An Iterative Scheme for Motion-Based Scene Segmentation

Alexander Bachmann[†] and Hildegard Kuehne[‡]
[†] Department for Measurement and Control
[‡] Institute for Anthropomatics
University of Karlsruhe (TH), 76 131 Karlsruhe, Germany
bachmann@mrt.uka.de, koehler@ira.uka.de

Abstract

We present an approach for dense estimation of motion and depth of a scene containing a multiple number of differently moving objects with the camera system itself being in motion. The estimates are used to segregate the image sequence into a number of independently moving objects by assigning the object hypothesis with maximum a posteriori (MAP) probability to each image point. Different to previous approaches in 3-dimensional (3D) scene analysis, we tackle this task by first simultaneously estimating motion and depth for a salient set of feature points in a recursive manner. Based on the evolving set of estimated motion profiles, the scene depth is recovered densely from spatially and temporally separated views. Given the dense depth map and the set of tracked motion estimates, the likelihood of each image point to belong to one of the distinct motion profiles can be determined and dense scene segmentation can be performed. Within our probabilistic model the expectation-maximization (EM) algorithm is used to solve the inherent missing data problem. A Markov Random Field (MRF) is used to express our expectations on spatial and temporal continuity of objects.

1. Introduction

This contribution addresses the detection of independently moving objects in a traffic scene as a stereo camera platform moves through it. Object detection is performed based on the *relative motion* of textured objects and the observer. To obtain a dense representation of the observed scene, object detection is formulated as an image segmentation task where each image point is tested for consistency with a set of possible hypotheses, each defined by its relative 3D motion. Although a field of active research for decades, a general and robust solution to this problem is still elusive, since the observable 2D image motion is generated by the combined effects of camera motion, the motion of independently moving objects and scene structure. As scene structure basically denotes the depth of a scene point relative to the observing camera, in the sequel we use the term scene depth likewise. Isolating these three factors proves to be a difficult task as depth discontinuities and independently moving objects both cause discontinuities in the image motion field. Therefore it is not possible to separate these factors without 3D motion and structure estimation. Methods for finding the motion and structure of an entire scene can be categorized into either (i) 'direct' methods [10], where the unknown motion and structure parameters are recovered simultaneously from measurable image quantities at each pixel in the image(s) or (ii) 'indirect' methods [14], which rely on a sparse set of distinct feature points that are extracted beforehand from the image(s). As for our application, the detection and segmentation of moving traffic participants, the principle requirements on our segmentation scheme are a sufficiently dense representation of the scene, *i.e.* a maximum number of scene points should be reconstructed, a reliable identification and correction of erroneous points, the ability to cope with general object motion and computational efficiency, we pursue the strategy of (ii) when estimating the motion components and then use the motion estimates to guide the computation of a dense depth and segmentation map in a direct manner – also in regions of the image where there is less information. Assuming rigidity of the individual scene objects, the 3D motion estimates can be used as a strong guide to stereo and temporal matching in recovering a dense scene structure and subsequently a dense scene segmentation. Object motion is estimated using an EM-based approach that consists of a multiple object tracking filter with probabilistic data association. In the E-step, the probability of an observation to belong to each of the object hypothesis is computed based on the current state estimates. In the M-step, the state and error covariance of each object is robustly updated in a recursive manner. Based on these motion estimates the scene structure is then recovered densely from spatially and temporally separated views. In the segregation step, we derive a global cost function that incorporates the motion and structure estimates while considering the fundamental property of spatial and temporal label consistency and segregate the image accordingly.

c) 2009 IEEE. Personal use of this material is permitted.
Permission from IEEE must be obtained for all other users, including reprinting/ republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

The remainder of the paper is organized as follows: Section 2 recapitulates the motion segmentation task and introduces notation and constraints used within this work. In Section 3 the segmentation filter framework is presented. The performance of our approach is illustrated in Section 4 on real and synthetic image data.

2. Problem Formulation

Following the derivations of Longuett-Higgins and Prazdny [12], for each object in the scene we consider the equivalent problem of a stationary object and a moving observer, *i.e.* we express the entire scene dynamics by a set of rigidly moving objects, each measured relative to a camera-fixed coordinate system. For each object, the instantaneous rigid body motion of this coordinate system is specified by the translational and rotational motion of its origin, $\mathbf{t} = (t_x, t_y, t_z)^T$ and $\Omega = (\omega_x, \omega_y, \omega_z)^T$, respectively. Formally this can be expressed by the 3D motion field $\omega(\mathbf{X}) \in \mathbb{R}^3$, with the parameterized motion of each point $\mathrm{X} = (X, Y, Z) \in \mathbf{X}$ being

$$\omega(\mathrm{X}) = \left(\dot{X}, \dot{Y}, \dot{Z}\right)^{\mathrm{T}} = -\mathbf{t} - \Omega \times \mathrm{X}\,. \tag{1}$$

The computation of the motion field is obviously underdetermined due to the projection $\mathrm{u}(\mathrm{x}) \in \mathbb{R}^2$ of $\omega(\mathrm{X})$ onto the image plane. The task becomes even more ambitious if the observer itself is moving and the scene consists of a multiple number of differently moving objects. Consequently, the flow field $\mathrm{u}(\mathbf{x})$ is generated by an unknown number of unknown object motions with the observer motion superimposing all motion vectors ($\mathbf{x}$ states the collection of all image points). A well studied area in this context is 2D motion estimation, which is concerned with the determination of the flow field $\mathrm{u}(\mathbf{x})$. Here, variational techniques based on the method of [8] yield the most accurate results as *e.g.* presented in [3]. The major limitation of 2D motion estimation is the fact that motion cues present in the 2D projection are insufficient to reconstruct the motion present in the 3D scene. The relatively young area of dense 3D motion estimation resolves this limitation by additionally estimating the depth of the scene. Here, good results have been achieved using the variational framework. Next to the computationally expensive joint estimation of motion and structure at the same time [9], in [15] a method is presented that separates the problem into the two sub-problem scene flow and depth estimation, resulting in an efficient computation of the individual problems.

