Stereo and Motion Analysis
of Long Stereo Image Sequences
for Vision-Based Driver Assistance

Reinhard Klette[+]

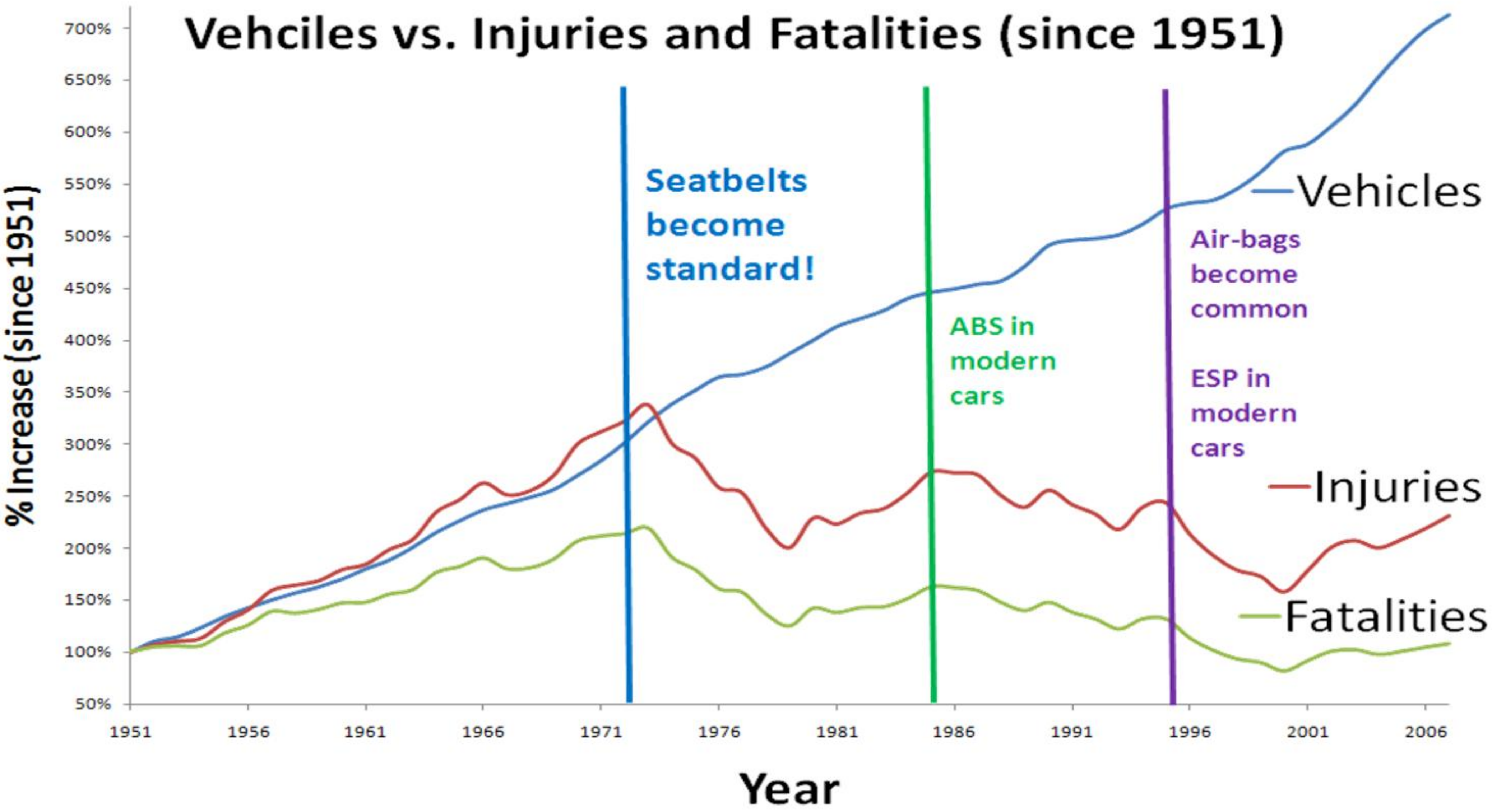
Sandino Morales[+] Tobi Vaudrey[+]

John Morris[+] Clemens Rabe[*] Ralf Haeusler[+]

[+] The University of Auckland, New Zealand

[*] Daimler AG, Sindelfingen, Germany

Accident statistics for New Zealand



Vision-based driver assistance systems (DAS)

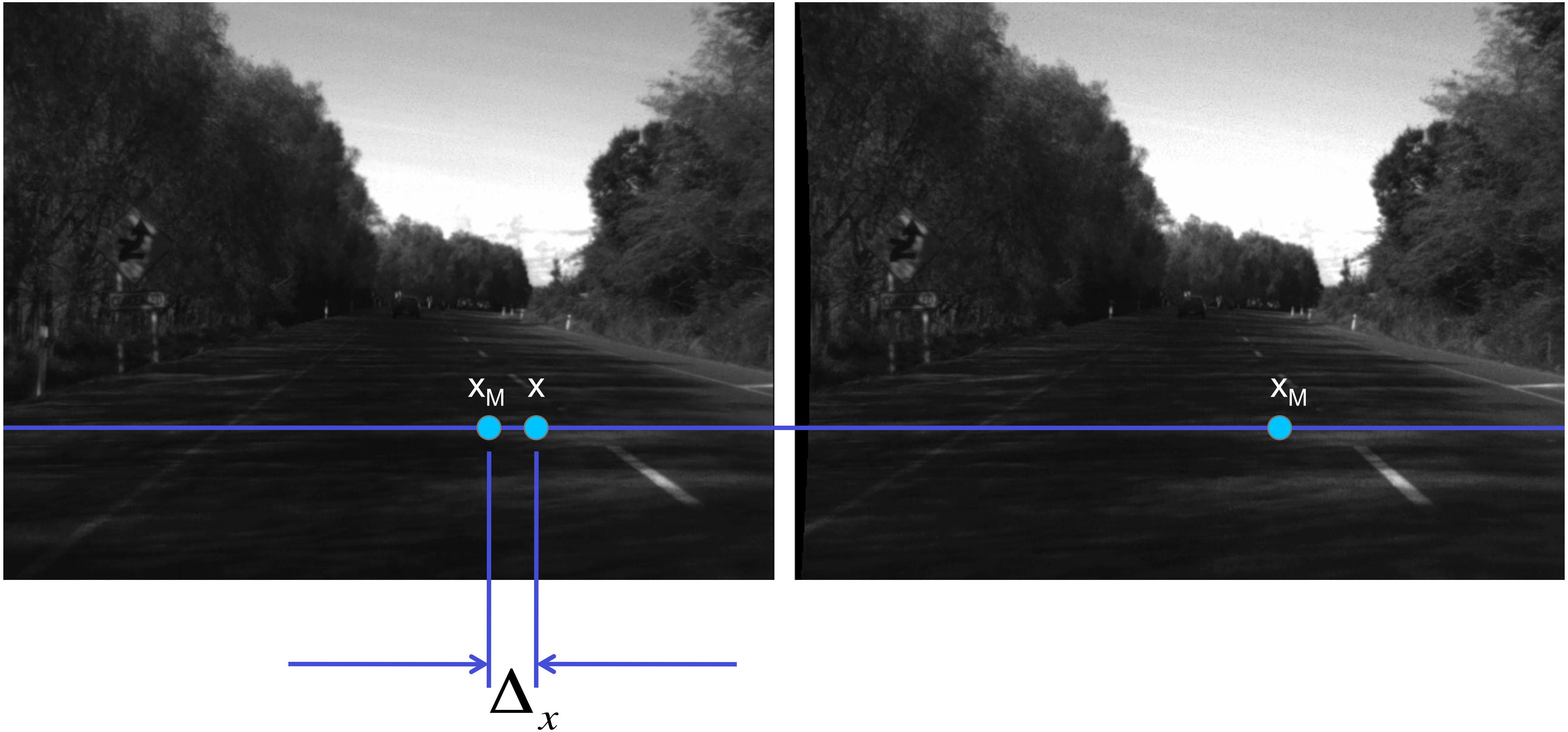
Rectified stereo frames, Auckland to Hamilton

Left camera

Right camera

Stero matching

is a 1D (along epipolar line) correspondence problem

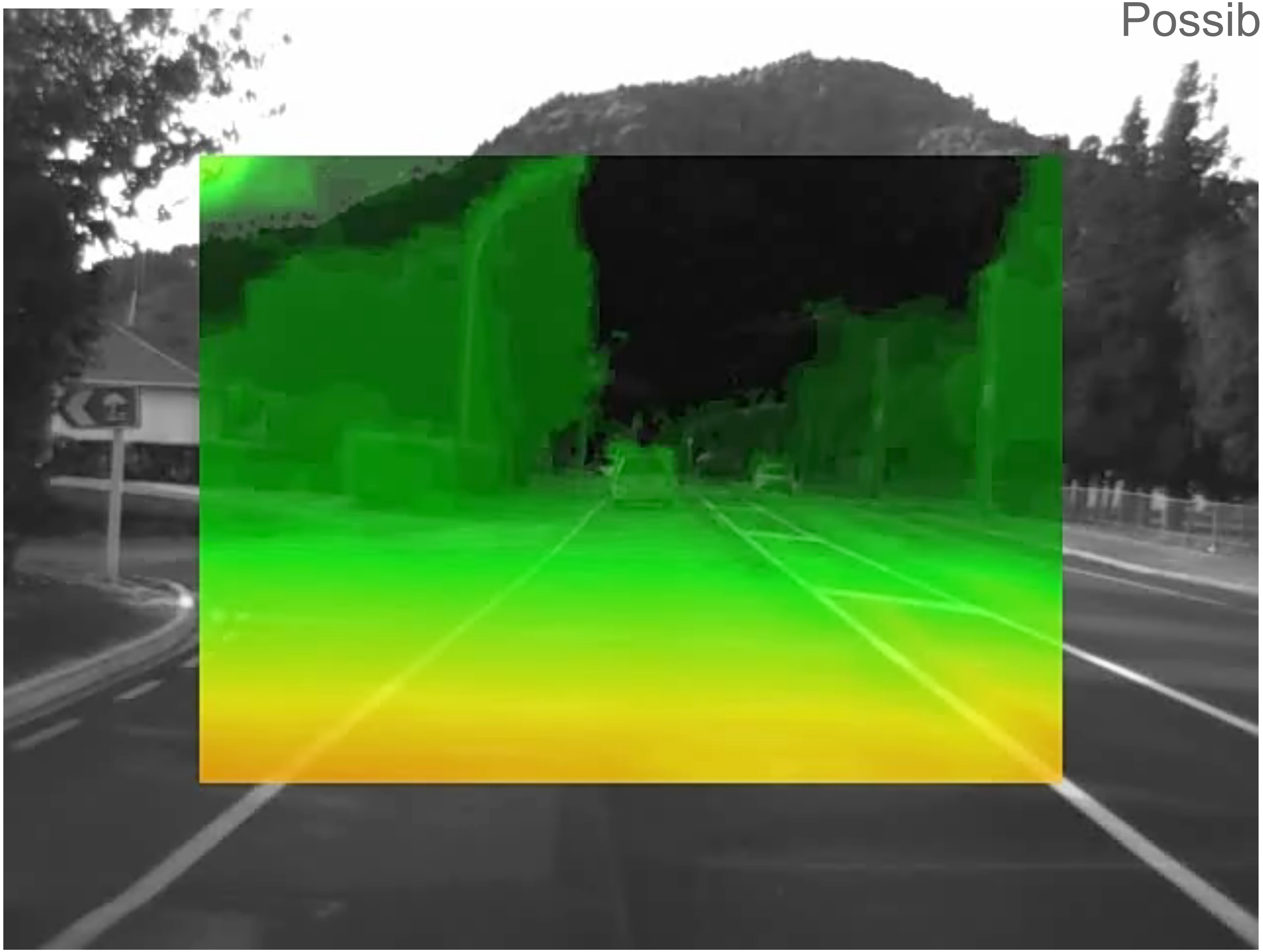


disparity - zero at infinity, finite range of disparities

Distance information

close ..... far

Possible key



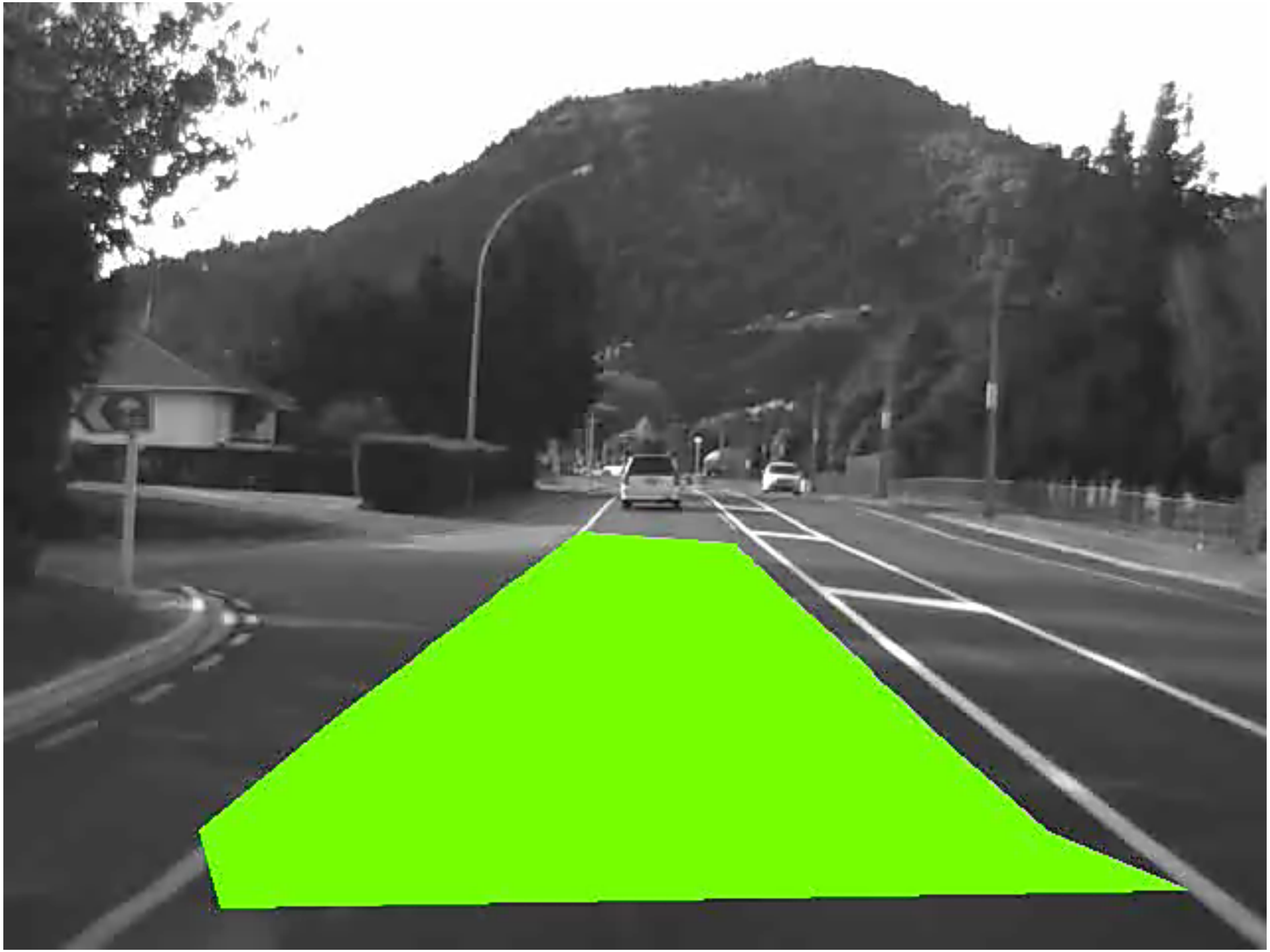
Common key (also in this talk): white ... gray ... black

The .*enpeda*.. Project

Environment Perception and Driver Assistance

Members of the *.enpeda..* project at Tamaki campus, The University of Auckland, June 09

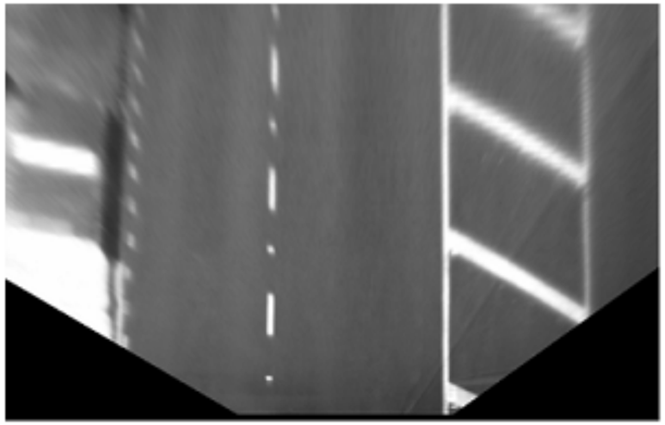
General Goal: predict - adapt - optimize for safety



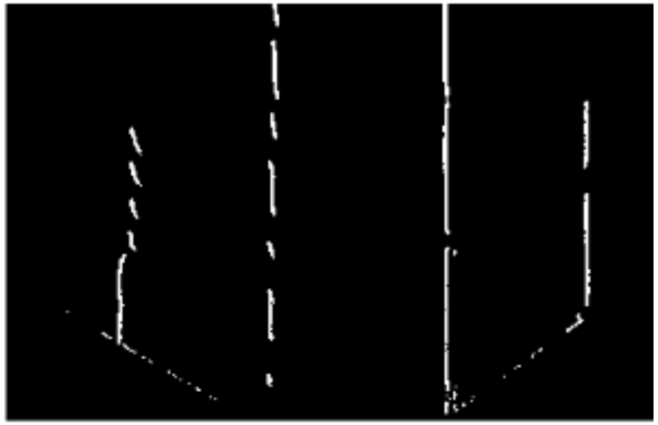
Corridor = predicted space the ego-vehicle will drive in the next few seconds

Workflow of lane detection (currently at 10 fps, 640x480)

Input image

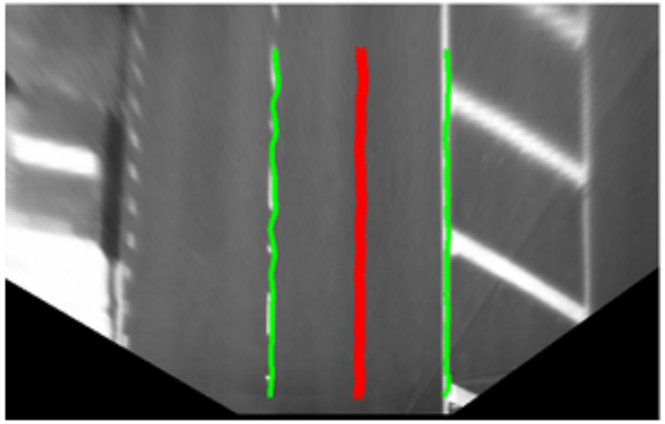


Bird's-eye View

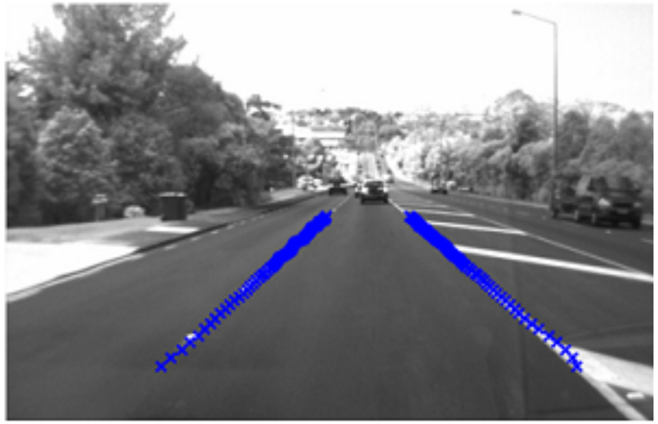


Edge detection

Row component of
Euclidean distance



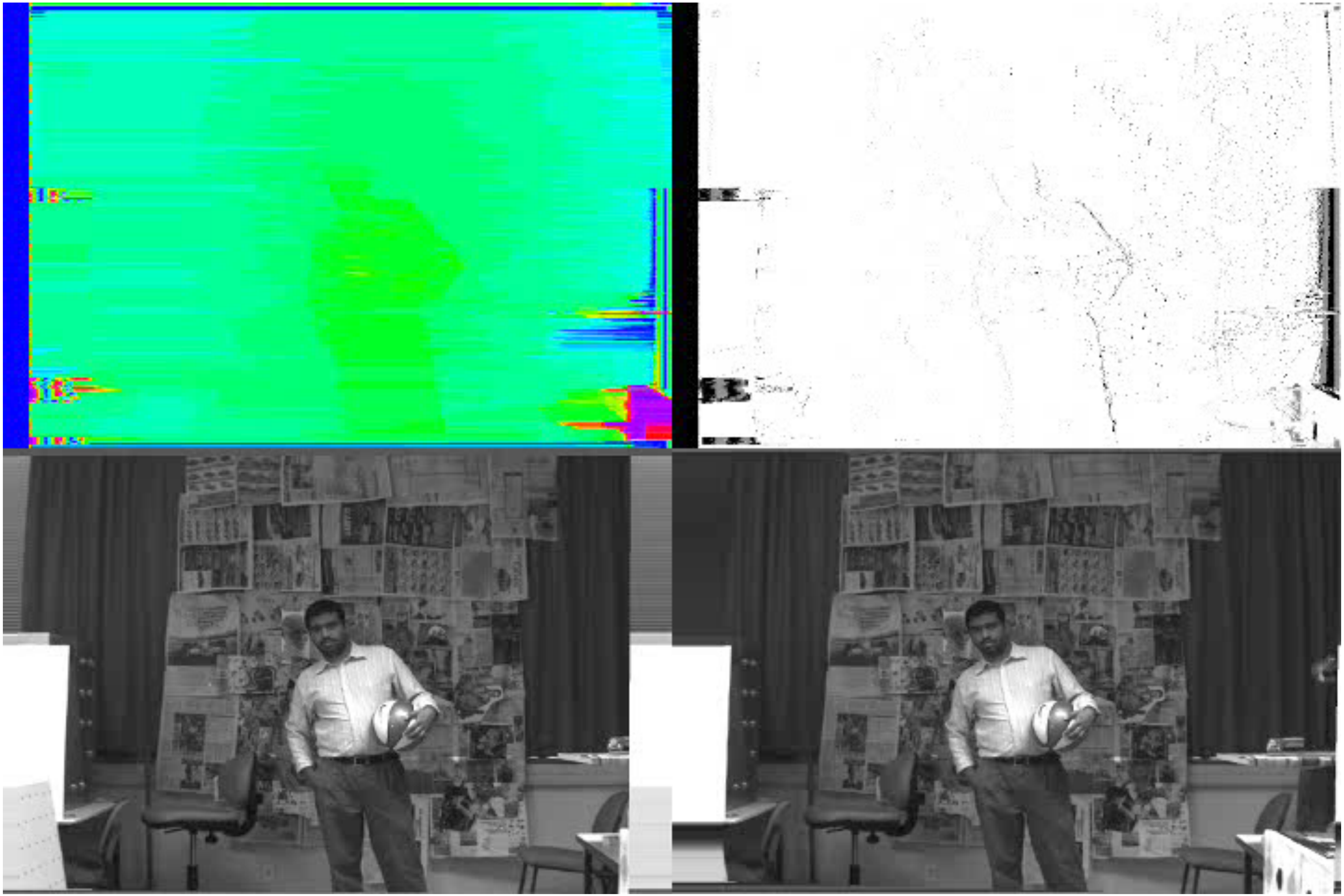
Lane detected in
bird's-eye view



Lane detected

Joint work with JiaTong University, Shanghai, China

Goal: all in real-time (here stereo analysis at 30 fps)



