Evaluation of SM Techniques
on Real-World Video Sequences

Reinhard Klette

The University of Auckland, New Zealand

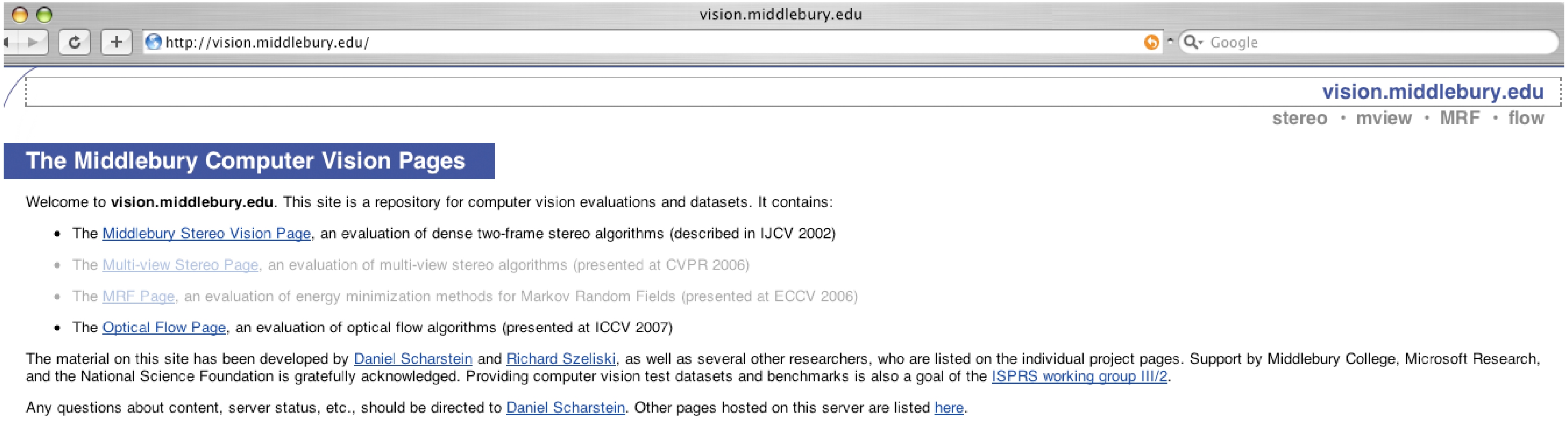
with contributions by

Shushi Guan, Ali Al-Sarraf and Zhifeng Liu

The University of Auckland, New Zealand

Jorge Sanchez

Universidad Tecnologica Nacional, Cordoba, Argentina



Middlebury

Microsoft Research

NSF

isprs

information from imagery

Evaluation of stereo and motion data

- engineered high-resolution, high contrast data
- synthesized or indoor color images
- each set only a few images
- images designed on purpose
- ground truth and evaluation method available

.enpeda.. Image Sequence Analysis Test Site

This web site of the *.enpeda..* (Environmental Perception and Driver Assistance) project offers sets of ego-motion corrected and geometrically rectified stereo image sequences for the purpose of comparative performance evaluation of stereo, motion, or 6D analysis techniques.

Mercedes-Benz New Zealand DAIMLER Giltrap NorthShore Hella Aglaia BLACKHAWK THE UNIVERSITY OF AUCKLAND

Set 1: Night Vision Stereo Sequences

These seven stereo night vision sequences (12 bit, between 220 and 300 pairs of frames each) have been provided by Daimler AG, Germany, in

Evaluation of stereo and motion data

- outdoor gray value images (low resolution and contrast ..)
- long sequences of rectified stereo images at 25 Hz
- real-world scenes, with “surprises”
- approximate ground truth and ideas about evaluation

www.citr.auckland.ac.nz/6D/

#1: Construction-site sequence

#2: Save-turn sequence

#3: Squirrel sequence

#4: Dancing-light sequence

#5: Intern-on-bike sequence

#6: Traffic-light sequence

Seven stereo sequences

Uwe Franke, Tobi Vaudrey et al., Daimler A.G. 2007

#7: Crazy-turn sequence

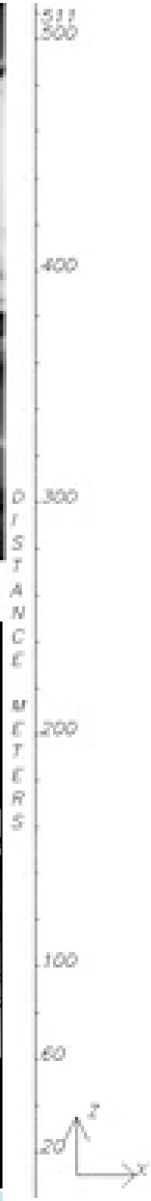
```
P5
# bigEndian
#[Units are rads, metres and seconds]
#[Inertial Sensor]
#YawRate= 0.004398
#Speed= 8.329330
#[Other Image Data]
#CycleTime: 0.080000
#ImageNumber: 111
#[Wheel Data]
#WheelRPM_FL: 240.500000
#WheelRPM_FR: 242.000000
#WheelRPM_RL: 235.500000
#WheelRPM_RR: 237.000000
#
640 481
3585
```

Part I - STEREO

#1: Construction-site sequence

DP with temporal propagation

with Zhifeng Liu 2008

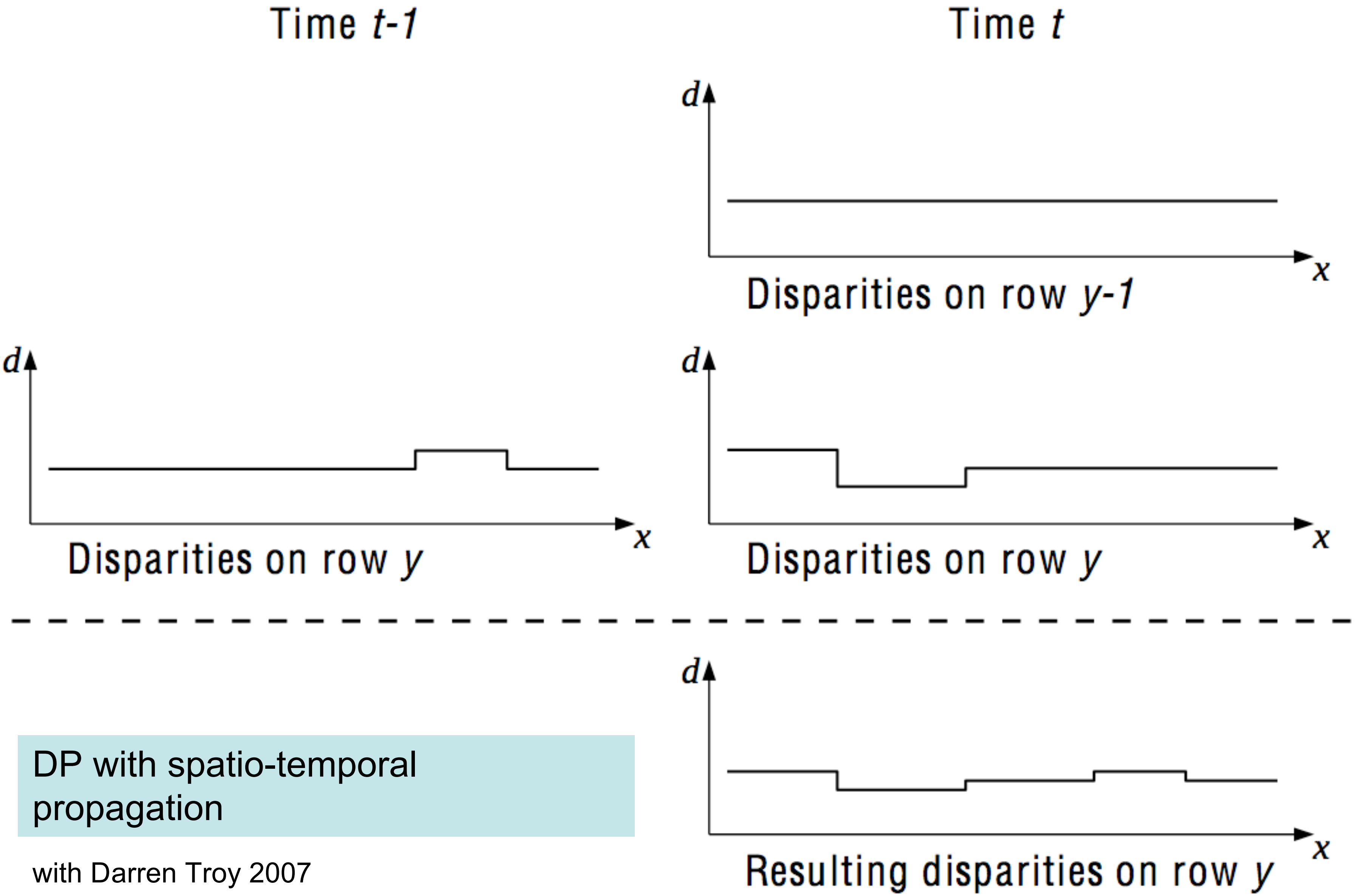


Trajectory of ego vehicle

with Ali Al-Sarraf 2008

DP with spatio-temporal propagation

with Darren Troy 2007

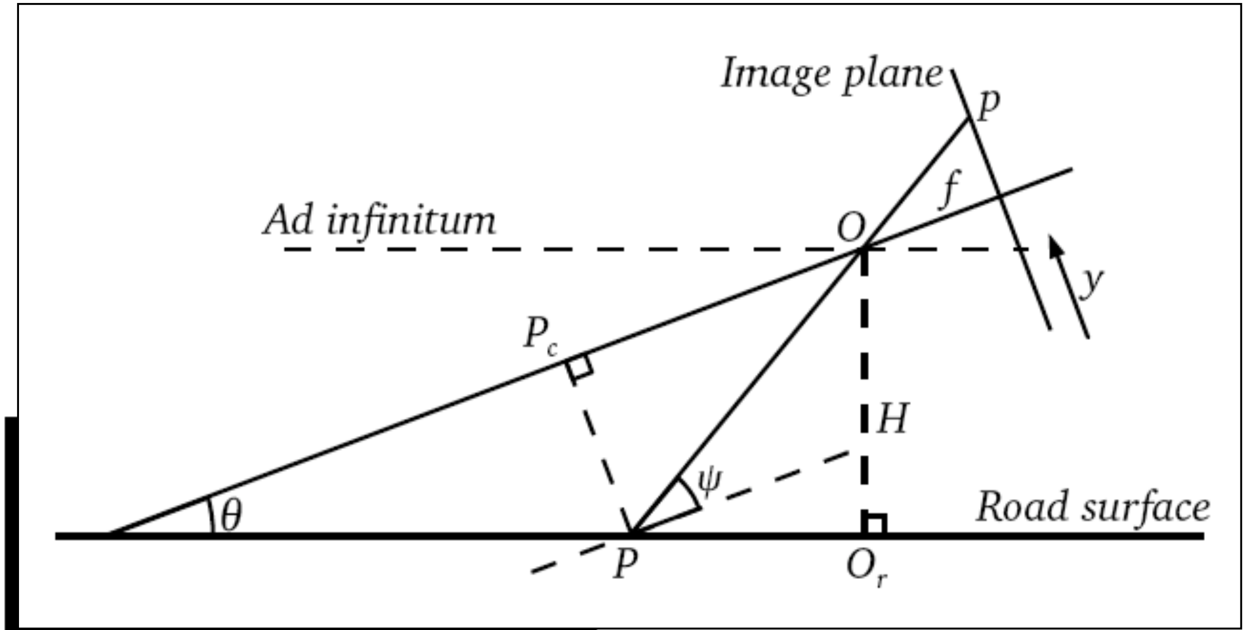


#2: Save-turn sequence

SGM + Mutual Information, 3 iterations

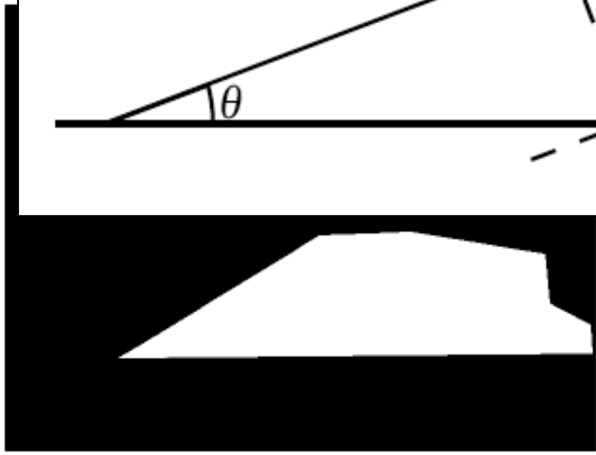
Heiko Hirschmüller 2005

Ground truth for disparities on road areas: *assume that the road is planar*

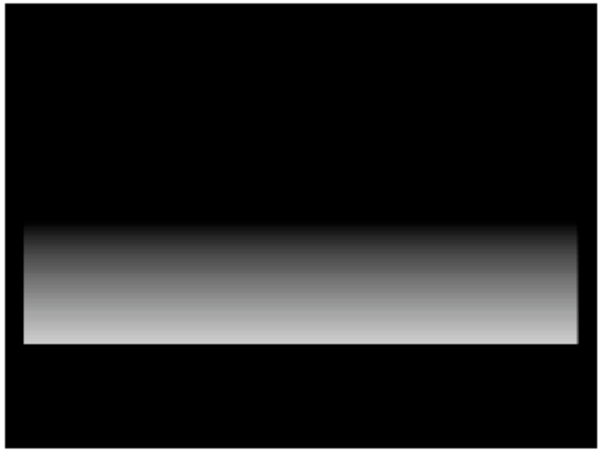


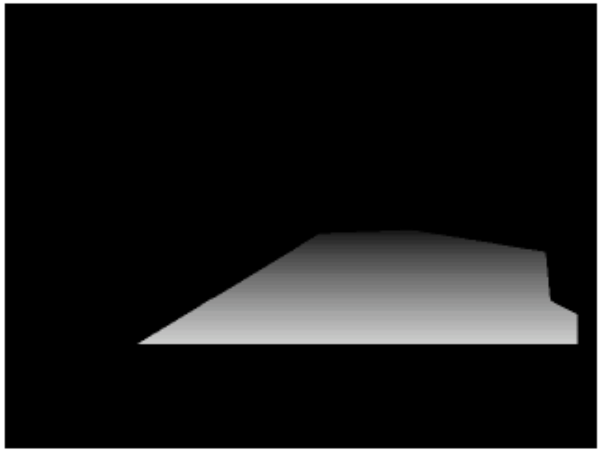
with Zhifeng Liu 2008


Original left image


Road surface mask

$$\psi = \arctan\left(\frac{(y_p - y_0)s_y}{f}\right)$$


Computed disparity

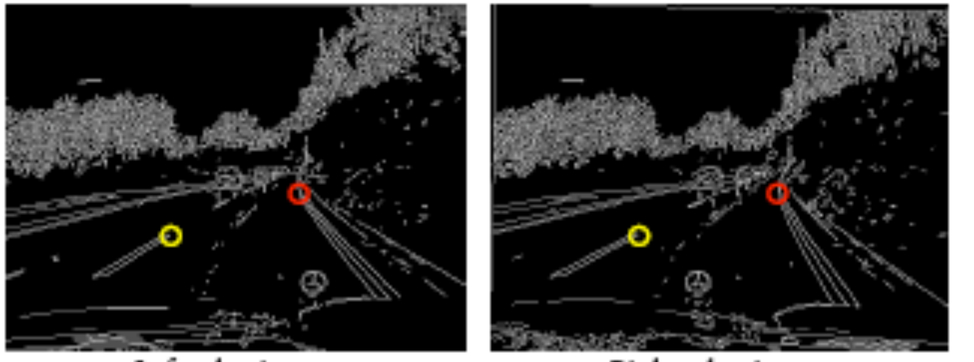

Disparity with mask

Tilt angle:

$$\theta = \arcsin\left(\frac{H\cos\psi \cdot d}{b \cdot f}\right) - \psi$$

Manual estimation of tilt angles

Selected pairs of corresponding points for 5 different pairs of frames in each of the 220 … 300 long sequences