In our approach we integrate 3D motion and depth estimation into one feature-based approach. Different to the approaches above, we apply a parametric motion model, expressing the motion field as a collection of of a rigidly moving objects with each object being specified by its motion parameters. Based on the joint estimate of depth and motion on the basis of a sparse set of features, we compute a dense depth map of the scene and derive dense scene flow. In the segmentation step, each image point is then assigned to the object hypothesis, expressed by a unique set of motion parameters, that explains the underlying motion best.

Object state. In our framework, the 3D motion for each object entity is expressed by the continuous-valued variable θ_t. We use a factorial representation of our state vector with $\theta_t = \{\theta_t^1, \ldots, \theta_t^j, \ldots, \theta_t^J\}$, representing the quantitative state information for each object $\theta_t^j = (\omega_x, \omega_y, \omega_z, t_x, t_y, t_z)^{\mathrm{T}}$, independently. The number of objects J within the current, discrete time instant t is assumed to be fixed. θ_t is estimated for each object j independently based on a distinct set of feature points that are tracked over time. More details will be given in Section 3.2.

Vision cues. In our implementation we use two vision cues to obtain dense 3D structure and motion information of a scene from multiple images. One is *stereo vision*, which drastically facilitates the problem stated above, as the scene is reconstructed from two views recorded at the same time, *i.e.* no constraints (*e.g.* rigidity, *etc.*) need to be imposed. The other is *visual motion*, which expresses the displacement of image points in temporally separated views caused by the relative scene motion in between.

We integrate stereo vision and visual motion into one probabilistic framework by introducing a depth related parametrization of the spatial and temporal (epipolar) constraints [13]. The approach is based on a depth estimate obtained by stereo vision, which greatly simplifies the complexity of the segmentation task. Using the initial information of the stereo reconstruction, the motion parameters can be obtained without scalar ambiguity. After obtaining the motion parameters, the initial depth estimates $\hat{\mathbf{z}}_t$ can be further improved by integrating the motion cue and therefore more epipolar geometry constraints can be used.

Our camera setup consists of a fully calibrated stereo rig with the world origin at the right camera (quantities that are related to the right camera are indexed with character 'r' in the sequel). In normalized image coordinates $\tilde{\mathrm{x}} = (x, y, 1)$, the mapping from an image point $\tilde{\mathrm{x}}_{\mathrm{r}}$ in the right camera to an image point $\tilde{\mathrm{x}}'$ in a second camera is

$$\tilde{\mathrm{x}}' = \mathbf{K}'\mathbf{R}\mathbf{K}_{\mathrm{r}}^{-1}\tilde{\mathrm{x}}_{\mathrm{r}} + \mathbf{K}'\mathbf{t}Z\,. \tag{2}$$

$\mathbf{K}_{\mathrm{r}}$ and $\mathbf{K}'$ state the camera calibration matrices of the two cameras respectively. $\mathbf{R}$ is the 3x3 rotation matrix and $\mathbf{t}$ the 1x3 translation vector specifying orientation and pose of the second camera relative to the right camera. It can be seen that the mapping is divided into a component that depends on the image position alone but not on the depth Z. This term takes account of the camera rotation. The second term depends on Z but not on the image position and scales with the amount of translation between the cameras, *i.e.* $\tilde{\mathrm{x}}'$

moves along the epipolar line as a function of the inverse depth, starting from $Z = \infty$ and going in the direction of the epipole $\mathbf{K}'\mathbf{t}$.

Concerning the *stereo cue*, $\tilde{\mathrm{x}}'$ represents the corresponding image point $\tilde{\mathrm{x}}_\mathrm{l}$ in the left image. Given the internal and external camera parameters, the only unknown in (2) is $d = 1/Z$ and the corresponding point in the left image, along the epipolar line, can be parameterized by

$$s\left(\mathrm{x}_\mathrm{r}, d\right) = s = \mathrm{x}_\mathrm{r} + d\mathbf{K}_\mathrm{l}\mathbf{t}\,. \tag{3}$$

Above, the correspondence search has been drastically simplified by rectifying the stereo images such that the epipolar lines coincide with the corresponding horizontal scan lines in the warped images, *i.e.* $\mathbf{t} = (t_x, 0, 0)^\mathrm{T}$.

For the *motion cue*, $\tilde{\mathrm{x}}'$ consists of the temporally corresponding image point $\tilde{\mathrm{x}}_{t+1,\mathrm{r}}$ in the right camera. The epipolar constraint here expresses the geometrical relationship of point correspondences between two views due to the motion of the stereo camera system. The feature-based estimation of object motion θ_t from point correspondences is explained in Section 3.2. An adequate description, which formally describes the displacement of an image point as a function of 3D motion and (inverse) depth, is given by

$$\begin{aligned} \mathrm{u}_{t,\mathrm{i}} = \mathbf{C}_\Omega \Omega_t + d\mathbf{C}_\mathbf{t}\mathbf{t}_t\,,\ \text{with}\ \mathbf{C}_{\mathbf{t}_t} &= \begin{bmatrix} -\mathrm{f} & 0 & x \\ 0 & -\mathrm{f} & y \end{bmatrix} \\ \text{and}\ \ \mathbf{C}_\Omega &= \begin{bmatrix} \frac{x_\mathrm{i} y_\mathrm{i}}{\mathrm{f}} & -\frac{\left(\mathrm{f}^2 + x_\mathrm{i}^2\right)}{\mathrm{f}} & y_\mathrm{i} \\ \frac{\left(\mathrm{f}^2 + y_\mathrm{i}^2\right)}{\mathrm{f}} & -\frac{x_\mathrm{i} y_\mathrm{i}}{\mathrm{f}} & -x_\mathrm{i} \end{bmatrix}. \end{aligned} \tag{4}$$

The corresponding image point in image G_{t+1} for image point x_r in image G_t, parameterized by d, then is

$$m\left(\mathrm{x}_{t,\mathrm{r}}, d, \theta_t\right) = m = \mathrm{x}_{t,\mathrm{r}} + \mathbf{C}_\Omega \Omega_t + d\mathbf{C}_\mathbf{t}\mathbf{t}_t\,. \tag{5}$$

The constraints introduced above are used within the motion and structure estimation process to efficiently guide the correspondence search and evaluate current estimates as presented in the sequel.