30 fps example on previous slide

(John Morris et al., looking for partners in the industry)

Symmetric dynamic programming stereo matching

The hardware (FPGA) can handle

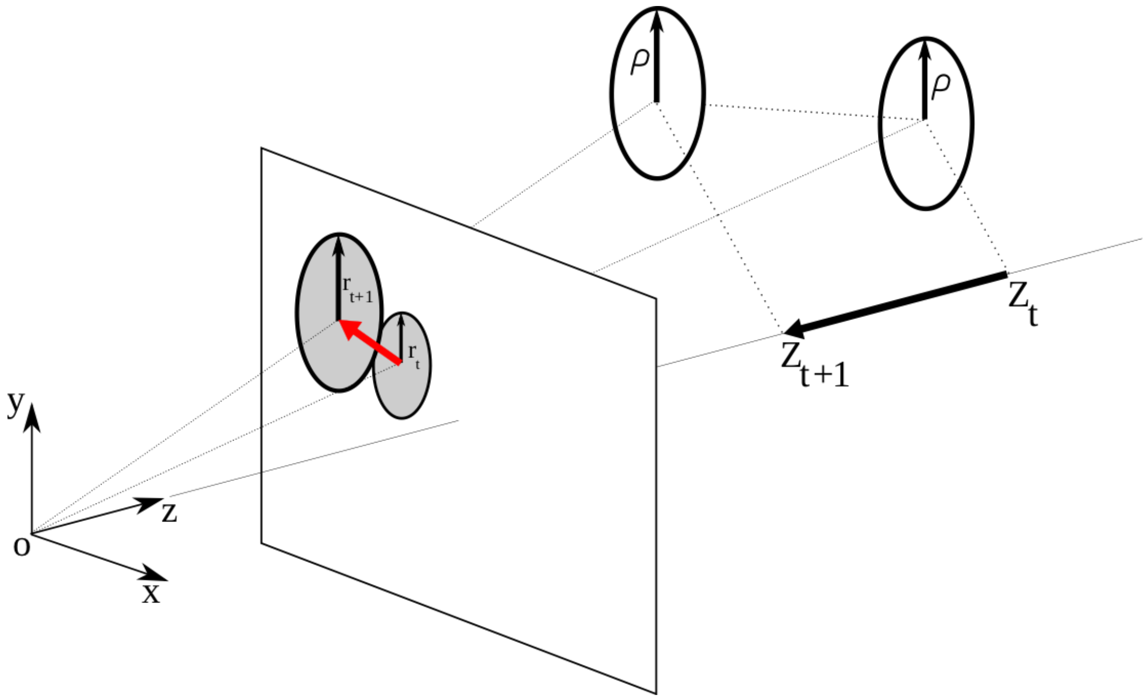
up to 1 Mpixel (1280x1024) frames
at 30 fps with
a disparity range of up to 100

Other time optimizations in .*enpeda*.. use CUDA or a playstation (joint work with Shandong University, Jinan, China).

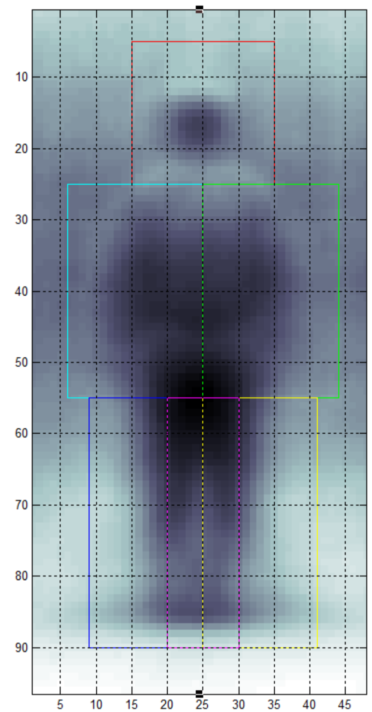
General goal:

from early vision (stereo, motion, ...) to a more advanced understanding of traffic events

Two examples of current work in *.enpeda..*



3D motion vector estimation in scale space
(with TU Cordoba, Argentina)



A "mean pedestrian" (of 10,000 pedestrians in the public Daimler data base)

HAKA1

High Awareness Kinematic Automobile no. 1
test vehicle in the .*enpeda*.. project

Sponsored by Mercedes-Benz New Zealand and Coutts North-Shore

2 x 1.3 MP 10 bit gray-value (fish-eye) cameras

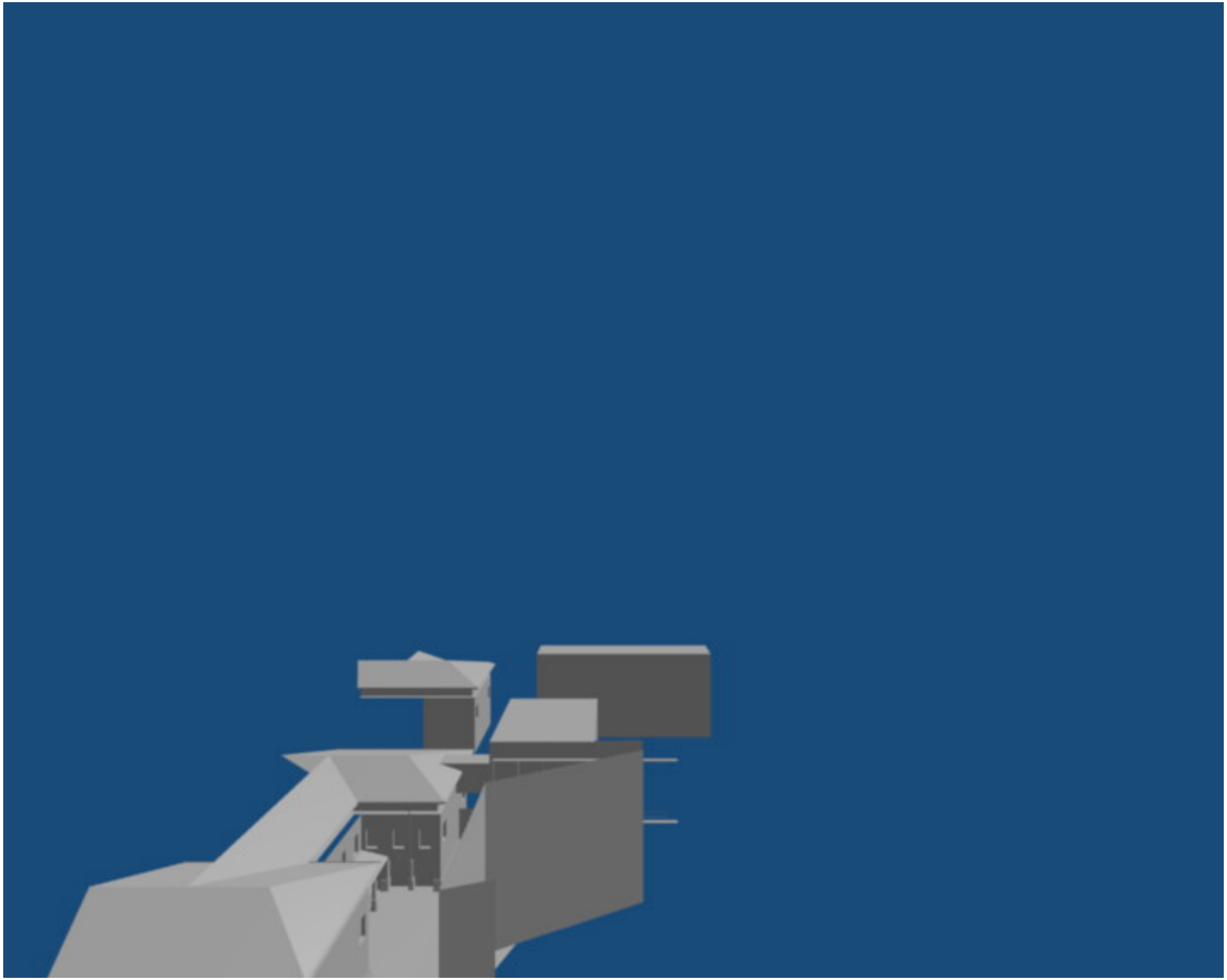
5 VGA (640 x 480) 10 bit gray-value cameras; **default:** three cameras

Currently: 28 students recording & analyzing sequences

Ground-truth ?

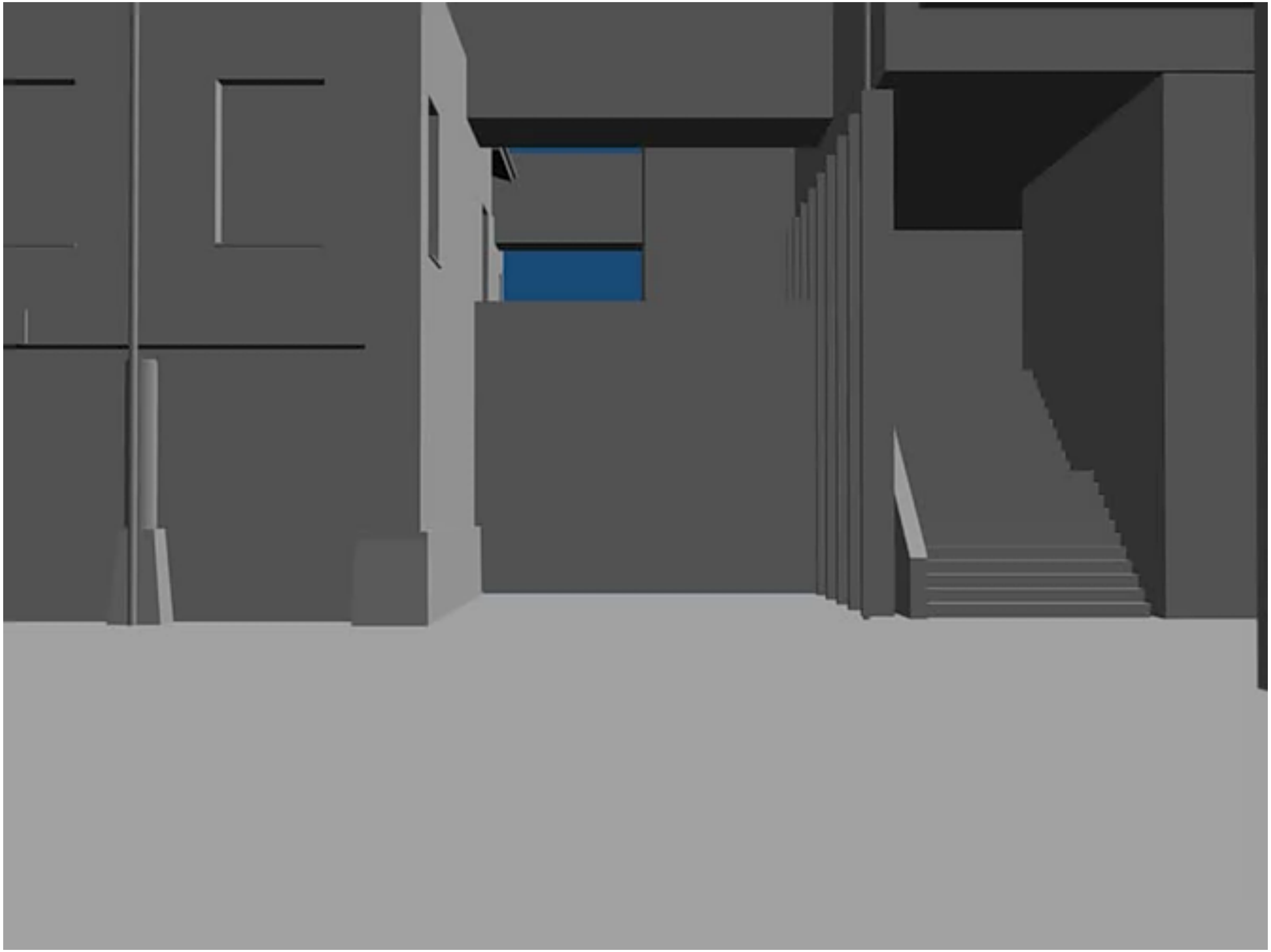
A “reasonable truth” we are able to provide - modulo measurement errors

3D model of a part of Tamaki campus

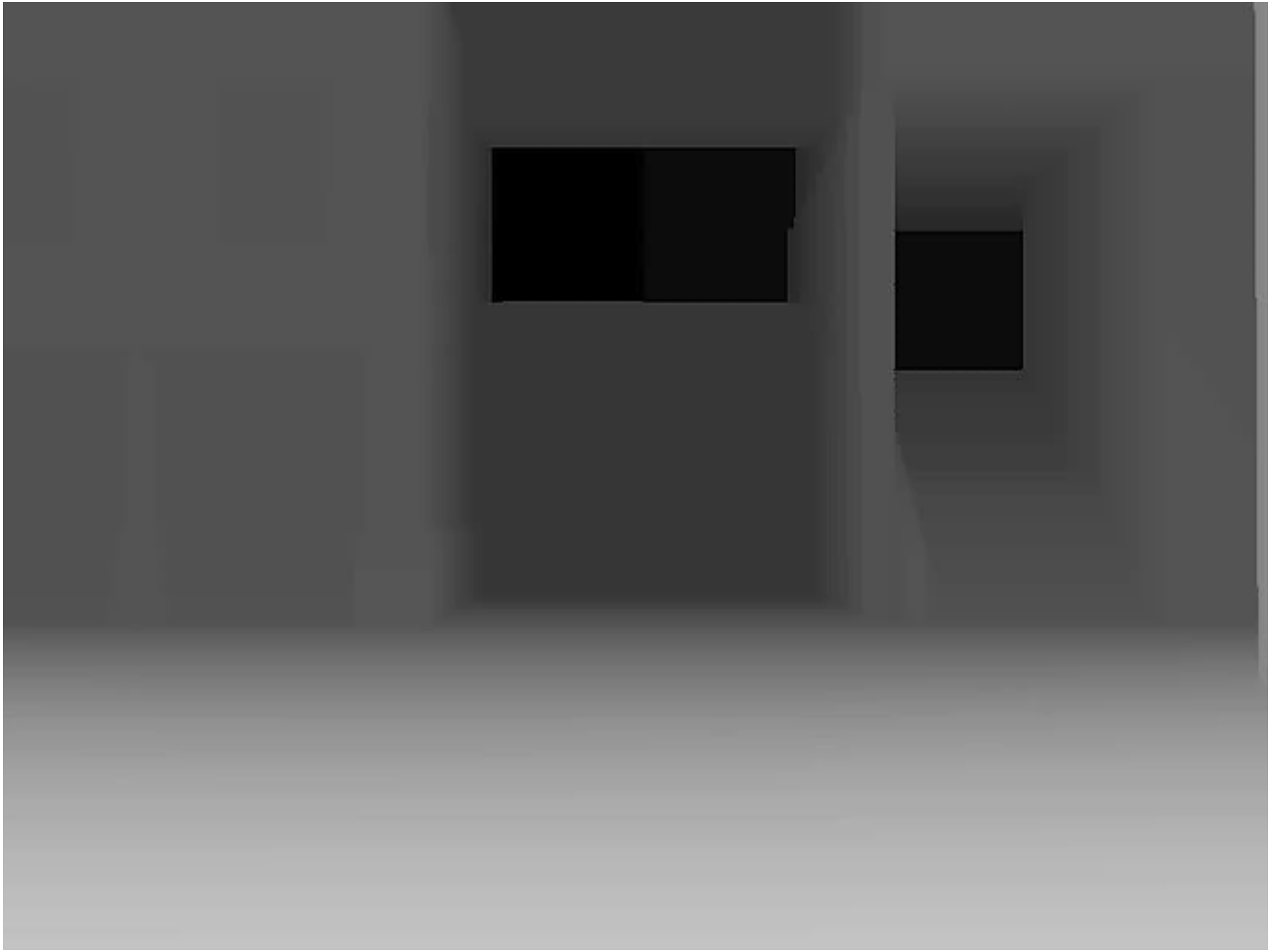


Rectified right camera sequence recorded with HAKA1

Corresponding sequence while driving into the 3D model



Ground-truth sequence (depth)



Actually, "corresponding sequence" is an unsolved issue here:

Evaluation of stereo & motion techniques would require an exact trajectory of the car, and this is not yet available to us.

Thus:

Search for alternative ways for evaluating (early vision for) vision-based DAS

Prediction error analysis

for stereo triples [R. Szeliski, 1999]

calculate disparities for base and match sequence

warp base intensities into **third camera view** T,
based on calculated disparities

compare those **virtual images** V with third images
(i.e., images of the third camera)
using the **normalized cross-correlation measure**

$$N(t) = \frac{1}{|\Omega_t|} \sum_{p \in \Omega_t} \frac{\left[T_t(p) - \mu_{T,t}\right]\left[V_t(p) - \mu_{V,t}\right]}{\sigma_{T,t}\sigma_{V,t}}$$

The sum is for available (e.g., non-occluded) pixels only.

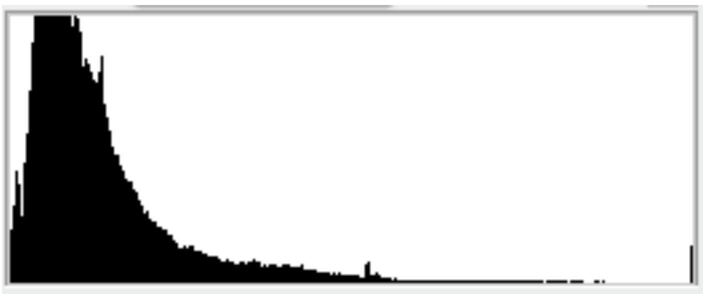
Root-mean squared error:

$$R(t) = \cancel{\frac{1}{|\Omega_t|} \sum_{p \in \Omega_t} \left(T_t(p) - V_t(p)\right)^2}$$

We found this measure misleading for real-world sequences.

