Monocular Rear-View Obstacle Detection Using Residual Flow

Jose Molineros[1], Shinko Y. Cheng[1], Yuri Owechko[1],
Dan Levi[2], and Wende Zhang[3]

[1] HRL Laboratories, LLC
3011 Malibu Canyon Road,
Malibu CA 90265
{jmmolineros,sycheng,yowechko}@hrl.com
[2] GM Advanced Technology Center-Israel
[3] GM Research
{dan.levi,wende.zhang}@gm.com

Abstract. We present a system for automatically detecting obstacles from a moving vehicle using a monocular wide angle camera. Our system was developed in the context of finding obstacles and particularly children when backing up. Camera viewpoint is transformed to a virtual bird-eye view. We developed a novel image registration algorithm to obtain ego-motion that in combination with variational dense optical flow outputs a residual motion map with respect to the ground. The residual motion map is used to identify and segment 3D and moving objects. Our main contribution is the feature-based image registration algorithm that is able to separate and obtain ground layer ego-motion accurately even in cases of ground covering only 20% of the image, outperforming RANSAC.

1 Introduction

Many automotive accidents involve cars backing up [1]. Existing backup warning systems based on sonar or radar sensors provide gross localization information, not enough to accurately determine if the vehicle will collide with an obstacle. Ranges go

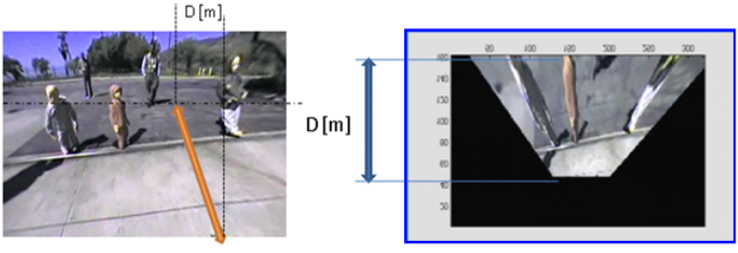

Fig. 1. (Left) An image taken from a camera rigidly mounted on a vehicle. (Right) Result of virtually rotating the camera view to bird-eye view.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 504–514, 2012.

up to 5m. Reaction times for humans are between .25s and .8s before brakes are even applied [2]. A car backing up at 15mi./h will travel 5m in .75s, which might be too close for comfort.

Cameras offer the capability to extend range and determine the nature of the alarm. Research work in rear obstacle detection using cameras include [3][4][5].

Previous systems for obstacle detection based on cameras sometimes utilize stereo and sometimes monocular input. Many utilize the idea of inverse perspective mapping (IPV) in polar histogram analysis, for example [5] and [6]. The paper in [6] uses stereo, IPV and polar histograms to detect close range obstacles. Work in [5] uses only one camera but coarsely uses polar histogram analysis due to not being able to obtain accurate ego motion. In [3] a system is presented that uses monocular input and ego-motion from odometry sensors. The authors in [4] present a system for backup aid that detects obstacles from three independent methods, fused at the output stage. It assumes road boundaries are always within the video image and has some difficulty with false positives due to shadows not being totally eliminated.

We introduce a method utilizing monocular camera input from a wide angle camera that is a valuable alternative to polar histogram analysis. A novel ground ego-motion recovery algorithm is developed. Obstacle segmentation is based on variational optical flow algorithms. We assume an urban setting with a ground layer that in a short range around the car is parallel to the car longitudinal axis.

The rest of the paper is organized as follows: section II describes our system. Section III describes its use in a child detection application. Section IV is implementation details and section V presents the conclusions.

2 System Overview

Our system termed VOFOD (variational optical flow obstacle detection) uses a combination of methods that to the best of our knowledge is novel. By registering the ground plane in consecutive image frames and exploiting the obtained ego-motion we generate an optical flow map where areas corresponding to the ground layer are set to zero. This map provides a motion saliency mechanism that directs attention to areas in the image containing obstacles.

