Probabilistic Multi-Class Scene Flow Segmentation for Traffic Scenes

Alexander Barth[1], Jan Siegemund[1], Annemarie Meißner[2], Uwe Franke[2], and Wolfgang Förstner[1]

[1] University of Bonn, Department of Photogrammetry, Germany
[2] Daimler AG, Group Research, Sindelfingen, Germany

Abstract. A multi-class traffic scene segmentation approach based on scene flow data is presented. Opposed to many other approaches using color or texture features, our approach is purely based on dense depth and 3D motion information. Using prior knowledge on tracked objects in the scene and the pixel-wise uncertainties of the scene flow data, each pixel is assigned to either a particular *moving object* class (tracked/unknown object), the ground surface, or static background. The global topological order of classes, such as *objects are above ground*, is locally integrated into a conditional random field by an ordering constraint. The proposed method yields very accurate segmentation results on challenging real world scenes, which we made publicly available for comparison.

1 Introduction

Traffic scene segmentation and categorization is an active field of research in the computer vision community. Remarkable results on monocular images using color, intensity, or texture features have been achieved, e.g., by [1], [2], or [3]. Additionally, structure from motion is used for labeling static scenes in [4]. Traffic scenes are highly challenging since the cameras are (quickly) moving through an unknown environment with uncontrolled illumination or weather conditions, highly dynamic interaction of multiple objects, and a variety of different object classes in the scene. In practice, reliable color information is often not available.

Recent advances in *scene flow* computation allow for the reconstruction of dense 3D motion fields from stereo image sequences in real-time [5], [6]. With such methods, depth and motion information is available at almost every pixel in the image, enabling new opportunities for object detection and scene segmentation. In [7], Wedel et al. use graphcuts to separate moving points from stationary points in the scene flow data (two class problem).

We extend this idea to a multi-class segmentation problem, replacing the threshold-based reasoning as in [7] by a probabilistic hypothesis competition. At this, we focus on traffic scenes where the vision sensor is mounted behind the windshield of the *ego-vehicle*, which moves in a mainly static, but unknown structured environment. In our model, the world consists of a ground surface (typically the road), static elevated obstacles on the ground surface (buildings,

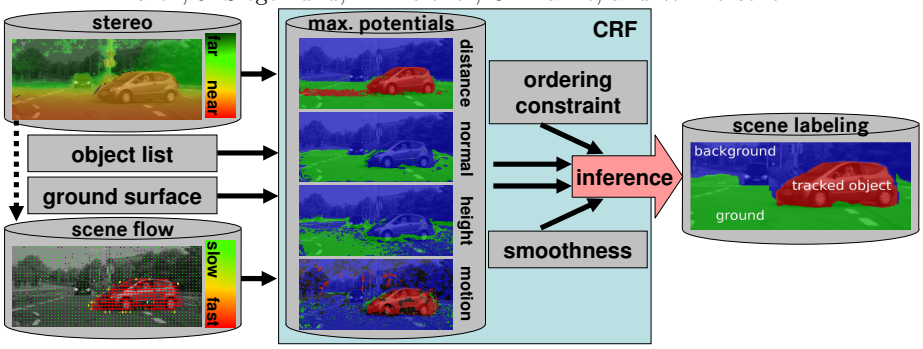


Fig. 1. System overview. Motion, depth, height, and surface normal features are extracted from 3D scene flow data and transferred into CRF class potentials for (known) moving objects, the ground surface, and the static background. Both smoothness and ordering constraints are integrated at the inference step.

traffic signs,...), as well as a finite number of independently moving objects (other cars, pedestrians, bicyclists,...). The objective of our approach is to provide a pixel-wise labeling, assigning each pixel in the current image to one of the disjunct classes *static background/obstacle*, *ground*, or *moving object*.

The moving object class is further separated into a set of *known objects*, which have been tracked before, and an *unknown moving object* class. This means, we directly exploit object information (position, orientation, velocity,...), available from previous time steps. The indiviual likelihood of each pixel belonging to a particular class based on the scene flow data is defined. The interaction of neighboring pixels is incorporated by modeling the problem as a Conditional Random Field (CRF), a widely used representation for segmentation problems. Beside requiring smoothness of the segmentation result, we integrate model knowledge on the scene topology such as **objects are above the ground** into our labeling. Fig. 1 gives an overview on the system.

