at position $\mathbf{x}'$. To obtain this position $\mathbf{x}'$ in the previous image, we assume that it is inferable from the flow field itself. Both assumptions read

$$\mathbf{r}^{t+1}_{\mathbf{x}} \sim \mathcal{N}(\mathbf{r}^{t+1}_{\mathbf{x}}|\mathbf{r}^t_{\mathbf{x}'}, \sigma_R) \ , \ \mathbf{x}' \sim \mathcal{N}(\mathbf{x}'|\mathbf{x} - \mathbf{r}^{t+1}_{\mathbf{x}}, \rho_R) \ . \tag{13}$$

Combining the two factors from (13) and integrating $\mathbf{x}'$ we get

$$P(\mathbf{r}^{t+1}_{\mathbf{x}}|\mathbf{R}^t) \propto \sum_{\mathbf{x}'} \mathcal{N}(\mathbf{x}'|\mathbf{x} - \mathbf{r}^{t+1}_{\mathbf{x}}, \rho_R)\mathcal{N}(\mathbf{r}^{t+1}_{\mathbf{x}}|\mathbf{r}^t_{\mathbf{x}'}, \sigma_R) \ . \tag{14}$$

We introduced new parameters ρ_R and σ_R for the uncertainty in spatial identification between two images and the transition noise between $\mathbf{R}^t$ and $\mathbf{R}^{t+1}$, respectively. The parameter ρ_R defines the spatial range of a flow-field patch, so we compare velocity vectors within flow-field patches at different times t and $t+1$.

For inference we need to propagate beliefs over the object flow field $\mathbf{R}^t$. The factored observation likelihoods and transition probabilities introduced in (8) and (9) ensure that the forward propagated beliefs

$$P(\mathbf{R}^t|\hat{\mathbf{E}}^{1:t}, \mathbf{I}^{1:t+1}) = \prod_{\mathbf{x}} P(\mathbf{r}^t_{\mathbf{x}}|\hat{\mathbf{E}}^{1:t}, \mathbf{I}^{1:t+1}) \tag{15}$$

will remain factored. Taking advantage of all the factorisation assumptions the belief propagation assembles to

$$\begin{aligned} P(\mathbf{r}^t_{\mathbf{x}}|\hat{\mathbf{E}}^{1:t-1}, \mathbf{I}^{1:t}) &\propto \sum_{\mathbf{x}'} \mathcal{N}(\mathbf{x}'|\mathbf{x} - \mathbf{r}^t_{\mathbf{x}}, \rho_R) \sum_{\mathbf{r}^{t-1}_{\mathbf{x}'}} \mathcal{N}(\mathbf{r}^t_{\mathbf{x}}|\mathbf{r}^{t-1}_{\mathbf{x}'}, \sigma_R) P(\mathbf{r}^{t-1}_{\mathbf{x}'}|\hat{\mathbf{E}}^{1:t-1}, \mathbf{I}^{1:t}) \\ &\approx \sum_{\mathbf{x}'} \mathcal{N}(\mathbf{x}'|\mathbf{x} - \mathbf{r}^t_{\mathbf{x}}, \rho_R) P(\mathbf{r}^{t-1}_{\mathbf{x}'} = \mathbf{r}^t_{\mathbf{x}}|\hat{\mathbf{E}}^{1:t-1}, \mathbf{I}^{1:t}) \ . \end{aligned} \tag{16}$$