Example: third and left view

Mean: 30.9
Standard deviation: 29.4



Mean: 13.2
Standard deviation: 18.7

Selected stereo matching algorithms

DP – dynamic programming   Ohta/Kanade 1985

DPt - temporal propagation (20% from same row in frame *t-1*)

BP – belief propagation   Felzenszwalb/Huttenlocher 2002

GC – graph cut   Boykov/Veksler/Zabih 2001

SGM – semi-global matching   Hirschmüller 2005

cost functions: MI (mutual information)
BT (Birchfield/Tomasi)

Default camera configuration in HAKA1

Third | Left (base) | Right (match)

40 cm left of left camera

30 cm apart from each other

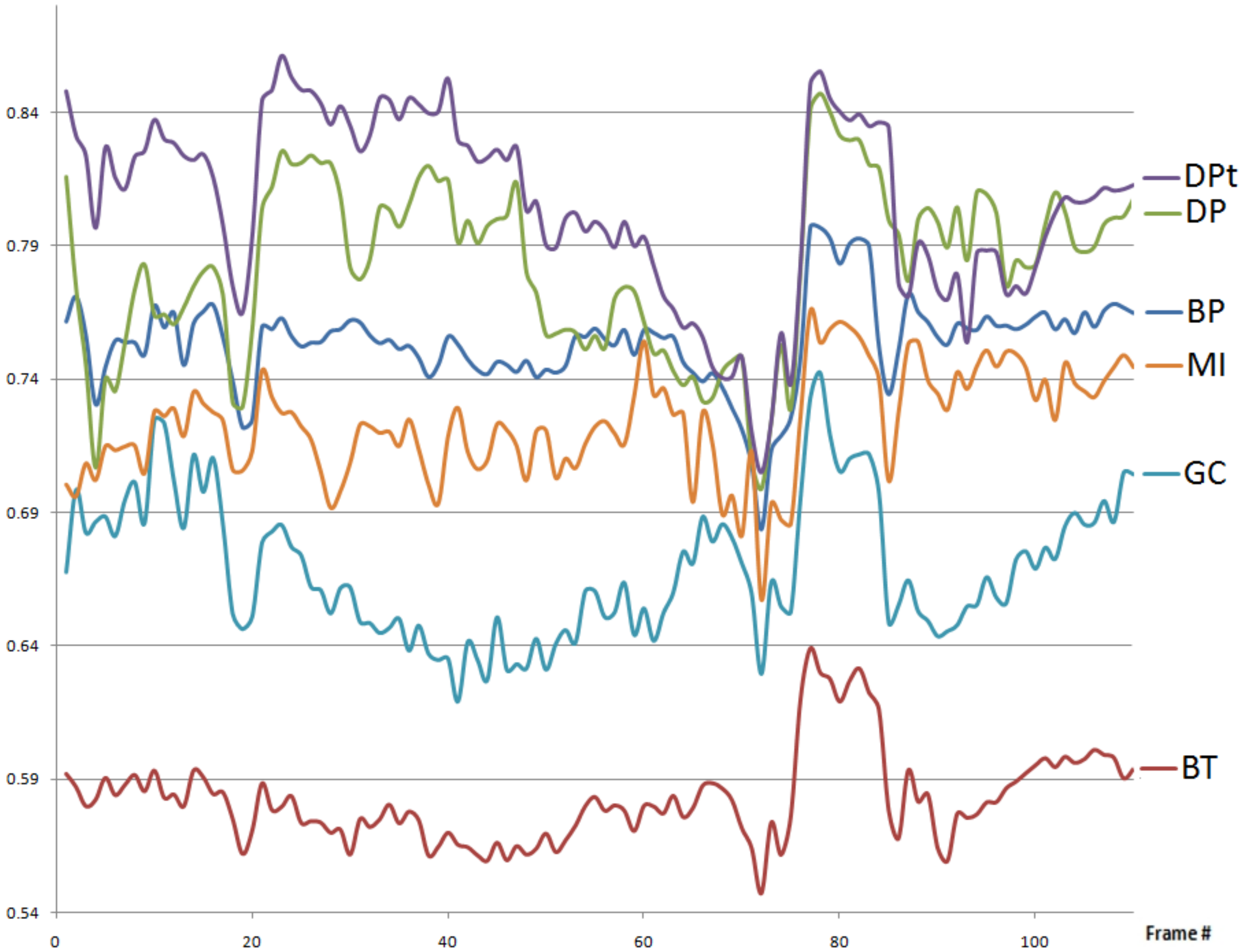
All three cameras on one bar behind windscreen

Left and right camera: rectified for stereo matching

Virtual view *V*
using DP
(Dynamic Programming)

Third view *T*
(reflections cannot be predicted
but are constant when comparing
different matching techniques)

120 NCC values for each method for this stereo sequence



Assume runners A and B on a 10,000 m distance track

Current world record: 26:17.53 min

Select 10 m out of those 10,000 m

Mean speed of A on those 10 m: 21.34 82 19 km/h

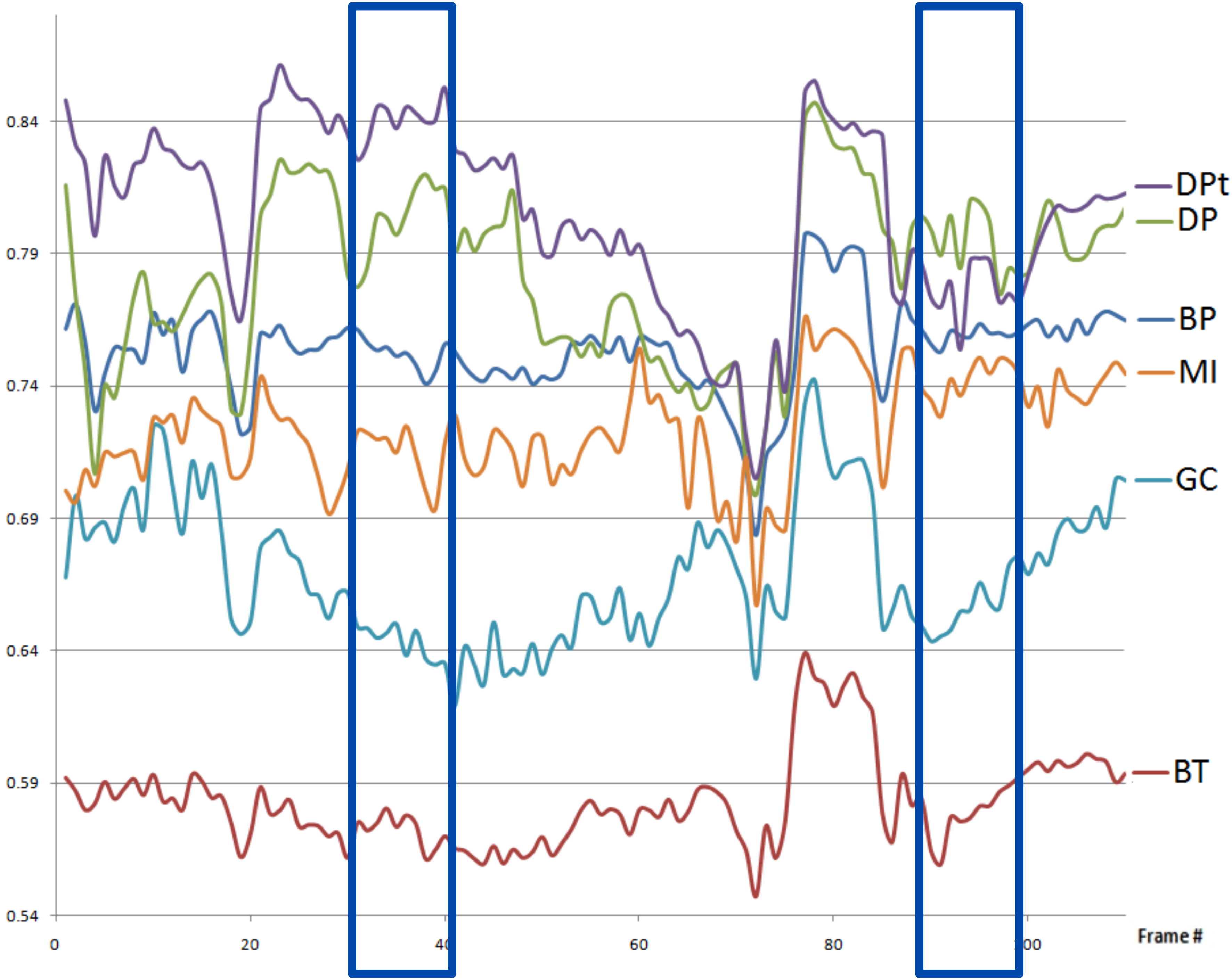
Mean speed of B on those 10 m: 21.34 77 78 km/h

B is thus better than A ?

Certainly not; this is in the range of measurement noise.
What counts:

final result (i.e., mean),
the steadiness (i.e., variance),
the robustness (e.g., other situations)

Note the changes in relative differences along the sequence



Situations

5 Examples

A situation (or scenario) is

a combination of circumstances for some sequence of recorded frames.

Situation 1: default driving conditions

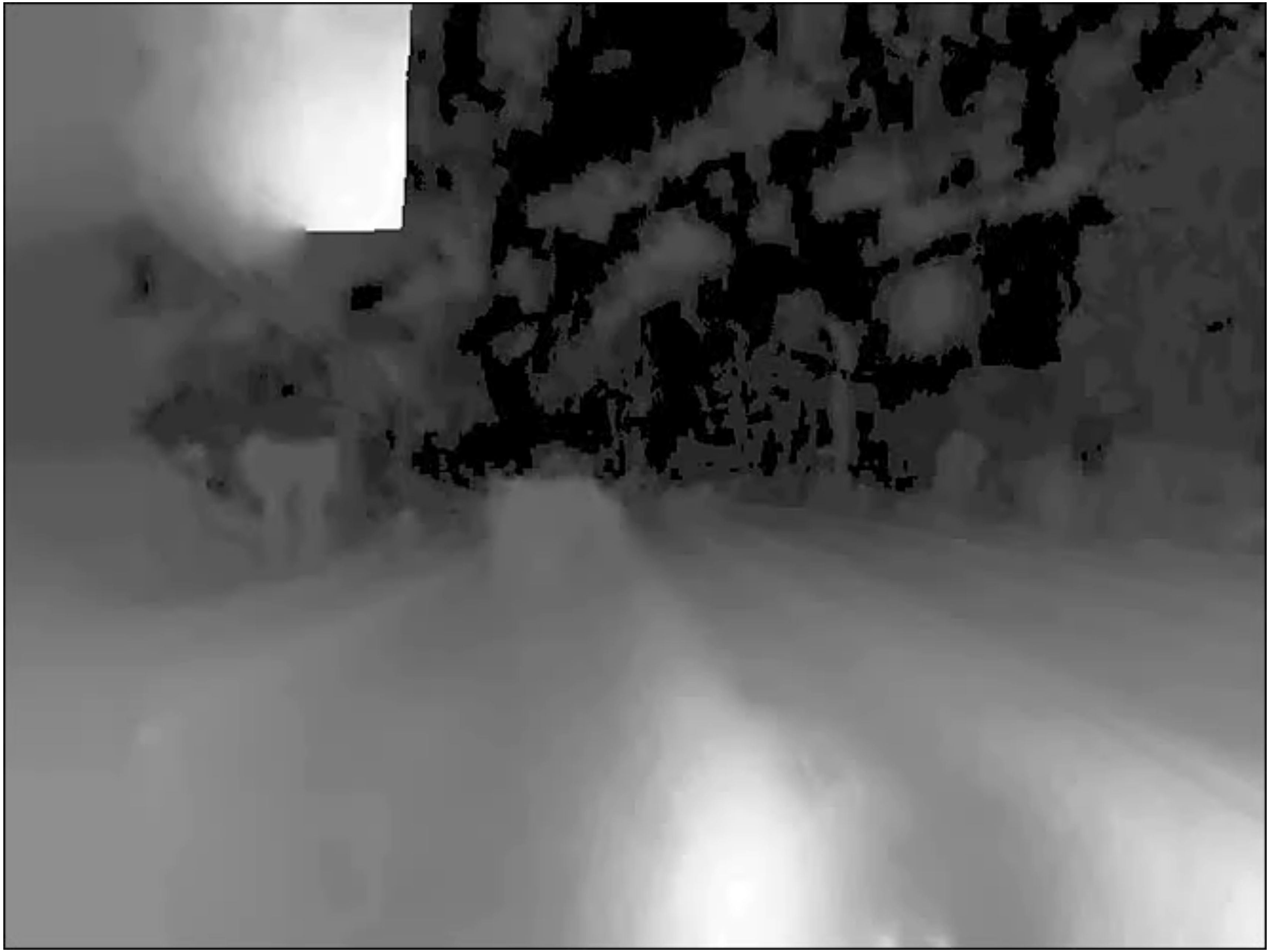
driving Auckland to Hamilton under “normal” conditions

Left Camera

Right Camera

Situation 1: default driving conditions

BP



Situation 2: close objects

stopping at a road construction site in Huntley

Left Camera

Right Camera

Situation 2: close objects

GC

Situation 3: inner-city at night

driving towards Mt. Wellington, Auckland

Left Camera

Right Camera

Situation 3: inner-city at night

SGM MI

Situation 4: brightness differences

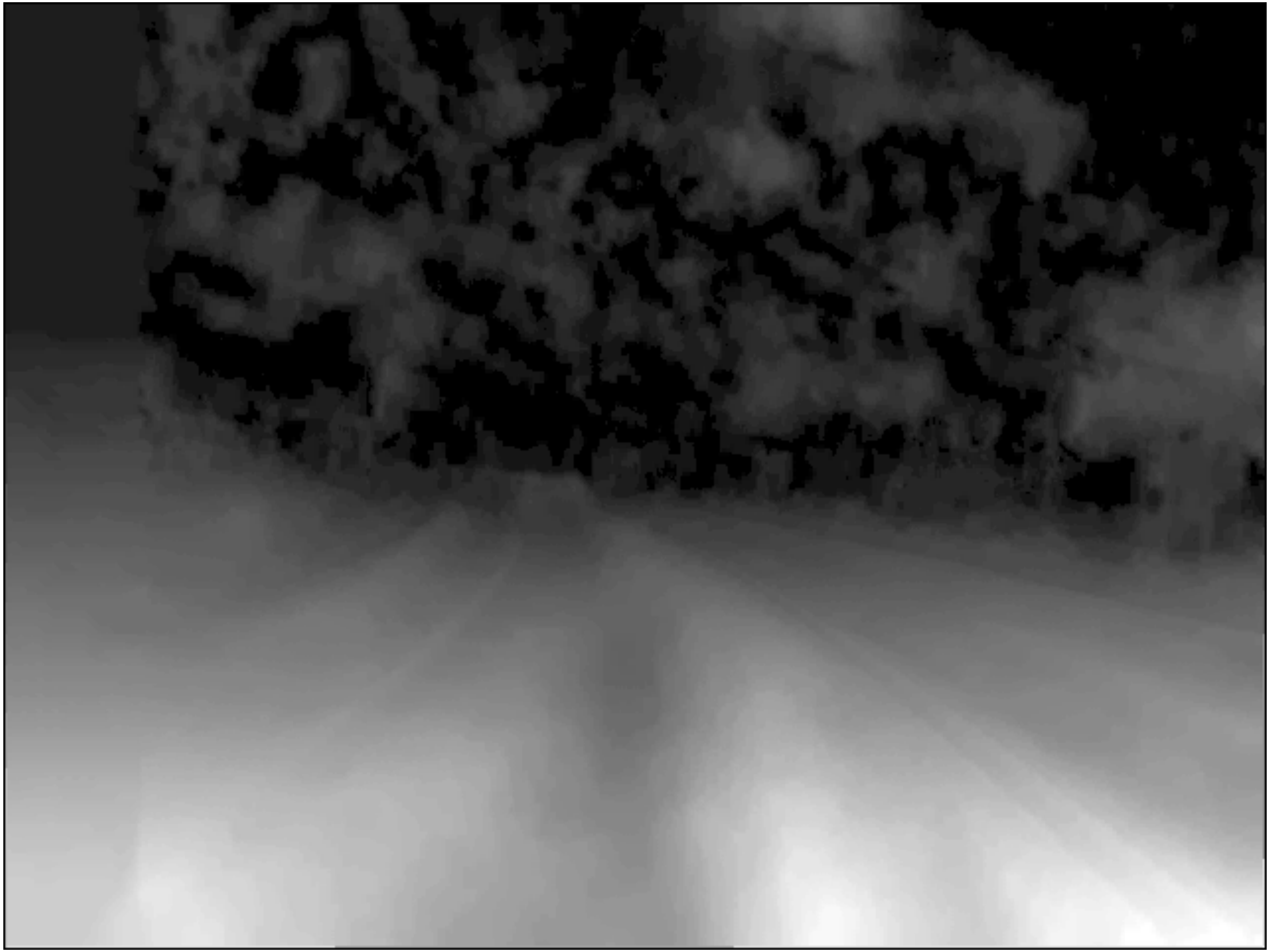
driving on a main road, Auckland to Hamilton

Left Camera

Right Camera

Situation 4: brightness differences

BP



Situation 5: illumination artifacts

driving through a sparsely forested area (Auckland)

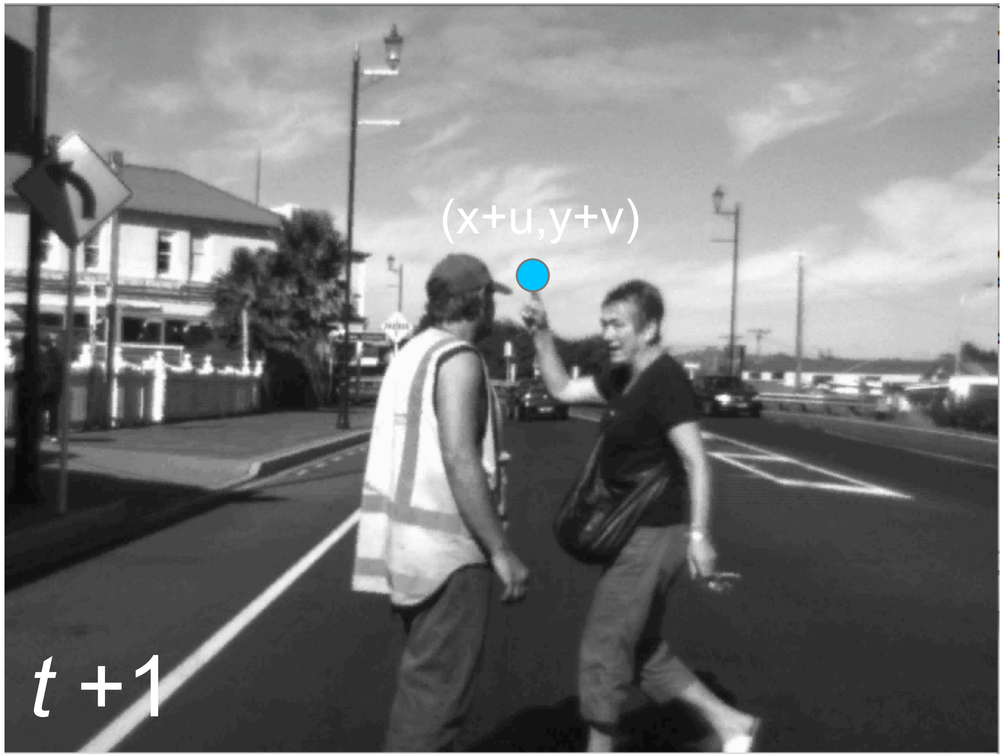
mean intensity: 84

mean intensity: 89

Optic flow

Motion analysis

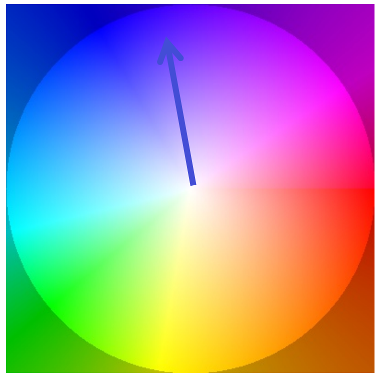
is a 2D (in image plane) correspondence problem



at 25 Hz

(u,v)

Color key



optic flow – aims at subpixel accuracy

Recording with only 25 Hz is still insufficient for using alternating frames for prediction error analysis.