Defining the potentials based on scene flow features has several advantages compared to similar stereo vision based approaches using gray value distances, for example, [8]. Issues such as robustness to varying illumination or denoising of the flow field are already addressed at the scene flow computation level. The segmentation directly benefits from all improvements at this level without changing the actual segmentation approach. Furthermore, we are able to apply the same segmentation algorithms to scene flow data provided by other sensors.

We will first introduce the generic mathematical framework in Section 2. Then, an exemplary realization of the CRF potential functions is given in Section 3. The system will be evaluated based on challenging real-world scenes in Section 4. Section 5 concludes the results and gives an outlook on future work.

2 General CRF Segmentation Framework

Let $L = \{l_1, \ldots, l_I\}$ denote a labeling for a given image, where the label $l_i \in \{C_1, \ldots C_J\}$ assigns a particular class C_j to the i-th pixel. The objective is to

find a labeling L^* from the set of all possible labelings, $\mathcal{L}$, that maximizes the conditional probability $p(L|\boldsymbol{z}, \boldsymbol{\Theta})$, i.e., $L^* = \arg\max_{L \in \mathcal{L}} p(L|\boldsymbol{z}, \boldsymbol{\Theta})$. Here, the feature vector $\boldsymbol{z}$, with $\boldsymbol{z} = [\boldsymbol{z}_1^\mathsf{T}, \ldots, \boldsymbol{z}_I^\mathsf{T}]^\mathsf{T}$, contains the pixel-wise input data for the segmentation process, and $\boldsymbol{\Theta}$ represents a set of global parameters. We model $p(L|\boldsymbol{z}, \boldsymbol{\Theta})$ as CRF [9] aligned to the pixel grid with a maximum clique size of two as

$$\log(p(L|\boldsymbol{z}, \boldsymbol{\Theta})) = \sum_{i=1}^{I} \Phi(l_i, \boldsymbol{z}_i, \boldsymbol{\Theta}) + \sum_{(s,t) \in \mathcal{N}} \Psi(l_s, l_t, \boldsymbol{z}_s, \boldsymbol{z}_t, \boldsymbol{\Theta}). \tag{1}$$

In our model, the positive function Φ defines the unary potentials for each class C_j. At this point it is assumed that the potential at pixel i depends only on the parameters and the feature data at this position. The potentials between neighboring pixels are given by the positive function Ψ, where $\mathcal{N}$ denotes the set of all pairs of neighboring pixels.

There exist several inference methods, such as graph cuts or loopy belief propagation (LBP) [10], to minimize the energy of a CRF. For a comparative study on these methods see [11]. The segmentation method proposed in this paper utilizes LBP, but is generic in a sense that it does not depend on the actual choice of the inference method. In the following, we will give a concrete realization of the potential functions.

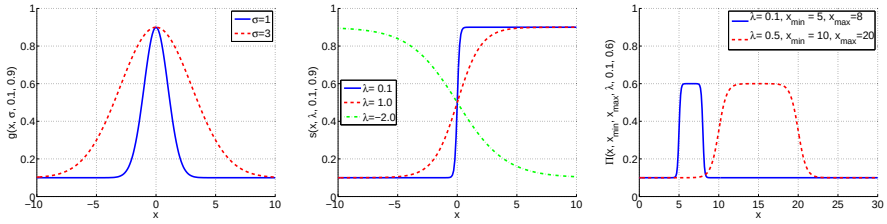
3 Scene Flow-based Traffic Scene Segmentation

In our approach, the feature vector $\boldsymbol{z}_i$ of the i-th pixel consists of a position and velocity vector of the corresponding 3D point with respect to a static world coordinate system, i.e., $\boldsymbol{z}_i = [X_i, Y_i, Z_i, \dot{X}_i, \dot{Y}_i, \dot{Z}_i]^\mathsf{T}$. For each $\boldsymbol{z}_i$ a covariance matrix is computed as in [7]. The parameter set $\boldsymbol{\Theta}$ includes the intrinsic and extrinsic parameters of the camera, ego-motion, as well as a ground surface model Ω with parameters $\boldsymbol{\Theta}_\Omega$, and a list of M tracked objects $\mathcal{O} = \{\mathbf{O}_1, \ldots, \mathbf{O}_M\}$.