Observations. Information from the environment is acquired through a sequence of observations $\mathbf{Y}_{0:t} = (\mathbf{Y}_0, \ldots, \mathbf{Y}_t)$, with $\mathbf{Y}_t = \{\mathrm{Y}_{t,1}, \ldots, \mathrm{Y}_{t,i}, \ldots, \mathrm{Y}_{t,N}\}$ being a set of random variables. Throughout the paper, $\mathbf{y}_t = \{\mathrm{y}_{t,1}, \ldots, \mathrm{y}_{t,i}, \ldots, \mathrm{y}_{t,N}\}$ represents a sample realization of $\mathbf{Y}_t$. N states the number of image points. The reconstruction and segmentation process are directly evaluated on the image gray values $g_t(\mathrm{x}_i) = g_{t,i} \in \mathrm{G}_t$, assuming observations to be i.i.d. Gaussians. Concerning the stereo cue, the similarity between gray value $g_{t,i}$ at image position $\mathrm{x}_{t,i}$ in the right camera ('r' is omitted in the sequel) and g_{t,s_i} at image position s_i in the left image, parameterized by $\mathrm{z}_{t,i}$, is expressed by

$$P(\epsilon^s_{t,i}|\mathrm{z}_{t,i}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(g_{t,i} - g_{t,s_i}\right)^2}{2\sigma^2}\right). \tag{6}$$

The similarity of the motion cue is evaluated along the temporal epipolar line and is written

$$P(\epsilon^m_{t,i}|\theta_t, \mathrm{z}_{t,i}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(g_{t,i} - g_{t,m_i}\right)^2}{2\sigma^2}\right), \tag{7}$$

with m_i stating the corresponding point in the next right image, parameterized by θ_t. σ states the error distribution of motion and stereo cue. With this, our observations consist of the combined error map $\mathbf{E}_t$, which is composed of the error ϵ^m_t and ϵ^s_t at each pixel position. For optimal motion and depth estimates, $\mathbf{E}_t$ should reach values close to or equal to zero.

Association process. The fact that the scene consists of a multiple number of differently moving objects is considered next by introducing a data association process. Obviously, this is a crucial task for motion estimating as proposed in this work. Different to methods that are operating on pixel level to determine the local motion field, our approach operates on a global scale in the entire image. A wrong or erroneous assignment of observations in the motion estimation process would lead to heavily corrupted reconstruction and segmentation results. Generally, the different approaches can be classified as either *hard* or *soft* data assignment method. Hard denotes the assignment of an observation to one (and only one) hypothesis, whereas soft means the assignment of an observation to an object hypothesis proportional to some weight. We pose the problem of multiple object tracking and scene segmentation as incomplete data problem with the observations being the incomplete data, whereas the object-associated observations state the complete data.

To capture the unknown relationship between an observation and the object that caused it, an association process $\mathbf{L}_t$ is introduced with its components being defined

$$l^j_{t,i} = \begin{cases} 1 \text{ if } \mathrm{y}_{t,i} \text{ originated from object } j \\ 0 \text{ else} \end{cases}. \tag{8}$$

A sample realization of the association process is defined as binary label field $\mathbf{l}_t = (\mathrm{l}_{t,1}, \ldots, \mathrm{l}_{t,N})$ with a label vector $\mathrm{l}_{t,i} = \mathrm{e}_j$ for each observation i. e_j states a unity vector of length J. This formalizes our assumption that an observation $\mathrm{y}_{t,i}$ originates from exactly one object j. For notational convenience we restrict our description to only one object instance within θ_t in the sequel and skip the hypothesis index j. If necessary, we will resort to it in the text. To account for observations with a low confidence, *i.e.* all hypotheses seem to be equally unlikely for that observation, we further introduce an *ambiguity* label (AMB) which will be expressed by $j = 0$. In our estimation scheme the association process is interpreted as missing data regarding the observations. Finally, we define the complete data through time as $\mathbb{E}_{0:t} = \{\mathbf{E}_{0:t}, \mathbf{L}_{0:t}\}$.

3. Segmentation Filter

The aim is now to find the optimal motion and depth estimates and derive from that a segmentation that segregates the scene into regions of similar 3D motion. In this contribution we carry on the work presented in [1], where an iterative scheme is proposed that splits the problem stated above into a set of sub-problems which are then solved separately. Concerning the estimation of θ_t the EM framework [5] is applied, which consists of iteratively computing the expected complete data in the E-step and afterwards estimating the state, based on the complete data, in the M-step. Different to the standard EM algorithm, a penalized maximum likelihood (ML) estimate is obtained (see *e.g.* [6]), leading to a MAP estimate of θ_t according to the Bayesian recursive update rule. The presented approach is time-recursive in that the motion estimates from the previous time instant are used as prior. Based on the motion estimates, the scene depth $\hat{\mathbf{z}}_t$ is then recovered densely from spatially and temporally separated views. Each EM-loop terminates with the segregation of the image into a set of disjoint, non-overlapping regions, resulting in a hard scene segmentation $\hat{\mathbf{l}}_t$. Though the approach presented in [1] incorporates temporal dependencies in the motion estimates, it ignores them in the segmentation and depth estimation process. Therefore, in this contribution, we extend our existing framework by temporal dependencies of the segmentation and depth estimation process, resulting in a spatially and temporally consistent scene segmentation scheme. Following the notation of [1], each iteration of the proposed algorithm consists of the following steps on the $(k+1)$-th iteration, with $\Theta_t = \{\theta_t, \mathbf{z}_t\}$ for notational convenience.

In the **E-step**, the conditional expectations of $\mathbf{l}_t$ are computed based on the actual observations and state estimates, which is equivalent to computing the probabilities of an image point to belong to each of the object hypotheses

$$\mathrm{Q}(\Theta_t|\hat{\Theta}_t^k) = \mathrm{E}[\log P(\mathbb{E}_t|\Theta_t^k, \mathbb{E}_{t-1})|\mathbf{E}_t, \hat{\Theta}_t^k] \,. \tag{9}$$

A segmentation

$$\hat{\mathbf{l}}_t^{k+1} = \arg\max_{\mathbf{l}_t} \left\{ \mathrm{Q}(\Theta_t|\hat{\Theta}_t^k) \right\} \,, \tag{10}$$

is derived from these conditional expectations in the image segmentation step. The segmentation is used to derive object-specific data, as *e.g.* the gray value/depth distribution within the object boundaries. The conditional probabilities are then used in the motion update of the **M-step**

$$\hat{\Theta}_t^{k+1} = \arg\max_{\theta_t, \mathbf{z}_t} \left\{ \mathrm{Q}(\Theta_t|\hat{\Theta}_t^k) + \log P(\Theta_t|\mathbb{E}_{t-1}) \right\} \,, \tag{11}$$

to weight the observations. Scene depth $\hat{\mathbf{z}}_t$ is recovered densely from spatially and temporally separated views. $k \in \{1, \ldots, K\}$ states the iteration index. The two steps are repeated until either the parameter estimates converge or some maximum number of iterations is reached.

