# WarpCut - Fast Obstacle Segmentation in Monocular Video

Andreas Wedel, Thomas Schoenemann, Thomas Brox, Daniel Cremers

Computer Vision Group University of Bonn
wedel, schoenemann, brox, dcremers@cs.uni-bonn.de

**Abstract.** Autonomous collision avoidance in vehicles requires an accurate seperation of obstacles from the background, particularly near the focus of expansion. In this paper, we present a technique for fast segmentation of stationary obstacles from video recorded by a single camera that is installed in a moving vehicle. The input image is divided into three motion segments consisting of the ground plane, the background, and the obstacle. This constrained scenario allows for good initial estimates of the motion models, which are iteratively refined during segmentation. The horizon is known due to the camera setup. The remaining binary partitioning problem is solved by a graph cut on the motion-compensated difference images.
Obstacle segmentation in realistic scenes a monocular camera setup has not been feasible up to now. Our experimental evaluation shows that the proposed approach leads to fast and accurate obstacle segmentation and distance estimation without prior knowledge about the size, shape or base point of obstacles.

## 1 Introduction

Year by year, thousands of people die in car accidents. Many of those accidents could be avoided or alleviated by autonomous collision avoidance systems providing for faster and more adequate reaction of the driver. In this paper we propose a key component for an assistance system, namely a framework for segmenting stationary distant obstacles in the direction of the moving vehicle. See Fig. 1 for an example of a stationary obstacle in the vehicle corridor. Stationary objects pose a particular challenge. Moving objects can easily be detected by optical flow based methods or - in vehicle application - by radar. Accurate segmentation allows for the verification of obstacle hypotheses and enables the driver assistance system to decide whether there is enough space to drive around the obstacle.

Three aspects are of critical importance for such an obstacle segmentation system. Firstly, the segmentations must be generic in the sense that they cannot rely on specific assumptions regarding the color or shape of the obstacles. Secondly, it needs to provide reliable segmentations in particular when objects are still far from the driving vehicle, i.e. where the obstacle is close to the focus of expansion (FOE), thus leaving enough time to induce obstacle avoidance strategies. This is typically a challenge, because at such an early stage the obstacle covers only a small portion of the image and the relative pixel motion is very small [10]. Thirdly, a useful collision avoidance system requires the segmentations results in real-time.

**Fig. 1. Motion segmentation of a stationary obstacle** in 36 m distance from monocular video. Notice the two cones, which are difficult to capture by means of their gray value.

In recent time, the graph cut has become very popular for fast computation of globally optimal solutions to binary partitioning problems [2,3,6]. The graph cut method gave rise to numerous interesting applications in computer vision. In [9], a stereo camera and ternary graph cuts are employed to separate a person in front of the camera from the background. In two successive works, the approach was modified to work also with monocular video by relying (predominantly) on the difference image of a moving person [5,13] in front of a static background. For our application, such approaches would not work as the entire scene is moving due to the strong ego-motion of the car. General motion segmentation with graph cuts, without a specific application in mind, has been suggested [1,12]. Mathematically, such unconstrained motion segmentation is a highly ill-posed problem. In addition to the partitioning also the motion fields in the regions have to be estimated. In contrast to segmentation based on difference images as used in [5] and [13], motion segmentation cannot be solved in a globally optimal manner anymore. The iteration of segmentation and motion estimation is likely to end up in unsatisfactory local minima.

It turns out that the obstacle segmentation task considered here actually does provide additional information and constraints. In the following, we will show which additional information is available and how it can be imposed in the graph cuts based segmentation scheme. Experimental results confirm that the integration of additional information will lead to reliable segmentations of obstacles from a driving vehicle.