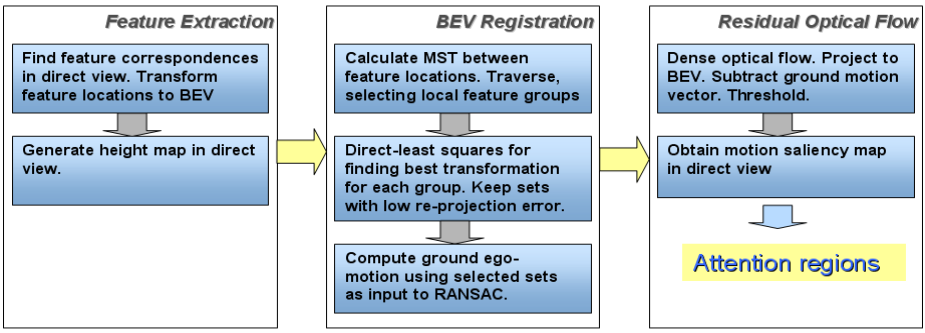

Fig. 2. System block diagram. There are three defined stages: feature correspondences, BEV registration and ground ego-motion estimation, and residual optical flow for obstacle detection.

We change the point-of-view to a virtual camera view looking straight down, termed bird-eye view, or BEV (figure 1). This allows us to run fast, accurate registration based on point features, and ultimately to obtain the ground motion vector.

We obtain optical flow information for each point in the image, and use the knowledge about ground motion to select the flow vectors corresponding to objects not part of the ground plane. We accomplish this by projecting the optical flow vectors to BEV. The obtained ground motion vector is then subtracted from projected dense optical flow, selecting moving areas not belonging to the ground plane.

In BEV, points at a height with respect to the ground plane will have a different difference in deformation than points in the ground plane when looking at consecutive images. Optical flow vectors in BEV are easily classified into ground/non-ground, and by assuming a planar ground, classified into obstacles and non-obstacles.

The algorithm consists of three stages (figure 2):

-*Stage 1*: Finding feature correspondences in consecutive images. Transforming point feature locations to BEV. A map is generated to obtain metric distance from the camera to a particular pixel location.

-*Stage 2:* Classifying feature points into ground/non-ground by means of our registration algorithm. Obtaining three-parameter homography (ground ego-motion).

-*Stage 3*: Estimating an 8-parameter homography transformation between consecutive frames, in original camera viewpoint. Obtaining residual optical flow and segmenting obstacles. Outputting metric distances from vehicle to obstacles.

2.1 Stage 1

Stage 1 is an algorithm to extract image feature points and virtually transform them to a virtual camera viewpoint looking straight down. Figure 1 shows an example image after virtual camera rotation. Changing viewpoint simplifies the problem of estimating ego-motion by reducing the number of unknown parameters from eight to three. This simplification allows us to estimate orthographic transformation parameters directly, with corresponding benefits in terms of registration accuracy and speed (see figure 5).

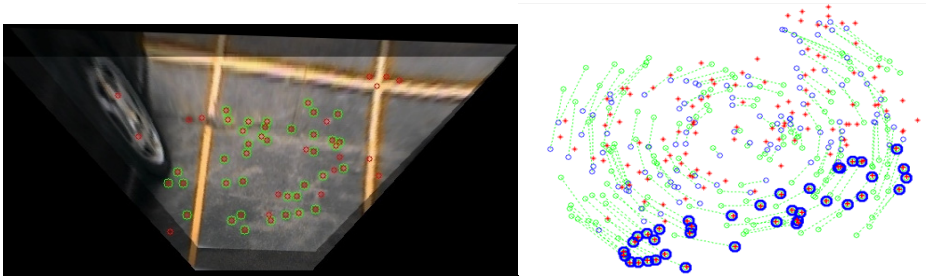

Fig. 3. Result of stage 2. Two bird-eye view images have been registered. (Left) The red dots are features in image 1. Green circles indicate features in image 2 that have been classified as ground. (Right) Simulation to obtain registration error: Ground points are localized in a small area. Red = image 1 points. Green = image 2 points. Blue = transformed image 2 ‘obstacle’ points. Thick blue = transformed ‘ground’ points.

Suitability to obstacle detection applications of a particular feature detector/descriptor will depend on video quality, feature matching accuracy and computation speed. In our application scenario, images are captured by a fish-eye rear production camera with relatively low video quality. Another important fact is that concrete surfaces tend to have low feature density, making it difficult to extract enough quality features for registration. A BEV camera viewpoint exploits the approximation of vehicle motion as planar motion and offers the advantages of no ambiguity between rotational and translational ego-motion parameters, and reliability in transformation parameter estimation due to the linearity of image motion.

