imaging solutions with free software and open hardware

elphel

High Resolution Wide FoV CNN System for Target Classification, Ranging and Tracking

Andrey Filippov | Olga Filippova | Oleg Dzhimiev

Responding to the challenge of the high resolution low cost image sensors revolution – enabling real-time 3D-aware machine learning systems

imaging solutions
with free software
and open hardware
elphel

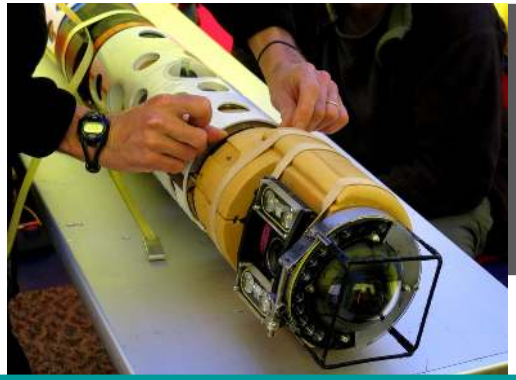
Elphel Expertise and Applications

Elphel high-performance open hardware cameras are used in scientific and robotic applications:

Google Streetview

Google Books

SCINI - robot



SOLO - robot

In addition to FPGA, software and mechanical design the cameras integrate advanced optical elements measurement, selection, alignment and calibration for high-resolution imaging

Machine Learning and High Resolution Images

Problem

Modern CNNs are ***very efficient*** for low resolution complex images

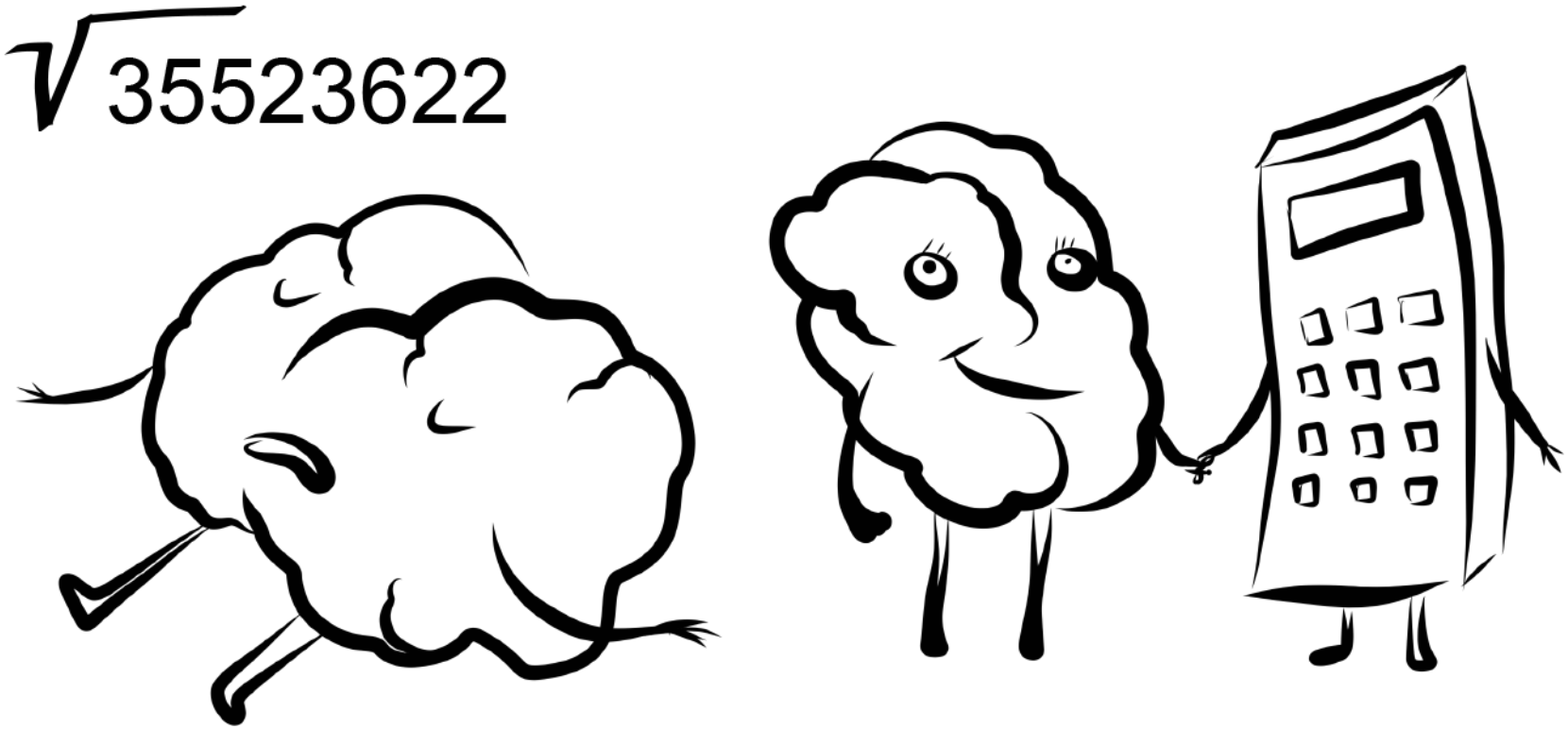
High resolution low SWaP-C image sensors *are* ***alredy available*** because of the camera phone industry