3.1. E-step

By only considering data from the present and previous time step, the likelihood term in (9) gets

$$\begin{aligned} \log P(\mathbb{E}_t|\Theta_t, \mathbb{E}_{t-1}) &= \log P(\mathbf{E}_t, \mathbf{l}_t|\Theta_t, \mathbb{E}_{t-1}) = \\ \log P(\mathbf{E}_t|\mathbf{l}_t, \Theta_t, \mathbb{E}_{t-1}) &+ \log P(\mathbf{l}_t|\Theta_t, \mathbb{E}_{t-1}) \,. \end{aligned} \tag{12}$$

Regarding our label prior, spatial dependencies are incorporated into the model to account for the natural notion that physical objects extend in space. The final Q-function gets [1]

$$\begin{aligned} \mathrm{Q}(\Theta_t|\hat{\Theta}_t^k) = \sum_{i=1}^{N} \mathrm{E}[\mathrm{l}_{t,i}|\mathbf{E}_t, \hat{\Theta}_t^k, \mathbb{E}_{t-1}] \\ \big(\mathrm{D}(\epsilon_{t,i}|\Theta_t) - \mathrm{V}_1(\Theta_t)\big) - \sum_{i,n \in c} \mathrm{E}[\mathrm{l}_{t,i}^{\mathrm{T}} \mathrm{V}_2(\Theta_t) \mathrm{l}_{t,n}|\hat{\Theta}_t^k] \,. \end{aligned} \tag{13}$$

$\mathrm{V}_2(\Theta_t)$ is a matrix of dimension $J \times J$ with element $\{j, u\}$ equals to $\log P(\mathrm{l}_{t,i} = \mathrm{e}_j, \mathrm{l}_{t,n} = \mathrm{e}_u) = \lambda(\mathrm{e}_j^{\mathrm{T}} \mathrm{e}_u)$. λ is a regularization constant rating the influence of neighbouring sites to the prior term. This model can be interpreted as the well known Potts model. A detailed mathematical derivation of the single terms can be found in [1].

To capture the strong statistical dependency of the assignment process in the temporal domain (as image points with coherent label value are expected to follow the underlying, true object along a smooth trajectory, parameterized by state variable θ_t) we extend the single elements of our first-oder clique $\mathrm{V}_1(\cdot)$ from above in such a way that

$$\begin{aligned} \mathrm{V}_1(\Theta_t, \mathbb{E}_{t-1}) &= [\log P(\mathrm{e}_1|\Theta_t, \mathbb{E}_{t-1}), \\ \ldots, \log P(\mathrm{e}_j|\Theta_t, \mathbb{E}_{t-1}), &\ldots, \log P(\mathrm{e}_J|\Theta_t, \mathbb{E}_{t-1})]^{\mathrm{T}} \,. \end{aligned} \tag{14}$$

We take the corresponding fixed label field estimate from the previous time step and evaluate the label consistency over time. Therefore, for each image point at time t, its expected location in the previous image must be determined. This is done by deriving the expected image coordinates from (5), given state estimate θ_t. With this simple image warping method, the corresponding label value in the previous label field can be determined. As a measure of label similarity along the expected object trajectory we propose to evaluate the similarity of the assumed object motion in the present and in the previous image. This is quantified by the translational motion component $\mathbf{t}_t \subset \theta_t$. The individual elements of $\mathrm{V}_1(\Theta_t, \mathbb{E}_{t-1})$ then are

$$\log P(\mathrm{e}_j|\Theta_t, \mathbb{E}_{t-1}) = \left(\mathbf{t}_t^j - \mathbf{t}_t^u\right)^{\mathrm{T}} \Sigma^{-1} \left(\mathbf{t}_t^j - \mathbf{t}_t^u\right) \,, \tag{15}$$

with u being the hypothesis index which has been extracted at the back-projected position in $\hat{\mathbf{l}}_{t-1}$. Equation (15) quantifies the dissimilarity of the two state vectors indexed j and u

[1] $\mathrm{D}(\epsilon_{t,i}|\Theta_t) = [\log P(\epsilon_{t,i}|\mathrm{e}_1, \Theta_t), \ldots, \log P(\epsilon_{t,i}|\mathrm{e}_J, \Theta_t)]^{\mathrm{T}}$
$\mathrm{V}_1(\Theta_t) = [\log P(\mathrm{e}_1|\Theta_t), \ldots, \log P(\mathrm{e}_J|\Theta_t)]^{\mathrm{T}}$

under the assumption that they are distributed Gaussian. Σ states a fixed, isotropic covariance matrix. With this measure we enforce a smooth propagation of our association process over time.

Finally, the posterior probability of the label variable at position x_i is $\mathrm{E}[\mathrm{l}_{t,i}|\epsilon_{t,i}, \hat{\theta}_t^k, \hat{\mathbf{z}}_{t,i}^k, \hat{\mathbf{l}}_{t-1}] = \pi_{t,i}$, with the j-th element

$$\pi_{t,i}^j = \frac{P(\epsilon_{t,i}|\mathrm{l}_{t,i} = \mathrm{e}_j, \hat{\Theta}_t^k) P(\mathrm{l}_{t,i} = \mathrm{e}_j|\hat{\Theta}_t^k)}{\sum_{s=1}^J P(\epsilon_{t,i}|\mathrm{l}_{t,i} = \mathrm{e}_s, \hat{\Theta}_t^k) P(\mathrm{l}_{t,i} = \mathrm{e}_j|\hat{\Theta}_t^k)} , \qquad (16)$$

expressing the probability that x_i is assigned to object hypotheses j.