For the labeling decision, each class provides a certain expectation on particular elements of the feature vector. For example, the ground surface gives a strong constraint on a point's height, while the motion state of a known tracked object forecasts the velocity of a point that belongs to this object. This means, we can extract different criteria based on the feature vector that each are discriminative for a subset of our target classes. Thus, we compose the total potential function Φ by the sum of K single potential functions Φ_k, $1 \leq k \leq K$, that incorporate these criteria:

$$\Phi(l_i, \boldsymbol{z}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \Phi_k(l_i, \boldsymbol{z}_i, \boldsymbol{\Theta}). \tag{2}$$

These functions could be learned from sufficiently large training data. Alternatively, this concept also allows for an intuitive modeling of the scene. Below we will propose $K = 4$ realizations of unary potentials for traffic scene segmentation based on scene flow, including motion, distance, height, and surface normal criteria. Other knowledge on the expected scene could be easily added accordingly.



(a) gaussian (b) sigmoidal (c) Π-shaped

Fig. 2. Base functions used for defining the potentials.

3.1 Basic Functions

The single potential functions Φ_k are defined based on different parametrization of three basic functions scaled to the range $\boldsymbol{\kappa} = [\kappa_{\min}, \kappa_{\max}]$ (see Fig. 2).
Gaussian: A bell-shaped, zero-mean, multi-dimensional Gaussian function $\mathcal{g}$ with covariance matrix C_x, defined as

$$\mathcal{g}(\boldsymbol{x}, \mathsf{C}_x, \boldsymbol{\kappa}) = (\kappa_{\max} - \kappa_{\min}) \exp\left(-1/2\ \boldsymbol{x}^\mathsf{T} \mathsf{C}_x^{-1} \boldsymbol{x}\right) + \kappa_{\min} \tag{3}$$

The function is scaled in a way that its maximum is $\kappa_{\max}$ and it converges towards a minimum value of $\kappa_{\min}$. For $\kappa_{\max} = (\sqrt{(2\pi)|\mathsf{C}_x|})^{-1}$ and $\kappa_{\min} = 0$ it corresponds to a normal distribution.
Sigmoidal: A one-dimensional sigmoidal function $\mathcal{s}$ with width λ and turning point at $x = 0$, scaled to the range $\boldsymbol{\kappa}$ with

$$\mathcal{s}(x, \lambda, \boldsymbol{\kappa}) = (\kappa_{\max} - \kappa_{\min})/(1 + \exp(-x/\lambda)) + \kappa_{\min}. \tag{4}$$

Π-shaped: A gating function Π that is composed of two opposite sigmoidal functions with slope λ

$$\begin{aligned}\Pi(x, x_{\min}, x_{\max}, \lambda, \boldsymbol{\kappa}) =& (\kappa_{\max} - \kappa_{\min})\,(\mathcal{s}(x - x_{\min}, \lambda, 0, 1) \\ & -\mathcal{s}(x - x_{\max}, \lambda, 0, 1)) + \kappa_{\min}\end{aligned} \tag{5}$$

It has its maximum value $\kappa_{\max}$ within $x_{\min}$ and $x_{\max}$, respectively, and converges towards $\kappa_{\min}$ outside this range. To limit the number of parameters, $\kappa_{\min}$ and $\kappa_{\max}$ will be assigned to one of three basic potential levels $\kappa_{\texttt{VL}}$, $\kappa_{\texttt{UL}}$, and $\kappa_{\texttt{DK}}$ for *very likely*, *unlikely*, and *don't know*. Each level can be increased by the constant offset $\kappa_{\texttt{SP}}$ to be able to *slightly prefer* a given class (notation: $\kappa^{+}_{\texttt{XX}} = \kappa_{\texttt{XX}} + \kappa_{\texttt{SP}}$).

3.2 Unary Potentials

In the following, the main ideas of the function design are presented. In our approach, the classes $C_1, \ldots, C_J$ are denoted as `BG` (static background/ obstacle), `GS` (ground surface), `O1` (tracked object no. 1), ..., `OM` (tracked object no. M), and `UO` (unknown moving object), i.e., $J = M + 3$. Each potential function Φ_k must be defined for all candidate classes C_j.