## 2 Obstacle Segmentation with Graph Cut

The system is continuously fed with live gray scale video data $I : \Omega \times [0, \infty) \rightarrow \mathbb{R}$ represented as 2-D gray value fields $I_t(x, y)$ at time $t$ and image points $\boldsymbol{x} = (x, y)^\top$ in camera coordinates. As soon as another frame becomes available, it is segmented into obstacle and non-obstacle regions, based on the last two frames and the previous segmentation. This is done by computing a binary labeling $L_t(\boldsymbol{x})$ of each pixel $\boldsymbol{x} = (x, y)^\top$ in a region of interest (ROI) $\Omega_s \subset \Omega$ of the image $I_t$ at time $t$, such that

$$L_t(\boldsymbol{x}) = \begin{cases} 1 \text{ if } \boldsymbol{x} \text{ obstacle} \\ 0 \text{ otherwiese.} \end{cases} \tag{1}$$

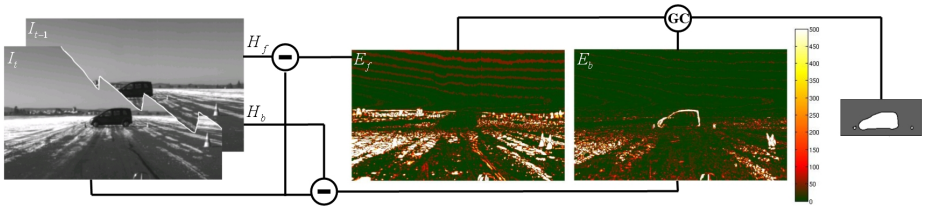



**Fig. 2. Motion-compensated difference images.** From left to right: original gray value input images and difference images for foreground ($E_f$) and background ($E_b$) based on the quadratic difference between the motion-compensated image $I_{t-1}$ and image $I_t$ after the last iteration. The camera translation was 1.9 m. $H_f$ and $H_b$ denote foreground and background motion.

The ROI is the area around the focus of expansion, where potential obstacles in the driving corridor are located. Its size corresponds to the image size of mapped obstacles. Approximate obstacle distance estimates are given from an obstacle detection system, which will be described in Section 4.

Segmentation by grouping similar gray values is not sensible in our context because the gray value of different obstacles is not fixed and may be similar to the gray value of the street. We therefore base base the labeling on motion information. The classical approach to segmentation minimizes an energy on the labeling of the form

$$E(L_t) = E_{Data}(L_t) + \alpha E_{Smooth}(L_t) \; . \tag{2}$$

### 2.1 WarpCut

In the following we show how to design the data term in Eq. 2 which is optimally suited for the segmentation of obstacles in the driving corridor of a moving vehicle. While traditionally the data term aims at segmenting the intensities [2] or the motion field [1,12], in this paper we propose to segment the warped image.

We assume that the scene is static and all image motion is caused by the camera installed in the moving vehicle. The camera motion is approximately known from odometric measurements of the vehicle. Due to the given scenario we impose the following assumptions:

1. The street is approximately planar. Hence, the image motion in this area is described by a homography $H_s$. The homography can be approximated from the known camera motion and the camera parameters.
2. Visible object points on distant obstacles have approximately the same depth. Applying the weak perspective camera model, the motion field in the obstacle region is affine, which can be expressed by another homography $H_o$.
3. Finally, the background region, i.e., the region above the horizon can be approximated as a plane at infinity, which leads to a third homography $H_b$.

Consequently, there are three regions, each with a different motion model. The separation of the obstacle region from the other two regions is done by the sought segmentation of the obstacle. The street and background region are separated a-priori by a horizontal line $y = y_{hor}$ that can be derived analytically from the camera parameters which leaves us with a binary partitioning problem.

The key idea of differentiating between obstacles and background is to penalize the difference between the current frame and motion-compensated (*warped*) previous frame. Separate motion predictions are computed for the obstacle and the non-obstacle regions. Notice that for the presented application this is much more sensible than the approaches described in [5,13] as it allows to drop the assumption of a static camera. The motion-compensated images are composed as follows:

$$I^{mc}_{0,t-1}(\boldsymbol{x}) = \begin{cases} I_{t-1}(H_b(\boldsymbol{x})) \; y < y_{hor} \\ I_{t-1}(H_s(\boldsymbol{x})) \; y \geq y_{hor} \end{cases} , \tag{3}$$

$$I^{mc}_{1,t-1}(\boldsymbol{x}) = I_{t-1}(H_o(\boldsymbol{x})) \; . \tag{4}$$

Values between grid points are determined by bilinear interpolation. Figure 2 shows the motion-compensated difference images of the introductory example in Figure 1 for $L_t = 1$ and $L_t = 0$, respectively. The data term evaluates the consistency between the warped previous image and the current image. It consists of the sum over the squared differences between both images:

$$E_{Data}(L_t) = \sum_{\boldsymbol{x} \in \Omega_s} \left(I_t(\boldsymbol{x}) - I^{mc}_{L_t(\boldsymbol{x}),t-1}(\boldsymbol{x})\right)^2 . \tag{5}$$