Soft data assignment. A probabilistic data association measure is obtained by applying pseudo-likelihood (PL) approximation. The PL is evaluated by restricting the statistical dependencies of the label field in above expression to the local neighborhood $\mathcal{G}_i$ of each point, *i.e.*

$$P(\mathrm{l}_{t,i}|\Theta_t) \approx P(\mathrm{l}_{t,i}|\pi_{t,u}^{k-1}, u \in \mathcal{G}_i, \Theta_t) , \qquad (17)$$

with $\pi_{t,u}^{k-1}$ being the estimates from the previous iteration step.

Scene segmentation. In the segmentation step the label that generates the highest probability is assigned to each image point. C quantifies the overall costs of a segmentation with each erroneously assigned image point producing the same costs. Our Bayesian decision rule assigns the hypotheses with MAP probability to each image point and therefore minimizes C, *i.e.* minimizes the number of segmentation errors. Based on our test statistic $\pi_{t,i}$, we formulate our segmentation problem as

$$\hat{\mathbf{l}}_t^{k+1} = \arg \min_{\mathbf{l}_t} \left\{ \sum_{i=1}^N \mathrm{D}(\epsilon_{t,i}|\hat{\Theta}_t^k) + \mathrm{V}_1(\hat{\Theta}_t^k, \mathbb{E}_{t-1}) + \sum_{i,n \in c} \mathrm{l}_{t,i}^{\mathrm{T}} \mathrm{V}_2(\hat{\Theta}_t^k) \mathrm{l}_{t,n} \right\} . \qquad (18)$$

An optimal labeling is found using a discrete energy minimization technique based on the well known graph-cut framework [2]. The optimal labeling $\hat{\mathbf{l}}_t^{k+1}$ is then used within a gating process in the update step of our tracking filter, restricting the number of valid observations for a given object hypothesis to the observations that have been assigned to the respective label. If a track has no support in the current segregation step, *i.e.* no image point is assigned to the respective label, the track is deleted from the list of tracked object hypotheses.

3.2. M-step

Motion estimation. Different to direct approaches as *i.e.* [15, 9], which determine the motion from the image gray value variations itself, we use an indirect feature-based approach in our current framework. This allows for efficient and robust motion and depth estimation simultaneously in one filter step. We show that, though we are minimizing a different error metric when applying feature-based motion estimation, also the overall error $\mathbf{E}_t$ converges to low values. For each object hypothesis independently, a state estimate is obtained based on a set of $M_t \subset N$ salient feature points $\mathbf{X}_{t,\mathrm{i}}$, $\mathrm{i} = (1, \ldots, M_t)$ of a rigidly moving object in 3D space. Following [4], we have applied the idea of the 'reduced-order observer' in order to reduce the dimension of $\mathbf{X}_t$ to one state for each tracked point, encoding its depth $\rho_{t,\mathrm{i}}$, *i.e.* $\mathrm{X}_{t,\mathrm{i}} = (\mathrm{x}_{t,\mathrm{i}}, \rho_{t,\mathrm{i}})$. It is assumed that the corresponding image points of $\mathrm{x}_{t,\mathrm{i}}$ can be determined exactly for all scene points in all views within the feature tracking scheme. Depth points propagate over time according to

$$\rho_{t,\mathrm{i}} = (0, 0, 1) \left[\mathbf{R} \left(\Omega_t \right) \mathrm{X}_{t-1,\mathrm{i}} + \mathbf{t}_t \right] . \qquad (19)$$

With this, the coordinates of a scene point at time t are $\mathrm{X}_{t,\mathrm{i}} = \Pi^{-1} \left(\mathrm{x}_{t,\mathrm{i}}, \rho_{t,\mathrm{i}} \right)$, where $\Pi^{-1} \left(\cdot \right)$ states the inverse projection function.

Given the set of tracked feature points, observations consist of the corresponding image points in the left and following right camera with image coordinates $\mathrm{x}_{t,\mathrm{l},\mathrm{i}}$ and $\mathrm{x}_{t+1,\mathrm{i}}$, *i.e.* $\mathrm{y}_{t,\mathrm{i}} = (\mathrm{x}_{t,\mathrm{l},\mathrm{i}}, \mathrm{x}_{t+1,\mathrm{i}})$.

Concerning our observation model, the image position in the following right image can be predicted from the current frame using (5), which states the instantaneous velocity field model. Given the current depth estimate $\rho_{t,\mathrm{i}}$, it is also possible to derive the corresponding image coordinate $s_{t,\mathrm{i}}$ in the current left image using (3). With this, our combined observation equation is defined as $h(\theta_{t,\mathrm{i}}, \rho_{t,\mathrm{i}}) = (m_{t,\mathrm{i}}; s_{t,\mathrm{i}}) + \mathrm{r}_t$. Observation noise r_t is assumed to be a zero-mean, white Gaussian with covariance matrix $\mathbf{R}_t = \mathrm{E}[\mathrm{r}_t \mathrm{r}_t^{\mathrm{T}}]$. The observation residual is then $\mathrm{v}_{t,\mathrm{i}}^{\mathrm{T}} \mathbf{R}_{t,\mathrm{i}}^{-1} \mathrm{v}_{t,\mathrm{i}}$, with $\mathrm{v}_{t,\mathrm{i}} = (|\mathrm{x}_{t+1,\mathrm{i}} - m_{t,\mathrm{i}}| + |\mathrm{x}_{t,\mathrm{l},\mathrm{i}} - s_{t,\mathrm{i}}|)$, expressing the projection error for each point $\mathrm{x}_{t,i}$ into the following right and current left camera image. We weight this residual with the posterior probability $\pi_{t,\mathrm{i}}$ of the label variable at the respective position i.