Motion Potential. The larger the distance of the velocity vector $\boldsymbol{V}_i = [\dot{X}_i, \dot{Y}_i, \dot{Z}_i]^\mathsf{T}$ to the expected velocity $\tilde{\boldsymbol{V}}_i(C_j, \boldsymbol{\Theta})$ at this position, the more unlikely belongs this point to class C_j. If it is very close to the expectation, we do not know whether this pixel belongs to the given class, since there might be another class of similar motion in the scene, but we want to slightly prefer this hypothesis. For all classes beside $\texttt{UO}$ we are able to specify $\tilde{\boldsymbol{V}}_i$. The background and ground are stationary and, thus, move only according to the known camera motion. The velocity vector of tracked objects is also assumed to be known. This yields

$$\Phi_1^{(\text{motion})}\left(l_i = C_j, \boldsymbol{z}_i, \boldsymbol{\Theta}\right) = \log \mathcal{g}\left(\boldsymbol{V}_i - \tilde{\boldsymbol{V}}_i(C_j, \boldsymbol{\Theta}), \mathsf{C}_{\Delta V}, \boldsymbol{\kappa}_1^{(\text{j})}\right) \quad , C_j \neg \texttt{UO} \tag{6}$$

where $\mathsf{C}_{\Delta V}$ denotes the covariance matrix of the velocity difference and $\boldsymbol{\kappa}_1^{(\text{j})} = [\kappa_{\texttt{UL}}, \kappa^+_{\texttt{DK}}]$. For $C_j = \texttt{UO}$, a constant potential of $\kappa^+_{\texttt{UL}}$ is defined.

Distance Potential. Assuming we have an idea on the m-th tracked object's pose and dimension in 3D space, we are able to specify an expected distance range $[\tilde{Z}_{\min,i}(\mathbf{O}_m), \tilde{Z}_{\max,i}(\mathbf{O}_m)]$ for the class $\texttt{Om}$. If Z_i lies outside this range, the i-th point does very unlikely belong to the given object class. On the other hand, if it is within the range, the likelihood for the object class increases. This is modeled by the $\mathit{\Pi}$-shaped basic function. For the class $\texttt{GS}$, we can directly predict the distance $\tilde{Z}_i(\boldsymbol{\Theta}_\Omega)$ of the i-th point based on the surface model Ω. As for the motion potentials, a Gaussian function is used to transform the distance into a potential. There is no expectation on the distance for the classes $\texttt{BG}$ and $\texttt{UO}$. However, we define points above a maximum distance $Z_{\max}$ to be very likely to belong to the background, and unlikely to belong to an unknown object. Points closer than $Z_{\max}$ are equally likely to belong to either background or an unknown object based on the distance, which is expressed by a sigmoidal function. The distance potential function $\Phi_2^{(\text{dist})}$ is thus defined as $\Phi_2^{(\text{dist})}\left(l_i = C_j, \boldsymbol{z}_i, \boldsymbol{\Theta}\right) =$

$$\begin{cases} \log \mathcal{s}\left(Z_i - Z_{\max}, \lambda_2^{(\text{j})}, \boldsymbol{\kappa}_2^{(\text{j})}\right) & , C_j \in \{\texttt{BG}, \texttt{UO}\} \\ \log \mathcal{g}\left(Z_i - \tilde{Z}_i(\boldsymbol{\Theta}_\Omega), \sigma^2_{\Delta Z}, \boldsymbol{\kappa}_2^{(\text{j})}\right) & , C_j = \texttt{GS} \\ \log \mathit{\Pi}\left(Z_i, \lambda_2^{(\text{j})}, \tilde{Z}_{\min,i}(\mathbf{O}_m), \tilde{Z}_{\max,i}(\mathbf{O}_m), \boldsymbol{\kappa}_2^{(\text{j})}\right) & , C_j = \texttt{Om} \end{cases} \tag{7}$$

with $\boldsymbol{\kappa}_2^{(\text{j})} : [\kappa_{\texttt{UL}}, \kappa^+_{\texttt{DK}}]^{(\texttt{Om},\texttt{GS})}, [\kappa_{\texttt{DK}}, \kappa_{\texttt{VL}}]^{(\texttt{BG})}, [\kappa_{\texttt{UL}}, \kappa_{\texttt{DK}}]^{(\texttt{UO})}$; $\lambda_2^{(\texttt{BG})} > 0$, $\lambda_2^{(\texttt{UO})} < 0$, and $\sigma^2_{\Delta Z}$ corresponding to the variance of the distance difference.