Left edge image Right edge image

Sequence Name	Tilt Angle (radian)
1: 2007-03-06_121807	0.01608
2: 2007-03-07_144703	0.01312
3: 2007-03-15_182043	0.02050
4: 2007-04-20_083101	0.06126
5: 2007-04-27_145842	0.06223
6: 2007-04-27_155554	0.06944
7: 2007-05-08_132636	0.05961

Estimated tilt angles

Calibrated parameters of stereo camera system, also with respect to car:

```
################################################################################
#              Camera parameter file for ts_StereoCamera class.               #
################################################################################

[INTERNAL]
F         = 820.428         # [pixel] focal length
SX        = 1.0             # [pixel] pixel size in X direction
SY        = 1.000283 # [pixel] pixel size in Y direction
X0        = 305.278  # [pixel] X-coordinate of principle point
Y0        = 239.826  # [pixel] Y-coordinate of principle point

[EXTERNAL]
B         = 0.308084 # [m] width of baseline of stereo camera rig
LATPOS    = -0.07    # [m] lateral position of rectified images (virtual camera)
HEIGHT    = 1.26            # [m] height of rectified images (virtual camera)
DISTANCE = 0.0              # [m] distance of rectified images (virtual camera)
TILT      = 0.06     # [rad] tilt angle
YAW       = -0.01    # [rad] yaw angle
ROLL      = 0.0             # [rad] roll angle
```

Estimated angle seems to be more accurate, may be improved based on calculated (correct) disparities

Root Mean Squared Error and Bad Matches (> 1 pixel difference)

Sequence Name	Num of Frames	RMS	Bad Match %
1: 2007-03-06_121807	300	0.020271	2.68%
2: 2007-03-07_144703	300	0.023257	8.51%
3: 2007-03-15_182043	300	0.023400	23.11%
4: 2007-04-20_083101	250	0.067744	21.40%
5: 2007-04-27_145842	250	0.063743	17.50%
6: 2007-04-27_155554	250	0.071799	44.78%
7: 2007-05-08_132636	220	0.056440	35.75%

Original DP (without propagation) on road areas

Root Mean Squared Error and Bad Matches (> 1 pixel difference)

Sequence Name	Num of Frames	RMS	Bad Match %
1: 2007-03-06_121807	300	0.026014	15.12%
2: 2007-03-07_144703	300	0.053607	51.87%
3: 2007-03-15_182043	300	0.025122	40.37%
4: 2007-04-20_083101	250	0.069820	46.37%
5: 2007-04-27_145842	250	0.064231	24.85%
6: 2007-04-27_155554	250	0.074456	58.16%
7: 2007-05-08_132636	220	0.061994	50.80%

DP with spatial propagation on road areas

Root Mean Squared Error and Bad Matches (> 1 pixel difference)

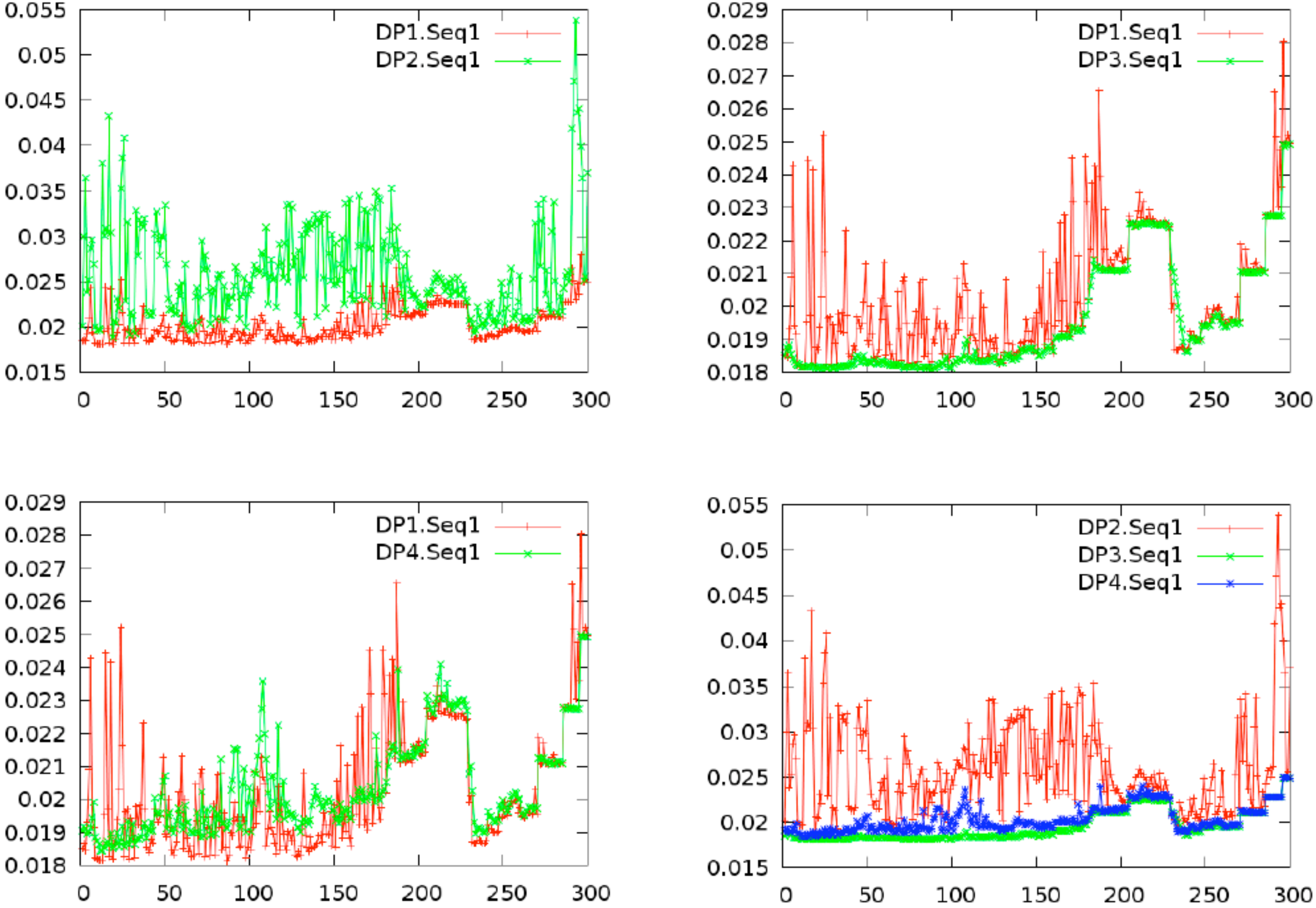
Sequence Name	Num of Frames	RMS	Bad Match %
1: 2007-03-06_121807	300	0.019513	1.88%
2: 2007-03-07_144703	300	0.018198	3.28%
3: 2007-03-15_182043	300	0.022127	17.75%
4: 2007-04-20_083101	250	0.067528	19.24%
5: 2007-04-27_145842	250	0.063678	16.37%
6: 2007-04-27_155554	250	0.071739	45.28%
7: 2007-05-08_132636	220	0.054376	32.87%

DP with temporal propagation on road areas

Root Mean Squared
Error and Bad Matches
(> 1 pixel difference)

Sequence Name	Num of Frames	RMS	Bad Match %
1: 2007-03-06_121807	300	0.020348	5.54%
2: 2007-03-07_144703	300	0.038693	18.98%
3: 2007-03-15_182043	300	0.022827	24.36%
4: 2007-04-20_083101	250	0.067902	31.79%
5: 2007-04-27_145842	250	0.063755	16.19%
6: 2007-04-27_155554	250	0.072373	51.01%
7: 2007-05-08_132636	220	0.058989	41.79%

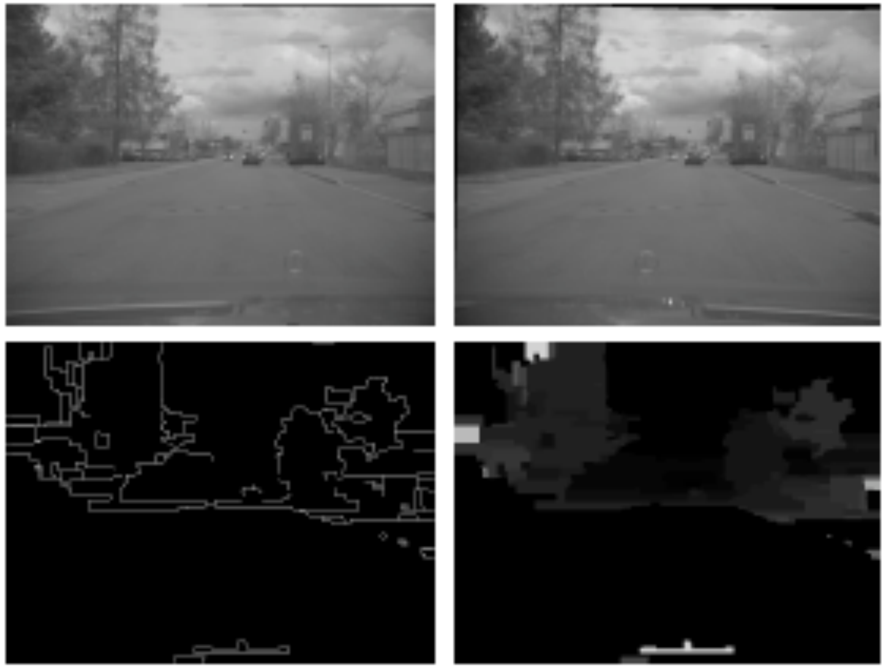
DP with spatio-temporal propagation on road areas

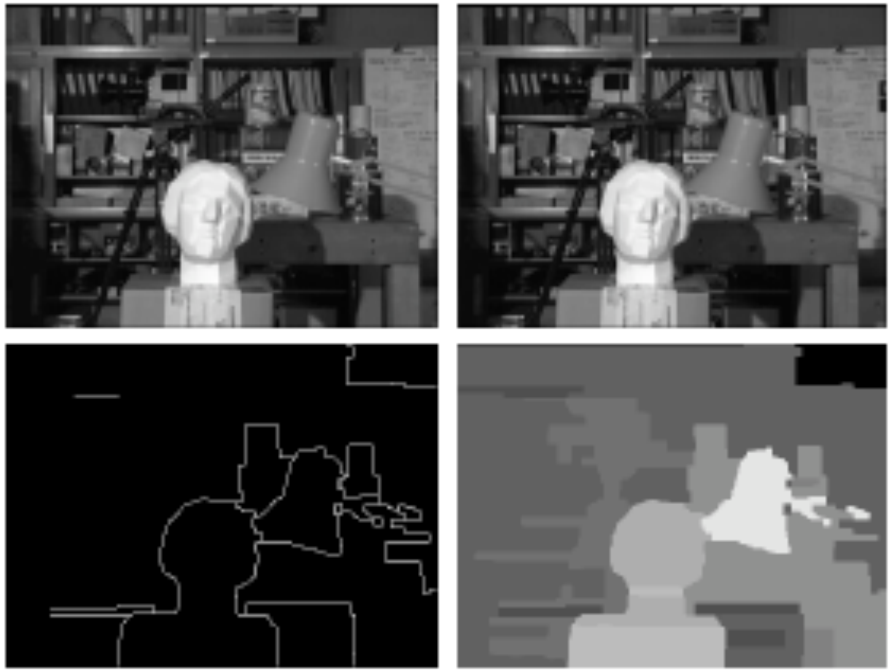


Best: DP3 (with temporal propagation only)