### 2.2 Spatio-temporal Regularity of the Labeling

Additionally to the data consistency term, our energy model incorporates assumptions on the spatial and temporal regularity of the labeling:

$$E_{Smooth} = E_{Spatial} + \beta E_{Temporal} \; . \tag{6}$$

The spatial regularity is measured by the geodesic length of the segmentation boundary. In particular, the boundary length is locally weighted by the gray value difference along the boundary. With $\mathcal{N}$ being the set of pairs of pixel neighbors (here we use $\mathcal{N}8$ neighborhood) the spatial regularity constraint reads

$$E_{Spatial}(L_t) = \frac{1}{2} \sum_{(p,q) \in \mathcal{N}} \frac{[L_t(p) \neq L_t(q)]}{\|p - q\|} \left(1 - \frac{|I_t(p) - I_t(q)|}{I_{max}}\right) \tag{7}$$

with $I_{max}$ being the maximum possible gray value. Given two boundary pixels, the energy takes its maximum for equal gray values and decreases linearly.

In addition to spatial regularity, we impose temporal regularity of the labeling setting

$$E_{Temporal}(L_t) = \sum_{\boldsymbol{x} \in I_t} [L_t(\boldsymbol{x}) \neq L_{t-1}(\boldsymbol{x})] \; . \tag{8}$$

Two aspects are considered here: the continuity of labels and the size of segments. For relatively small camera movement in stationary scenes one expects the current segmentation to be close to the most recent one. Additionally, we set the parameter $\beta$ according to the validity of the most recent segmentation. As we explain in the next section, the scale of the foreground region is used to determine the obstacle distance. With known distance, segmentation size, and calibrated camera an obstacle size is deduced. $\beta$ in our case can be seen as a switch. The parameter is set to zero for unrealistic obstacle size or distance from a given prior (for example in the beginning $\beta$ is set to zero as no prior segmentation exists). However, $\beta$ could be continuously changed if other post-processing algorithms are used to evaluate the current segmentation result.

The total energy can be minimized globally via the graph min cut method [6,4].

## 3 Adaptation of the Motion Fields

The segmentation above was solely based on pre-computed motion fields, derived from the approximate camera motion and assumptions on the planarity of the involved structures. In order to improve the segmentation, we propose to iteratively refine these motion fields. This is related to estimating the camera motion (ego-motion) from the image data [8] but aims at estimating the scene depth for static scenes. Based on the gray value constancy in 5, one can apply an incremental warping technique as originally proposed in [11] and later extended to non-translatory motion. This is detailed for our motion model in the following.

For the homographic motion model $H_*, \quad * \in \{o, b, s\}$, a point $\boldsymbol{x}$ in a given frame is associated with the point

$$H_*(\boldsymbol{h}, \boldsymbol{x}) = \begin{pmatrix} \frac{h_{1,1} \cdot x + h_{1,2} \cdot y + h_{1,3}}{h_{3,1} \cdot x + h_{3,2} \cdot y + 1} \\ \\ \frac{h_{2,1} \cdot x + h_{2,2} \cdot y + h_{2,3}}{h_{3,1} \cdot x + h_{3,2} \cdot y + 1} \end{pmatrix}$$

in the previous frame, where $\boldsymbol{h} \in \mathbb{R}$ is a parameter vector. Given an estimate $\boldsymbol{h}^0$ of these parameters, one can generate an estimate of the motion-compensated frames for the parameters $\boldsymbol{h}^0 + \Delta\boldsymbol{h}$:

$$I^{mc}_{*,t-1}(\boldsymbol{h}^0 + \Delta\boldsymbol{h}, \boldsymbol{x}) \approx I_{t-1}(H_*(\boldsymbol{h}^0, \boldsymbol{x})) + \nabla I_{t-1}(H_*(\boldsymbol{h}^0, \boldsymbol{x})) \left. \frac{\mathbf{d}H_*(\cdot, \cdot)}{\mathbf{d}\boldsymbol{h}} \right|_{\boldsymbol{x}, \boldsymbol{h} = h^0} \Delta\boldsymbol{h} \; .$$