Selected optic flow algorithms

BBPW – accurate optic flow from warping

Brox, Bruhn, Papenberg and Weickert 2004

HS – pyramid Horn/Schunck algorithm

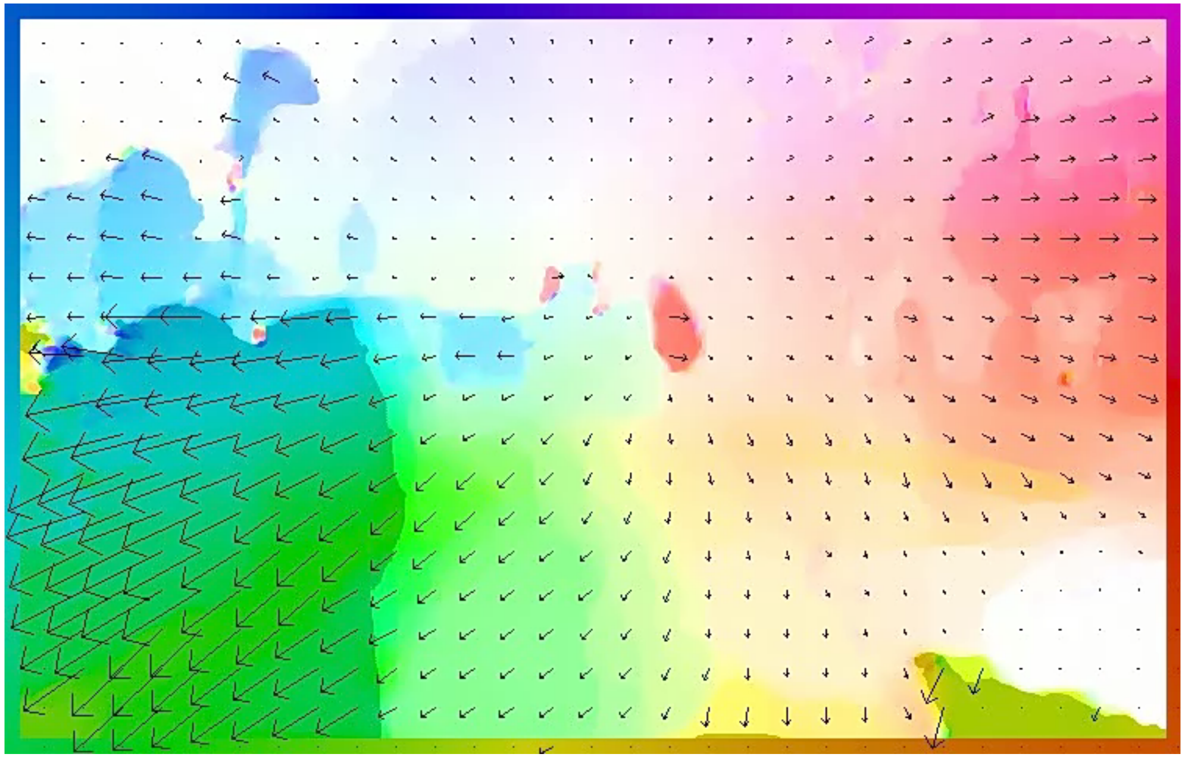
TV-L^1 – duality-based optic flow

Zach, Pock and Bischof 2007

TV L^1 on 10-bit data

Some early interaction between optic flow techniques (often TV) and stereo matching, e.g.

[N. Slesareva, A. Bruhn, J. Weickert DAGM 2005]

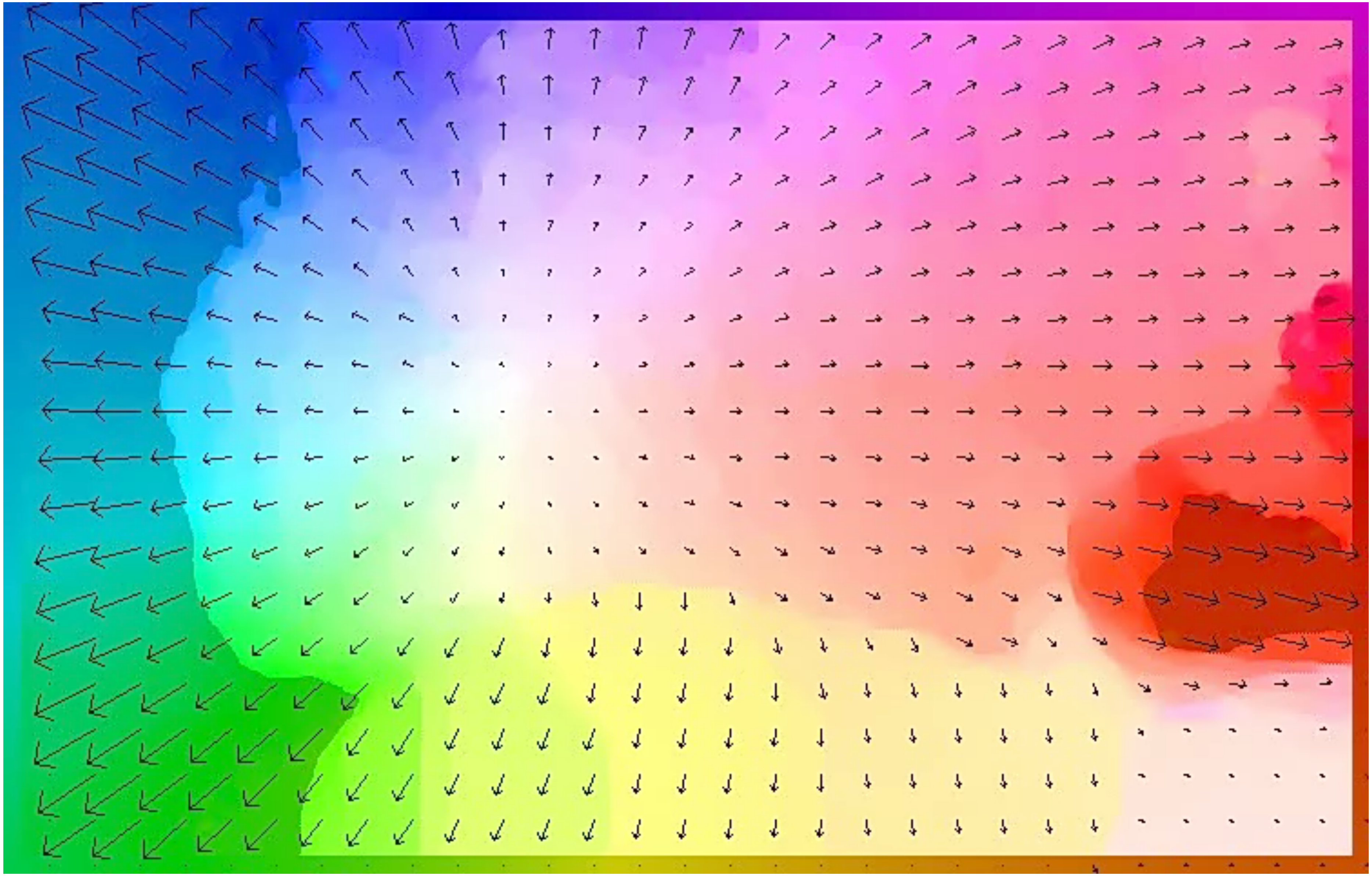


Situation 5: illumination artifacts

driving through forested areas (Waitekere, Auckland)

Situation 5: illumination artifacts

TV-L^1 optic flow (10 bit data)



Rendered Sequences

not yet photo-realistic, not yet physics-realistic

Test of correspondence algorithms on rendered or engineered sequences (with ground truth) is very useful for

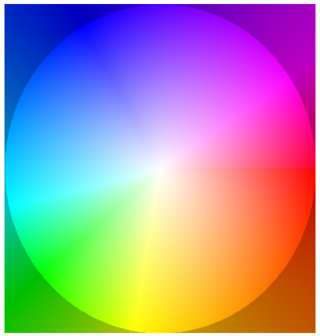
testing particular situations

(esp. for optic flow where we still are missing an evaluation scheme on real-world sequences) but (at least so far) insufficient (or misleading) for any ranking of algorithms for vision-based driver assistance

Rendered sequences on www.mi.auckland.ac.nz/EISATS

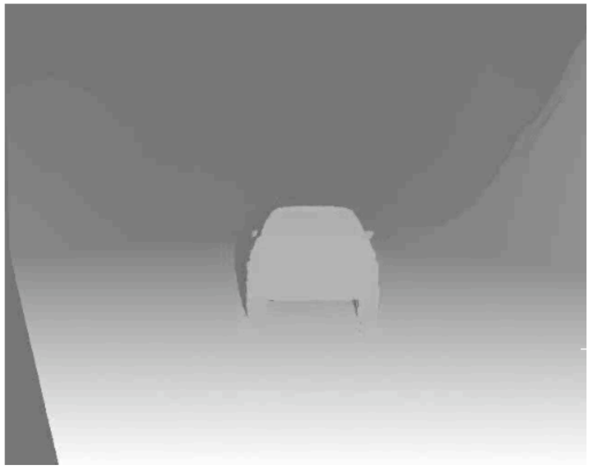
In gray values (left view)

In color (right view)



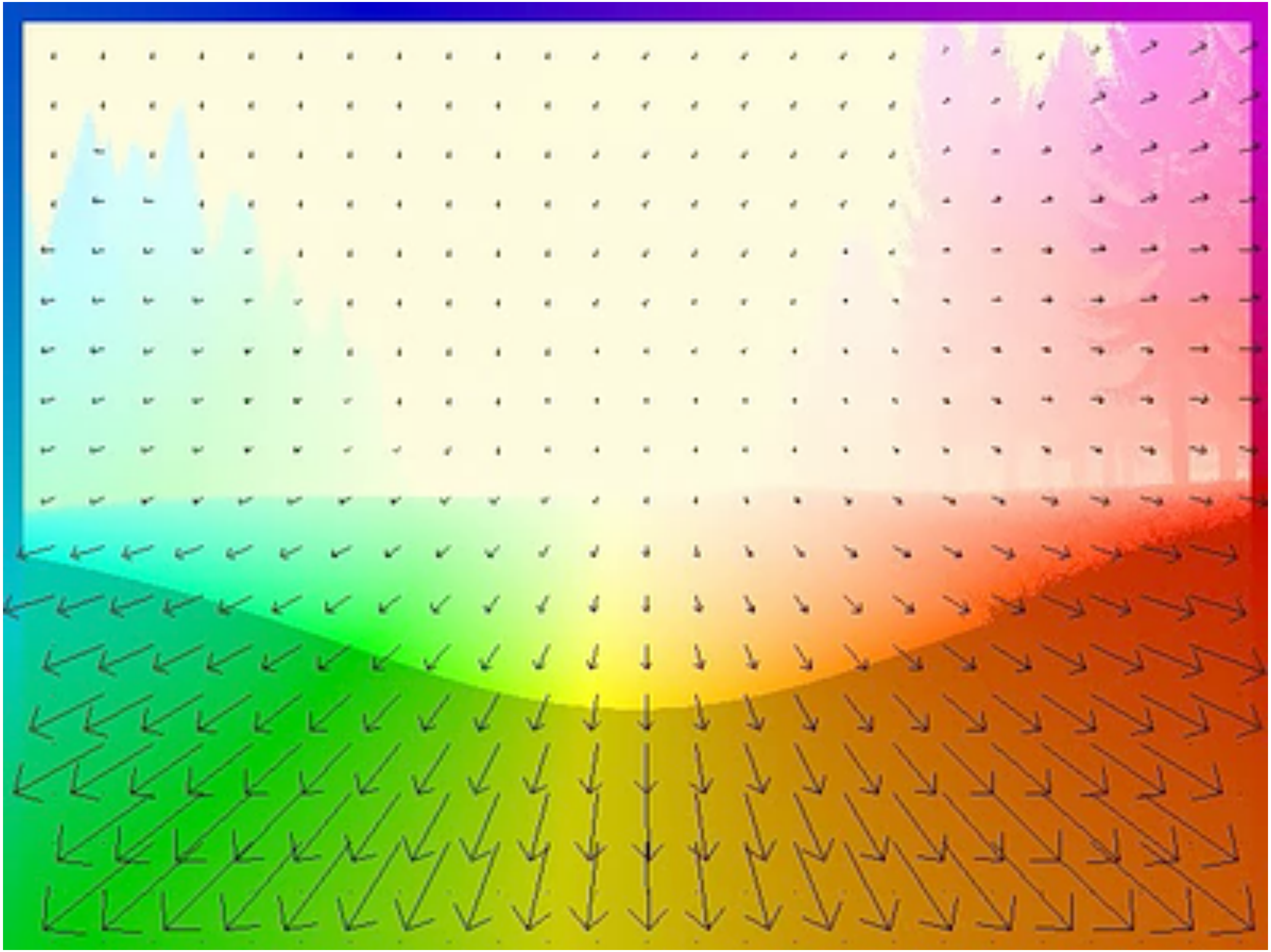
Flow key

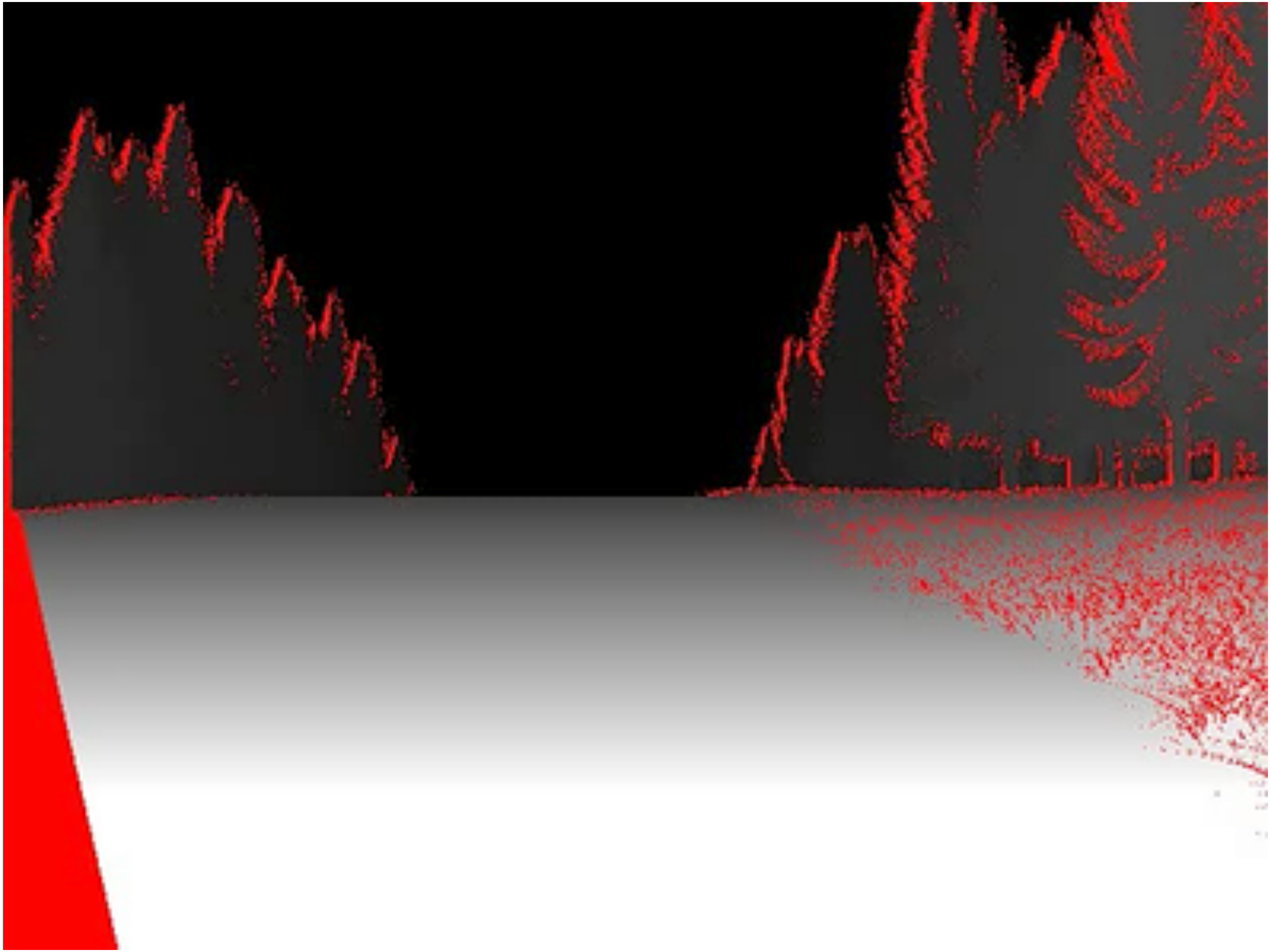
GT optical flow



GT depth map

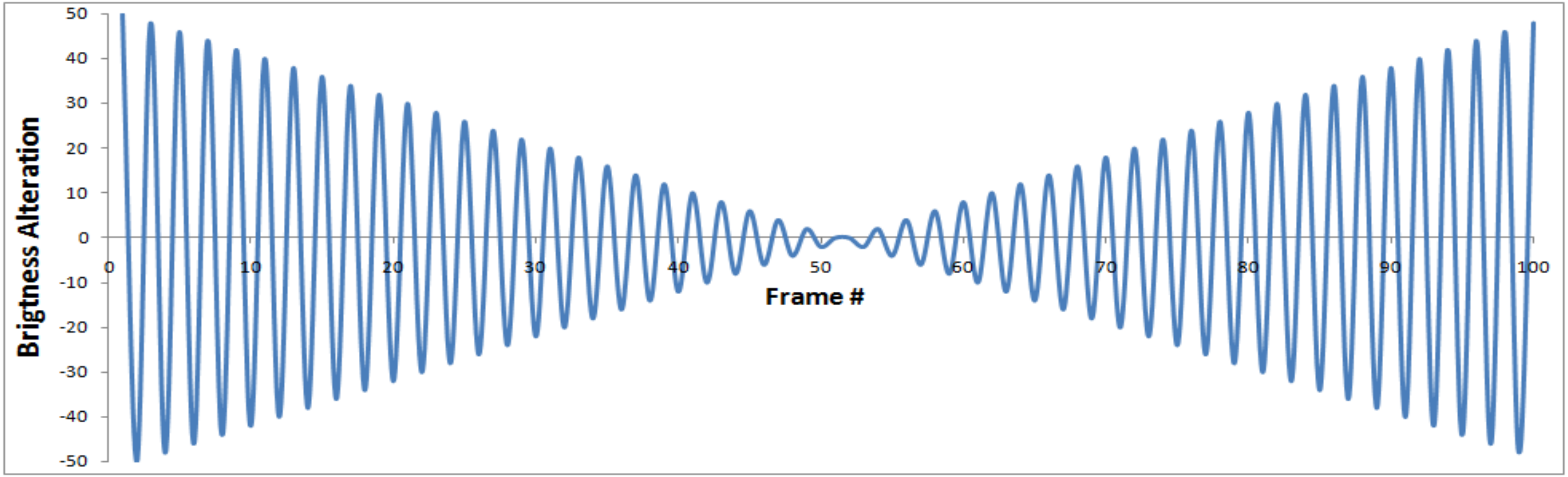
Depth key





Simulation of Situation 5 (illumination artifacts)

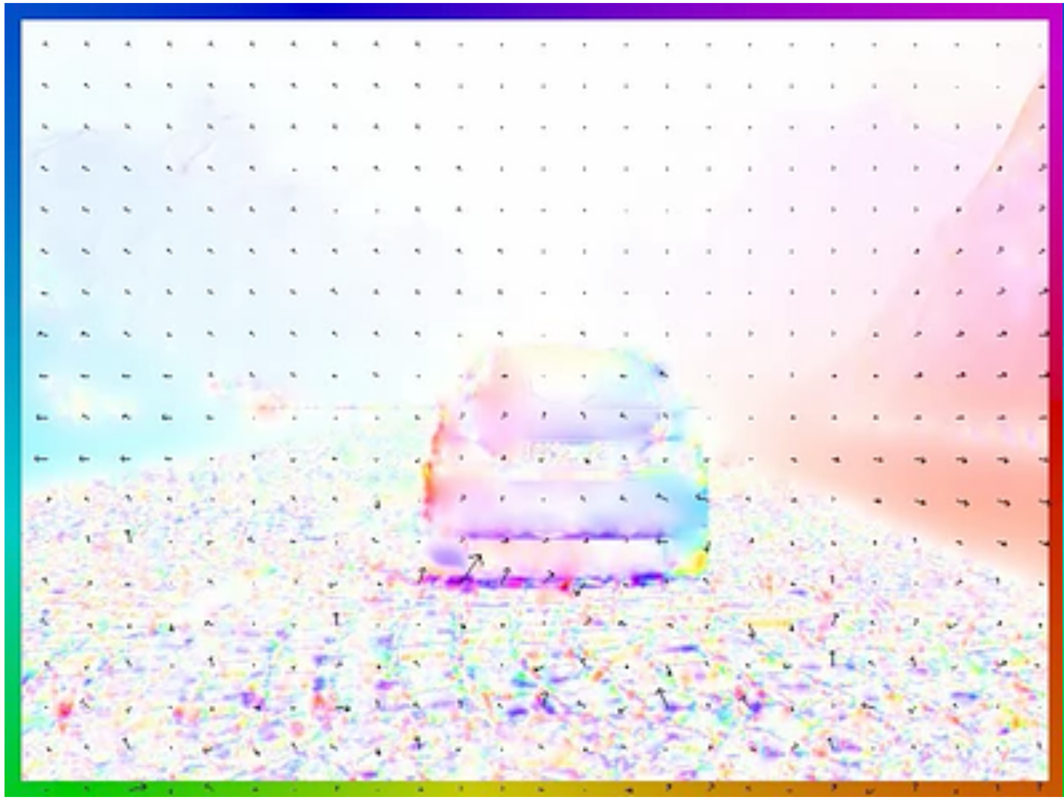
for motion analysis



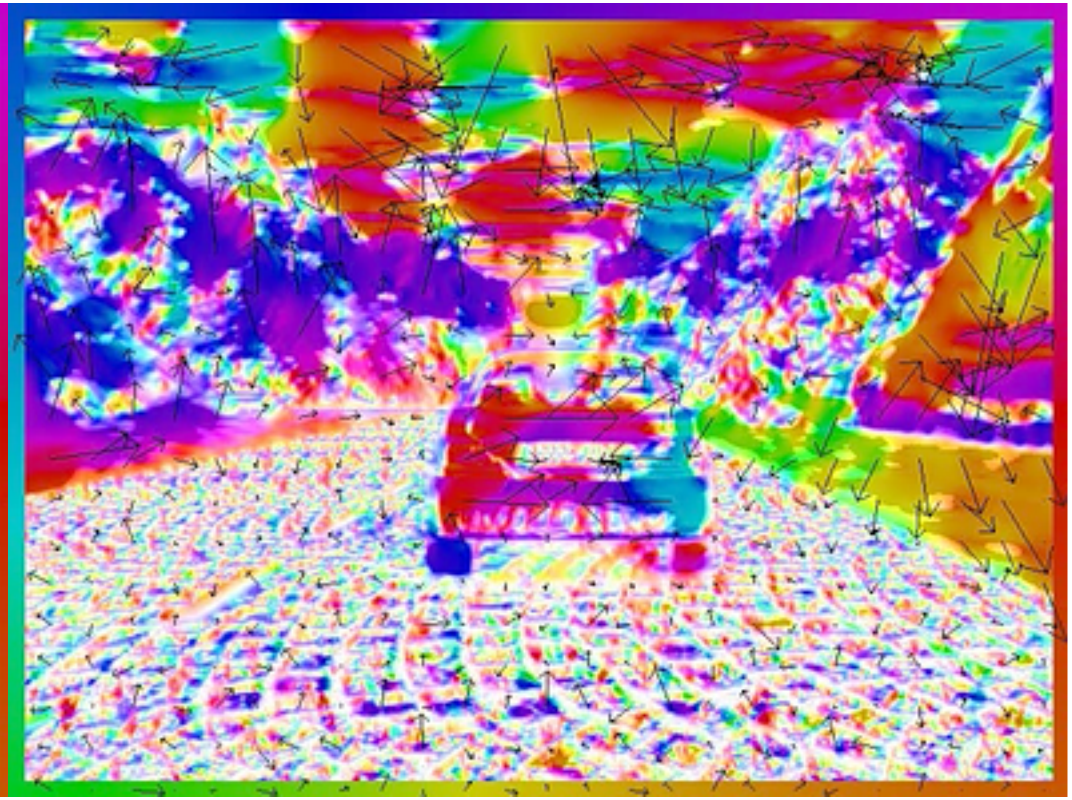
Brightness altered EISATS Sequence #1

HS results (original & brightness altered Sequence #1)