Birchfield-Tomasi on road areas

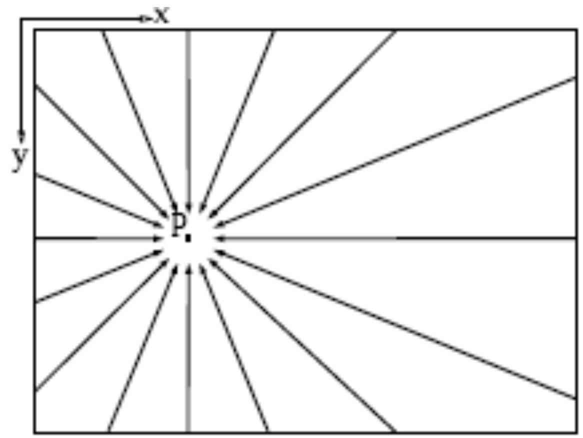
Sequence Name	Num of Frames	RMS	Bad Match %
1: 2007-03-06_121807	300	0.089806	60.98%
2: 2007-03-07_144703	300	0.105662	96.80%
3: 2007-03-15_182043	300	0.109850	81.04%
4: 2007-04-20_083101	250	0.125842	99.21%
5: 2007-04-27_145842	250	0.116894	94.63%
6: 2007-04-27_155554	250	0.135165	99.82%
7: 2007-05-08_132636	220	0.104936	99.43%





Sequence Name	Num of Frames	RMS	Bad Match %
1: 2007-03-06_121807	300	0.036927	44.88%
2: 2007-03-07_144703	300	0.076011	77.97%
3: 2007-03-15_182043	300	0.063756	71.58%
4: 2007-04-20_083101	250	0.077724	60.57%
5: 2007-04-27_145842	250	0.080806	65.85%
6: 2007-04-27_155554	250	0.083163	73.90%
7: 2007-05-08_132636	220	0.067442	64.65%

SGM[3] MI[16]
on road areas



Original left input sequence

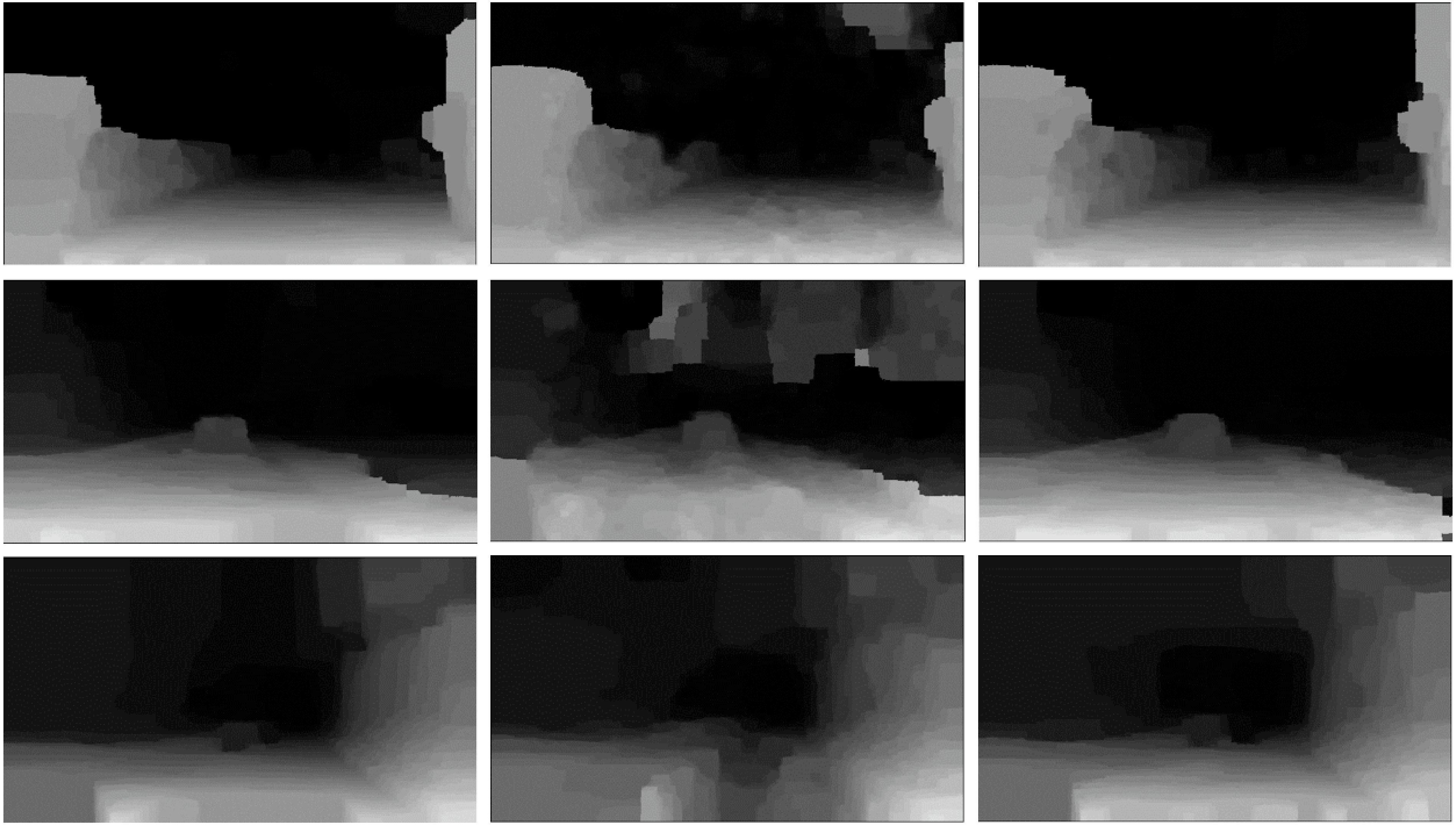
BP on original input sequences

Sobel of left input sequence

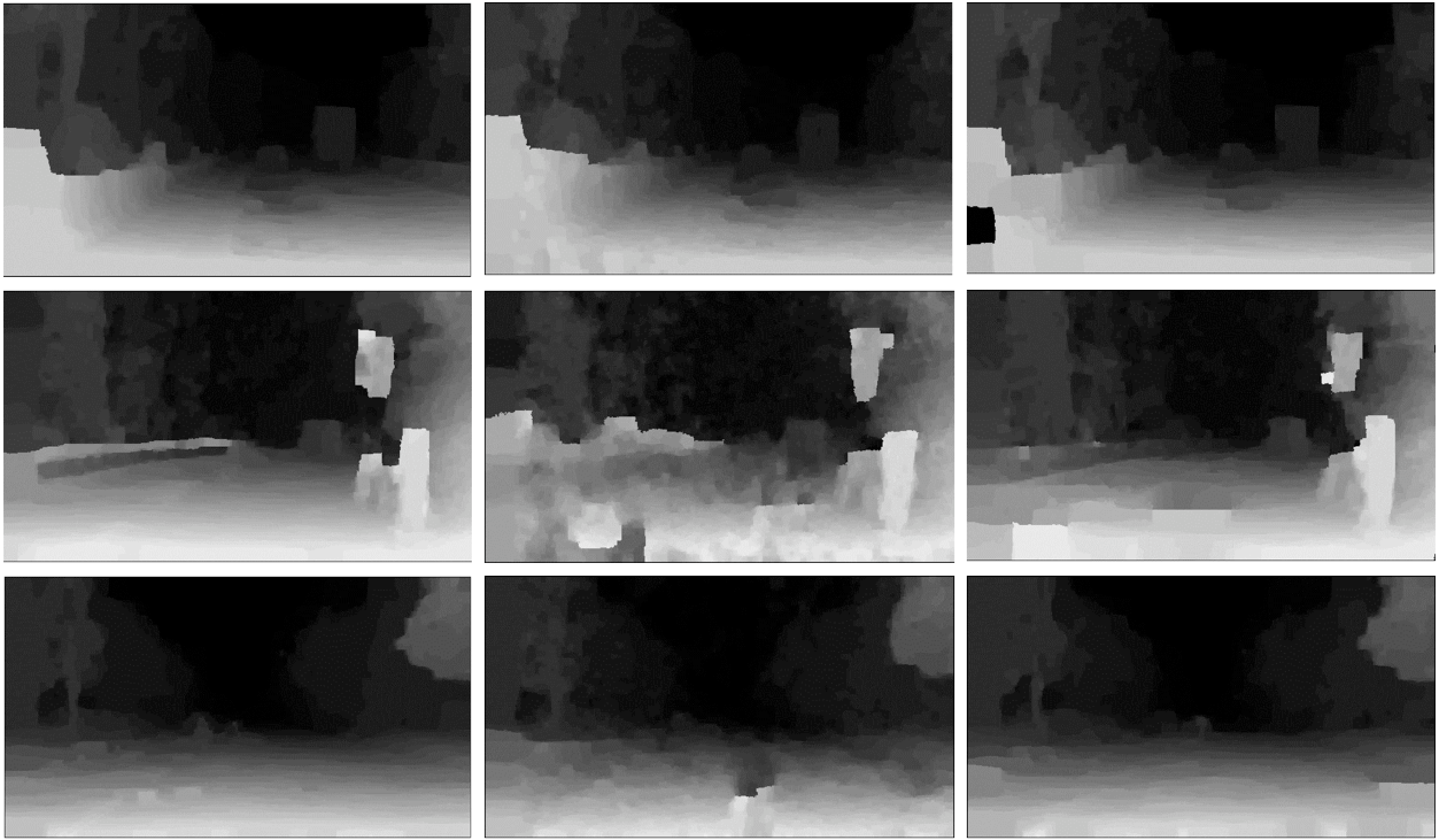
BP on Sobel input sequences

with Shushi Guan 2007

Kovesi-Owen max
Kovesi-Owen min
Sobel
Canny (high thresholds) OpenCV



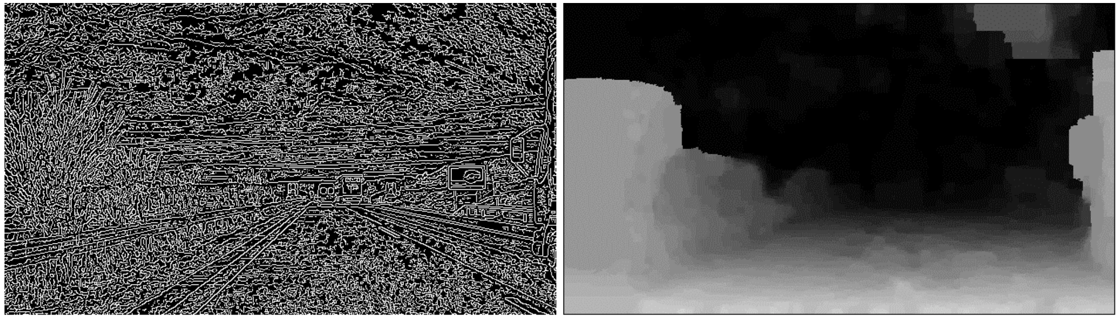
Sobel Canny Kovesi-Owen max



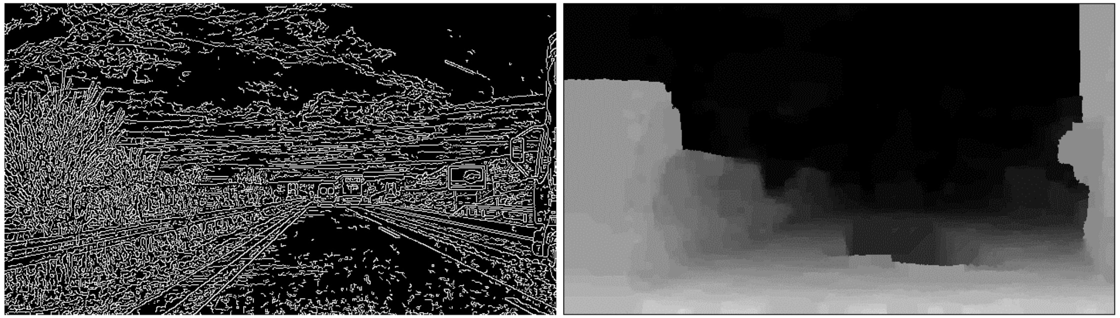
Sobel | Canny | Kovesi-Owen max

Canny, different thresholds:

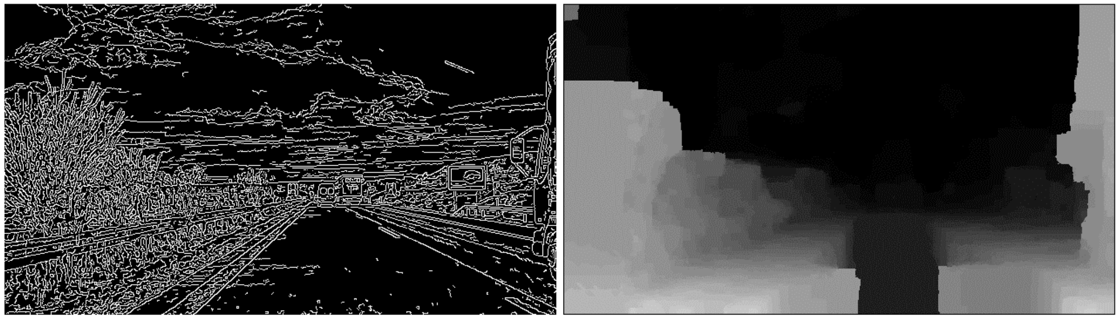
Upper: 12
Lower: 5



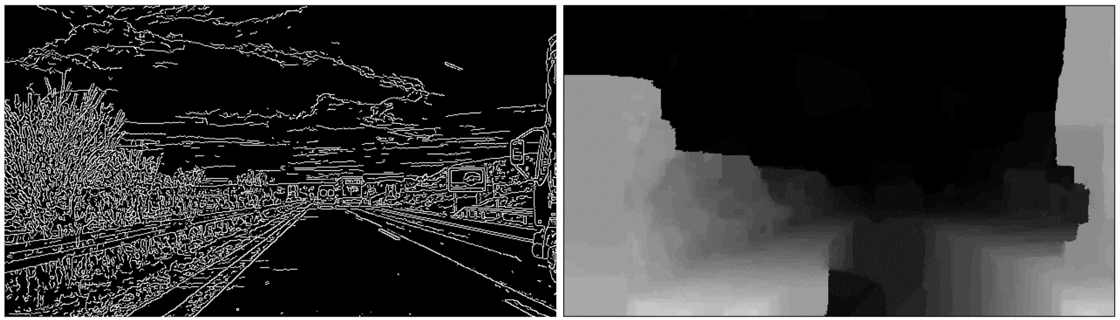
Upper: 20
Lower: 12

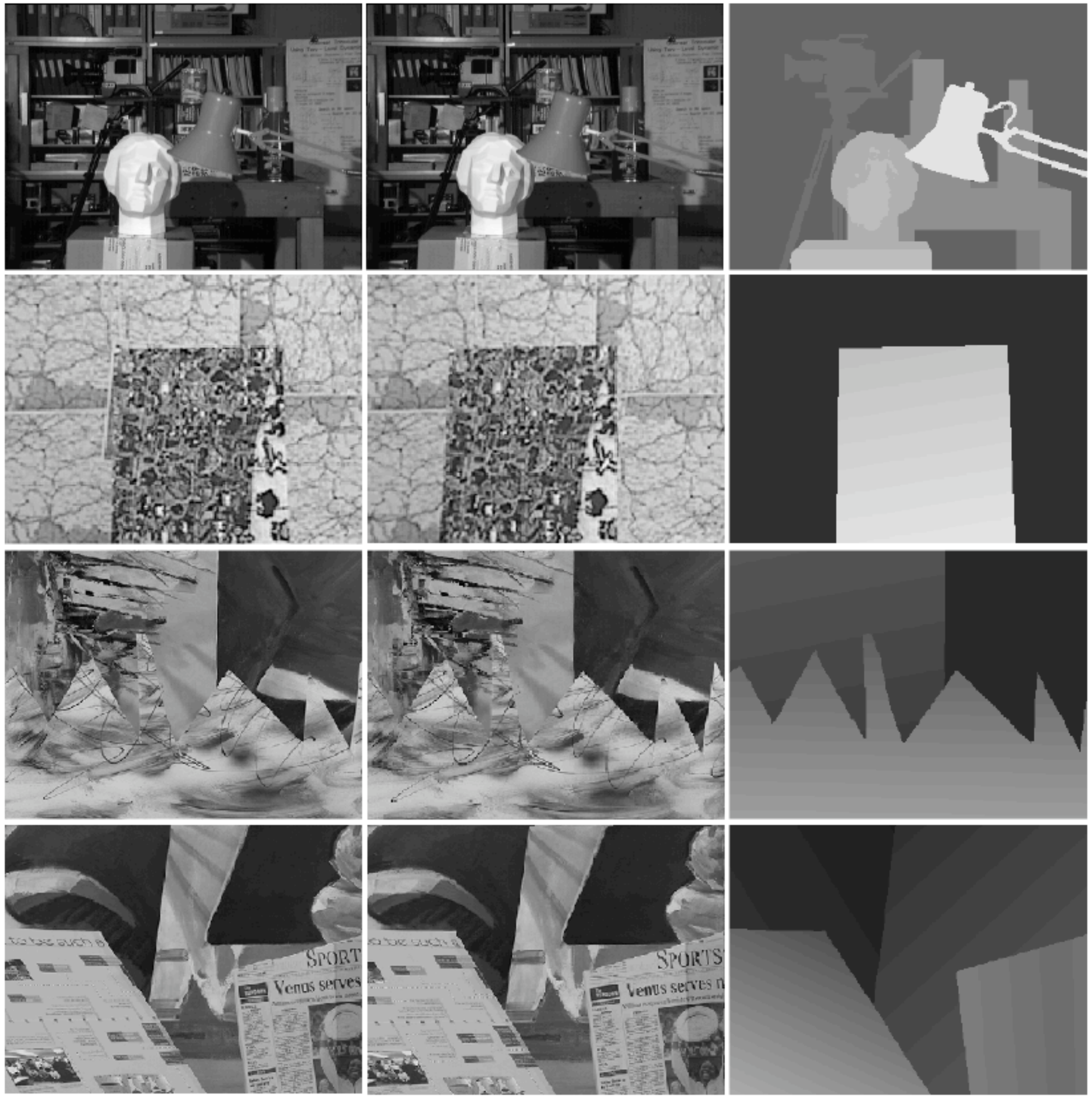


Upper: 28
Lower: 19



Upper: 36
Lower: 26





BP on Middlebury stereo image pairs

Table below: results get slightly worse for edge (Sobel) images of those stereo pairs

Image pair	Tsukuba	edge	Map	edge	Sawtooth	edge	Venus	edge
error	1.75	1.81	0.31	0.33	0.94	0.95	0.99	1.02

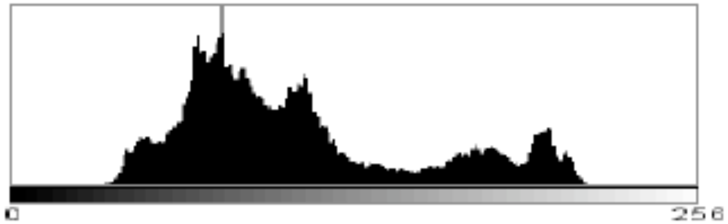
Specification of (finally) used BP algorithms

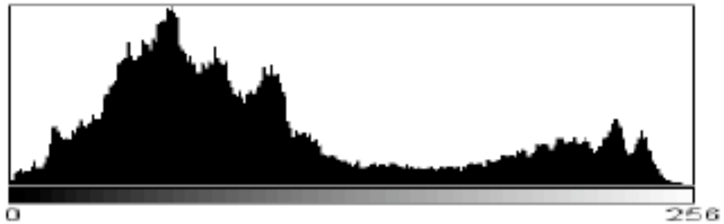
Number	Max-disparity	Iterations	Image size	Running time	Truncation of discontinuity cost	Truncation of data cost
1	30 *pixel*	7	640 × 360 *pixel*	2.9 *s*	11	30
2	35 *pixel*	7	640 × 360 *pixel*	3.1 *s*	11	25
3	40 *pixel*	5	640 × 360 *pixel*	2.9 *s*	23	20
4	30 *pixel*	7	640 × 360 *pixel*	2.9 *s*	20	60
5	30 *pixel*	5	640 × 360 *pixel*	2.7 *s*	11	30
6	35 *pixel*	6	640 × 360 *pixel*	3.1 *s*	10	30
7	40 *pixel*	5	640 × 360 *pixel*	2.9 *s*	11	30
				(for one pair of images)	(penalty for intensity differences)	(allows to handle occlusions)

Sobel preprocessing
max-product
4-adjacency
quadratic cost function
red-black speed-up method
coarse to fine for more reliable matching (5 to 7 layers; reduces #iterations)

(no initialization with disparities at time t-1, for t>0)

Brightness differences between left and right image





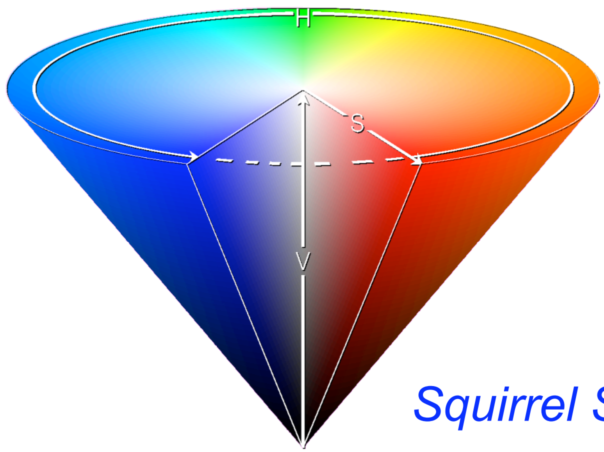
causes BP to fail (in difference to DP, SGM MI, or BT) - so far not discussed on Middlebury stereo page

Part II - MOTION

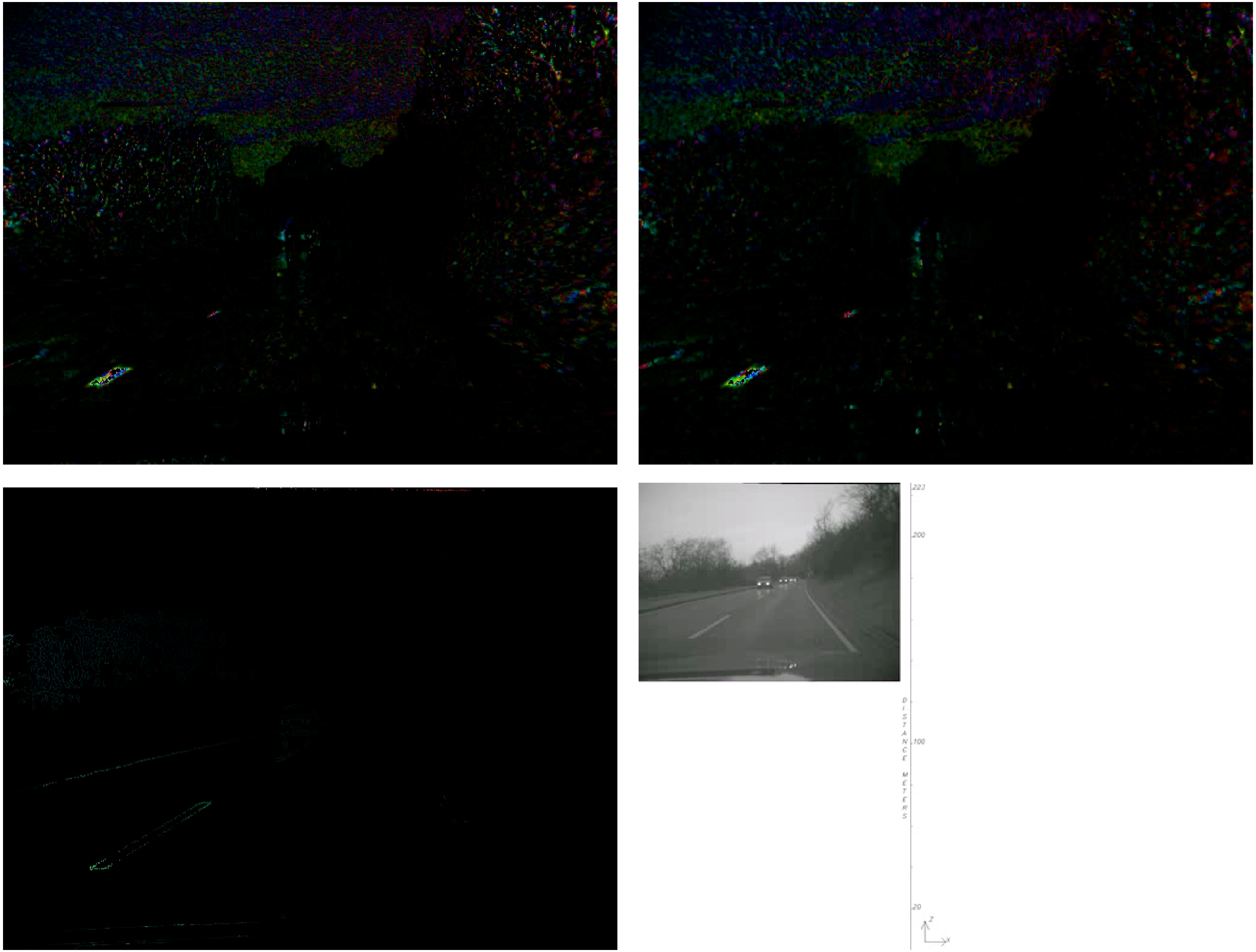
1. Horn-Schunck OpenCV

2. Lucas-Kanade OpenCV

3. Lucas-Kanade with Pyramids OpenCV



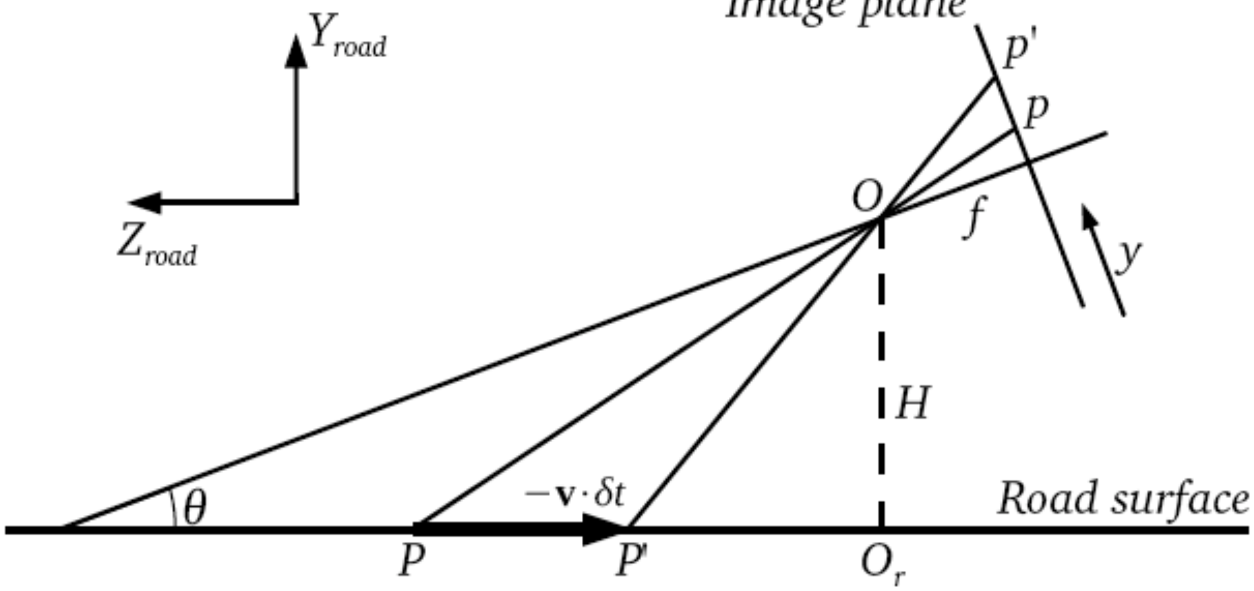
Squirrel Sequence (#3)