Height Potential. Analog to the distance potential we can define an expected height range $[\tilde{Y}_{\min,i}(\mathbf{O}_m), \tilde{Y}_{\max,i}(\mathbf{O}_m)]$ for a given known object class as well as for the expected ground height $\tilde{Y}_i(\boldsymbol{\Theta}_\Omega)$. For the unknown object class a constant height range is assumed. We do not have an expectation on the height of the background class. However, what we know is that points above a maximum height $Y_{\max}$ are unlikely to belong to moving objects or the ground surface and, thus, are likely to belong to the background class. The height potential function

$\Phi_3(\text{height})$ is given by $\Phi_3^{(\text{height})}\left(l_i = C_j, \boldsymbol{z}_i, \boldsymbol{\Theta}\right) =$

$$\begin{cases} \log \mathcal{s}\left(Y_i - Y_{\max}, \lambda_3^{(\text{j})}, \boldsymbol{\kappa}_3^{(\text{j})}\right) & , C_j = \texttt{BG} \\ \log \mathcal{g}\left(Y_i - \tilde{Y}_i(\boldsymbol{\Theta}_\Omega), \sigma_{\Delta Y}^2, \boldsymbol{\kappa}_3^{(\text{j})}\right) & , C_j = \texttt{GS} \\ \log \mathit{\Pi}\left(Y_i, \lambda_3^{(\text{j})}, \tilde{Y}_{\min,i}, \tilde{Y}_{\max,i}, \boldsymbol{\kappa}_3^{(\text{j})}\right) & , C_j \in \{\texttt{Om, UO}\} \end{cases} \tag{8}$$

with $\boldsymbol{\kappa}_3^{(\text{j})} : [\kappa_{\texttt{DK}}, \kappa_{\texttt{VL}}]^{(\texttt{BG})}$, $[\kappa_{\texttt{UL}}, \kappa_{\texttt{DK}}]^{(\texttt{GS})}$, $[\kappa_{\texttt{UL}}, \kappa_{\texttt{DK}}^{+}]^{(\texttt{Om,UO})}$; and $\lambda_2^{(\texttt{BG})} > 0$.

Surface Normal Potential. In traffic scenes, the class GS differs from all other modeled classes by its surface normal. The predicted surface normal of the ground surface at a given position i is defined by $\tilde{\boldsymbol{n}}_i(\boldsymbol{\Theta}_\Omega)$. The expected normal of any other class is assumed to be perpendicular to the ground surface normal. Thus, we can formulate a separation criteria based on the angle α between $\tilde{\boldsymbol{n}}_i(\boldsymbol{\Theta}_\Omega)$ and the measured surface normal $\boldsymbol{n}_i$ by a sigmoidal function as

$$\Phi_4^{(\text{normal})}\left(l_i = C_j, \boldsymbol{z}_i, \boldsymbol{\Theta}\right) = \log \mathcal{s}\left(\alpha\left(\tilde{\boldsymbol{n}}_i(\boldsymbol{\Theta}_\Omega), \boldsymbol{n}_i\right) - 45°, \lambda_4^{(\text{j})}, \boldsymbol{\kappa}_4^{(\text{j})}\right), \forall C_j \tag{9}$$

with $\boldsymbol{\kappa}_4^{(\text{j})} = [\kappa_{\texttt{UL}}, \kappa_{\texttt{VL}}]$ for all classes, and $\lambda_4^{(\texttt{GS})} < 0$, $\lambda_4^{(\texttt{BG,Om,UO})} > 0$.

At pixels with no scene flow data available, e.g., at stereo occlusions, a constant potential is added for all classes that slightly prefers the BG class above the horizon and GS below.

3.3 Binary Potentials

The binary terms Ψ in (1) define the interaction of two neighboring pixels concerning the labeling decision, where the neighborhood structure is defined by the four neighborhood of the image grid. In this contribution the modeling of the binary terms is based on two assumptions. First, we claim smoothness for the labeling result by defining neighboring pixels to be assigned to the same class with a high likelihood τ_1 and to be labeled different with a low likelihood τ_2 (Potts model). Second, prior knowledge on the global topological order of classes in the image is locally integrated by an *ordering constraint*.