but,

End-to-end CNNs are ***not efficient*** for the real time high-resolution 3-D image processing

Augmenting CNN with Efficient Training-Free Image Processing Algorithms

Solution

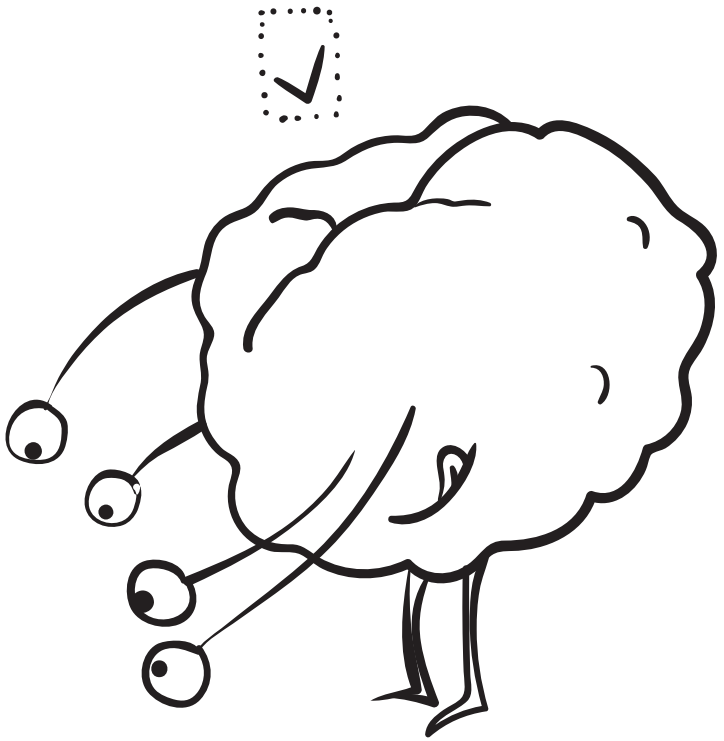


Modern CNNs are ***not efficient*** for the problems where known training-free, high data bandwidth algorithms exist

Efficiency of the CNN can be greatly improved when ***augmented*** with the proposed high performance Tile Processor that still delegates “decision making” to the network

End-to-End CNN vs. Augmented CNN

End-to-End



CNN Augmented with Training-Free, Linear Image Processing

Passive 3-D Reconstruction is a Difficult Task

- Recovering depth information from the images is an old idea, inspired by the human binocular vision
- Unfortunately most implementations so far fell short of the expectations, yielding to modern active ranging methods (LIDARs, ToF sensors), but for some applications active ranging has the same weakness as the old night vision devices with IR illumination - detectability
- Most successful among the passive methods is currently Structure from Motion (SfM), but it is not suitable for the dynamic scenes as the individual images are acquired at different times

Elphel approach enables true **passive long range** system capable of acquiring and processing **dynamic scenes**, highly **scalable**, and **noise-resilient**. It allows uniform processing of the **stereo** and **optical flow** data.