HS on original sequence

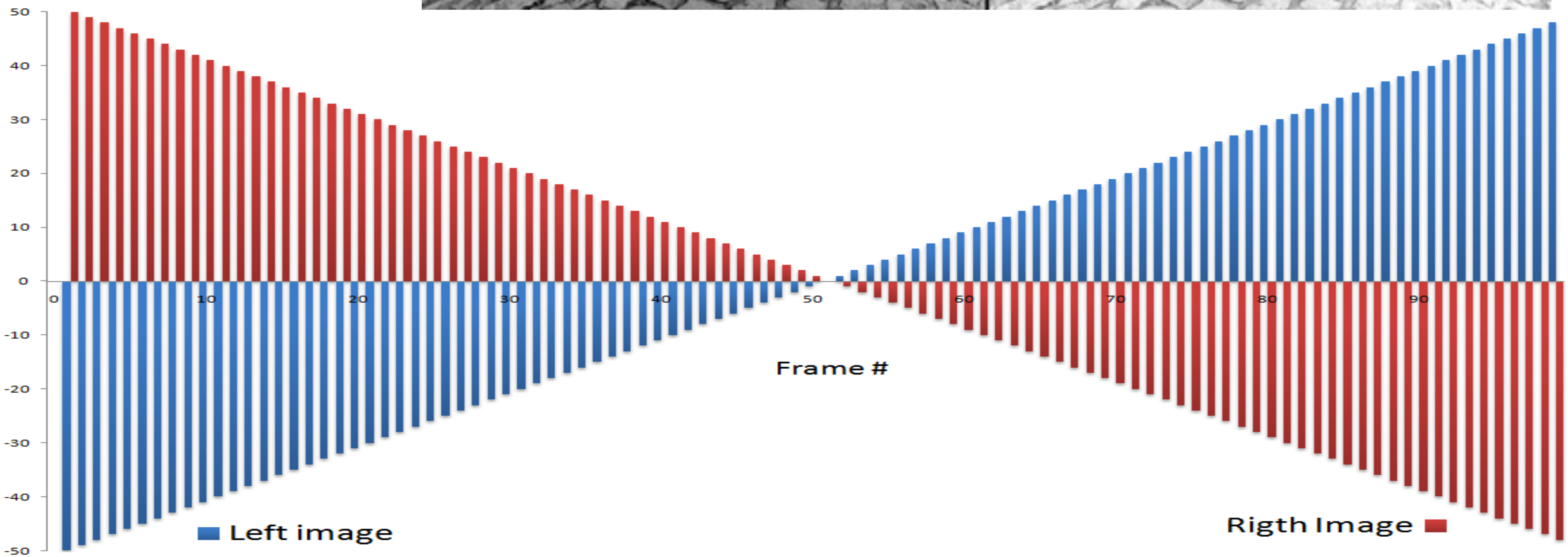


HS on brightness altered sequence



Simulation of Situation 5 (illumination artifacts) for stereo analysis

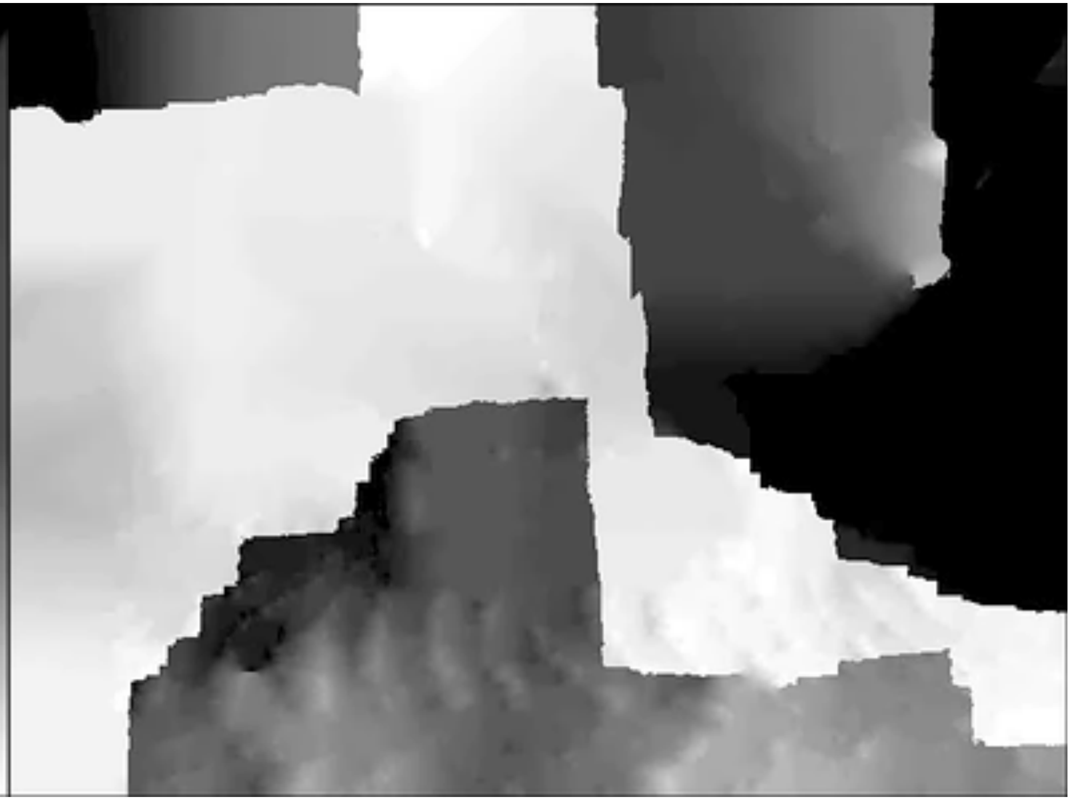
Brightness altered EISATS stereo Sequence #1



BP results (original & brightness altered Sequence #1)

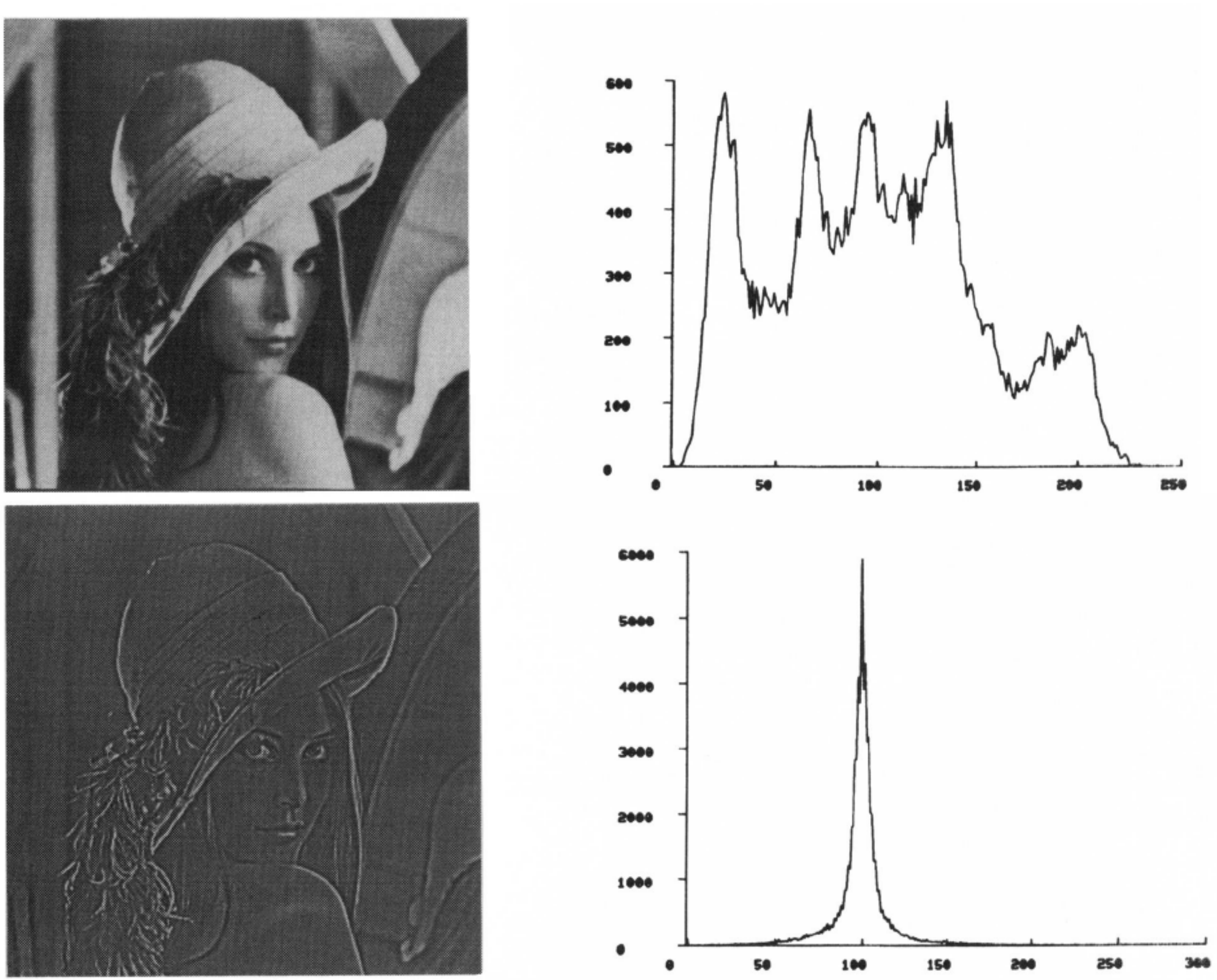
BP original

BP brightness altered

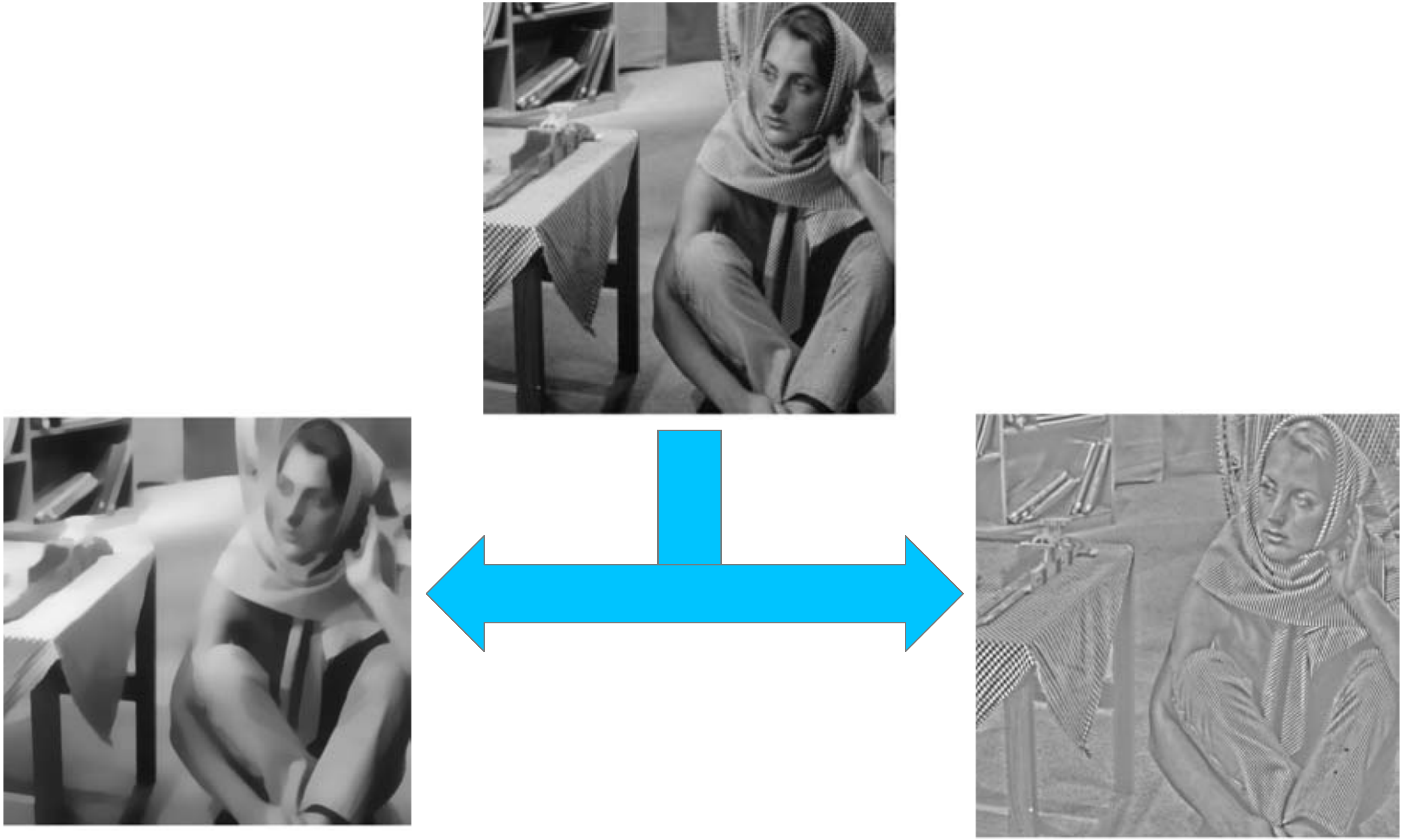


Residual images

Kuan et al. (1985) introduced the concept



Rudin, Osher & Fatemi (1992) introduced structure-texture decomposition using TV-L^1 minimisation



A residual image is effectively the result of a high-pass filter.

Possible process:

1. Use a smoothing filter *n* times
2. Subtract smoothed image from original
3. Rescale "residual" into an image

Residual images represent `texture' or `structure' of images.

Iteration scheme

Let f be any frame of a given image (or stereo) sequence.

$s = S(f)$ denotes the smooth component (of image f).