Point feature types found in literature include SIFT, SURF, Harris, KLT, and CenSurE. Perhaps not surprisingly, SIFT[10] produced the best quality registration in terms of average re-projection error. However, unless accelerated by GPU, it might be unsuitable for hard real-time applications. A second best is Center Surround features (CenSurE) [12], a feature type designed in the context of visual odometry in rough outdoor terrain over long periods. We found its reliability to be close to SIFT with speed improvements of four to five times. It produces the best compromise between match quality and speed. In all of our video test sequences, both CenSurE and SIFT have produced enough quality correspondences in production fish-eye cameras currently installed in automobiles. Other detectors we tested, such as SURF and KLT, have not.

2.2 Stage 2

The main novelty of our obstacle detection approach lies in this registration algorithm. It estimates orthographic transformation parameters, minimizing re-projection error and eliminating outliers.

Stage 2 takes as input BEV feature locations and their matched correspondences and calculates ground ego-motion. Ground points are found by first finding image features lying on a plane parallel to the virtual camera plane and refined by eliminating features lying on surfaces 'at height' with respect to the ground.

Methods to determine the ground plane include 2D dominant motion estimation and layer extraction. These methods can be top-down or bottom-up. Top-down

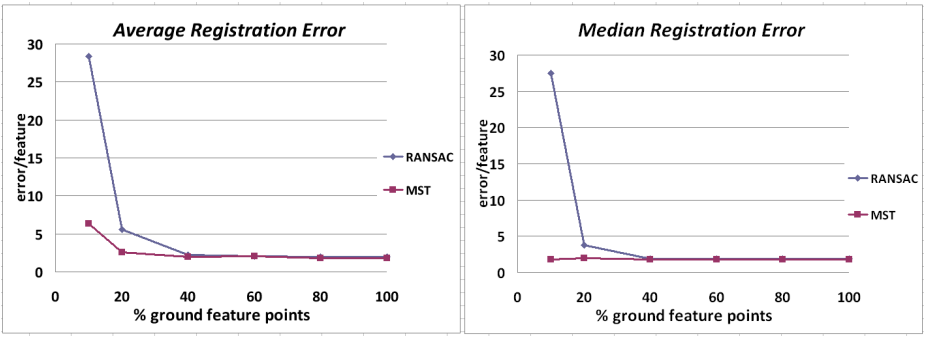

Fig. 4. Euclidean error per feature correspondence (pixels). RANSAC-based BEV registration (blue) vs. MST-based method (red). The *X*-axis represents the percentage of 'ground points'.

approaches assume the ground plane is dominant in the scene, or alternatively that the scene can be modeled with a few planar layers. In car backing up scenarios the ground plane might not be dominant and the scene is not necessarily well modeled with a few planar layers.

Bottom up approaches divide the image into small patches where a local patch transformation is calculated. These transformation calculations are then grouped together to form layers. In [7] measurements are combined with a robust weighting scheme for global ego-motion determination and ground plane segmentation. A regularization step is later applied to recover small non-planar motions. Our own preliminary implementation of the scheme in [7] yielded performance and speed that convinced us point features are still the way to go.

We developed a point feature-based registration scheme that does not assume a dominant ground layer. Our approach produces good quality registration even in cases of 80% of the image being background clutter and obstacles.

We utilize a two-tiered algorithm to obtain matches between two BEV clusters of points according to an orthographic projection model. The two tiers are a Minimum-Spanning Tree (MST) [17] followed by RANSAC or one of its variants [8][9]. A feature selection strategy outputs a "bag" of candidate ground features, after traversing an MST generated between the point correspondences in the first image. We test each vertex plus its immediate connected neighbors (a local set) against the corresponding local set in the second image. A direct least-squares estimation technique [12] is used to compute the transformation parameters between local sets, as well as to choose sets with low average re-projection error. These sets of features are added to a global "bag" of candidate ground points. Finally, we use RANSAC to obtain the final global orthographic transformation between bag points. Although the MST stage is enough most of the time, local sets of points in surfaces (at height and parallel to the ground) will pass the least-squares re-projection error test. RANSAC is applied to eliminate these outlier sets.

MST allows us to exploit naturally occurring clusters of ground features and produces more accurate registration in the presence of higher number of outliers than RANSAC. It also avoids having to unnaturally segment the image into coarse blocks for analysis that a two-tiered RANSAC scheme would require. Exploiting the fact that ground features tend to cluster together becomes crucial in cases where it is difficult to localize repeatable feature points, such as concrete surfaces in roads and parking lots. Figure 3 shows registration of two BEV frames after obtaining 2-D ground motion parameters.