Since cars and pedestrians move on the ground surface and are not assumed to fly, pixels representing one of the object classes are likely to be above GS labeled pixels, while BG pixels are likely to be above all other classes with respect to the image rows. Instead of *learning* the order of labels, as for example in [12], our ordering assumption is directly modeled by the relation '$\prec$', defining the strict topological ordering of the class labels $\texttt{GS} \prec \{\texttt{O1}, ..., \texttt{OM}, \texttt{UO}\} \prec \texttt{BG}$ from bottom to top in the image. For two neighboring pixels at image rows v_s and v_t, assuming w.l.o.g. $v_s \leq v_t$, the binary terms are given by

$$\Psi(l_s = C_{j_s}, l_t = C_{j_t}, \boldsymbol{z}_s, \boldsymbol{z}_t, \boldsymbol{\Theta}) = \begin{cases} \tau_1 \ , j_s = j_t \\ \tau_2 \ , j_s \neq j_t \wedge (j_s \prec j_t \vee v_s = v_t) \\ \tau_3 \ , j_s \neq j_t \wedge j_s \not\prec j_t \wedge v_s < v_t \end{cases}, \tag{10}$$

with $C_{j_s}, C_{j_t} \in \{\texttt{BG}, \texttt{GS}, \ldots, \texttt{UO}\}$, and $\tau_1 > \tau_2 \gg \tau_3 > 0$, i.e., τ_3 represents the very small likelihood that the ordering constraint is violated.

(a) INTERSECTION (b) STROLLER (c) LEAD_VEHICLE

Fig. 3. The test scenes (mask: pixels w/o scene flow data, e.g. due to stereo occlusions).

4 Experimental Results

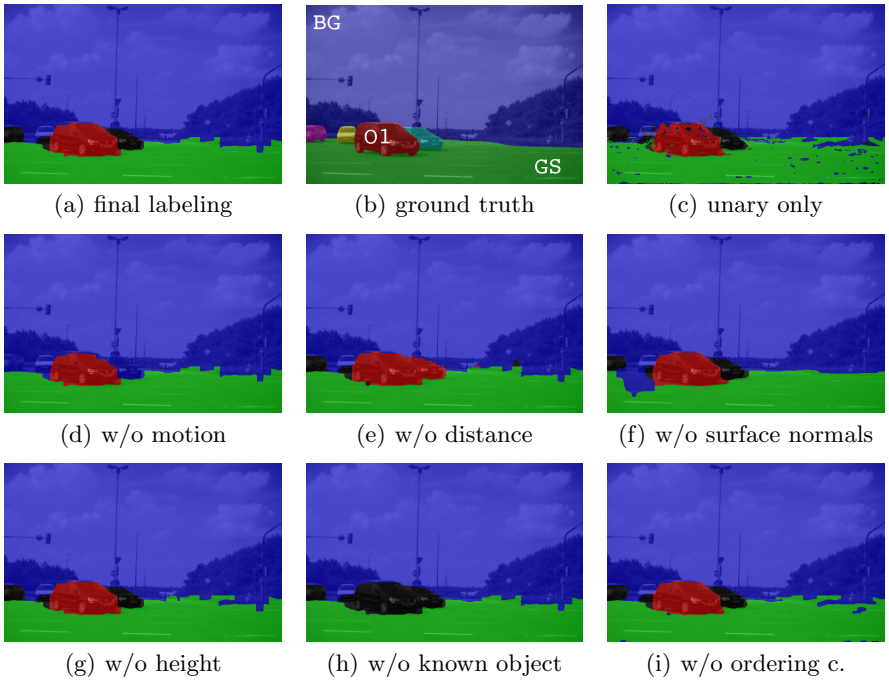
The proposed segmentation method is tested based on representative traffic scenes with manual ground truth available. The rectified stereo image pairs at two consecutive time steps together with the camera parameters, ego-motion information, as well as the ground truth labeling and prior information on moving objects in the scene is made publicly available.[3] We encourage other researchers in the field of scene flow segmentation to compare their methods based on these examples.

4.1 Data Set and Experimental Setup

Three classes of scenes have been selected (see Fig. 3).
INTERSECTION: An intersection scene with four oncoming cars. This scene contains partial occlusions, distant objects, as well as two nearby objects that move in the same direction with approximately equal velocity.
STROLLER: A pedestrian with a stroller is walking in front of a crossing car. The pedestrian casts a strong *stereo shadow* on the object, i.e., there are large regions that can only be seen from one camera.
LEAD_VEHICLE: The ego-vehicle follows the lead vehicle at approximately the same velocity through dense urban traffic, including two oncoming cars, a slow moving trailer ahead, and one car entering the visible field from the right.