$r = R(f) = f - S(f)$ is the residual, with

$$s^{(0)} = f$$

$$s^{(n+1)} = S(s^{(n)}) \qquad \text{for} \quad n \geq 0$$

$$r^{(n+1)} = f - s^{(n+1)}$$

Some of the smoothing filters tested

Mean Filter, 3 x 3

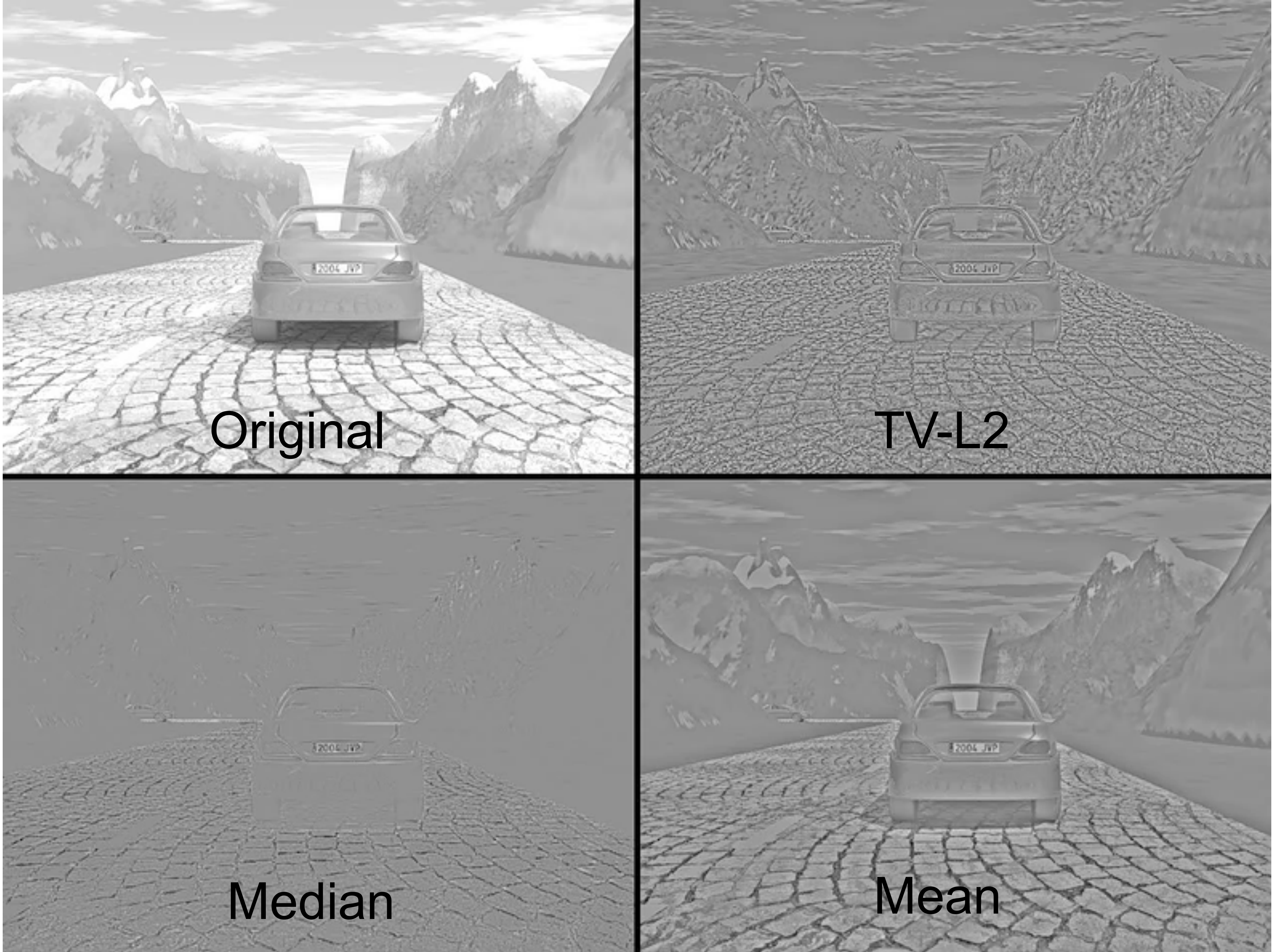
Median Filter, 3 x 3

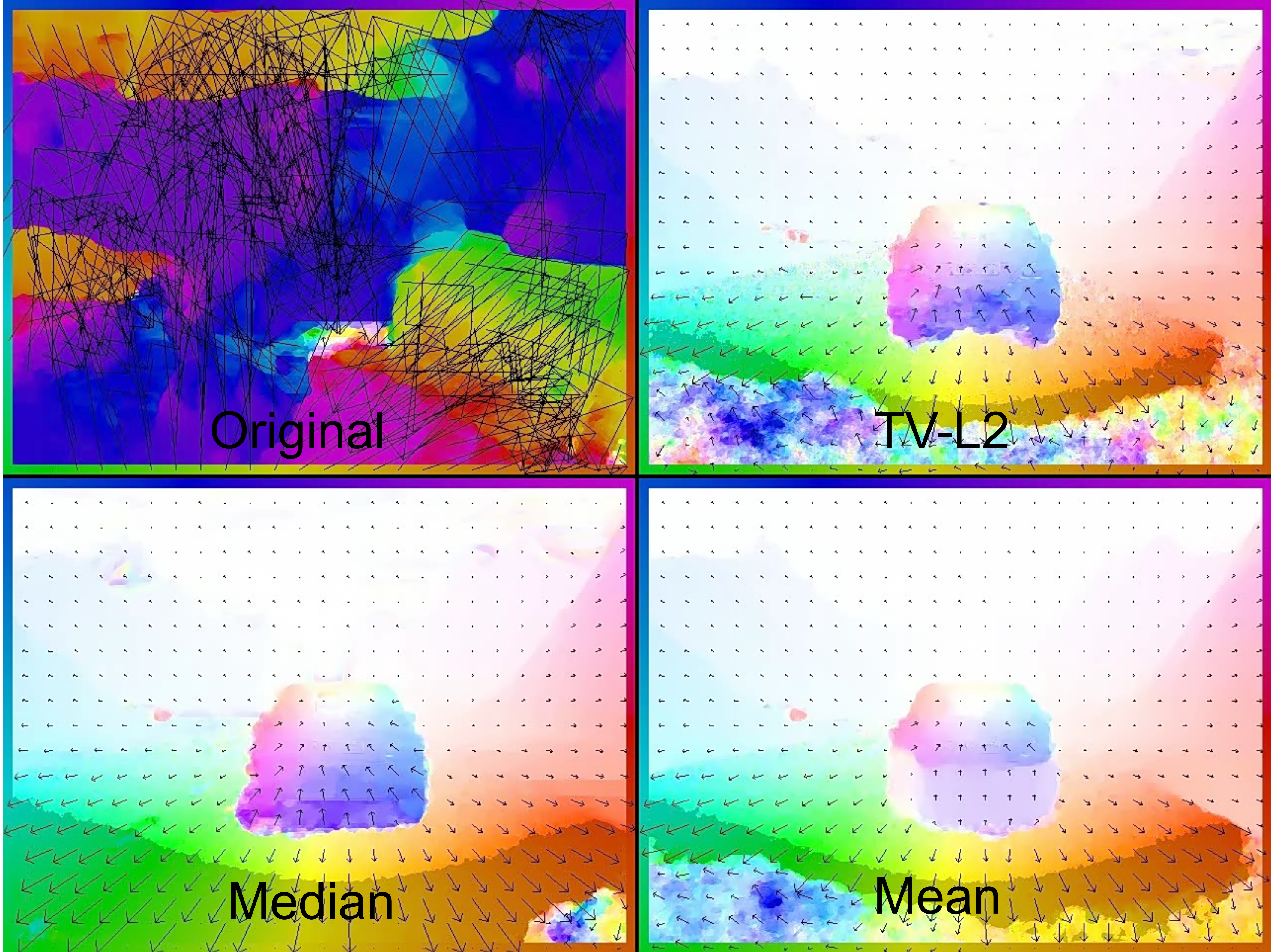
Sigma Filter (Lee, 1983)

TV-L^2 Filter (Rudin, Osher & Fatemi, 1992)

Bilateral Filter (Tomasi & Manduchi, 1998)

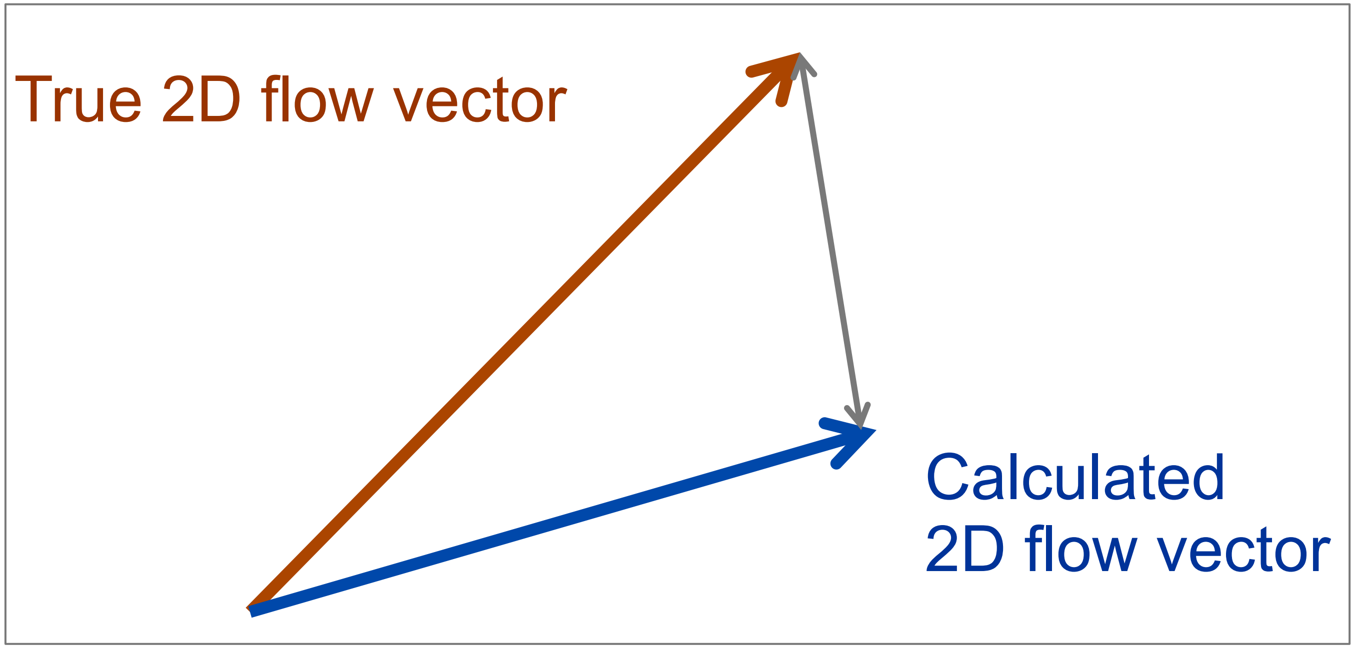
Trilateral Filter (Choudhury & Tumblin, 2003)


Original
TV-L2
Median
Mean


Original
TV-L2
Median
Mean

Quality measures on rendered sequences

Optic flow: endpoint error (EPE) (mean over all pixels)



Stereo analysis: RMS (between GT depth and calculated depth)

Decision for 40 iterations and 3x3 mean (use of TV-L^1)

n		TV-L^2	Sigma	Mean	Median	Bilateral	Trilateral
1	mean	7.6	7.7	7.7	7.4	6.8	6.3
	std	0.5	0.5	0.5	0.6	0.7	0.6
2	mean	7.4	7.7	7.4	6.8	6.2	5.0
	std	0.6	0.5	0.5	0.9	0.8	0.8
10	mean	6.9	7.5	5.6	4.7	3.3	1.7
	std	0.6	0.6	0.8	1.4	1.0	0.9
40	mean	5.4	6.1	2.8	3.9	1.6	-
	std	0.9	0.9	1.6	1.6	1.1	-

Conclusions from tests on rendered sequences

For relaxing the incorrectly posed intensity constancy in cost functions, we may do some

preprocessing

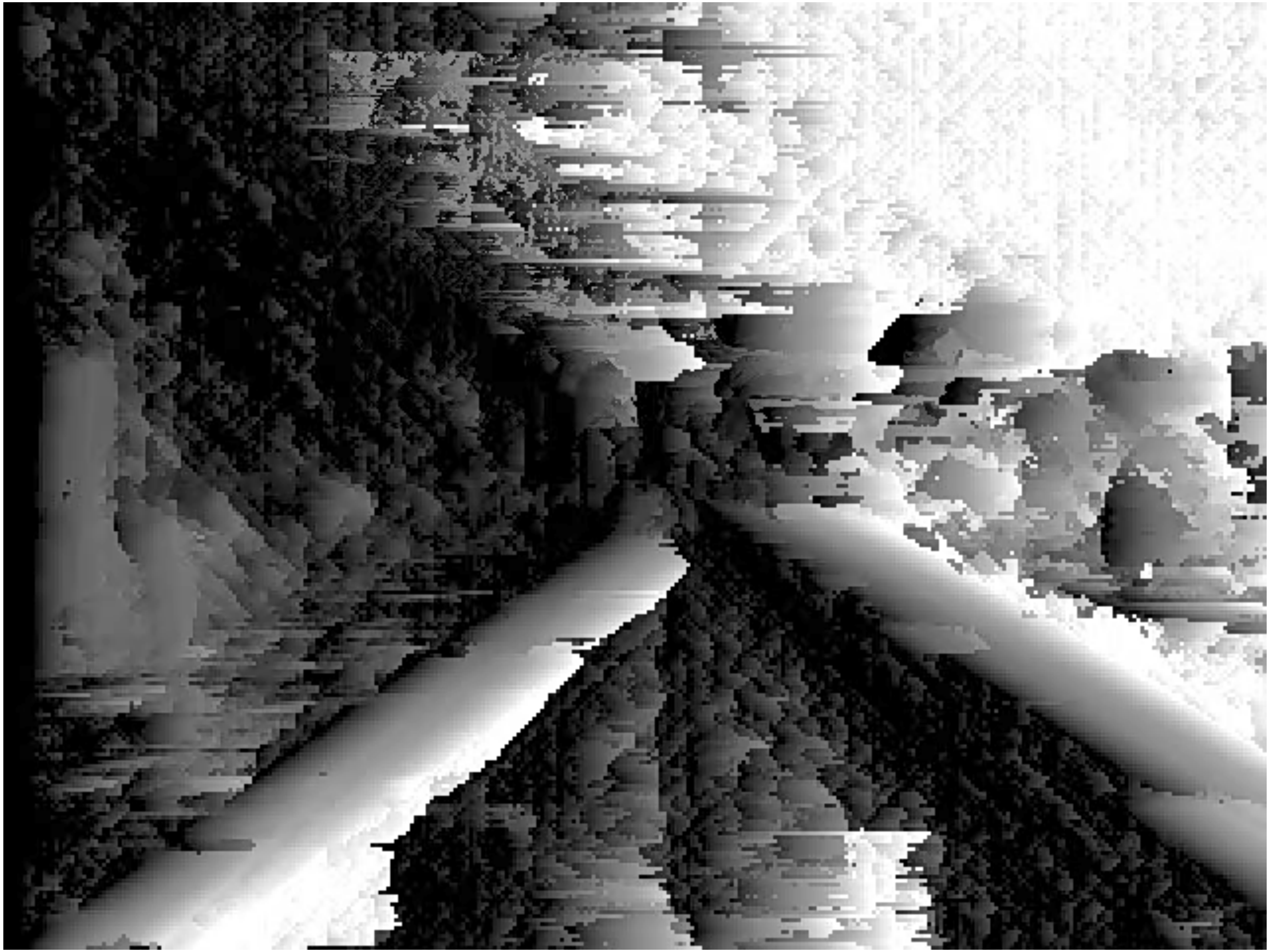
(residual images or edge detection) first, before entering into the correspondence analysis step.

No ranking of methods on rendered sequences (would be inconsistent with ranking on real-world sequences; photorealistic and physics-realistic sequences needed for particular situations).

Stereo on preprocessed sequences

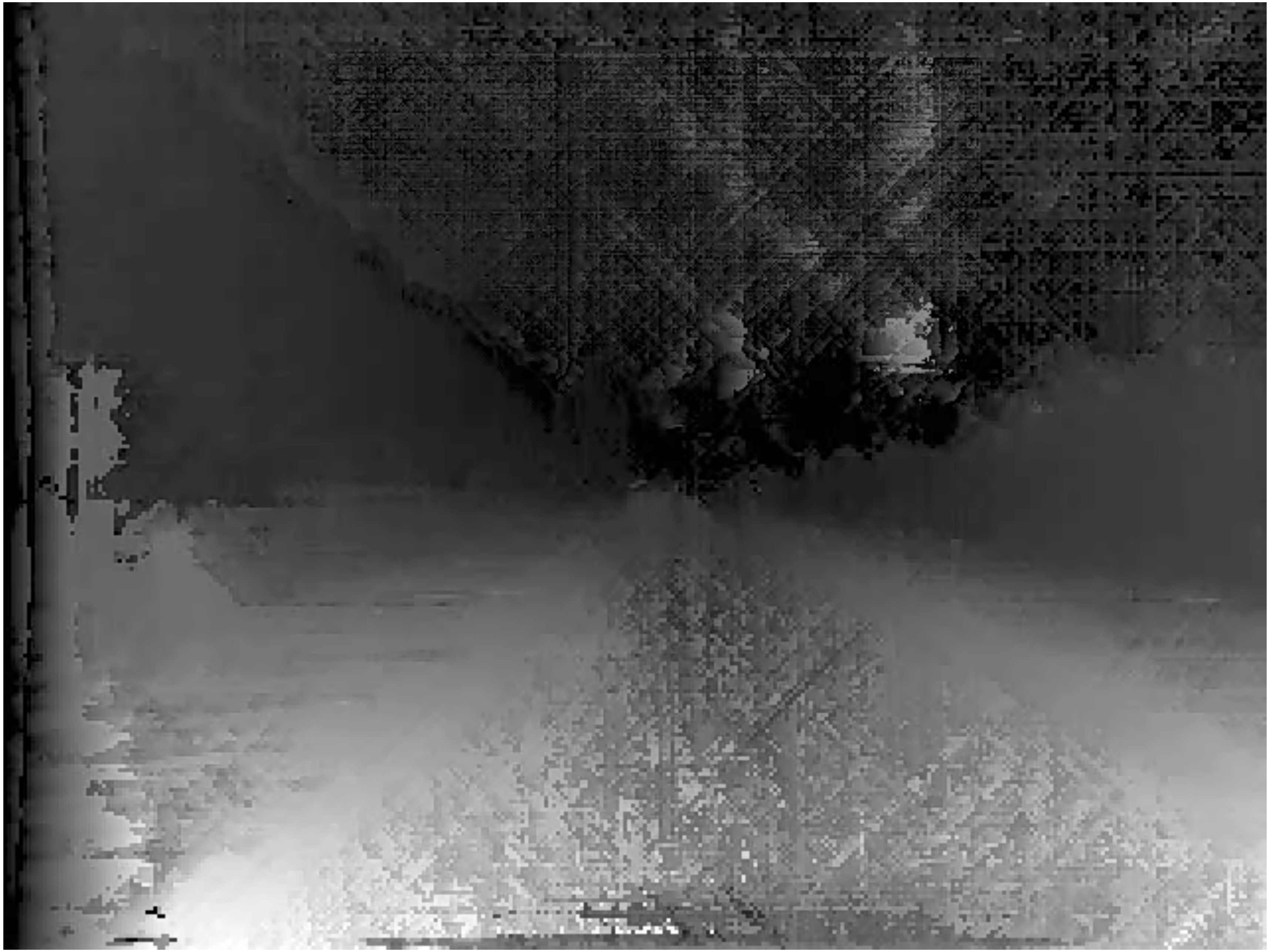
Situation 5: illumination artifacts

SGM BT

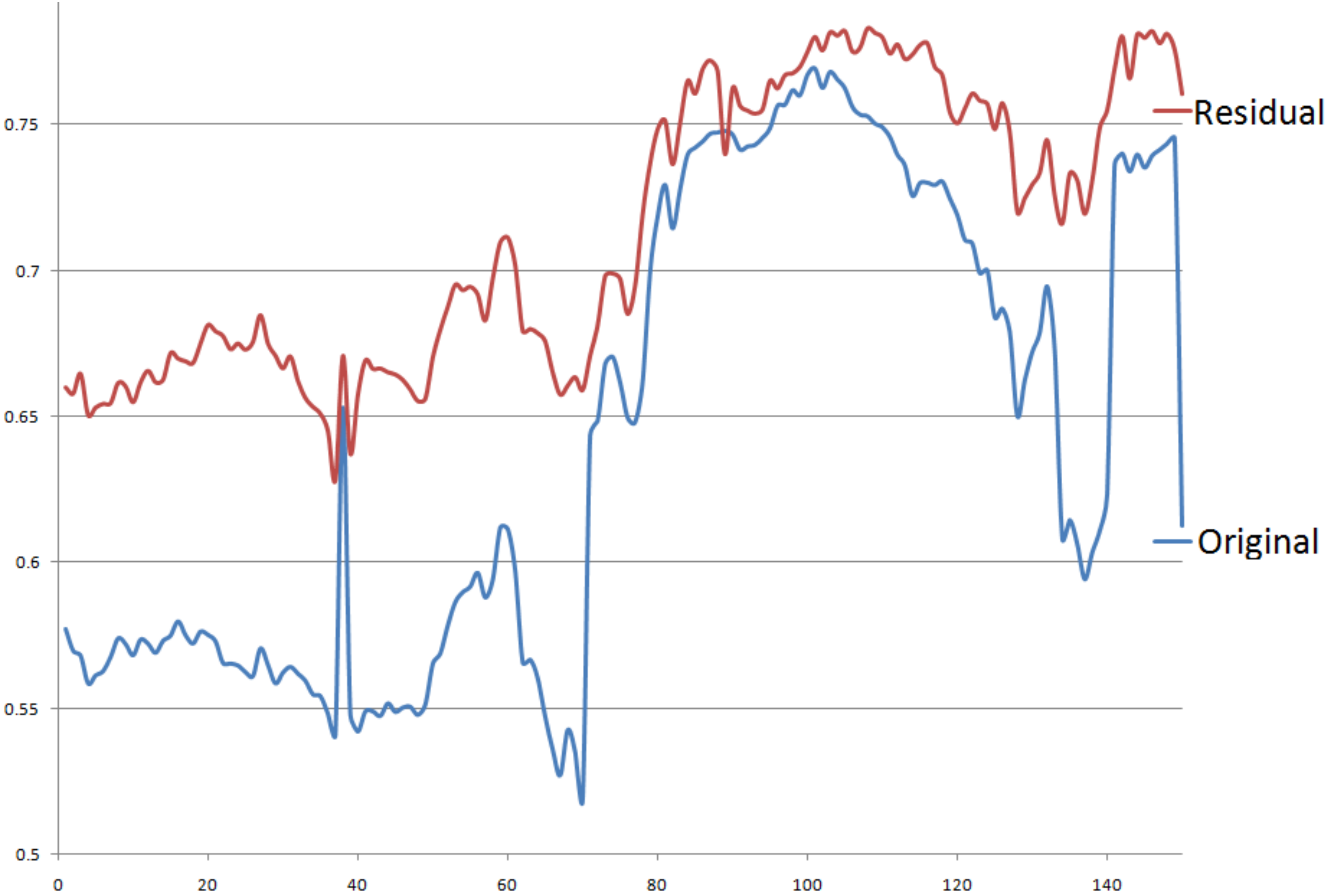


Situation 5: illumination artifacts

SGM BT (on residual sequence)