In order to compare registration quality between MST and RANSAC, we tested both algorithms on simulated data over 100+ trial runs. We generated a number of random points in one set, belonging to ground and non-ground (fig. 3, right). Ground points in a second set correspond to the first set, transformed according to known rotation and translation. Uncertainty in position is added to points in the second set. Ground points have a Gaussian spatial error in pixels of $(\mu=2.5, \sigma=1.45)$ and non-ground points follow a Gaussian distribution with $(\mu=25, \sigma=14.5)$. The proportion of ground points is varied from 10% to 90% and re-projection error for both algorithms is obtained. We added a tendency of ground points to cluster together spatially.

Figure 4 compares registration quality of MST vs. RANSAC [8]. X-axis represents percentage of true ground points in the set. MST efficiently exploits the tendency of points of the same class to cluster together in order to register images with a minority of ground points.

Our registration algorithm provides an improvement over other conventional methods for ground registration [16] (Fig.5).

Description of stage 2 is as follows:

1. Generate an *MST graph* between BEV point locations.
2. Obtain groups of features that are spatially close, by traversing MST.
3. Classify clusters as ground/non-ground according to conformance to an orthographic transformation model. Classification is based on average re-projection error [12].
4. Calculate 2-D ground motion using all selected ground point clusters. Global orthographic transformation is obtained by a RANSAC-style algorithm [9] to eliminate entire clusters that might have low average re-projection error but are non-ground.
5. Classifying original feature points obtained *in direct view*, into ground/non-ground, by their corresponding BEV re-projection error.
6. Obtain ground motion vector.

Once ground has been registered, obstacles can be detected by exploiting misregistration that objects at height or moving objects exhibit. We accomplish segmentation by obtaining a residual dense optical flow map in stage 3.

2.3 Stage 3

Stage 3 produces a residual motion map for non-ground plane objects. With a moving camera detected obstacles can be stationary or moving. If the camera is stationary, only moving objects will be detected.

To find obstacles it is not enough to cluster point features that have been classified as non-ground. More exhaustive image analysis is needed. We utilize dense optical flow for that purpose. Optical flow obtains regions of interest in the image, as well as provides an interest score to a particular pixel locations, according to the magnitude of the flow vector.

Fig. 5. Comparison of ground registration quality. (Left) Using a top-down method from literature to directly estimate 8-parameter homography. (Right) Using stage 2 algorithm, re-projected to normal camera view.

Recently, variational numerical schemes have been used successfully to obtain high accuracy and real-time dense optical flow. We use a bidirectional multi-grid method for accelerated variational optical flow computations (VOF) [13]. This method enables generation of our obstacle likelihood map.

The approach found in [13] can be considered a generalization of the traditional Horn-Schunck algorithm, but allowing for multi-grid schemes with non-dyadic grid hierarchies, as well as fast variational numerical methods for solving systems of equations.

Residual motion maps indicate motion different to ground plane motion. After obtaining optical flow vectors from two consecutive images, we project the vectors to BEV and subtract them from ground motion obtained in stage 2 (figure 6, center). Resulting BEV residual flow is used to obtain flow in direct view without barrel distortion. We use the mapping between BEV coordinates and direct view coordinates to create a binary mask in direct view that in combination with the optical flow map produces residual motion (fig. 6).

As the ground plane has been obtained, the camera has been calibrated, and the relationship between bird-eye views and original views are known, distances (in feet or meters) from the vehicle to the base of obstacles are straightforward to calculate. They vary linearly with height (fig. 1, right).

While varying the point-of-view to estimate ground motion and VOF computation are methods found in literature, their combination and application to obstacle detection is new to the best of our knowledge. Our registration algorithm plus VOF methods provide a powerful, novel alternative (and perhaps complement) to other simpler real-time algorithms in automotive applications.

Calculation of dense optical flow and BEV ground ego-motion estimation are currently implemented as parallel threads. Once the ground motion vector is obtained, we merge the results to produce the residual motion map and obstacle segmentation.

In summary, stage 3 consists of:

- Calculating dense variational optical flow for the image.
- Transforming VOF to bird-eye view (BEV).
- Subtracting BEV-VOF from ground plane motion vector
- Determining VOF of obstacle pixels in direct view.