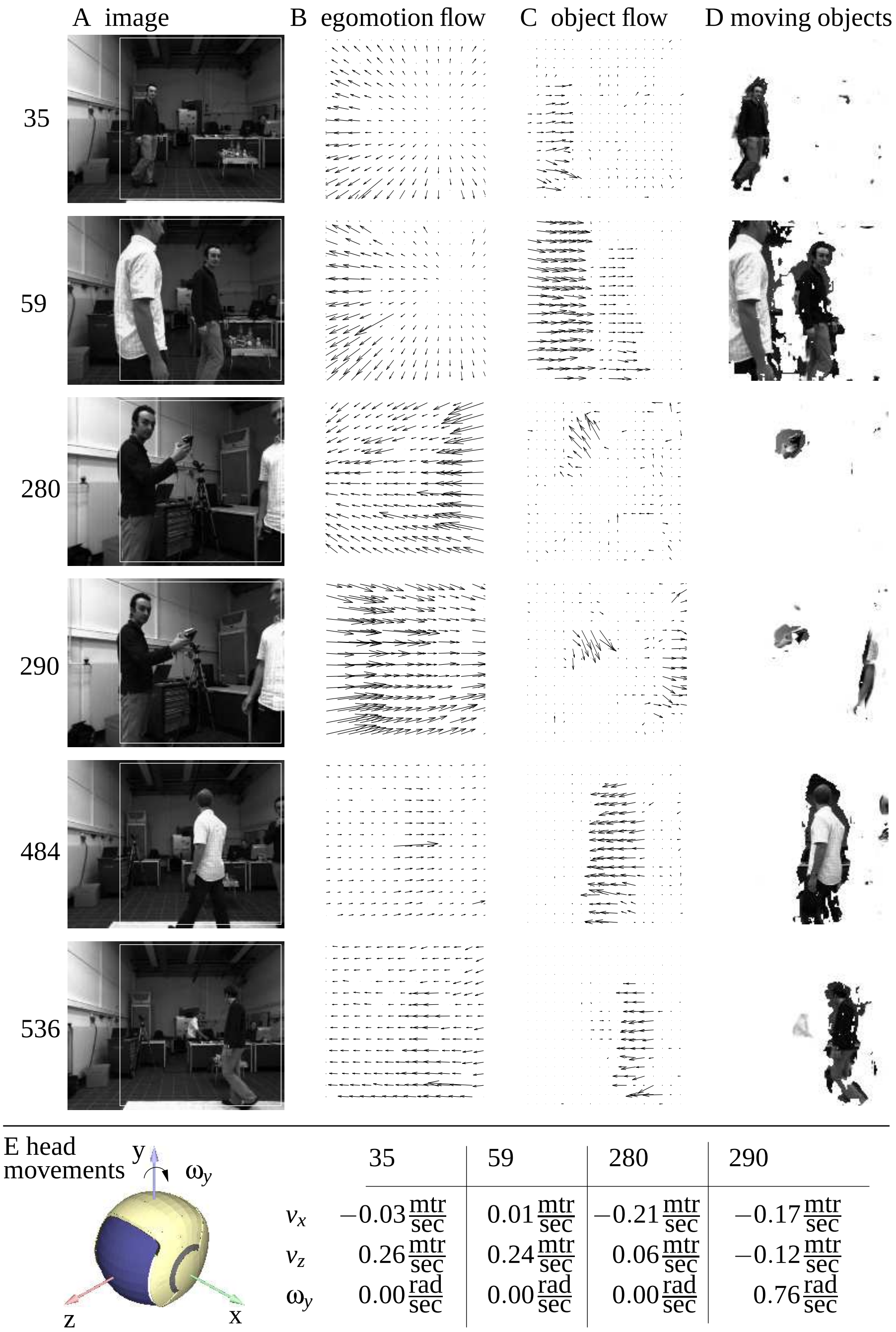
To speed up the computation, we simplified the filtering equation applying the limit $\sigma_R \rightarrow 0$ and thereby eliminating the sum of $\mathbf{r}^{t-1}_{\mathbf{x}'}$ and the factor $\mathcal{N}(\mathbf{r}^t_{\mathbf{x}}|\mathbf{r}^{t-1}_{\mathbf{x}'}, \sigma_R)$. The final inference step is the combination of the propagated belief with the actual observation using Bayes' theorem

$$P(\mathbf{r}^t_{\mathbf{x}}|\hat{\mathbf{E}}^{1:t}, \mathbf{I}^{1:t+1}) \propto \ell(\mathbf{I}^{t+1}|\mathbf{I}^t, \hat{\mathbf{E}}^t, \mathbf{r}^t_{\mathbf{x}}) P(\mathbf{r}^t_{\mathbf{x}}|\hat{\mathbf{E}}^{1:t-1}, \mathbf{I}^{1:t}) \ , \tag{17}$$

which can be done in parallel because it is a local operation for every location $\mathbf{x}$.

3 Results

The experiments are carried out with a Honda ASIMO robot. The computation is performed on a Pentium 4 single-core with 3.4 GHz and the images are captured with a constant framerate of 12 Hz. The whole scene consists of 550 images with an image resolution of 150×200 pixels and was recorded while Asimo was walking forward on an S-shaped path, superimposed with a rotation of the body about 45° in the second half. From the end-point he walked backwards to its starting position, turning his body straight



	35	59	280	290
v_x	$-0.03\frac{\mathrm{mtr}}{\mathrm{sec}}$	$0.01\frac{\mathrm{mtr}}{\mathrm{sec}}$	$-0.21\frac{\mathrm{mtr}}{\mathrm{sec}}$	$-0.17\frac{\mathrm{mtr}}{\mathrm{sec}}$
v_z	$0.26\frac{\mathrm{mtr}}{\mathrm{sec}}$	$0.24\frac{\mathrm{mtr}}{\mathrm{sec}}$	$0.06\frac{\mathrm{mtr}}{\mathrm{sec}}$	$-0.12\frac{\mathrm{mtr}}{\mathrm{sec}}$
ω_y	$0.00\frac{\mathrm{rad}}{\mathrm{sec}}$	$0.00\frac{\mathrm{rad}}{\mathrm{sec}}$	$0.00\frac{\mathrm{rad}}{\mathrm{sec}}$	$0.76\frac{\mathrm{rad}}{\mathrm{sec}}$