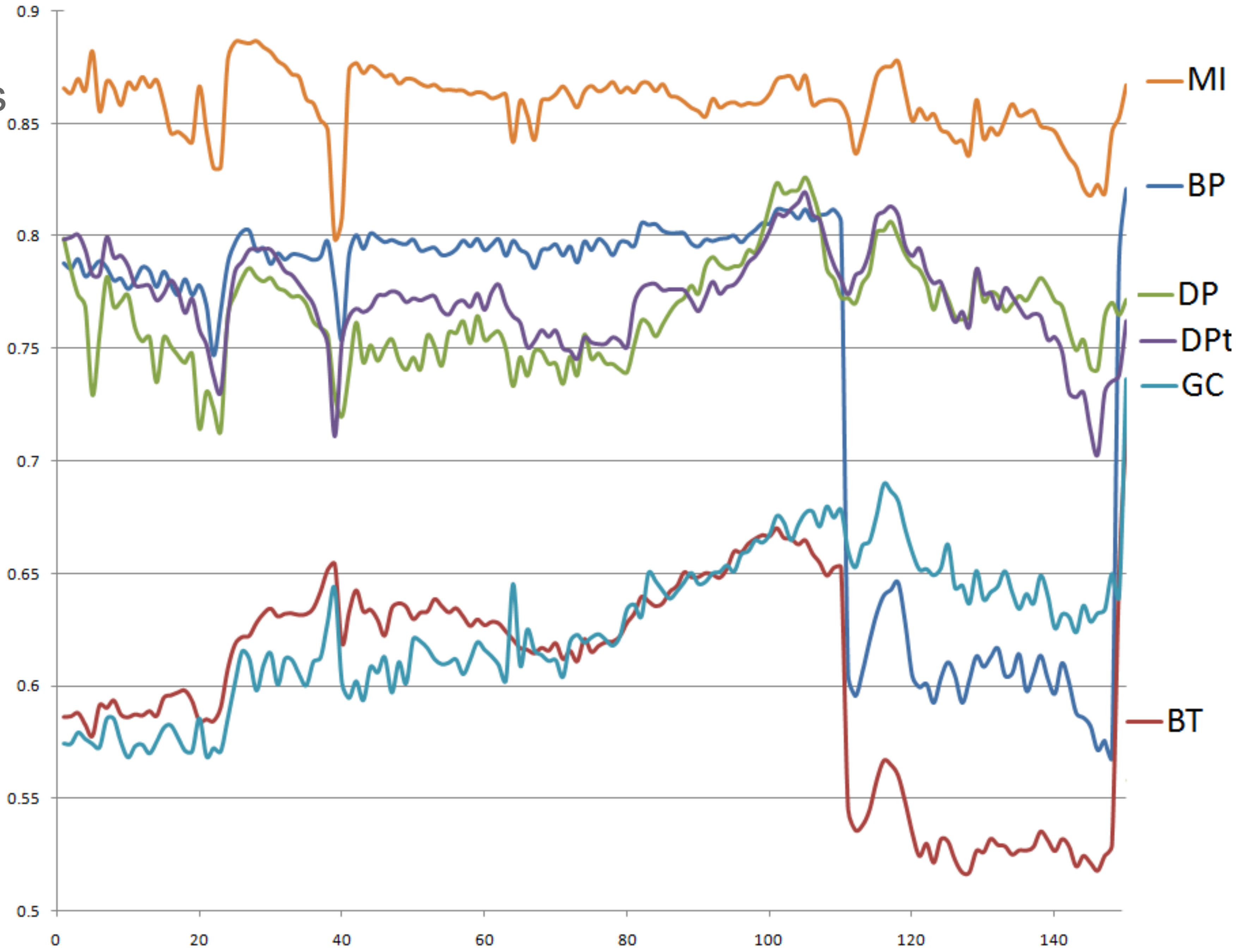
Performance of SGM BT on this sequence (w/o or with preprocessing)



NCC prediction for 150 frames of Situation 4

Brightness differences

original sequence



Sums of differences in NCC values

150 frames
Situation 4:
Brightness differences

original sequence

	BP	BT	DP	DPt	GC	MI		
BP	-						15.3	4
BT	-21.4	-					-113.0	6
DP	**3.0**	24.4	-				33.2	3
DPt	**4.1**	25.5	1.1	-			39.8	2
GC	-18.0	**3.4**	-21.0	-22.1	-		-92.6	5
MI	17.0	38.4	14.0	12.9	35.0	-	117.4	1

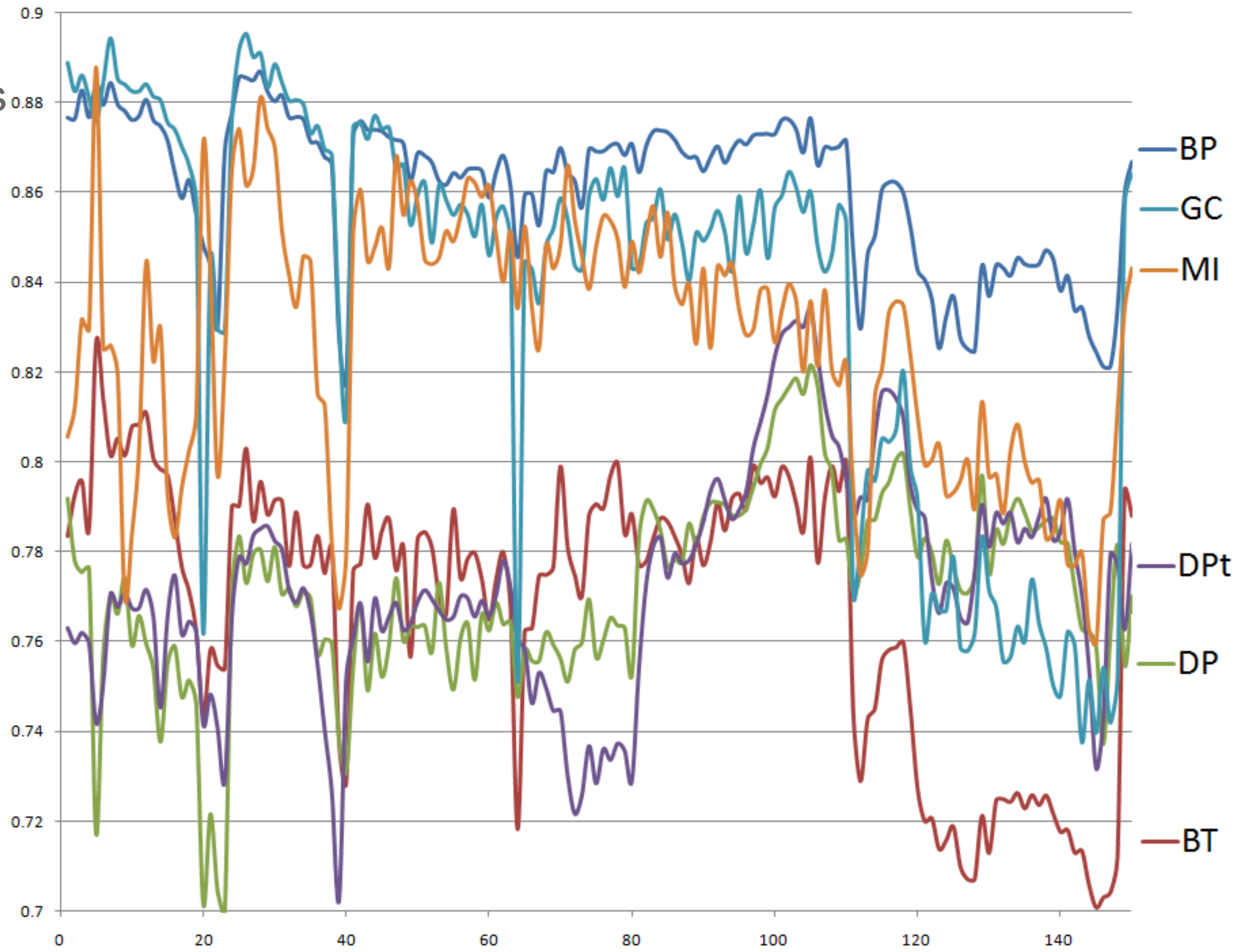
Sums of direct comparisons (sums of +1, 0, or -1)

	BP	BT	DP	DPt	GC	MI		
BP	-						174	2
BT	-150	-					-584	6
DP	**-58**	150	-				28	4
DPt	**-42**	150	64	-			172	3
GC	-74	**-16**	-150	-150	-		-540	5
MI	150	150	150	150	150	-	750	1

NCC prediction for 150 frames of Situation 4

Brightness differences

residual sequence



150 frames
Situation 4: Brightness differences

residual sequence

Sums of differences in NCC values

	BP	BT	DP	DPt	GC	MI		
BP	-						49.5	1
BT	-14.0	-					-34.2	6
DP	-13.5	0.5	-				-31.5	5
DPt	-13.3	0.7	0.2	-			-30.2	4
GC	-3.6	10.3	9.9	9.7	-		27.7	2
MI	-5.1	8.8	8.4	8.2	1.5	-	21.7	3

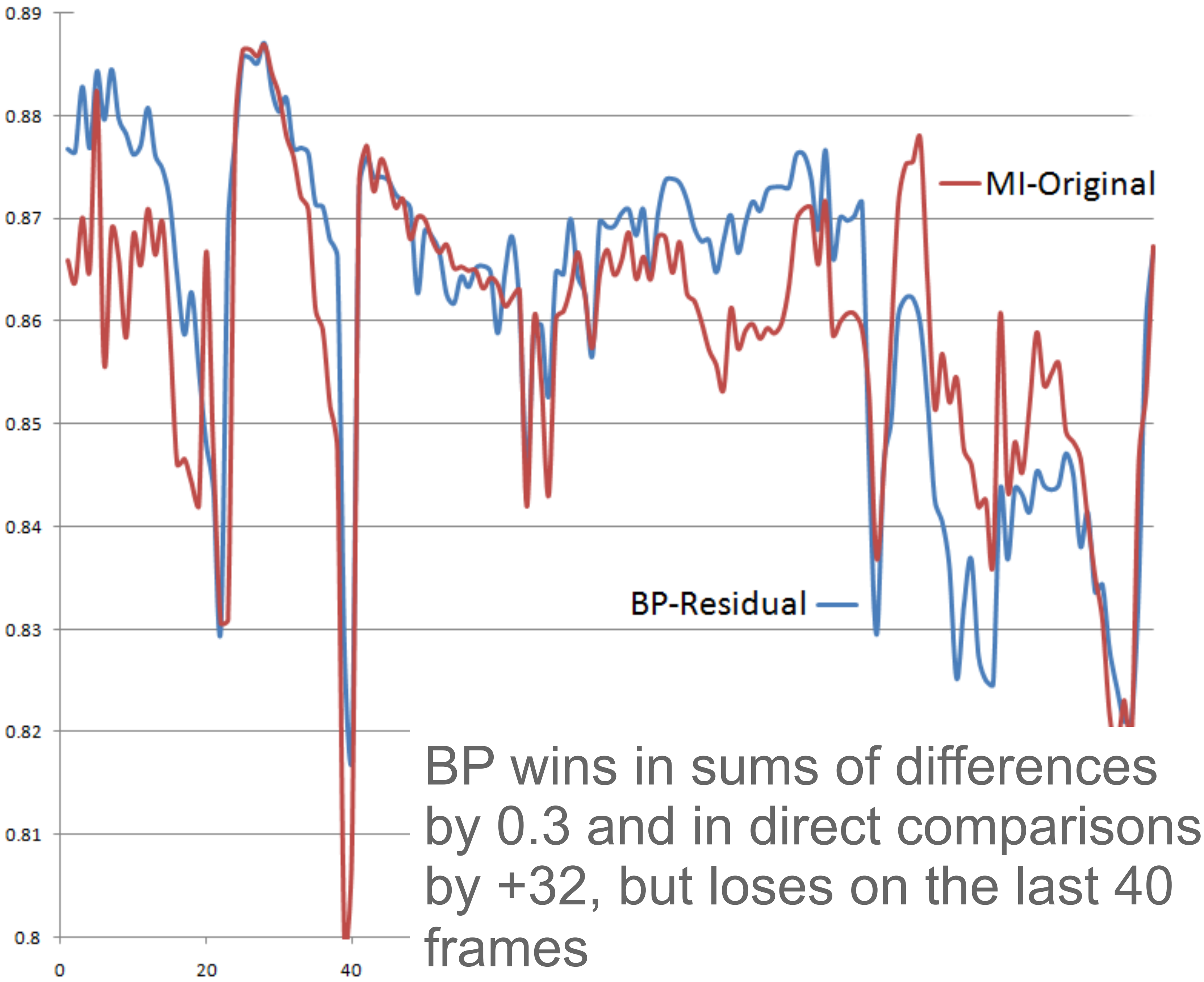
Sums of direct comparisons (sums of +1, 0, or -1)

	BP	BT	DP	DPt	GC	MI		
BP	-						666	1
BT	-150	-					-388	5
DP	-150	-26	-				-450	6
DPt	-150	-26	40	-			-358	4
GC	-74	150	96	88	-		302	2
MI	-142	140	138	134	-42	-	228	3

Best original versus best preprocessed

150 frames
Situation 4:
Brightness
differences

original
and
residual
sequence



BP wins in sums of differences by 0.3 and in direct comparisons by +32, but loses on the last 40 frames

Situation 4: Brightness differences
(original and residual sequences)

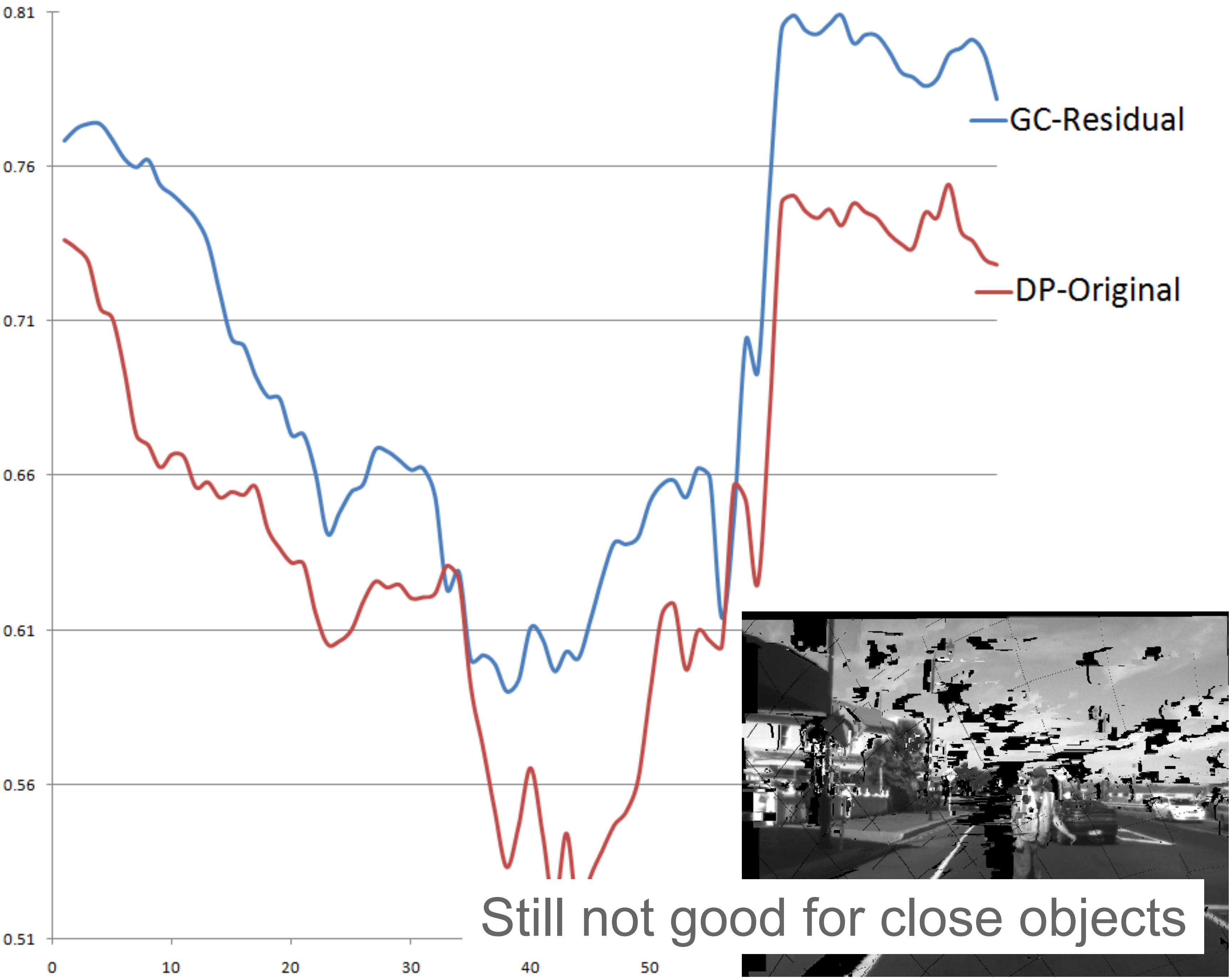
Virtual view for residual BP

Virtual view for original SGM MI

Best original versus best preprocessed

80 frames
Situation 2:
Close
object

original
and
residual
sequence

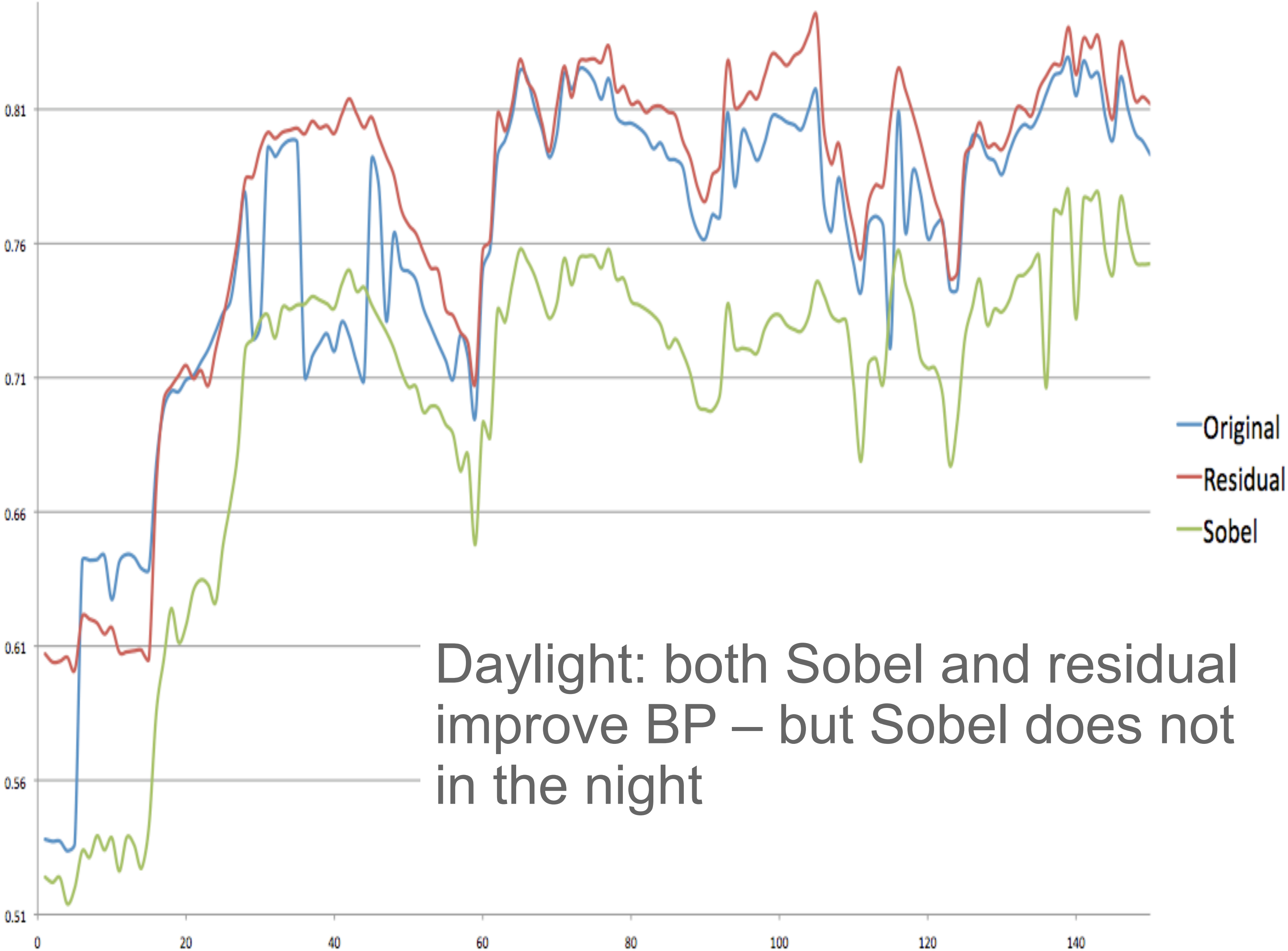


Still not good for close objects

BP: original versus Sobel versus residual

150 frames
Situation 3:
Inner city
at night

original
and
residual
sequence



Daylight: both Sobel and residual improve BP – but Sobel does not in the night

GC residual for a default driving situation

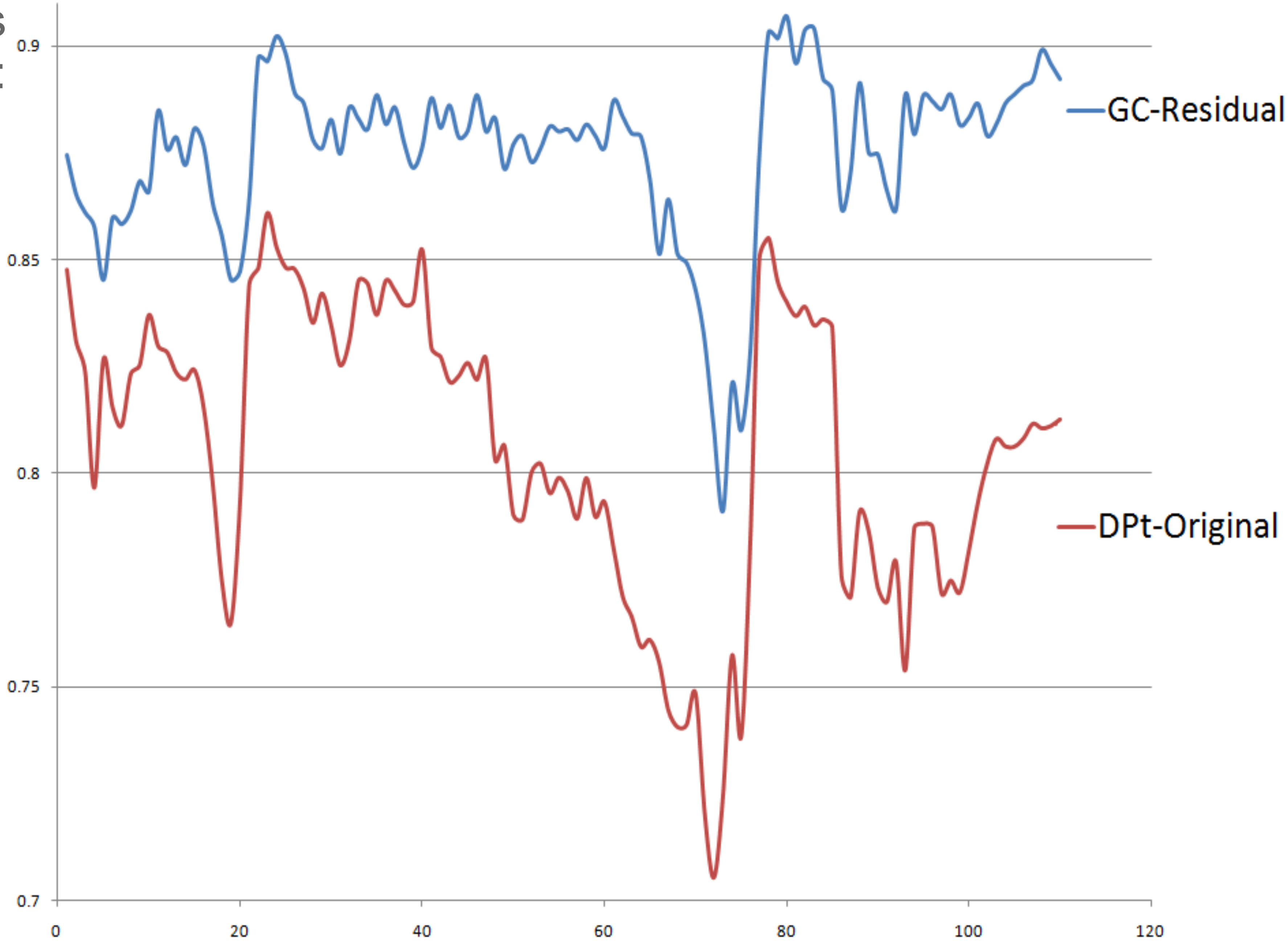
120 frames
Situation 1:
Default
conditions

residual
sequence

Best original versus best preprocessed

120 frames
Situation 1:
Default
conditions

original
and
residual
sequence



Currently: 28 students are recording trinocular sequences in HAKA1

Situations are still `manually' identified

Below: results for 5 situations as illustrated in this talk,

with 2 sequences each,

each sequence with 110 frames or more

Winner and steadiness (mean, std)