This is introduced into our objective function $E_{data}(\cdot)$

$$\sum_{\boldsymbol{x} \in R_*} \left( I_t(\boldsymbol{x}) - I^{mc}_{*,t-1} \left( \boldsymbol{h}^0 + \Delta\boldsymbol{h}, \boldsymbol{x} \right) \right)^2$$

where the region $R_*$ is given by all points associated with the respective model. When setting the derivative w.r.t. $\Delta\boldsymbol{h}$ to zero, one can solve for the update (with simplified notation):

$$\Delta \boldsymbol{h} = \sum_{\boldsymbol{x} \in R_*} \left( \frac{\mathrm{d}H_*}{\mathrm{d}\boldsymbol{h}}(\boldsymbol{x})^\top \nabla I_{t-1}(\boldsymbol{x})^\top \nabla I_{t-1}(\boldsymbol{x}) \frac{\mathrm{d}H_*}{\mathrm{d}\boldsymbol{h}}(\boldsymbol{x}) \right)^{-1} \cdot \sum_{\boldsymbol{x} \in R_*} \left( I_t(\boldsymbol{x} - I^{mc}_{*,t-1}(\boldsymbol{h}^0, \boldsymbol{x})) \right) \nabla I_{t-1}(\boldsymbol{x}) \frac{\mathrm{d}H_*}{\mathrm{d}\boldsymbol{h}}(\boldsymbol{x}) \, .$$

Such warping schemes are also known as the Gauss-Newton method. In our case, they allow the estimation of homographies without the knowledge of point-correspondences.

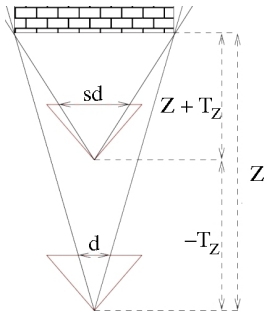
In contrast to the segmentation with fixed motion fields, the iteration of graph cuts and motion field adaptation usually does not result in a global optimum anymore. A prior for the homography parameters is given by the car odometer.

## 4 Initial Obstacle Detection and Depth Estimation

Our segmentation model is based on the restriction of the labeling domain to a region of interest around the focus of expansion. Moreover, the initial motion field in the obstacle region depends on the obstacle's distance to the camera. Although the detection of obstacles is not the focus of this paper, we briefly review a method that has recently been proposed in [14] and which we adopted in order to trigger the segmentation. Alternatively, one could use active sensors, such as radar or lidar, for this purpose.

Assume an image point $\boldsymbol{x}_t$ belonging to a static world point at $(X, Y, Z)^\top$. The camera translates by $(T_X, T_Y, T_Z)^\top$ in camera coordinates from frame $I_t$ to $I_{t+1}$. Then the world point at $t+1$ will be projected to

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \tfrac{f}{Z+T_Z} \begin{pmatrix} X + T_X \\ Y + T_Y \end{pmatrix} \\ &= \underbrace{\frac{Z}{Z+T_Z}}_{s} \underbrace{\frac{f}{Z} \begin{pmatrix} X \\ Y \end{pmatrix}}_{\boldsymbol{x}} + \tfrac{f}{Z+T_Z} \begin{pmatrix} T_X \\ T_Y \end{pmatrix} . \end{aligned}$$




Hence, the distance $Z$ of the point can be inferred from the scaling $s$ of $\boldsymbol{x}$ with respect to the focus of expansion. For obstacle detection, we track a number of points over multiple frames using the region tracker in [7]. Distance estimates at locations that are not consistent with the ground plane are considered as potential obstacle points. This way, stationary obstacles within $50$m are detected at interactive frame-rates. For a comparative test we refer to [14]. Given the location and distance of potential obstacle points allows to define the region of interest in which we compute the segmentation. The interest region is chosen large enough to capture obstacles up to a size of $10$ m $\times$ $3$ m.

In the same manner, one can derive an accurate depth estimate for the obstacle from the scaling of the obstacle region segmented by our approach. Such estimates are then used to verify the depth estimate from region tracking. Notice that the initial region tracking step can be replaced by other sensors such as radar. The segmentation is verified by comparing its distance estimate with the distance predicted by the region based tracker. If the deviation is smaller than $5\%$, the segmentation is considered trustworthy.