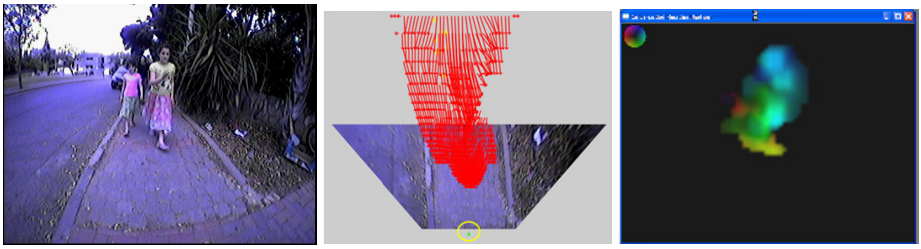

Fig. 6. (Left) Original image. (Center) BEV residual flow. Ground motion is shown as a small green arrow enclosed in the yellow circle. (Right) Residual optical flow map: color represents flow vector direction, intensity represents flow magnitude.

Figure 6 (right) shows the color-coded map obtained from stage 3. Distances from the camera to the corresponding position in the ground plane are known, varying according to *Y* coordinate in the undistorted image. When looking for obstacles that have an expected height range, this information allows us to estimate the size in pixels such an obstacle would have.

In the optical flow map in figure 6 (right), intensity represents magnitude of the flow vector, whereas color represents vector direction. This image can be used as a saliency map or it can be combined with other static-image saliency algorithms [14] to indicate regions of interest. In our example application, normalized magnitude of the optical flow vector at a pixel represents a likelihood of it belonging to an obstacle.

3 Application: Child Detection

VOFOD is used in combination with a state-of-the-art parts-based person classifier termed FSM (feature synthesis [15]) to detect children in arbitrary poses. Processing times for state-of-the-art parts-based classifiers is still far from real-time unless the number of input windows is drastically reduced. VOFOD is used to reduce the typical number of input windows from >50K to less than 500 after non-maximal suppression (nms).

Figure 7 is the result of exhaustively searching for windows of height approximately expected of children. Boxes were generated with a fixed aspect ratio, and their height is adjusted according to *Y*-coordinate to fall within an acceptable range. We sum the values of the VOFOD saliency map inside each box and obtain a child detection likelihood. Only boxes with a score above a predefined threshold are colored. Lighter colors represent higher box scores.

Although it is making great strides, state-of-the-art in human arbitrary pose detection still needs to improve. At the same time, in a collision warning application it is important to eliminate false positives as they desensitize the driver to ignore alerts. Under such conditions, it might be desirable to operate under low probability of *child* detection but very low false positives. Meanwhile, moving and close obstacles are still detected. VOFOD excels under such conditions.

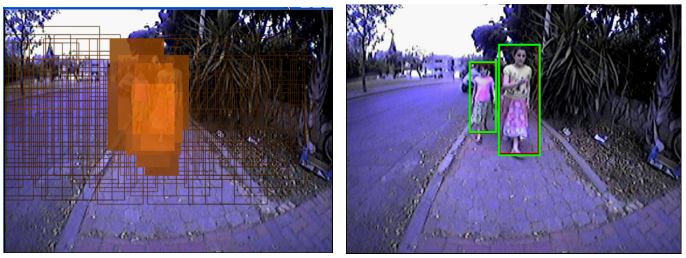

Fig. 7. (Left) Using residual flow map for child detection: boxes were generated according to expected child height and scored according to average flow intensity. Only boxes scoring above a threshold are colored (Right) Detections with a parts-based classifier.

Our video sequences include children at all distance ranges. Camera is moving less than 50% of the time. With a moving platform, static children at close and mid ranges can be detected by residual optical flow. However, non-moving children standing far away will not be detected. By using VOFOD in combination with an image-based saliency algorithm, such as frequency-tuned saliency (FTS [18]), we operate in high probability of child detection with approximately 50% window reduction over nms. Figure 8 shows the probability of child detection by FSM parts-based classifier, according to the reduction in number of input windows. We compare VOFOD, VOFOD + FTS, FTS itself, and random scoring. We detail our experiments to determine system performance in [14].

4 Implementation

We feel the VOFOD implementation can be greatly optimized, perhaps up to an order of magnitude. Current execution times, in C++ are 3.3 FPS for stage 1 + 2 (combined, in 640 x 480 resolution) and 2.5 FPS for stage 3 in 320 x 240 resolution. Timings are in an Intel Xeon CPU W3550 @ 3.07 MHz.

Many optimizations are possible in our proof-of-concept code, including avoiding processing in parts of the image and making the code more efficient. The slowest components are point feature extraction and dense optical flow. The complete system runs at video rate by processing only two or three pairs of consecutive images per second of video.