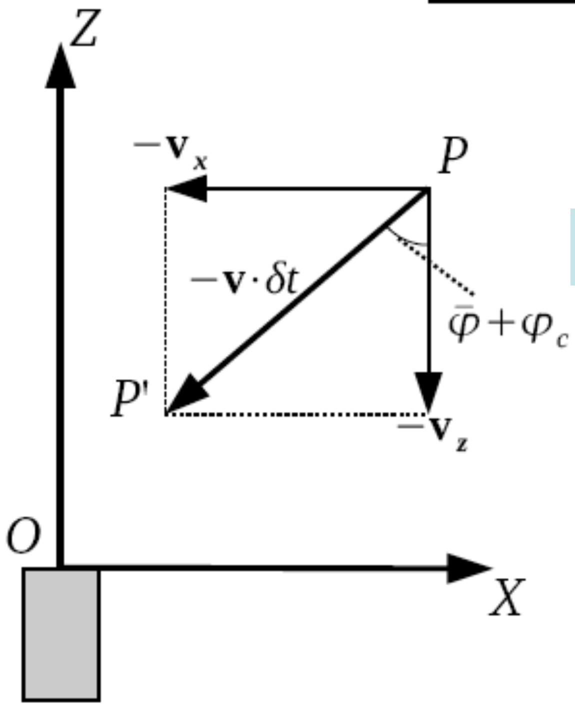


200
DISTANCE METERS
100
20

Ground truth for Optic flow on road areas

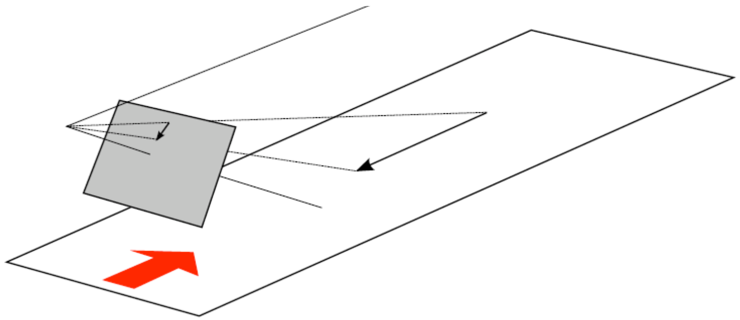
Side view





Top view

3D view



Mean Angular Error and End Point Error

$$E_{AE} = \frac{1}{n} \sum \arccos \left(\frac{\mathbf{u} \cdot \mathbf{u}_T}{|\mathbf{u}||\mathbf{u}_T|} \right)$$

Sequence Name	Num of Frames	Angular Error	End Point Error
1: 2007-03-06_121807	300	89.52	39.37
2: 2007-03-07_144703	300	84.96	8.87
3: 2007-03-15_182043	300	84.61	13.73
4: 2007-04-20_083101	250	86.34	27.74
5: 2007-04-27_145842	250	87.48	16.01
6: 2007-04-27_155554	250	46.29	25.22
7: 2007-05-08_132636	220	73.14	10.58

Horn-Schunck on road areas

Mean Angular Error and End Point Error

$$E_{AE} = \frac{1}{n} \sum \arccos\left(\frac{\mathbf{u} \cdot \mathbf{u}_T}{|\mathbf{u}||\mathbf{u}_T|}\right)$$

Sequence Name	Num of Frames	Angular Error	End Point Error
1: 2007-03-06_121807	300	89.31	34.90
2: 2007-03-07_144703	300	81.82	8.82
3: 2007-03-15_182043	300	83.02	13.80
4: 2007-04-20_083101	250	85.33	27.59
5: 2007-04-27_145842	250	85.40	16.19
6: 2007-04-27_155554	250	45.13	25.03
7: 2007-05-08_132636	220	69.60	10.38

Lucas-Kanade on road areas

Mean Angular Error and End Point Error

$$E_{AE} = \frac{1}{n} \sum \arccos\left(\frac{\mathbf{u} \cdot \mathbf{u}_T}{|\mathbf{u}||\mathbf{u}_T|}\right)$$

Sequence Name	Num of Frames	Angular Error	End Point Error
1: 2007-03-06_121807	300	72.69	20.72
2: 2007-03-07_144703	300	97.49	8.90
3: 2007-03-15_182043	300	64.12	9.50
4: 2007-04-20_083101	250	45.19	14.37
5: 2007-04-27_145842	250	65.88	13.38
6: 2007-04-27_155554	250	31.80	20.94
7: 2007-05-08_132636	220	32.36	6.46

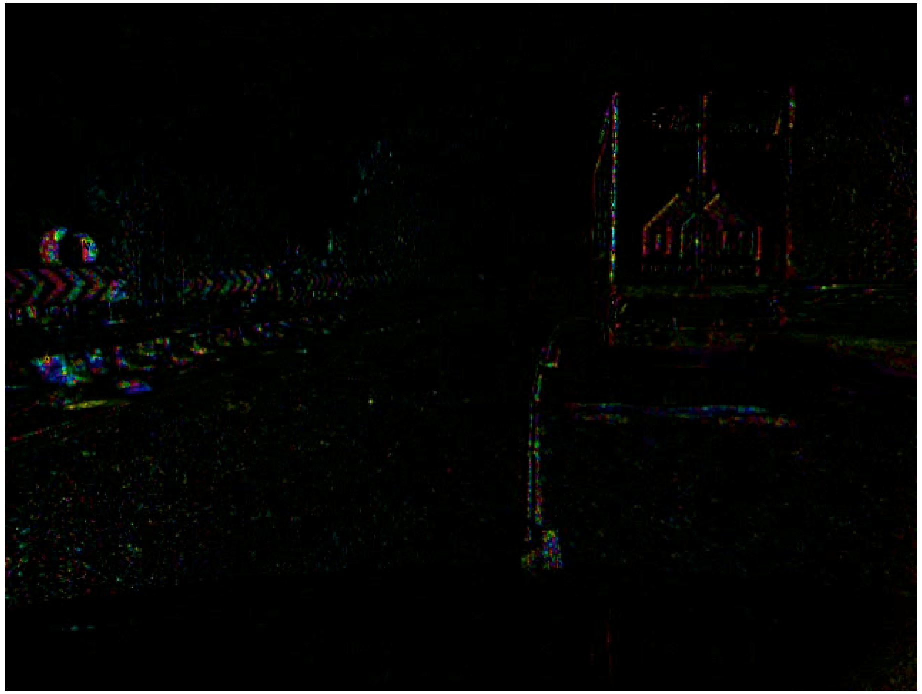
Pyramid Lucas-Kanade on road areas

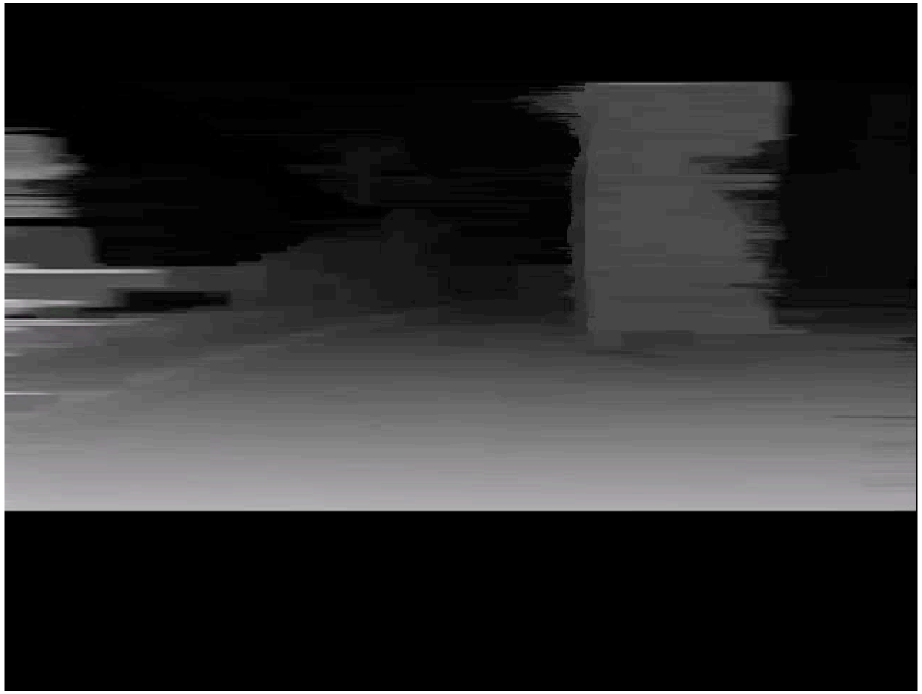
Dancing-Light Sequence (#4)