It combines **flexibility** of the **ML systems** and **high resolution image processing** making possible **real-time intelligent** object classification, ranging and tracking.

imaging solutions with free software and open hardware

elphel

Current Results: Long Range Photogrammetry and 3D Reconstruction

	position, m			orientation, °		
	x	y	z	azimuth	elevation	roll
mouse	0.81	-0.10	0.58	125.7	-6.0	0
camera	0.74	-0.54	0.00	87.5	0.6	-0.0

Marker 1 (Satellite vs 3D model)

x	y	z
963.14	127.85	-87.47

d_{map} drag over map

d_{3d} 967.1 m

Δ - m

∞ m

- **Long Range Multi-view Stereo Camera with 4 sensors**
- **Tile Processor** software implementation
- **3D models** generated from quad image sets, web x3d viewer available
- Integrated with the satellite images/maps as the ground truth data
- **Source code, hardware design**, and data are released under **FLOSS** licenses

4-sensor High Resolution Stereo Camera

967 meters

1011 meters

Map as ground truth data

imaging solutions with free software and open hardware

elphel

Muti-View Stereo (MVS) 3-D Scene Reconstruction

Regular approach:

- Binocular (dual lens) camera
- 1-D image matching along epipolar lines
- Destructive image resampling
- None or low image super-resolution
- Fixed goal processing

Does not meet expectations

Elphel Approach:

- At least 4 non-colinear sensors
- Native 2-D image matching
- Lossless rectification in the frequency domain
- Deep sub-pixel super-resolution
- Adaptable to the higher level goals processing

Long Range High Resolution MVS system measuring distances **thousands times** exceeding the baseline over **wide (60° x 45°) FoV**

imaging solutions with free software and open hardware

elphel

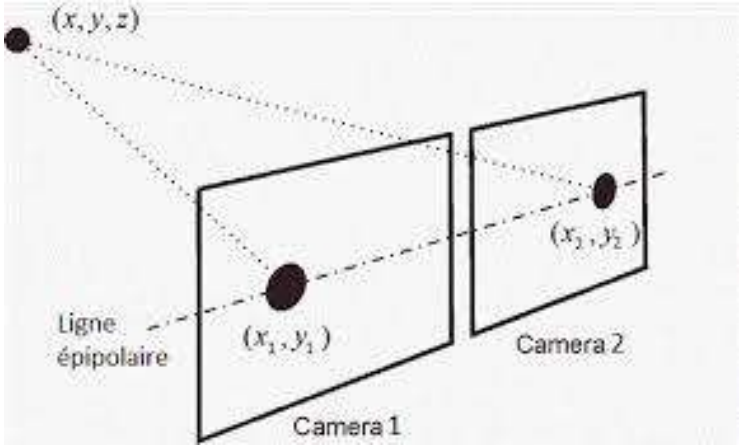
1-D vs. 2-D Frequency Domain (FD) Image Matching

1-D Matching

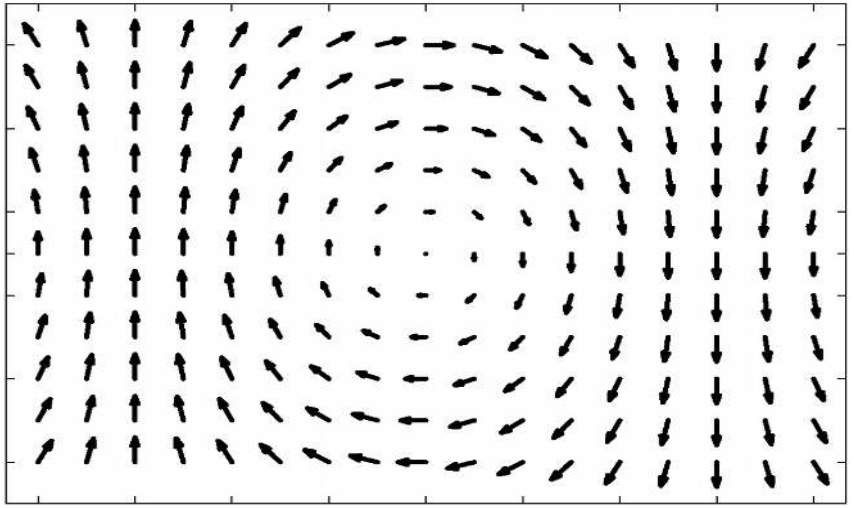
- Matching images along the epipolar lines
- Efficient for a single pair of images
- Requires separate aberration correction and rectification for high-res images
- Not compatible with the optical flow measurement
- Needs post-averaging for noise reduction