By additionally evaluating the second term of (11), we obtain a MAP state estimate considering state information from time $(t-1)$, *i.e.* we integrate the state evolution through time into our estimation scheme. This is formulated by the Chapman-Kolmogorov equation $P(\theta_t|\mathbb{E}_{t-1}) = \int P(\theta_t|\theta_{t-1}) P(\theta_{t-1}|\mathbb{E}_{t-1}) d\theta_{t-1}$, which expresses the predicted state distribution from time instant $(t-1)$ to t based on the a priori state distribution $P(\theta_{t-1}|\mathbb{E}_{t-1})$ and an appropriate model that accounts for the system dynamics. The model accounts for uncertainties and model errors through white, zero-mean Gaussian process noise q_t with error covariance matrix $\mathbf{Q}_t = \mathrm{E}[\mathrm{q}_t \mathrm{q}_t^{\mathrm{T}}]$. We assume the prior distribution being embodied in the probability statement $P(\theta_{t-1}|\mathbb{E}_{t-1}) = \mathcal{N}(\hat{\theta}_{t-1}, \mathbf{P}_{t-1})$, with $\hat{\theta}_{t-1}$ and $\mathbf{P}_{t-1}$ being the mean estimate and covariance of a Gaussian. Given the above equations, the best choice for θ_t then is $P(\theta_t|\mathbb{E}_{t-1}) = \mathcal{N}(\theta_{t|t-1}, \mathbf{P}_{t|t-1})$, with $\theta_{t|t-1} = f\left(\hat{\theta}_{t-1}\right)$ stating the predicted state based on the previous estimate $\hat{\theta}_{t-1}$. The same holds for the predicted error covariance $\mathbf{P}_{t|t-1}$. In [1], the iterated extended Kalman filter is presented to find the MAP estimate to this formulation. Nonlinearities in the system model are handled by iterative relinearization of the model equations within the update step. Outlier detection is performed based on a significance test of the error distribution of v_t.

Besides yielding a robust motion estimate, the output of the proposed method is also used to initialize new object hypotheses. This is done by analyzing the outliers for pattern of similar motion, as distinct moving objects that are not contained in the tracking process yet, produce coherent groups in the outlier vector. To identify these groups, an iterative RANSAC approach [2] is chosen. It partitions the outliers into point sets with the members of each specific set following a motion that can be approximated by the same constant motion model over n frames. For any stable outlier point set, 4 random points are selected from the set of outliers and the 2d homography $\mathbf{H}$ is calculated for all $n-1$ correspondences. The distance for each putative correspondences in the outlier set is calculated using the squared symmetric transfer error $d^2_{\text{transfer}} = (\mathbf{x}_{\text{out}} - \mathbf{H}^{-1}\mathbf{x}'_{\text{out}})^2 + (\mathbf{x}'_{\text{out}} - \mathbf{H}\mathbf{x}_{\text{out}})^2$ as proposed in [7]. To enforce local connectivity and motion similarity, the distance function is also weighted with the spatial distance and motion vector distance of outliers. As the number of matching points is unknown in this context, outliers are assumed to belong to the same group when their distance d^2_{transfer} to the estimated homography is below a predefined threshold t. The group that fulfills $d^2_{\text{transfer}} < t$ is the new consensus set. For the cost function of a consensus set the points of the consensus set are scored according to their distance d^2_{transfer} to the model while the outliers are given a constant weight. The steps are repeated until the number of correspondences is stable and the costs of the consensus are minimal. The resulting consensus set is assumed to be the largest group in the remaining outlier set that can be approximated by homographic projection. After storing the group members for later initializiation, they are removed from the outlier set. The algorithm is repeated with the remaining outliers until either the size of the last found consensus set or the number of remaining feature points is below a predefined thresh. The result is a set of groups which represent the new motion segments. To avoid an over segmentation, a maximum number of groups is defined. The resulting groupings are used to initialize a new object hypotheses if the number of spatially clustered image points exceeds a certain threshold.

Scene depth estimation. The final step within one EM-cycle is the dense estimation of the scene depth. For each scene point, the depth with MAP probability is determined

$$\hat{\mathbf{z}}_{t,i,\text{MAP}} \propto P(\epsilon_{t,i}|\hat{\mathrm{l}}_{t,i}, \hat{\theta}_t, \mathbf{z}_{t,i}) P(\mathbf{z}_t|\hat{\theta}_t, \mathbb{E}_{t-1}) \tag{20}$$

Similar to the notation introduced above, we formulate this as a discrete, combinatorial optimization problem with the likelihood being expressed as data term

$$\begin{aligned} &\log P(\epsilon_{t,i}|\hat{\mathrm{l}}_{t,i}, \hat{\theta}_t, \mathbf{z}_{t,i}) = T(\epsilon_{t,i}|\hat{\mathrm{l}}_{t,i}, \hat{\theta}_t) \text{ , with} \\ &T(\epsilon_{t,i}|\hat{\mathrm{l}}_{t,i}, \hat{\theta}_t) = [\log P(\epsilon_{t,i}|\mathbf{z}_{t,i} = 1, \hat{\mathrm{l}}_{t,i}, \hat{\theta}_t), \ldots \\ &\ldots, \log P(\epsilon_{t,i}|\mathbf{z}_{t,i} = \mathrm{T}, \hat{\mathrm{l}}_{t,i}, \hat{\theta}_t))]^{\mathrm{T}} \,. \end{aligned} \tag{21}$$

Like in the motion estimation step, a MAP estimate considering information from time $(t-1)$ is obtained by evaluating the second term of (11)

$$\begin{aligned} &\log P(\mathbf{z}_{t,i}|\hat{\theta}_t, \mathbb{E}_{t-1}) = \mathrm{V}_1(\hat{\theta}_t, \mathbb{E}_{t-1}), \text{ with} \\ &\mathrm{V}_1(\hat{\theta}_t, \mathbb{E}_{t-1}) = [\log P(\mathbf{z}_{t,i} = 1|\hat{\theta}_t, \mathbb{E}_{t-1}), \ldots \\ &\ldots, \log P(\mathbf{z}_{t,i} = \mathrm{T}|\hat{\theta}_t, \mathbb{E}_{t-1})]^{\mathrm{T}} \,. \end{aligned} \tag{22}$$

Temporal coherence is evaluated by predicting the MAP estimate $\hat{\mathbf{z}}_{t-1,i}$ from $(t-1)$ to t according to (5) and (19), resulting in the predicted depth estimate $\hat{\mathbf{z}}_{t|t-1,i}$. Equation (22) is evaluated according to the temporal compatibility term $\mathrm{V}_1(\hat{\theta}_t, \mathbb{E}_{t-1}) = \min(|\mathbf{z}_{t,i} - \hat{\mathbf{z}}_{t|t-1,i}|, a)$, assigning low cost to values that are close to the prediction and high values otherwise. a states the maximum value of our robust, truncated linear cost function.