In all scenes, the distance range and velocity of object O1 is known from tracking using a similar method as proposed in [13]. The velocity of the ego-vehicle and the intrinsic and extrinsic camera parameters are also known. The scene flow is computed based on [5], however, any other method could be used alternatively. A flat ground plane surface model is used here for simplicity. The only parameter of this model is the pitch angle of the camera relative to the ground, which is estimated from the measured 3D points. The constant camera height over ground is known in advance. The parameterization of the unary base potential levels is $\kappa_{\mathtt{VL}} = 0.9$, $\kappa_{\mathtt{UL}} = 0.1$, $\kappa_{\mathtt{DK}} = 0.5$, and $\kappa_{\mathtt{SP}} = 0.05$. We further use $Z_{\max} = 50$ m, $Y_{\max} = 3$ m, and $\tau_1 = 0.95$, $\tau_2 = 0.05$, and $\tau_3 = 0.0001$ for the binary terms in all experiments. However, the actual choice of these parameters is uncritical. Even larger changes influence the result only marginally.

[3] http://www.mi.auckland.ac.nz/EISATS



(a) final labeling (b) ground truth (c) unary only

(d) w/o motion (e) w/o distance (f) w/o surface normals

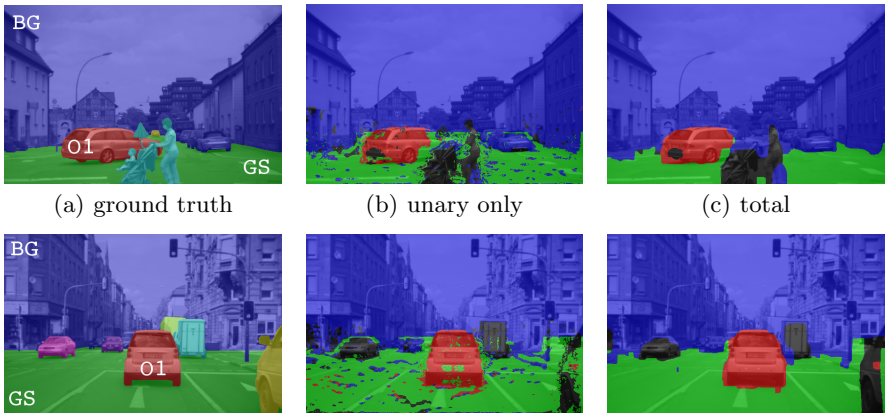
(g) w/o height (h) w/o known object (i) w/o ordering c.

Fig. 4. Labeling results at different system configurations. The colors encode the maximum class potentials at a given pixel (blue=static background, green=ground surface, red=tracked object, black=unknown moving object).

4.2 Labeling Results

The segmentation results for the `INTERSECTION` scene are depicted in Fig. 4 for different configurations. In (a) the final labeling after 40 iterations of message passing is shown, including all proposed unary and binary potentials. The manual ground truth labeling is depicted in (b). As can be seen, the resulting labeling correctly assigns most pixels on the first object to the tracked object class `O1`. Two of the three remaining cars are correctly assigned to the class `UO` (non-colored regions). Segments of this class can be used to initialize new object tracks. The white car behind the tracked object is too slow in this scene to be separated from the stationary background and, thus, is labeled as `BG`. The road surface is reconstructed very well. Only ground regions close to the objects are wrongly identified as `O1` or `UO` due to moving shadows on the ground. The confusion matrices for all investigated scenes are given in Table 1.

In (c), only the unary potentials are considered, yielding several background blobs within the ground surface region and the objects. From (d) to (g) the effect of skipping single unary potentials is demonstrated. Without motion information, the unknown moving objects are assigned to the background, while without the distance information, the two nearby objects are merged to one *tracked ob-*



(a) ground truth (b) unary only (c) total

Fig. 5. Segmentation results of `STROLLER` (top) and `LEAD_VEHICLE` (bottom) scene. Middle: Result if data is evaluated for each pixel independently. Right: Result if smoothness and global ordering constraints are incorporated via local neighborhood inference (result after 40 iterations of loopy belief propagation).

ject due to the similarity in motion. The missing surface normal potential in (f) leads to a degradation for a larger ground region at the left-hand side that is wrongly assigned to background, however, it also indicates that the surface normal is responsible for the discontinuities between ground and background at the horizon in the final labeling. The absence of the height potential alters the segmentation result only marginally in this scene, since there is not much structure about 3 m in the considered distance range. Without the information on the tracked object, all objects are assigned to the `UO` class in (h) as expected. The ordering constraint eliminates implausible background blobs that would occur within the road surface without this constraint as shown in (i).