Situation 1 (def. driving):	GC residual	(.87, .01)
Situation 2 (close object):	GC residual	(.70, .07)
Situation 3 (inner city night):	GC original	(.79, .05)
Situation 4 (brightn. diff.):	BP residual	(.86, .01)
Situation 5 (illumin. artif.):	GC residual	(.89, .02)

Robustness (across all those 5 situations)

On original data

3	Best mean:	DPt	(.75, .07)
4	Second:	BP	(.74, .08)
5	Third:	SGM-MI	(.73, .08)

On residual sequences (3x3 mean, 40 iterations)

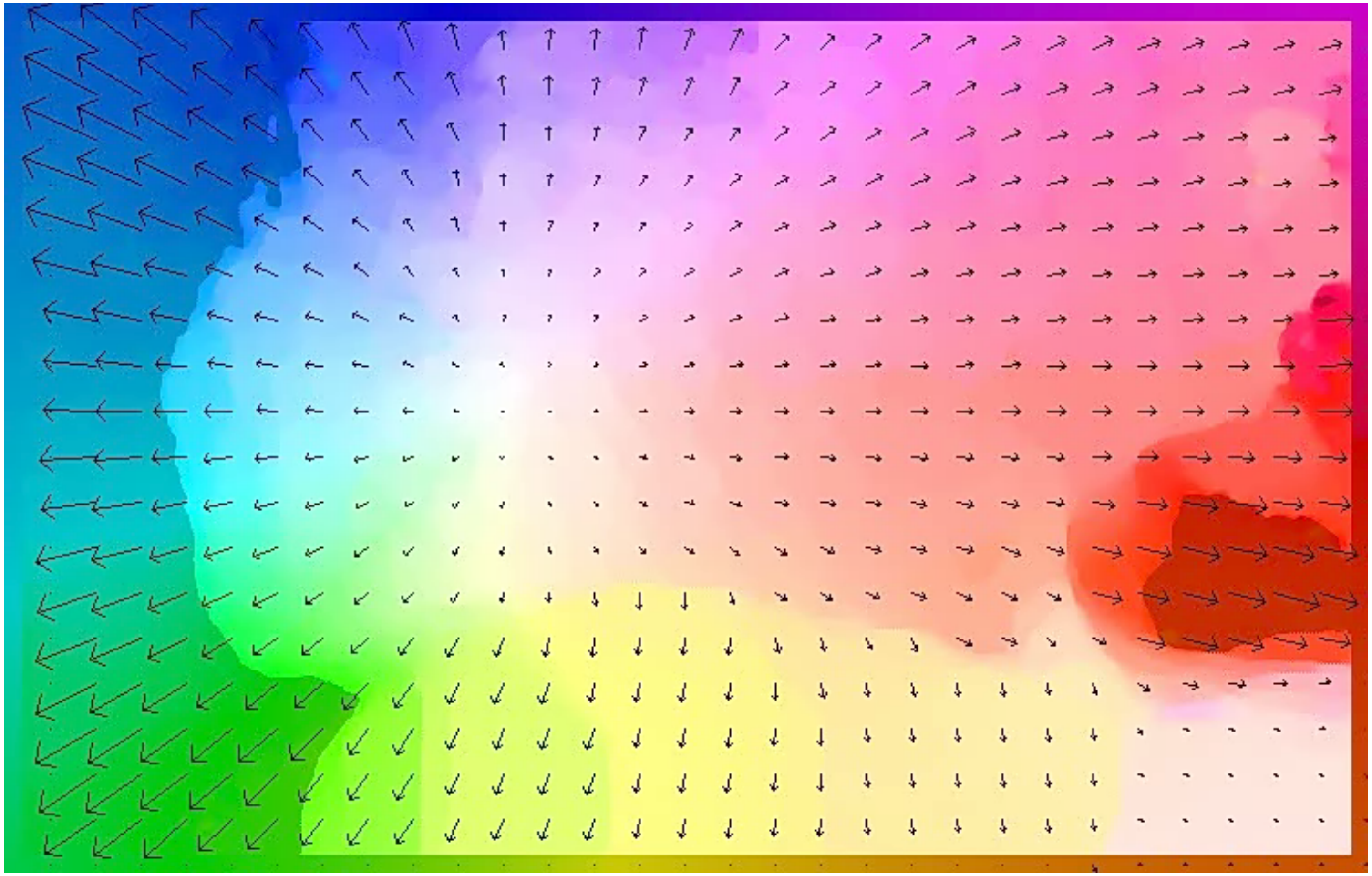
1	Best mean:	GC	(.829, .087)
2	Second:	BP	(.827, .085)
6	Third:	SGM-MI	(.70, .11)

Optic Flow on Preprocessed Sequences

Sequence and TV-L^1 on original sequence

120 frames
Situation 5:
Illumination
artifacts

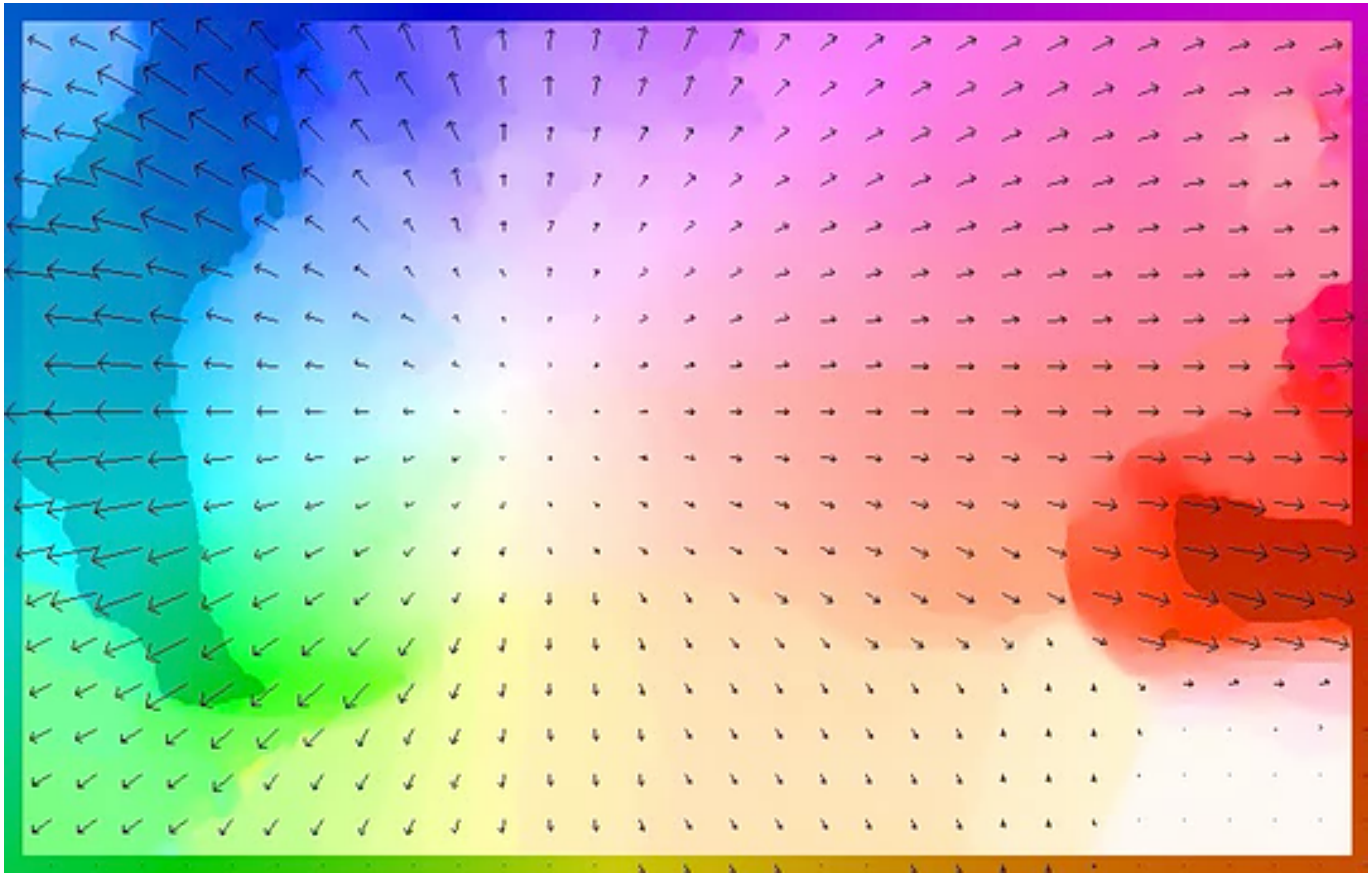
original
sequence



Sequence and TV-L^1 on residual sequence

120 frames
Situation 5:
Illumination
artifacts

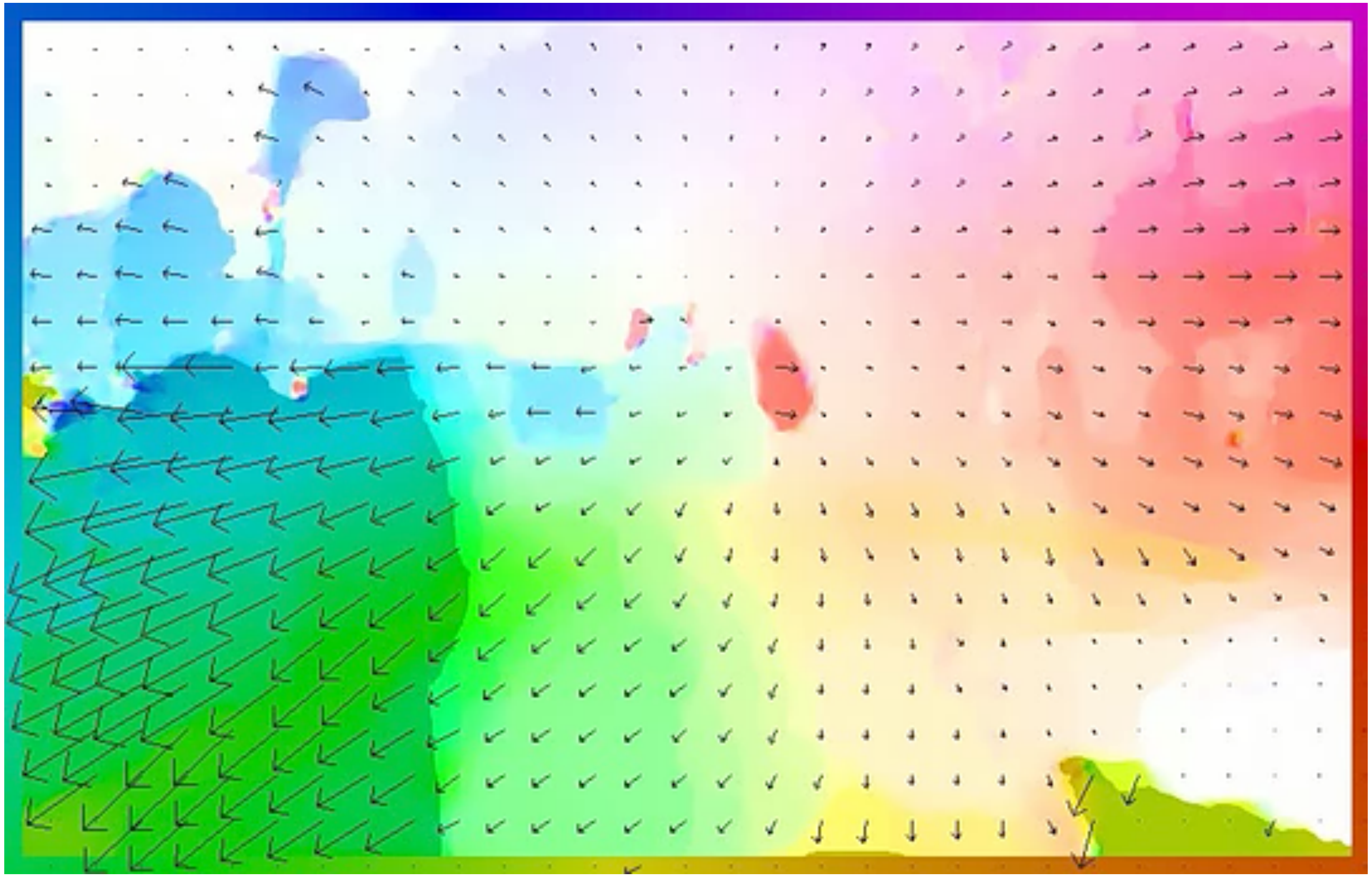
residual
sequence



Sequence and TV-L^1 on original sequence

150 frames
Situation 1:
Default
conditions

original
sequence



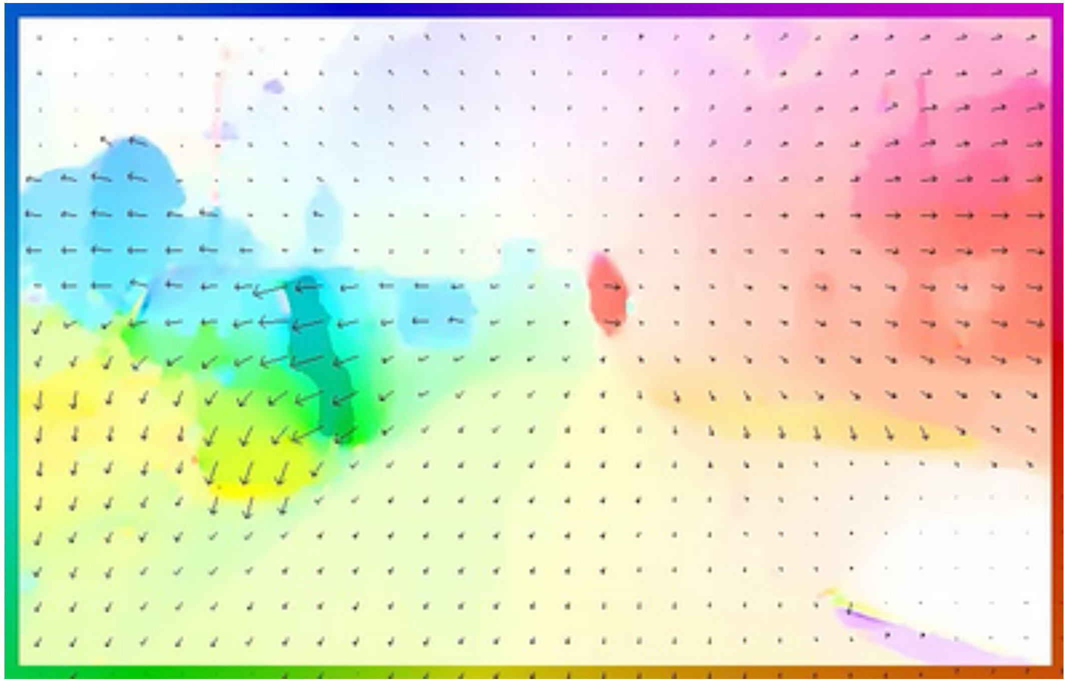
150 frames Situation 1: Default conditions

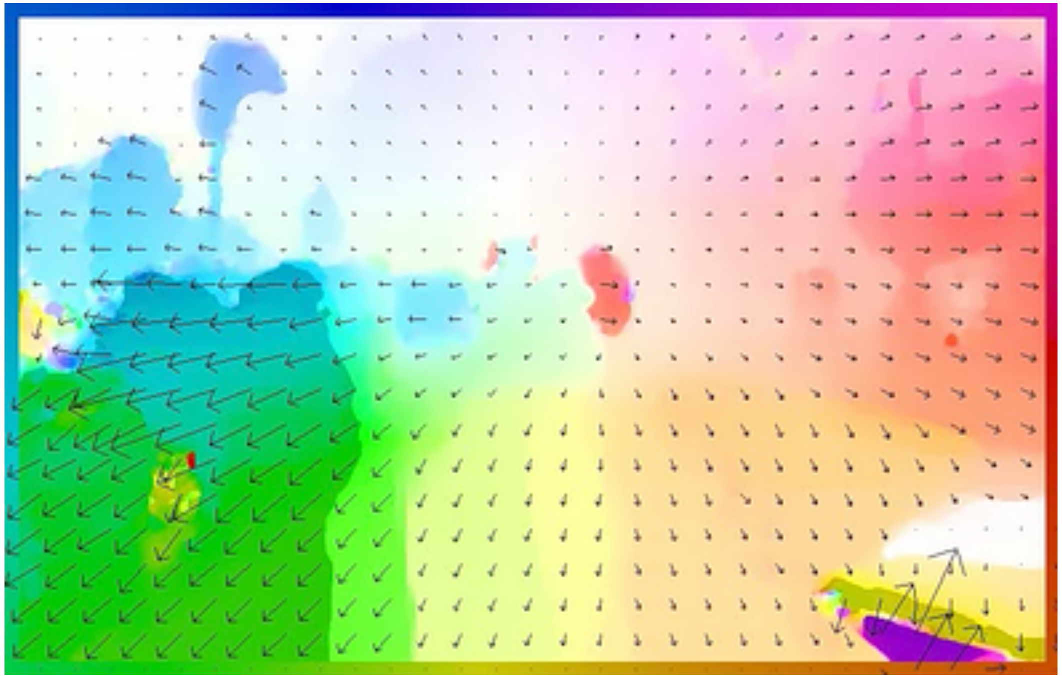
residual sequence

TV-L¹

3 iterations of 3x3 mean

40 iterations





Conclusions

Even with thousands of people evaluating soon stereo and motion algorithms, lane detection, pedestrian tracking, ... in their cars, the car industry and the vision community will have to verify the fulfillment of standards based on defined tests (for situations or scenarios).

Stereo: already a reasonable "toolbox"

Motion: we need to understand motion better than so far

"Crash tests" for vision-based DAS

identify situations of traffic scenes

calculate winner (mean) and steadiness (variance)
for highly ranked methods for those situations

calculate robustness by mean and variance across
identified situations

adaptation while driving:

(1) real-time situation recognition

(2) select method for the given situation

A few open problems along that way

exact trajectory calculation for the ego-vehicle

definition and identification of situations

improvement of correspondence techniques for "close objects", "rain in the night", "sun strike", ...; verified on long sequences

camera technology: inter-camera-communication, resolution, dynamic range, ...

wide-angle stereo and motion ... high-level vision-based DAS

Test sequences: see EISATS-link on

www.mi.auckland.ac.nz

A joint project with

- Environment perception group, Daimler A.G.,
- Hella Aglaia Mobile Vision GmbH, and the
- European "Drivsco" project.