Spatial smoothness of the reconstruction result is guaranteed by modeling $\mathbf{z}_t$ as a MRF with the same configuration as the label field. Disparity estimation is achieved to pixel accuracy by finding the optimal configuration of $\mathbf{z}_{t,i}$ for each image point, which is equivalent to minimizing the following energy functional

$$\begin{aligned} \hat{\mathbf{z}}_{t,\text{MAP}} = \arg\min_{\mathbf{z}_t} \Bigg\{ &\sum_{i=1}^{N} T(\epsilon_{t,i}|\hat{\mathrm{l}}_{t,i}, \hat{\theta}_t) + \\ \mathrm{V}_1(\hat{\theta}_t, \mathbb{E}_{t-1}) + &\sum_{i,n \in c} {\mathbf{z}_{t,i}}^{\mathrm{T}} \mathrm{V}_2(\hat{\Theta}_t^k) \mathbf{z}_{t,n} \Bigg\} \end{aligned} \tag{23}$$

Each image point can be assigned to one value $\mathbf{z}_{t,i} = z$ out of a set of candidate depth values $z \in (1, \ldots, \mathrm{T})$. Similar to the label term introduced above, we enforce spatial and temporal smoothness on the evolving 3D structure which is also modeled through an MRF. At initialization, depth map $\hat{\mathbf{z}}^1_{t,i}$ and segmentation $\hat{\mathrm{l}}^1_{t,i}$ consist of the predicted values from $\hat{\mathbf{z}}^K_{t-1}$ and $\hat{\mathbf{l}}^K_{t-1}$. (23) is computed using a belief propagation framework similar to the approach of Larsen *et al.* [11].

[2] Based on the *RANSAC Toolbox for Matlab* by M. Zuliani

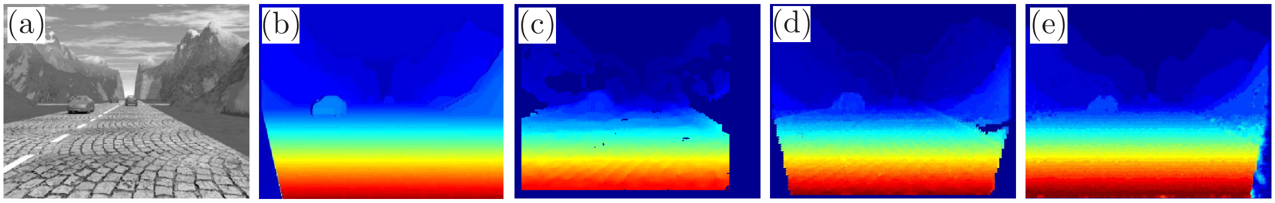

Figure 1. (a) Right camera image. (b) Ground truth disparity map. (c) Sub-pixel accurate correlation-based SVS stereo matcher ($\Gamma = 0.56, \Lambda = 0.57$) (`http://www.ai.sri.com/software/SVS`). (d) Sub-pixel accurate block-based stereo matcher ($\Gamma = 0.65, \Lambda = 0.68$). (e) Pixel accurate belief-propagation framework as presented above ($\Gamma = 1.03, \Lambda = 0.87$).

4. Experiments

The performance of our approach has been evaluated on real and synthetic [3] image data. For each track, the observations within the M-step are extracted from the respective set of tracked feature points based on a correlation-based block matching technique. The filter output, representing the time-smoothed object parameters of object j, is then fed back into the image segmentation process. In a subsequent step, points that are not conform with the labeling are deleted from the list of tracked feature points and replaced by new points, sampled from the labeled region in the image. The system state for each track is initialized with zero velocity and depth $\rho_{1:M_0}$, which is extracted from the initially estimated depth map. The number of possible hypotheses J is defined by the momentary number of distinct 6-DoF motion profiles in the scene. Regarding the relative importance of data and smoothness term the regularization factor has been adapted empirically to values between $\lambda = 0.05 - 0.5$. In Figure 1, the output of different stereo reconstruction methods is shown. The estimation quality of the different approaches has been determined quantitatively by computing the average disparity error $\Gamma = 1/N \sum_{i=1}^{N} |\Delta_{\text{gt}} - \Delta_{\text{est}}|$ (gt:ground truth, est:estimate). As we are interested in a dense scene representation, also the total number of reconstructed scene points must be considered when evaluating the different approaches. Therefore we define $\Lambda = N_{est}/N_{gt}$ as a measure of the density of our depth map. Our examination of the different stereo matching algorithms showed, that the block-based approach (Figure 1-(d)) produced slightly better estimation results ($\Gamma < 0.6, \Lambda \approx 0.6$) compared to the matcher based on belief-propagation ($\Gamma < 0.9, \Lambda \approx 0.85$) (Figure 1-(e)). The crucial benefit of latter approach is the much higher reconstruction density. See caption for details. Figure 2 depicts the mean segmentation error per pixel, according to (6) and (7), over time when initialization of new object hypothesis is suppressed (solid blue) and when new object hypotheses are added to the filter bank for the standard case (dashed red). The colored, vertical bars indicate the time instant when an object hypotheses has been added to the filter bank. Figure 3 illustrates the segmentation pipeline. After estimating scene motion (a) and scene structure (b), the image is partitioned accordingly (c+d). Figure 4 shows segmentation results of typical traffic scenarios with differently moving objects.