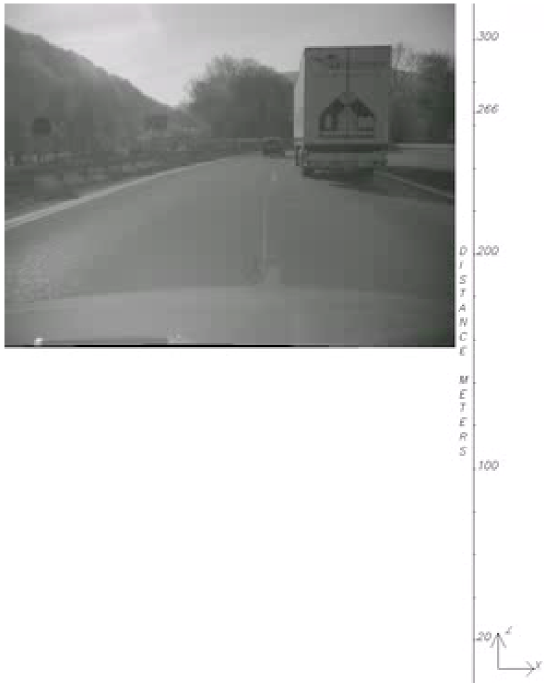
1. Horn-Schunck OpenCV

2. DP with temporal propagation

3. Lucas-Kanade with Pyramids OpenCV

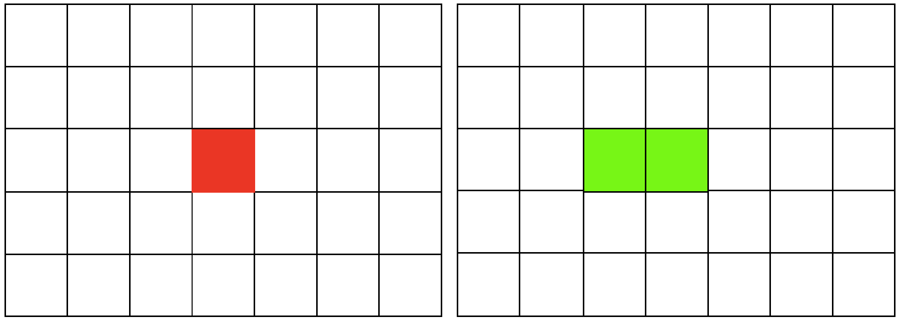





300
266
200
DISTANCE
METERS
100
20

BP for Optical Flow

Maximum disparities =2



Left image　Right image

One-dimensional Search window

Maximum displacement =2

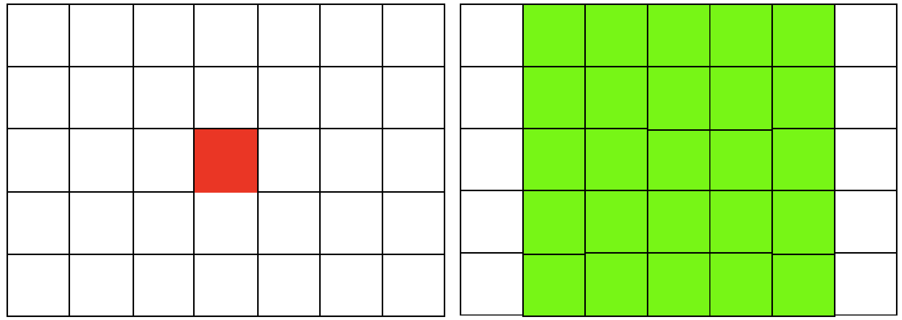


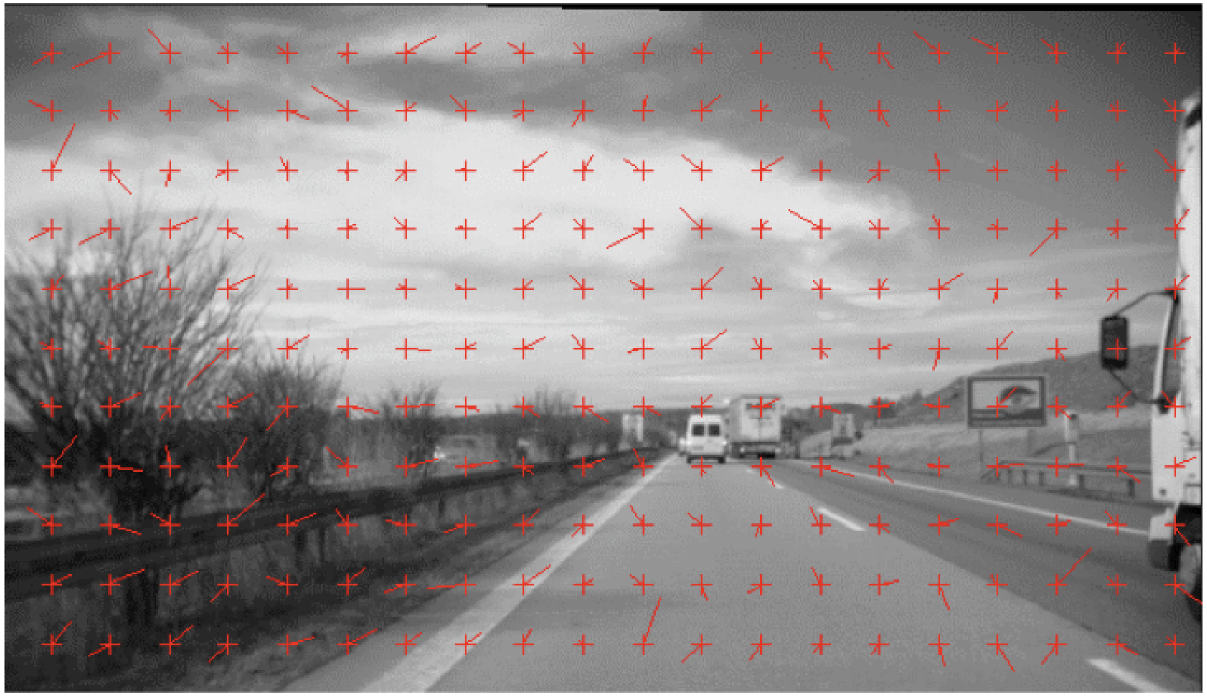
Image at time t　Image at time t+1

Two-dimensional Search window

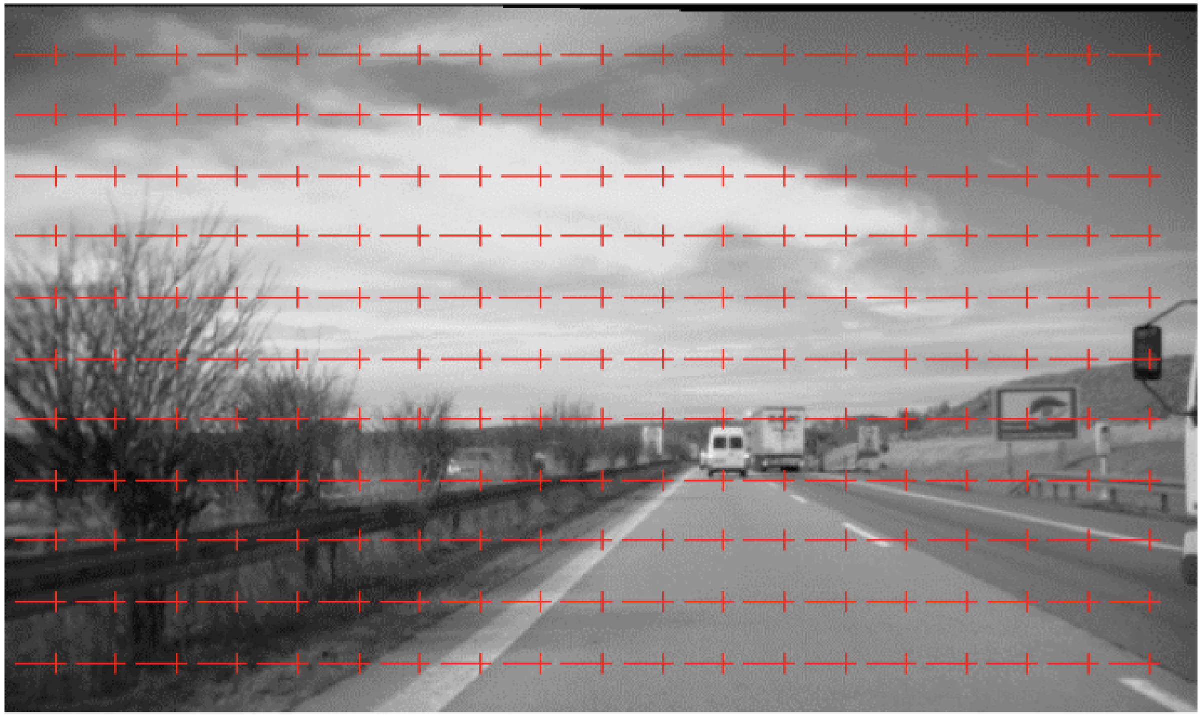
2 labels = 2 arrays

25 labels = 25 arrays

with Shushi Guan 2008



2D search in input sequences: insufficient time (space) for accurate results



1D search in a simulated input sequence: BP leads to a perfectly accurate result !

KLT tracker, max-response disks in scale space for 3D vector estimation

www.ces.clemson.edu/~stb/klt/

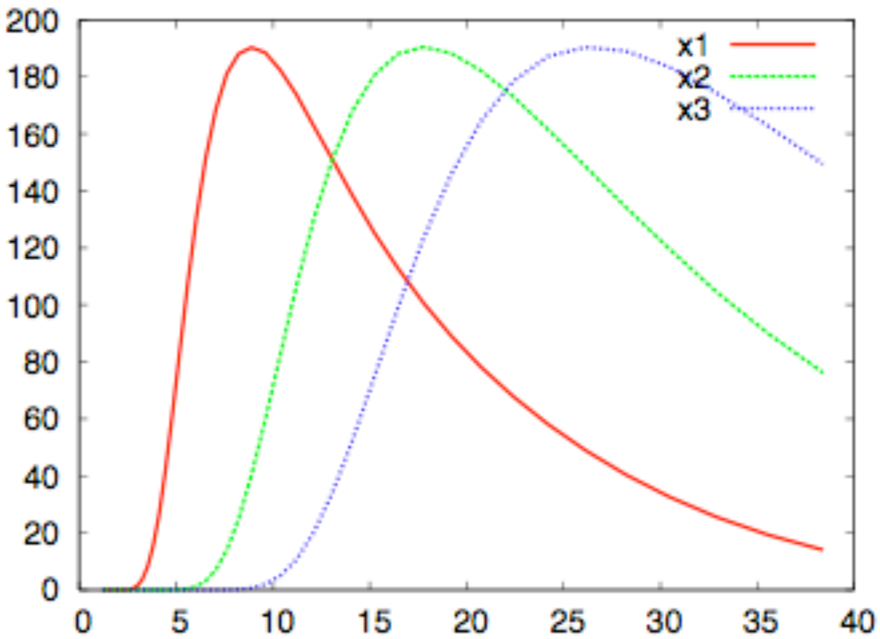
with Jorge Sanchez 2008

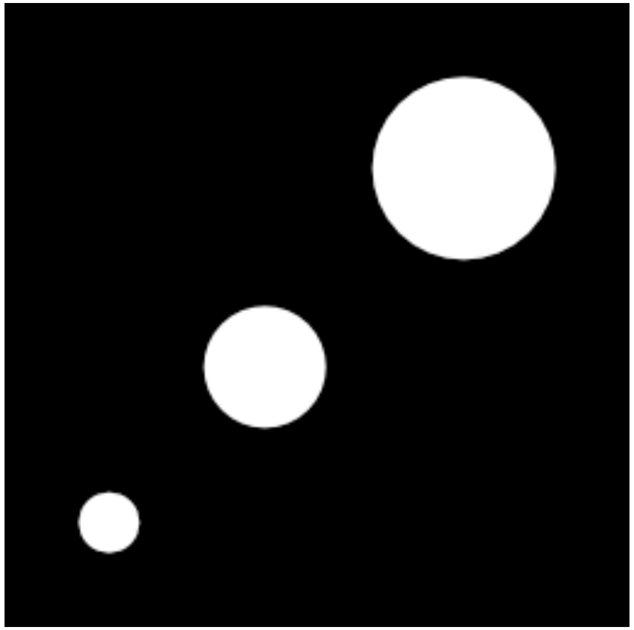
Feature detection with automatic scale selection

Tony Lindeberg 1998

$$\nabla_{norm} L(\mathbf{x},\sigma) = \sigma^2 \left| (D_x^2 L)(\mathbf{x},\sigma) + (D_y^2 L)(\mathbf{x},\sigma) \right|$$ for pixel **x**

with magnitude $A = c(\sigma_1)\sigma_1^{-p} e^{\sigma_{max}/\theta}$ at local maxima

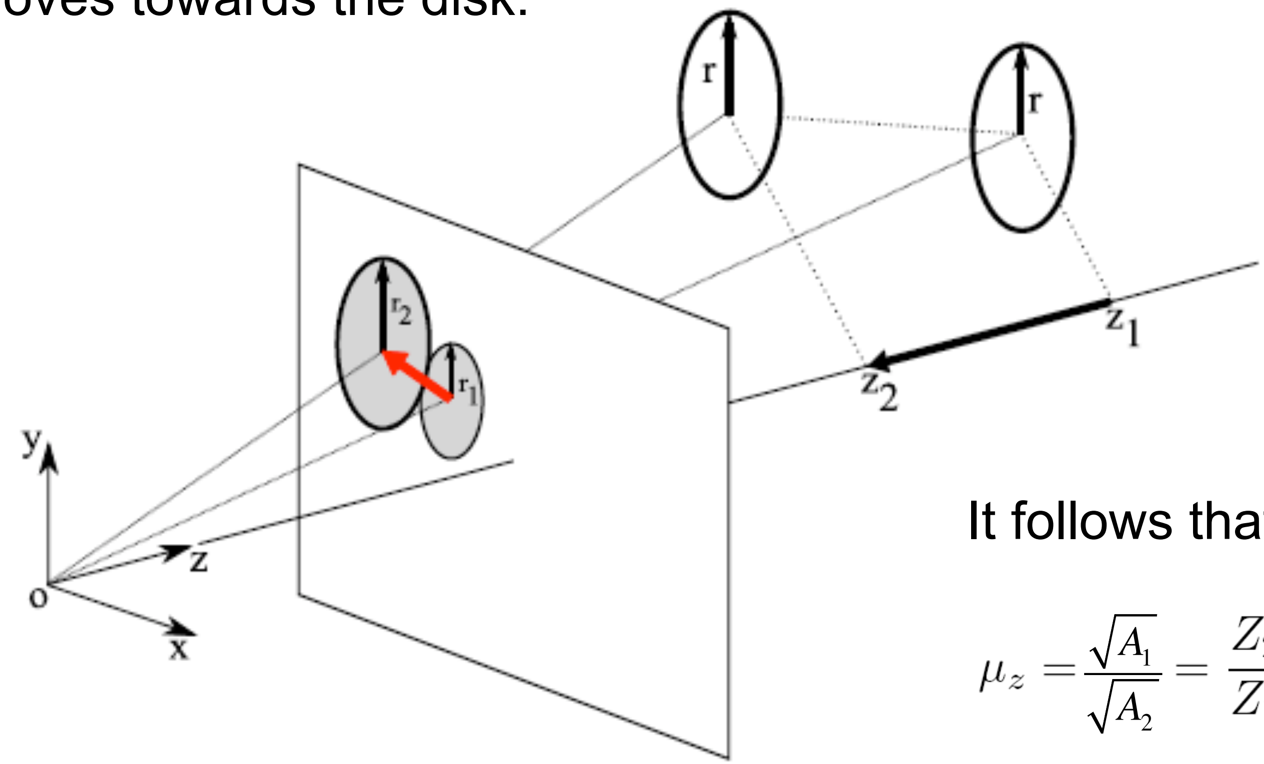




Scale evolutions of the centers of the three white blobs:

The ratio between the location of the extrema in scale equals the ratio between the areas of corresponding white disks.

A camera moves towards the disk:



It follows that

$$\mu_z = \frac{\sqrt{A_1}}{\sqrt{A_2}} = \frac{Z_2}{Z_1}$$

Let $\mu_x = \frac{X_{t+1}}{X_t}$ and $\mu_y = \frac{Y_{t+1}}{Y_t}$. Then $\begin{pmatrix} X_{t+1} \\ Y_{t+1} \\ Z_{t+1} \end{pmatrix} = \begin{pmatrix} \mu_x & 0 & 0 \\ 0 & \mu_y & 0 \\ 0 & 0 & \mu_z \end{pmatrix} \begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix}$

Note: only corresponding image points and μ_z (i.e., A_1 and A_2) needed.

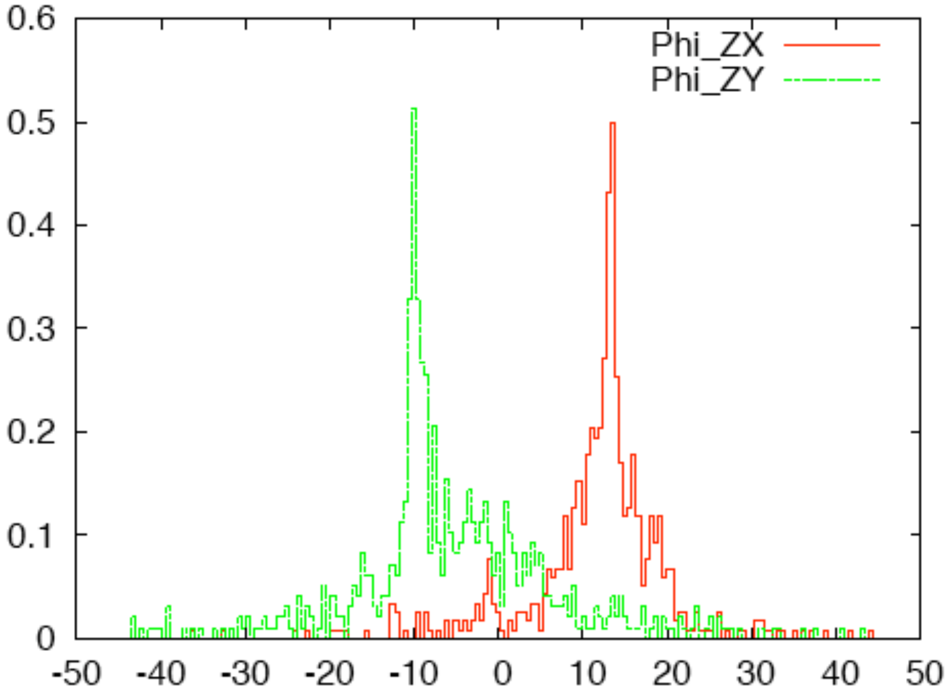
1. Use KLT tracker and calculate pairs of a tracked 3D point between L_t and L_{t+1}.
2. Calculate scale-space representations for set of predefined scales.
3. Select both local maxima for each pair of points.
4. Compute scale ratio for each pair of points and thus its μ_z-factor.
5. Obtain the 3D motion angles as arctan of the ratios

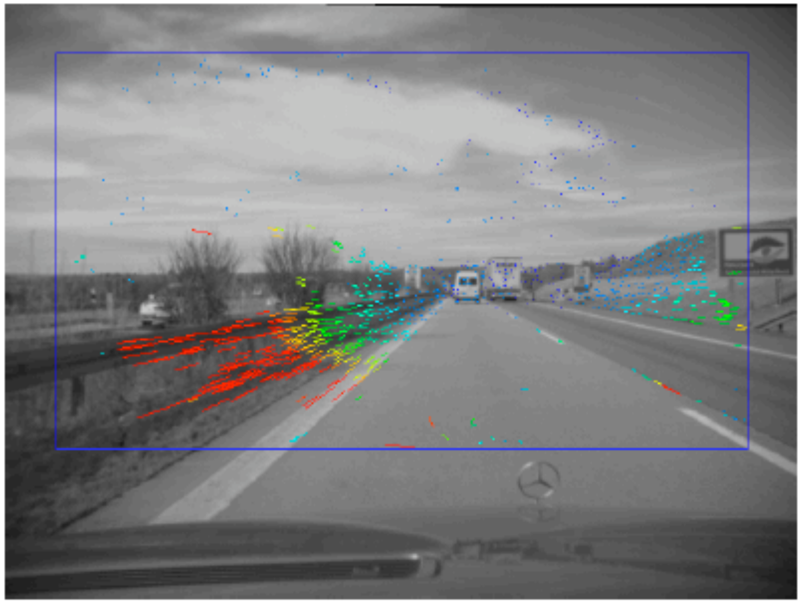
$$\frac{\Delta X}{\Delta Z} = \left(\frac{\mu_x - 1}{\mu_z - 1}\right)\frac{X_t}{Z_t}$$