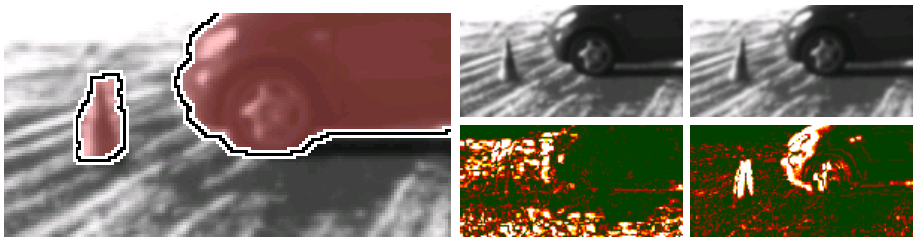


**Fig. 3. Closeup of segmentation for an obstacle in** $17$ **m distance.** The middle image shows the warped foreground with the foreground energy. The right image shows the warped ground plane and according energy. Color warmth denotes higher gray value difference compared to the current image. Notice the correct segmentation across the shadow boundary and the incorrect segmentation of the traffic cone due to occlusion.

## 5 Results

We evaluated the method in some real world scenarios. For all the experiments we show in this section, the parameters have been kept fixed. In particular, we set $\alpha = I_{max}$ and $\beta = \frac{5}{I_{max}}$ with $I_{max} = 255$. The camera had a focal length of 8mm, which corresponds to approximately $800$ pixels.

Figure 3 demonstrates the accuracy of the segmentation even in areas close to the base point of the obstacle. In these areas the motion model of the street is almost identical to the motion field of the obstacle. As the segmentation is based on differences between those models, the segmentation is much more sensitive to noise here. The correct segmentation even along the bottom of the car reveals the robustness of the overall method even in these critical areas. Another reason for inaccuracies are occlusions of the ground plane by the obstacle. The traffic cone, for instance, is not perfectly segmented due to this fact. Apart from occlusion artifacts, however, the segmentation result is very precise. Moreover, the algorithm runs at interactive frame rates of $5$ fps including obstacle detection and segmentation.

Figure 4 shows another result for a scenario with two differently colored obstacles. The color of the gray car actually fits very well to large parts of the background region. Clearly, an intensity based segmentation with graph cuts, as shown in the Figure, is not appropriate here. On the other hand, the motion cues used in the proposed


**Fig. 4. Motion segmentation and distance estimation for different color obstacles** in 53 m distance from monocular vision alone. The camera translation was 2.6 m between the frames. The *right plots* show the region of interest with motion segmentation (*top*) and gray value segmentation (*bottom*) for the same frames. Clearly, gray value segmentation is not suitable for the segmentation of different colored obstacles in scenes with arbitrary background.

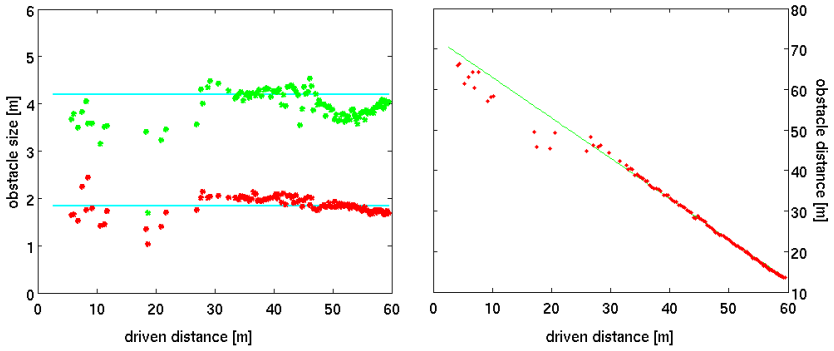



**Fig. 5. Distance and obstacle size estimation** for the example with one obstacle (ground truth: 4.19 m×1.83 m) in Figure 1.


**Fig. 6. Detection of a gap between obstacles.** Taking the center of mass for the distance measurements results in one detected object. The gap between the two trucks is ignored.

approach can segment the two obstacles very well, though they are still 53 m away. However, with the general motion segmentation approach the obstacles are not segmented from the background and, hence, distance and size estimates are not possible. The motion parallax (motion difference between ground plane and obstacle) decreases non-linearly with increasing distance. Thus, it is quite small in this case. Nevertheless, there is enough difference to outline the shape of the obstacles without implying any prior shape knowledge using our WarpCut algorithm.