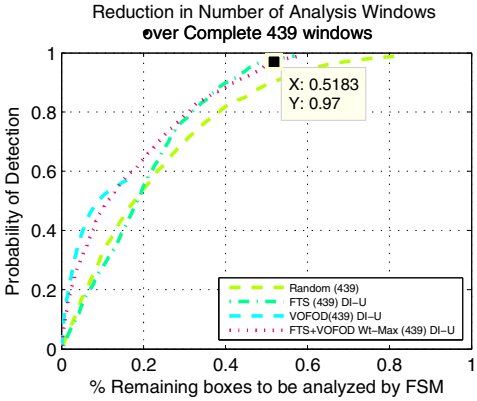

Fig. 8. Probability of child detection by FSM, according to number of input analysis windows. VOFOD+FTS eliminates 49% of all windows over exhaustive approach after nms. 25% more windows over random scoring.

5 Conclusion

We have introduced an obstacle detection system based on BEV registration for estimating ground motion, and variational dense optical flow for detecting obstacles.

Our approach of using an MST for ground point selection resulted in acceptable separation of ground regions in cases with very few ground points, the instances in which RANSAC by itself tends to fail. This in turn allows for better registration and ego-motion estimation. Fast variational optical flow allows us to segment obstacles by analysis of a residual motion map. One example application of our algorithm is as an attention mechanism for child detection. Highlighted areas can be later processed with a parts-based classifier to identify children. Utilization of this attention mechanism greatly reduces computation time allowing for real-time performance.

References

[1] http://www.osha.gov/dcsp/success_stories/compliance_assistance/motor_vehicle_case_study.html

[2] Sens, M.J., Cheng, P.H., Wiechel, J.F., Guenther, D.A.: Perception/Reaction Tim Values for Accident Reconstruction. SAE 890732, Society of Automotive Engineers, pp. 79–94 (1989)

[3] Vestri, C., et al.: Real-Time Monocular 3D Vision System. In: 16th World Congress on Intelligent Transport Systems, Stockholm (2009)

[4] Ma, G., et al.: A Real-Time Rear View Camera Based Obstacle Detection. In: 12th IEEE International Conference on Intelligent Transportation Systems, pp. 1–6 (September 2009)

[5] Yankun, Z., Chunyang, H., Norman, W.: A Single Camera Based Rear Obstacle Detection System. In: Proc. IEEE Intelligent Vehicles Symposium (2011)

[6] Broggi, A., Medici, P., Porta, P.P.: StereoBox: A Robust and Efficient Solution for Automotive Short-Range Obstacle Detection. EURASIP Journal on Embedded Systems (2007)

[7] Ke, Q., Kanade, T.: Transforming Camera Geometry to a Virtual Downward-Looking Camera: Robust Ego Motion Estimation and Ground-Layer Detection. In: Proceeding of the IEEE Computer Vision and Pattern Recognition, CVPR 2003 (2003)

[8] Fischer, M.A., Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated cartography. Comm. of the ACM 24, 381–396 (1981)

[9] Torr, P.H.S., Zisserman, A.: MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. In: IEEE Conference on Computer Vision and Image Understanding, CVPR 2000 (2000)

[10] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)

[11] Konolige, K., Agrawal, M., Sola, J.: Large Scale Visual Odometry for Rough Terrain. In: Proc. International Symposium on Robotics Research (2007)

[12] Umeyama, S.: Least-squares Estimation of Transformation Parameters Between Two Point Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 13(4), 376–380 (1991)

[13] Bruhn, A., et al.: Variational Optical Flow Computation in Real Time. IEEE Trans. on Image Processing 14(5) (May 2005)

[14] Cheng, S., et al.: Parts-based Object Recognition Seeded by Frequency Tuned Saliency for Child Detection in Active Safety. In: Intelligent Transportation Systems Conference-ITSC (2012)

[15] Bar-Hillel, A., Levi, D., Krupka, E., Goldberg, C.: Part-Based Feature Synthesis for Human Detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 127–142. Springer, Heidelberg (2010)
[16] Zhou, J., Li, B.: Homography-based Ground Detection for a Mobile Robot Platform Using a Single Camera. In: IEEE Conference on Robotics and Automation, ICRA 2006 (2006)
[17] Chazelle, B.: A Minimum Spanning Tree Algorithm with Inverse-Ackermann Type Complexity. Journal of the Association for Computing Machinery 47(6), 1028–1047 (2000)
[18] Achanta, R., et al.: Frequency-tuned Salient Region Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1597–1604 (2009)