Figure 4: Results produced by the proposed motion estimation. It is shown the sequence with the area that is processed by the algorithm marked with a white rectangle (A), the egomotion flow (B), the object flow (C), the moving objects (D) and some corresponding head movements (E).

forward again. Along the way, two persons crossing several times in front of ASIMO and handle objects in front of him. Once the parameters are chosen, no adaptation to the scene is needed. The most critical parameters in terms of runtime are the disparity range U and the velocity range W which have to cover the minimum depth and the maximum speed of the objects you want to detect in the scene. The larger these ranges the higher the computational costs. We set the parameters as follows: $U = 50$, $W = 5 \times 21$, $\rho_I = 3$ with a filter length of $l_I = 7$, $\rho_V = 5$ with a filter length of $l_V = 11$, $\alpha = 0.05$, $\varepsilon = 0.05$. Since all filters are 2D Gaussians they can be separated which results in a computational complexity for the whole algorithm of $O(U2l_IX)$ for the egomotion flow and $O(W2l_IX) + O(W2l_VX)$ for the object flow. In Fig. 4 several snapshots at certain times t of the sequence (A), the egomotion flow (B) and the object flow (C) are presented. Additionally, image segments (D) are shown that exceed a velocity amplitude of one pixel per frame in the object flow field. The egomotion flow results (B) comprise typical motion patterns, e.g. mainly divergent flows ($t = 35, 59$) if the robot moves straight ahead or mainly translating flows ($t = 280, 290, 484$) if the robot swings because of stepping from one foot to the other or rotates his head. In Fig. 4 (E) some components of the head movements are shown, like the velocity in x- and z-direction v_x and v_z and the angular velocity about the y-axis ω_y. As long as the egomotion measurements are correct and the predictive assumptions of the filter hold the object flow results (C) indicate what kind of movements the objects carry out in front of the robot. As depicted in (D) object movements, like e.g. walking persons, or body parts of persons handling objects, like e.g. arms, can be extracted. The results have been computed offline with a framerate of 3 Hz using an optimized C implementation. With a reduced velocity range of $W = 25$ we achieve realtime performance with a framerate of 12 Hz (which was the capturing framerate of the sequence).

4 Conclusion

The proposed motion estimation system allows for a separation of egomotion flow and object flow. Beside some minor errors mainly because of wrong disparity measurements in the stereo algorithm we achieve quite smooth object flow fields, which is not the case without spatiotemporal filtering. Nevertheless, problems arise if the egomotion flow is wrong or imprecise. Both, disparity and the optical flow measurements become unreliable at object boundaries because the underlying generative model cannot handle overlapping regions. In order to improve the existing algorithm also the egomotion flow should be spatiotemporally filtered. However, if the camera movements are rapidly changing a prediction of the movement is difficult to realize.

The presented algorithm can serve as a good starting position for applications, like 2D segmentation and/or 3D-motion estimation of moving objects. In particular, we consider an active visual scene tracking system by setting up a control loop that tries to compensate the movement of segregated objects and fulfills smoothness constraints on the robot movements.

Acknowledgments

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

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science+Business Media, 2006.

[2] D. Chekhlov, M. Pupilli, W. Mayol, and A. Calway. Robust real-time visual slam using scale prediction and exemplar based feature description. In *Proc. IEEE Conf. on CVPR*, pages 1–7, June 2007.

[3] D. Comaniciu and B. Xie. Real-time obstacle detection with a calibrated camera and known ego-motion. In *US patent 20040183905, to Siemens Corp.*, pages 1–18, February 2004.

[4] M. Irani and P. Anandan. A unified approach to moving object detection in 2d and 3d scenes. *IEEE Trans. on PAMI*, 20(6):577–589, 1998.

[5] G. Overett and D. Austin. Stereo vision motion detection from a moving platform. In *Proc. Australasian Conf. on Robotics and Automation*, December 2004.

[6] C. Rabe, U. Franke, and S. Gehrig. Fast detection of moving objects in complex scenarios. In *Proc. IEEE Symp. on Intelligent Vehicles*, pages 398–403, June 2007.

[7] S. Soatto and P. Perona. Reducing structure from motion part 1: modeling. *IEEE Trans. on PAMI*, 20(9):933–942, 1998.

[8] T.Y. Tian, C. Tomasi, and D.J. Heeger. Comparison of approaches to egomotion computation. In *Proc. IEEE Conf. on CVPR*, pages 315–320, June 1996.

[9] V. Willert, J. Eggert, J. Adamy, and E. Körner. Non-gaussian velocity distributions integrated over space, time, and scales. *IEEE Trans. on SMCB - Part B*, 36(3):482–493, 2006.

[10] V. Willert, M. Toussaint, J. Eggert, and E. Körner. Uncertainty optimization for robust dynamic optical flow estimation. In *Sixth Int. Conf. on Machine Learning and Applications (ICMLA)*, pages 450–457, December 2007.