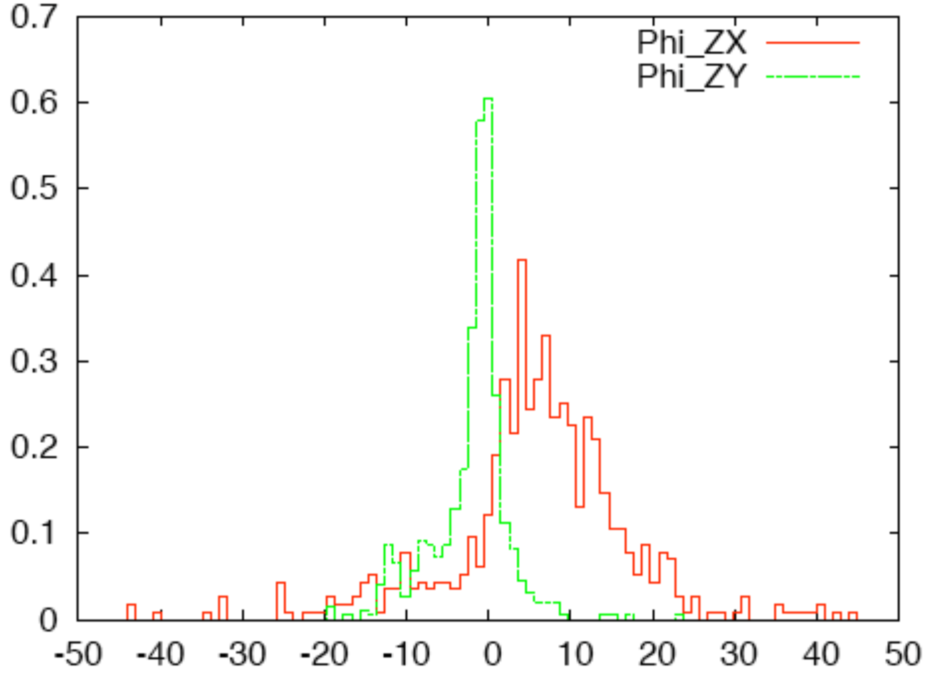
$$\frac{\Delta Y}{\Delta Z} = \left(\frac{\mu_y - 1}{\mu_z - 1}\right)\frac{Y_t}{Z_t}\ :$$

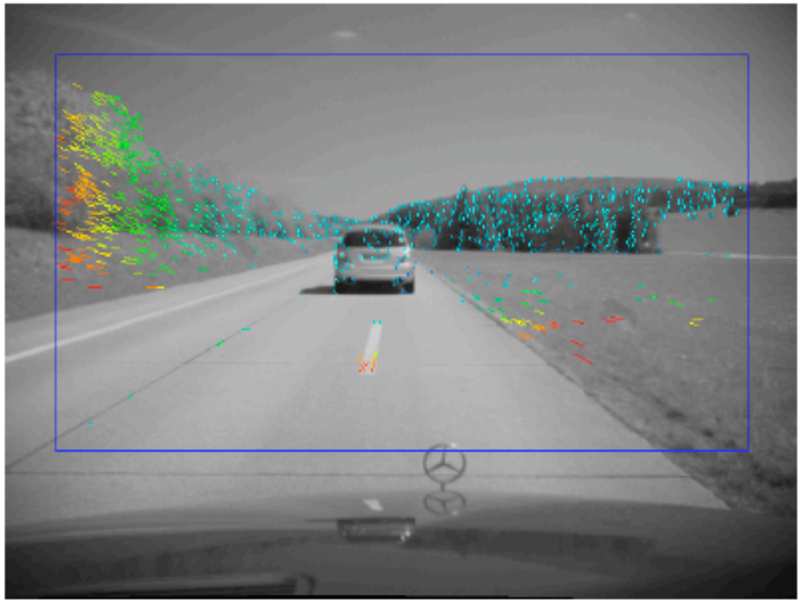
with
$$\begin{aligned}\Delta X &= X_{t+1} - X_t = (\mu_x - 1)X_t\\ \Delta Y &= Y_{t+1} - Y_t = (\mu_y - 1)Y_t\\ \Delta Z &= Z_{t+1} - Z_t = (\mu_z - 1)Z_t\end{aligned}$$

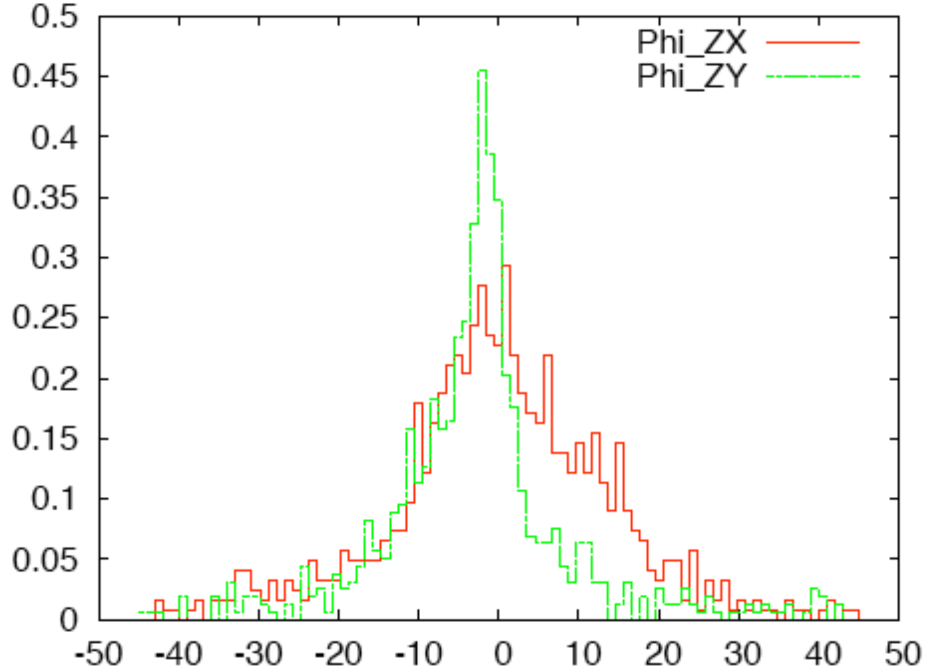
Evaluation for known ground truth: angles of about -10 and 12 for a translated calibrated camera and a desk-top scene:

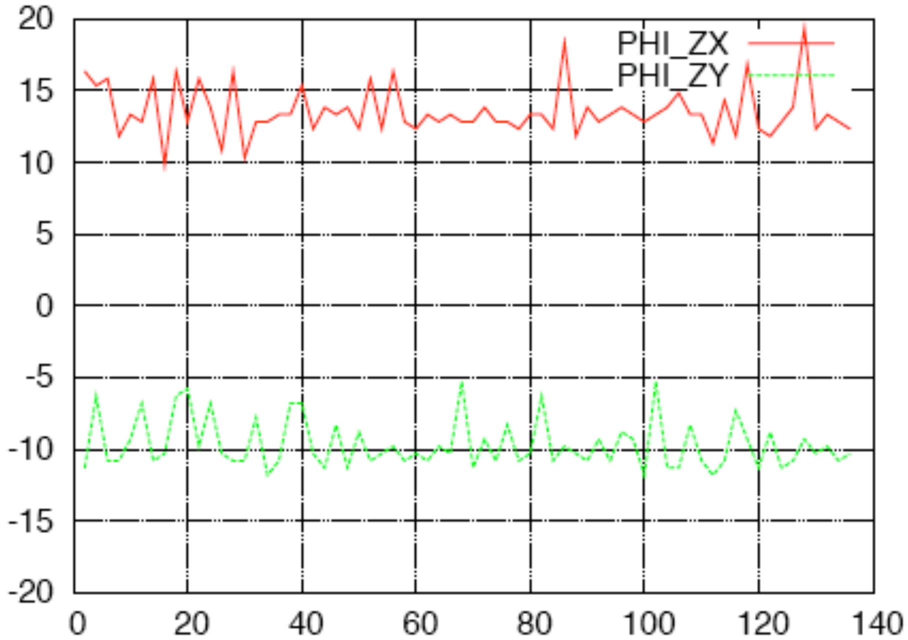




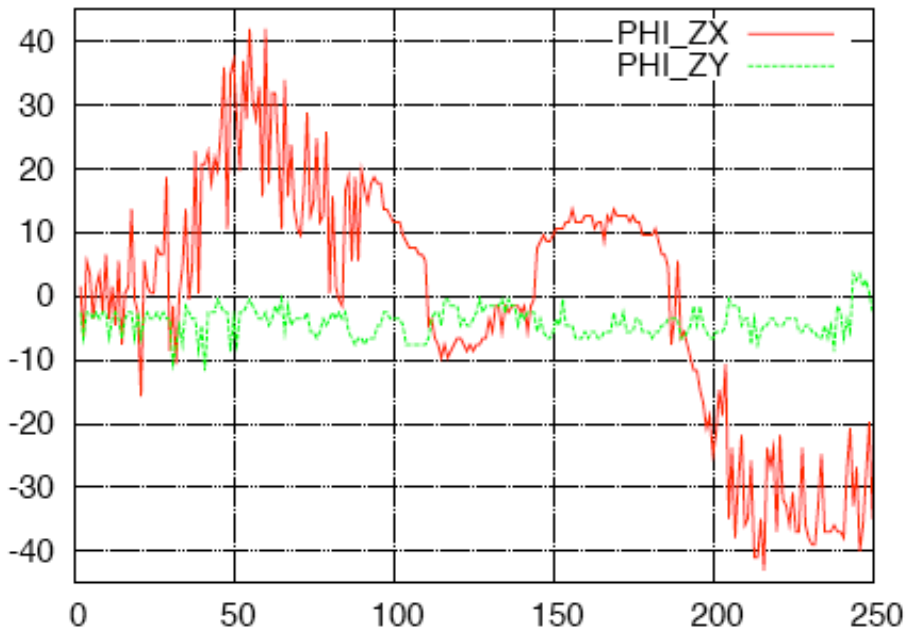

Phi_ZX
Phi_ZY
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-50 -40 -30 -20 -10 0 10 20 30 40 50




Phi_ZX
Phi_ZY
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-50 -40 -30 -20 -10 0 10 20 30 40 50



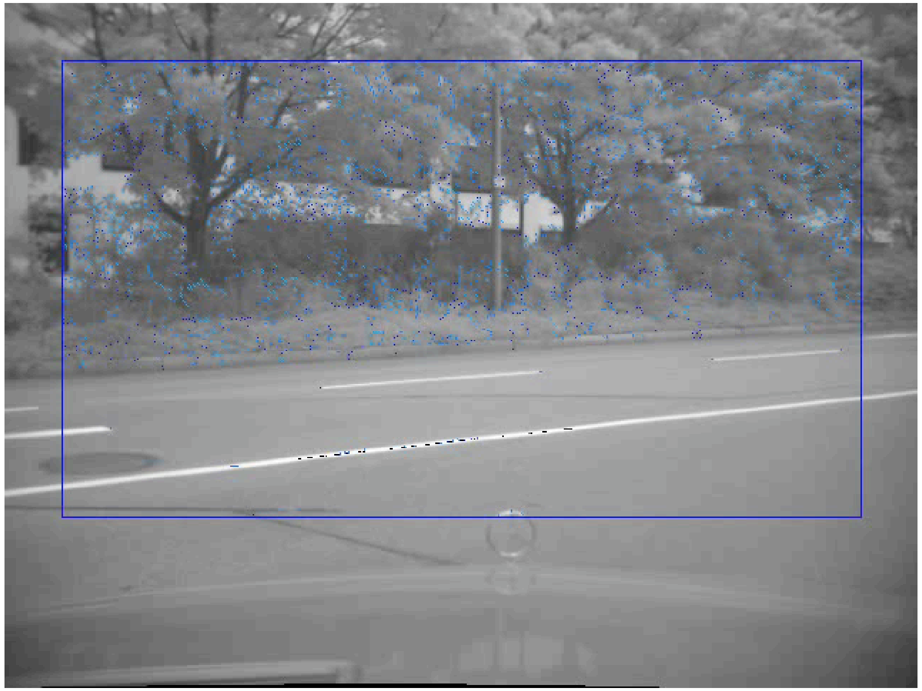
Mean 3D direction for a video sequence (top: constant translation, -10 and 12)