The plots in Figure 5 show the size and distance estimates of the approaching obstacle from Figure 7 by means of segmentation . The ideal values are indicated by the straight lines. The estimates by the segmentation are very good. This emphasizes the precise segmentation of the obstacle throughout the video sequence, pictured with extracted frames in Figure 7.

Figure 6 shows that obstacle *segmentation* is more than just obstacle *detection*. The segmentation allows to detect gaps between obstacles and to measure the size of these gaps in order to decide whether it is possible to drive through this gap. Common radar sensors, for instance, would only consider the center of mass and detect a single object. This example demonstrates the relevance of segmentation for autonomous collision avoidance. Similar scenarios appear in robot navigation.

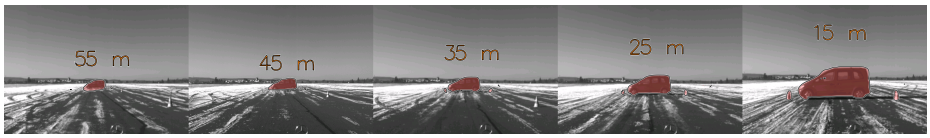



**Fig. 7. Segmentation and distance estimation from monocular video**. The segmentation and distance estimation of the stationary obstacle proves to be precise throughout the video sequence. The early detection and localization of the obstacle leaves time to induce obstacle avoidance strategies.

## 6 Conclusions

We presented a method for accurate stationary obstacle segmentation from motion in monocular video. In particular, we propose to obtain segmentations based on intensity differences of the current frame and motion-compensated versions of the previous frame. As spatially regularized segmentations are desired in a real-time context, energy minimization via graph cuts on such warped images proved to be very useful. For the motion segmentation to be robust, we exploited a number of assumptions that are reasonable in the context of obstacle segmentation. Experimental results confirmed the validity of these assumptions in several scenes and demonstrated the robust and accurate segmentation of obstacles. Moreover, we showed that from the scaling of the obstacle region in time, one can accurately estimate the obstacle's distance. Also conclusions about obstacle dimensions can be deduced.

## References


1. S. Birchfield and C. Tomasi. Multiway cut for stereo and motion with slanted surfaces. In *International Conference on Computer Vision*, pages 489–495, 1999. 2, 3
2. Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *International Conference on Computer Vision*, volume 1, pages 105–112 vol.1, 2001. 2, 3
3. Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 359–374, 2001. 2
4. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 23(11):1222–1239, 2001. 5
5. A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov. Bilayer segmentation of live video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 53–60, Washington, DC, USA, 2006. IEEE Computer Society. 2, 4
6. D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *Royal Journal on Statistical Society*, 51(2):271–279, 1989. 2, 5
7. G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1025–1039, 1998. 6
8. Q. Ke and T. Kanade. Transforming camera geometry to a virtual downward-looking camera: Robust ego-motion estimation and ground-layer detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 390–397, 2003. 5
9. V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. Bi-layer segmentation of binocular stereo video. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 407–417, 2005. 2
10. H. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. In *Proceedings of the Royal Society of London*, volume 208 of *Series B, Biological Sciences*, pages 385–397, July 1980. 1
11. B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc.7th International Joint Conference on Artificial Intelligence*, pages 674–679, Vancouver, 1981. 5
12. T. Schoenemann and D. Cremers. Near real-time motion segmentation using graph cuts. In *Pattern Recognition (Proc. DAGM)*, volume 4174 of *LNCS*, pages 455–464, Berlin, Germany, September 2006. Springer. 2, 3
13. J. Sun, W. Zhang, X. Tang, and H.-Y. Shum. Background cut. In *European Conference on Computer Vision (ECCV)*, pages 628–641, 2006. 2, 4
14. A. Wedel, U. Franke, J. Klappstein, T. Brox, and D. Cremers. Realtime depth estimation and obstacle detection from monocular video. In *Pattern Recognition (Proc. DAGM)*, volume 4174 of *LNCS*, pages 475–484, Berlin, Germany, September 2006. Springer. 6