2-D Matching

- Full 2-D matching with epipolar constraints
- Extra computational penalty reversed by sharing FD transformation for aberration correction, stereo, and optical flow measurements for motion detection and objects tracking
- Noise resilience through implicit 2-D filtering and accumulating data from multiple high resolution image sensors



Binocular stereo matching along an epipolar line



Optical flow with arbitrary directions of the motion vectors

imaging solutions with free software and open hardware

elphel

Current Technology: Sensors vs. Lenses

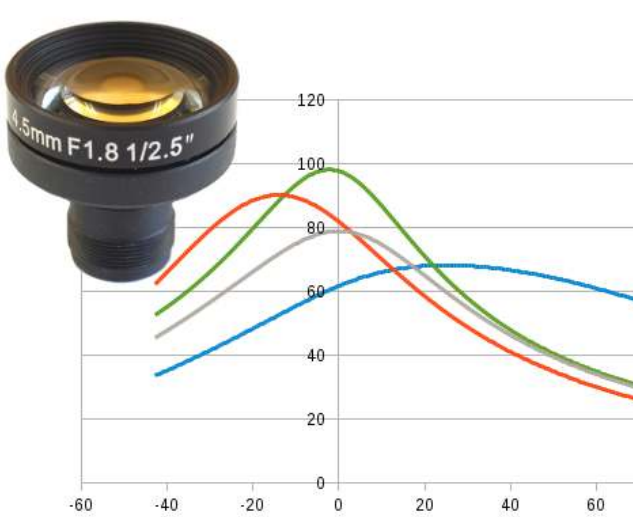
Small Format Sensors:

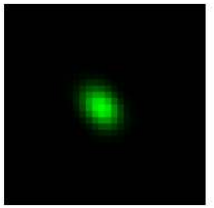
- Low SWaP-C
- High resolution (10-20 MPix)
- High quantum efficiency
- Low noise

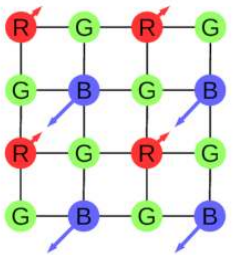
Modern High-resolution image sensors perfected by cell phone industry

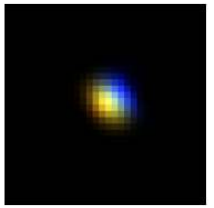
Small Format Lenses:

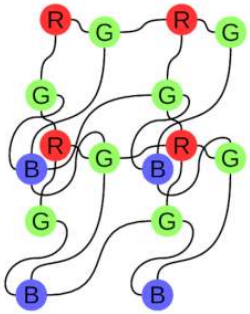
- Full frame resolution breaks even with sensors at ~1-3 MPix
- Need calibration + aberration correction to match sensor quality









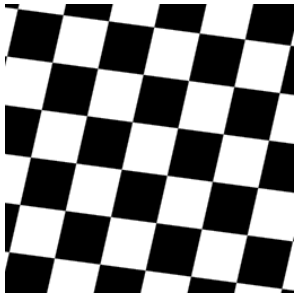


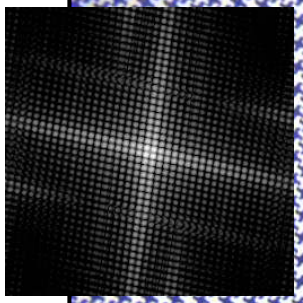
imaging solutions with free software and open hardware

elphel

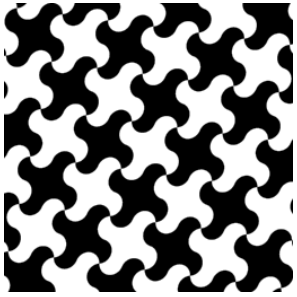
Current Results: Lens Calibration for Optical Aberrations Correction