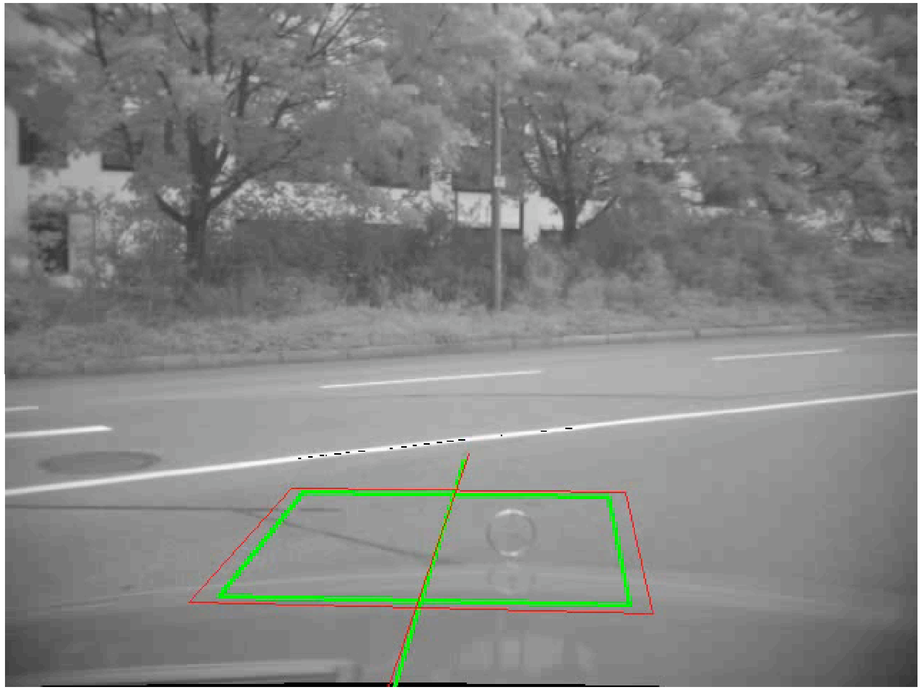
Mean direction of estimated 3D directions:

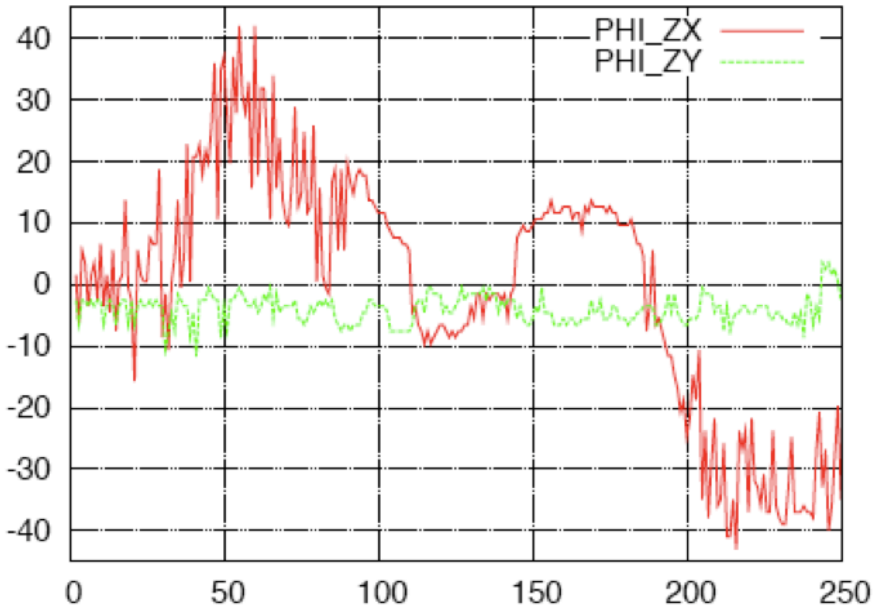
red mark = instantaneous estimation

green mark = smoothed (Kalman-filtered) version



Crazy-Turn Sequence (#7)





Conclusions

Significant differences in evaluation for Middlebury data and used data

Stereo: (1) Tilt angle calibration based on accurate road disparity method

(2) DP3 on road better than DP1, DP2, DP4, BT (!), SGM^{3} MI^{16} (!)

(3) BP on Sobel fine if both images of equal brightness

(4) Edge operator prior to BP should not filter out any `structure'

(NEXT) Combining DP3 on road and SGM or BP on no-road ?

Motion: (1) PyrLK better than HS, LK, but actually - still not usable

(2) BP potentially fine - but we need faster (or: parallel) computers

(3) Scale ratio between tracked points: depends on scale estimation

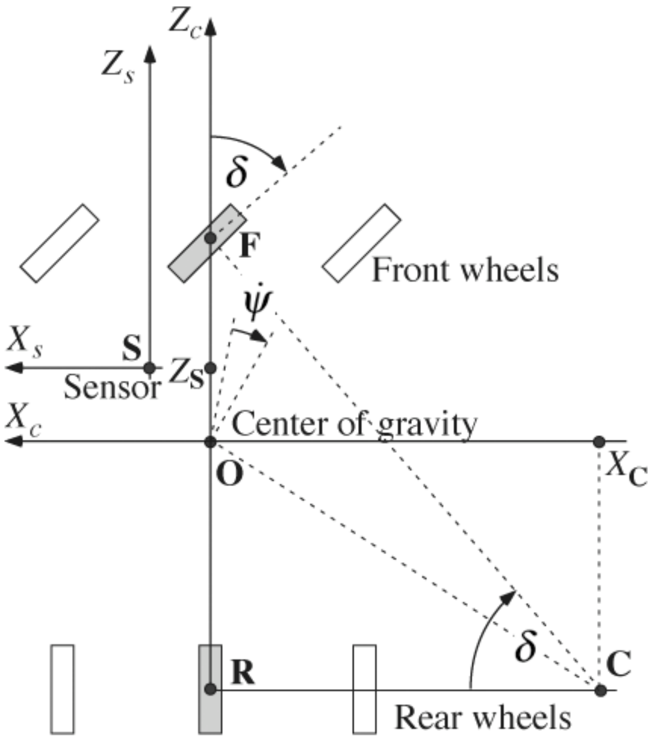
(4) Use left and right sequence for 3D direction + location

(NEXT) MSER region extractor (Matas et al., 2002) possibly more robust

.enpeda.. Project at The University of Auckland

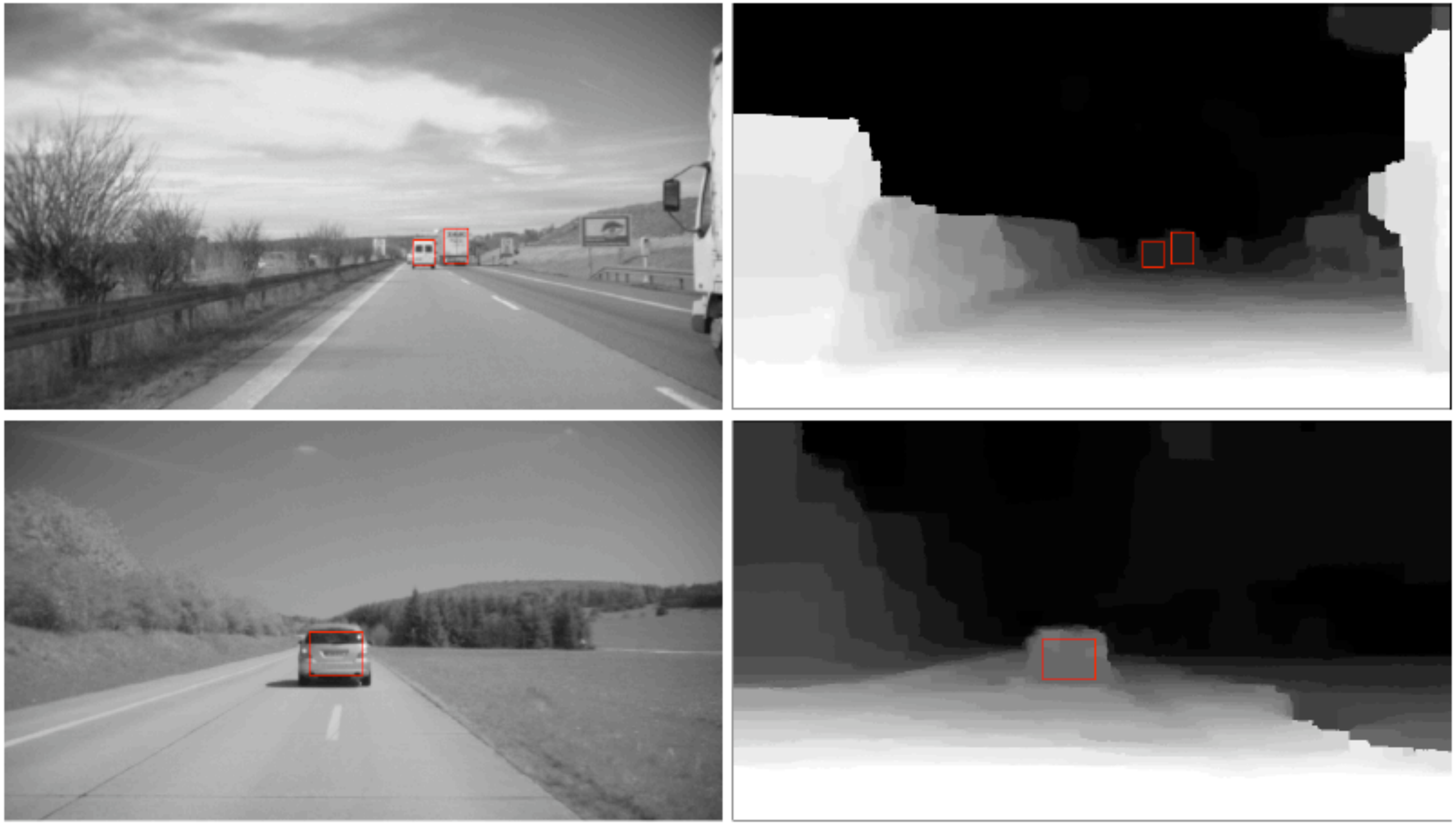
(**En**vironment **Pe**rception and **D**river **A**ssistance)

HAKA1

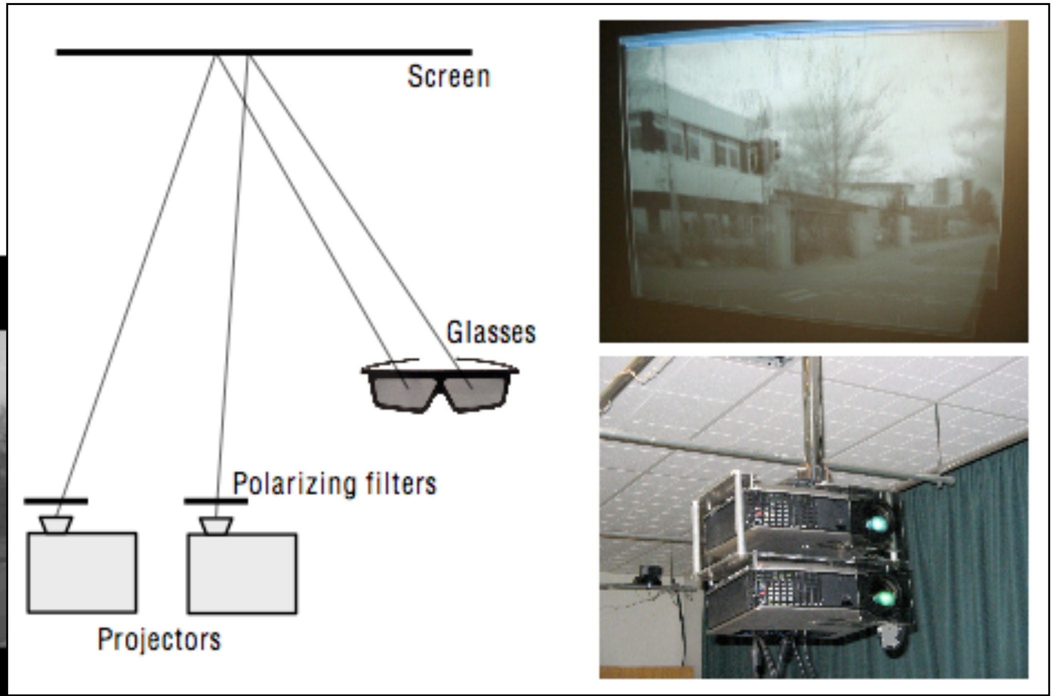


Various stereo cameras, GPS, …

Ground truth for
rectangular regions:
assume that those are parallel to the image plane

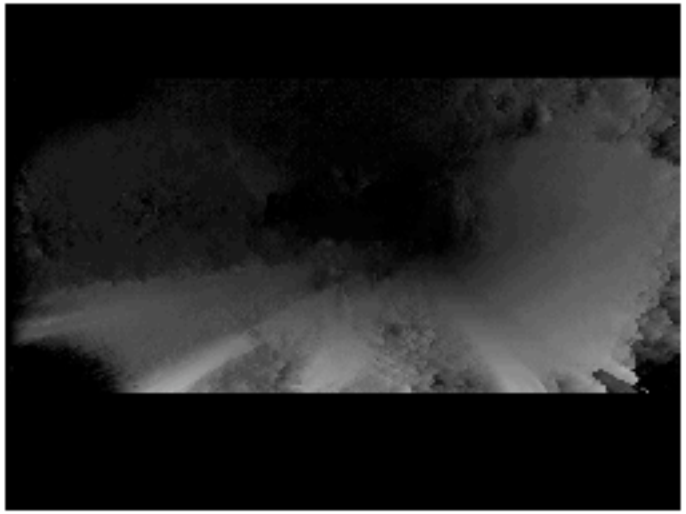


Performance evaluation: subjective, using a "Mini-IMAX"

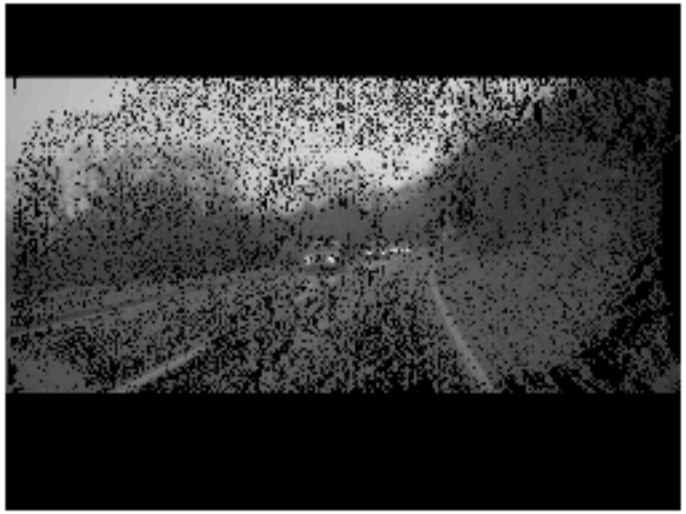


Original left image

Original right image



Disparity result



Generated right image

Warped right image and polarized light

with Zhifeng Liu 2008