In Fig. 5, large parts of the tracked car and the pedestrian with the stroller are correctly labeled as `O1` and `UO`, respectively. Note that the currently stationary leg is assigned to the background, since it is a non moving obstacle. The stereo occlusion is filled with `GS` from below and `BG` from the top. The `LEAD_VEHICLE` results show a very good reconstruction of the ground surface (freespace) and the moving objects in the scene, although the ego-vehicle is also moving.

GT\Est.	%	BG	GS	O1	UO	%	BG	GS	O1	UO	%	BG	GS	O1	UO
BG	72.9	**99.3**	0.7	0	0	75.5	**99.8**	0.1	0.1	0	58.7	**98.8**	0.5	0.2	0.5
GS	21.3	1.9	**94.3**	2.2	1.6	15.4	11.0	**80.5**	3.1	5.4	28.0	5.3	**88.7**	4.2	1.8
O1	3.6	4.9	0.1	**94.9**	0.1	4.8	14.0	4.4	**74.5**	7.1	5.2	0.8	0	**99.2**	0
UO	2.2	29.1	0.1	3.3	**67.5**	4.3	29.4	0.2	0	**70.4**	8.1	27.0	7.0	1.6	**64.4**

Table 1. Confusion matrices for `INTERSECTION`, `STROLLER`, and `LEAD_VEHICLE` scene. `BG`=background, `GS`=ground surface, `O1`=tracked object, `UO`=unknown moving object.

5 Conclusion

In this contribution a generic framework for precise segmentation of traffic scenes based on scene flow data and object priors has been proposed. This framework is generic in a way that it is independent of the actual scene flow implementation, CRF inference method, or object tracking algorithm. The proposed potential functions represent an intuitive model of traffic scenes, including four class types as well as ordering constraints for these classes. The model can be easily extended by more complex features, other class types, or sophisticated surface models.

The experimental results have shown that the proposed segmentation method performs very well on the considered test scenes. The main problems arise at pixels with missing or error-prone scene flow data. In such situations, appearance features, such as intensity edges or texture information, could provide useful information to further improve the segmentation results, especially at the object boundaries. Appearance potentials could be easily integrated into our framework.

Based on our segmentation algorithm and the published ground truth, it is possible to evaluate and compare different scene flow implementations in future. We are excited to see how other methods perform on our test scenes.

References

1. Wojek, C., Schiele, B.: A dynamic conditional random field model for joint labeling of object and scene classes. In: ECCV (4). (2008) 733–747
2. Ess, A., Müller, T., Grabner, H., van Gool, L.: Segmentation-based urban traffic scene understanding. In: BMVC. (2009)
3. Brox, T., Rousson, M., Deriche, R., Weickert, J.: Colour, texture, and motion in level set based segmentation and tracking. Image & Vision Comp. **28** (2010)
4. Sturgess, P., Alahari, K., Ladicky, L., Torr, P.: Combining appearance and structure from motion features for road scene understanding. In: BMVC09. (2009)
5. Wedel, A., Rabe, C., Vaudrey, T., Brox, T., Franke, U., Cremers, D.: Efficient dense scene flow from sparse or dense stereo data. In: ECCV (1). (2008) 739–751
6. Rabe, C., Müller, T., Wedel, A., Franke, U.: Dense, robust, and accurate 3D motion field estimation from stereo image sequences in real-time. In: ECCV. (2010)
7. Wedel, A., Meißner, A., Rabe, C., Franke, U., Cremers, D.: Detection and segmentation of independently moving objects from dense scene flow. 7th International Conference EMMCVPR 2009, Bonn, Germany (2009) 14–27
8. Bachmann, A.: Applying recursive EM to scene segmentation. In: 31th DAGM Symposium on Pattern Recognition, Springer (2009) 512–521
9. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML. (2001) 282–289
10. MacKay, D.J.C.: Information Theory, Inference, and Learning Algorithms. Cambridge University Press (2003)
11. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M.F., Rother, C.: A comparative study of energy minimization methods for markov random fields. In: ECCV (2). (2006) 16–29
12. Ramalingam, S., Kohli, P., Alahari, K., Torr, P.H.S.: Exact inference in multi-label CRFs with higher-order cliques. In: CVPR. (2008) 1–8
13. Barth, A., Franke, U.: Estimating the driving state of oncoming vehicles from a moving platform using stereo vision. IEEE Trans. on ITS **10**(4) (2009) 560–571