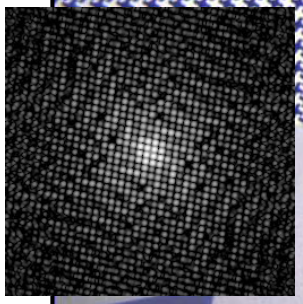
Elphel uses high resolution sensors, calibrates lenses and uses space-variant deconvolution to maintain full sensor resolution over the camera FoV





Standard checkerboard pattern and its power spectrum: large spectral areas are not probed





Elphel curved pattern - rich spectrum permits probing at all spatial frequencies

imaging solutions with free software and open hardware

elphel

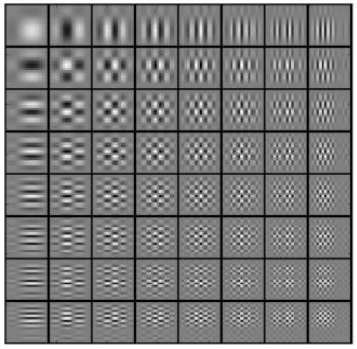
Technical Approach: Tile Processor

Tile Processor (TP) is a Swiss Army knife of Frequency Domain (FD) image processing, providing the following functionality:

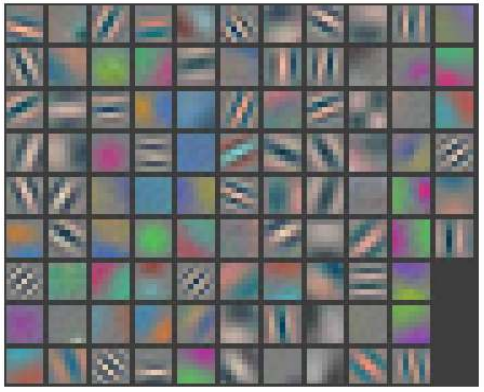
- Direct and inverse FD transformation
- Optical aberrations correction
- Lossless image rectification
- Image 2-D phase correlation

TP modules are currently under active development and 50% of them are already implemented as RTL code and tested in FPGA; code is available at GitHub

When used with the CNN, **TP** offers two orders of magnitude reduction of the number of input features without sacrificing universality of the end-to-end processing, making it feasible for the real-time applications with plurality of the high resolution images.



a) 2-d MDCT N=8 basis functions features



b) Trained CNN 1st layer features M. D. Zeiler and R. Fergus, 2014

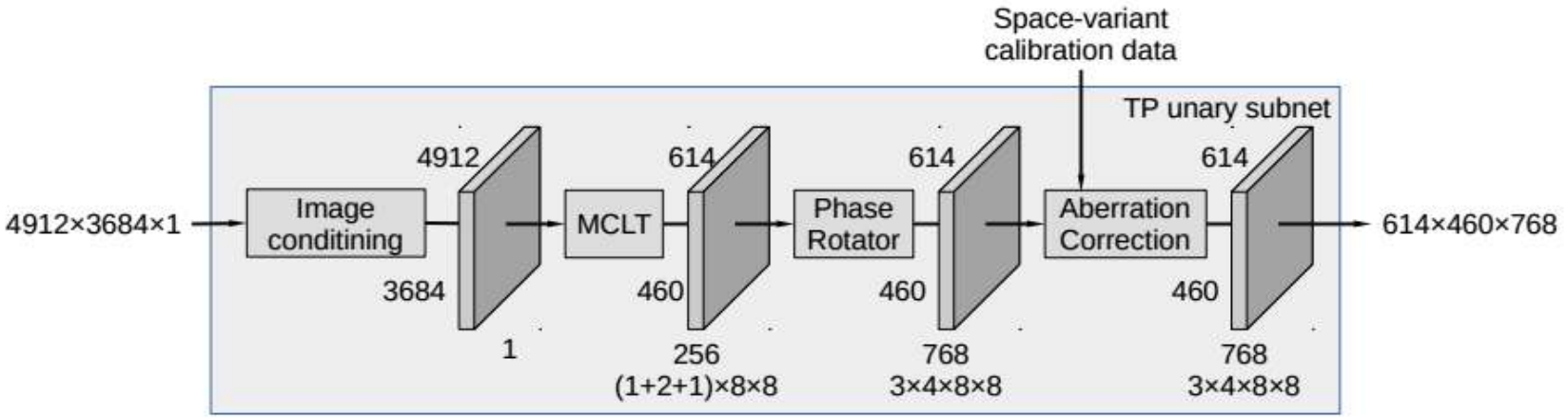
TP efficiently implements basis functions that are similar to the first layers' features of the trained CNN.