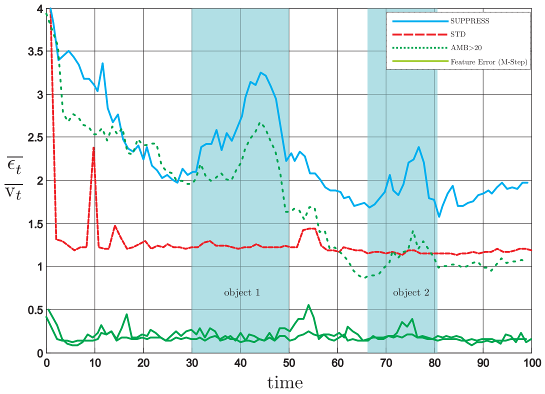

Figure 2. Mean segmentation error per pixel $\overline{\epsilon_t}$ and mean reconstruction error per feature point $\overline{\mathrm{v}_t}$ ('Feature Error (M-Step)') for a traffic scene where two objects are entering the field of view at frame 30 and 65, respectively. 'SUPPRESS' shows the error propagation if the initialization of objects with a distinct motion is suppressed and the ambiguity label (AMB) is switched off. 'AMB>20' shows the error propagation if AMB is considered in the segmentation. Image points with an error $\epsilon_t^i > 20$ are then assigned to AMB. 'STD' shows the standard case, where object hypotheses have been initialized at frame 32 and 68.

5. Conclusion

In this contribution we have presented an iterative procedure for dense estimation of motion and depth of a scene containing a multiple number of differently moving objects with the camera system itself being in motion. The data association problem has been solved using the EM framework. Within the association process, which has been implemented as labeling problem, a MRF has been used to ac-

[3] Downloaded from the *.enpeda..* web site at `http://www.mi.auckland.ac.nz`

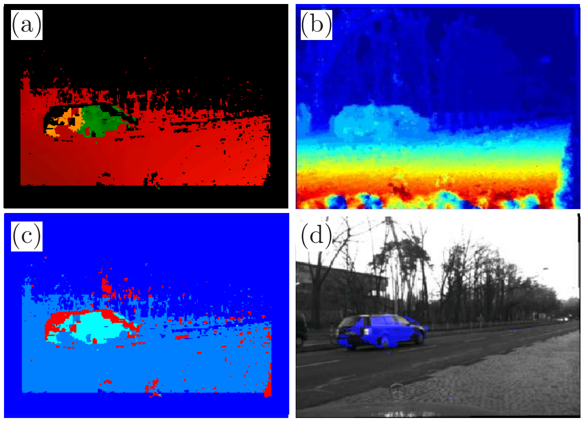

Figure 3. (a) Estimated image flow field. (b) Estimated depth map. (c) Segmentation result. Points that are coloured red are assigned the ambiguity label. (d) Original image with image points assigned to the object hypothesis being highlighted in blue.

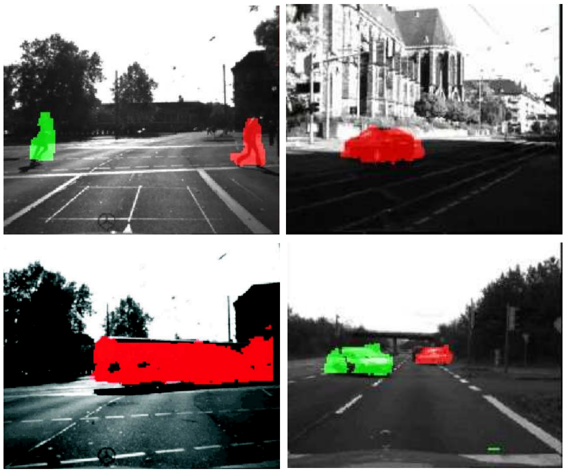

Figure 4. Segmentation results with detected objects being coloured differently.

count for spatial and temporal relationships. The EM framework has been adapted for time-recursive tracking of a multiple number of objects. Based on the motion estimates, the scene structure is recovered densely from spatially and temporally separated views. In the segregation step, the image is segregated into a set of non-overlapping regions which represent independently moving objects.

In ongoing work we integrate relational classification based on Markov logic into our segmentation scheme. We assume that the interaction of segmentation/tracking with the results from the classification step can be exploited to drive low-level object detection schemes tending towards more human-like scene perception.

Acknowledgement

The authors gratefully acknowledge the contribution of the German collaborative research center 'SFB/TR 28 – Cognitive Automobiles' granted by Deutsche Forschungsgemeinschaft (DFG).

References

[1] A. Bachmann. Applying recursive EM to scene segmentation. In *Deutsche Arbeitsgemeinschaft fuer Mustererkennung DAGM e.V.*, LNCS 5748, pages 512–521, Jena, Germany, September 2009. Springer-Verlag.

[2] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, Nov 2001.

[3] A. Bruhn, J. Weickert, T. Kohlberger, and C. Schnoerr. Discontinuity-preserving computation of variational optic flow in real-time. In *Scale-Space*, pages 279–290, 2005.

[4] A. Chiuso and S. Soatto. Motion and structure form 2d motion causally integrated over time: Anlaysis. In *IEEE Trans. Robotics and Automation*, 2000.

[5] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statistical Society*, B 39(1):1–38, 1977.

[6] P. Green. On use of the EM algorithm for penalized likelihood estimation. *Journal Royal Statistical Society*, 52(3):443–452, 1990.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[8] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[9] F. Huguet and F. Devernay. A variational method for scene flow estimation from stereo sequences. In *Proc. Intl. Conf. on Computer Vision*, Rio de Janeiro, Brasil, 2007.

[10] M. Irani and P. Anandan. About direct methods. In *ICCV '99: Proceedings of the International Workshop on Vision Algorithms*, pages 267–277, London, UK, 2000. Springer-Verlag.

[11] S. Larsen, M. Philippos, M. Pollefeys, and H. Fuchs. Simplified belief propagation for multiple view reconstruction. In *3DPVT '06*, pages 342–349, Washington, DC, USA, 2006. IEEE Computer Society.

[12] H. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. *Proceedings of Royal Society of London*, 208:385–397, 1980.

[13] C. Strecha and L. Gool. Motion - stereo integration for depth estimation. In *In Eur. Conf. on Computer Vision*, pages 170–185. SpringerVerlag, 2002.

[14] P. Torr and A. Zisserman. Feature based methods for structure and motion estimation. In *Vision Algorithms: Theory and Practice, number 1883 in LNCS*, pages 278–295. Springer-Verlag, 2000.

[15] A. Wedel, C. Rabe, T. Vaudrey, T. Brox, U. Franke, and D. Cremers. Efficient dense scene flow from sparse or dense stereo data. In *European Conference on Computer Vision*, pages 739–751, Berlin, Heidelberg, 2008. Springer-Verlag.