TP is free of any training, all the application-specific processing and "decision making" is delegated to the CNN. It is analogous to how a "hardwired" multilayer retina offloads the human brain.

imaging solutions with free software and open hardware

elphel

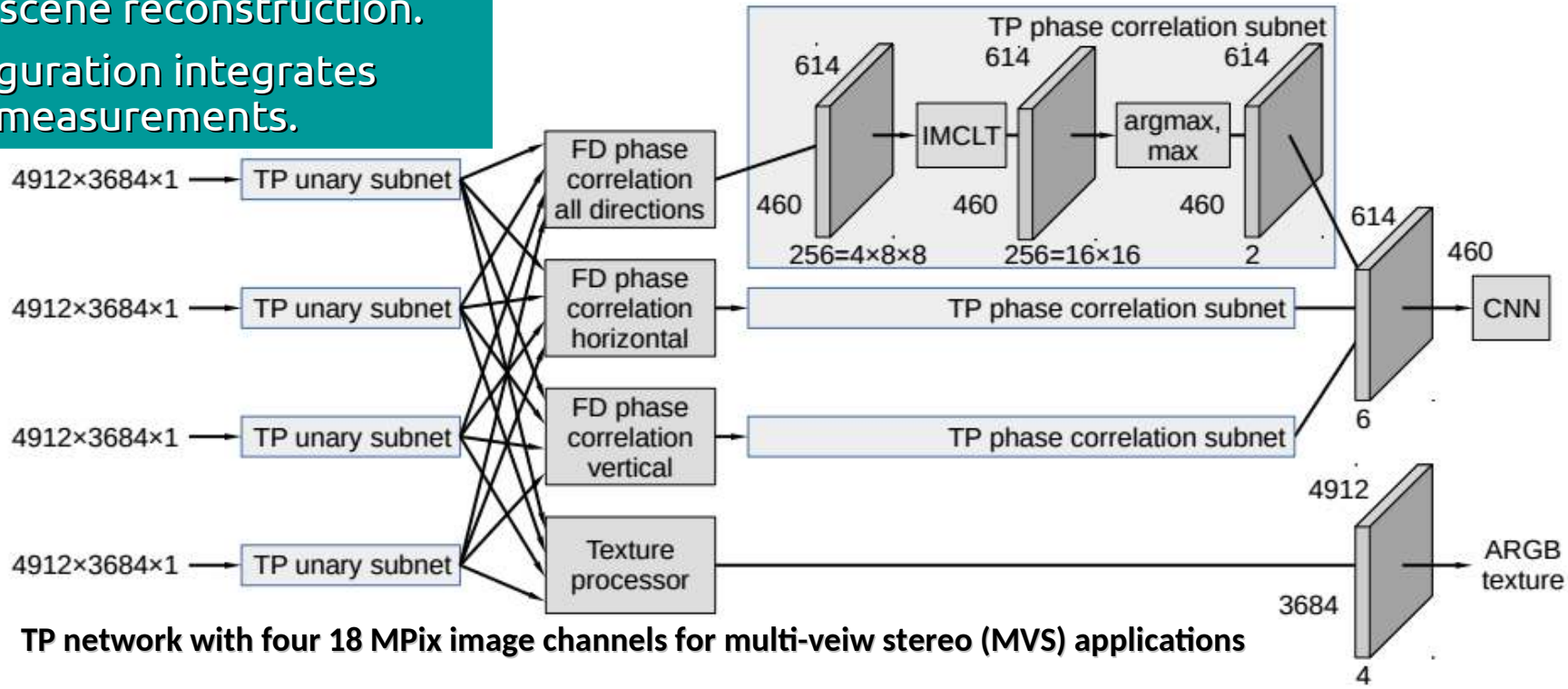
Technical Approach: Tile Processor Data Flow for Four Imagers



TP unary subnet for a single 18 MPix image sensor

TP as an efficient CNN pre-processor for MVS 3-D scene reconstruction.

Similar configuration integrates optical flow measurements.



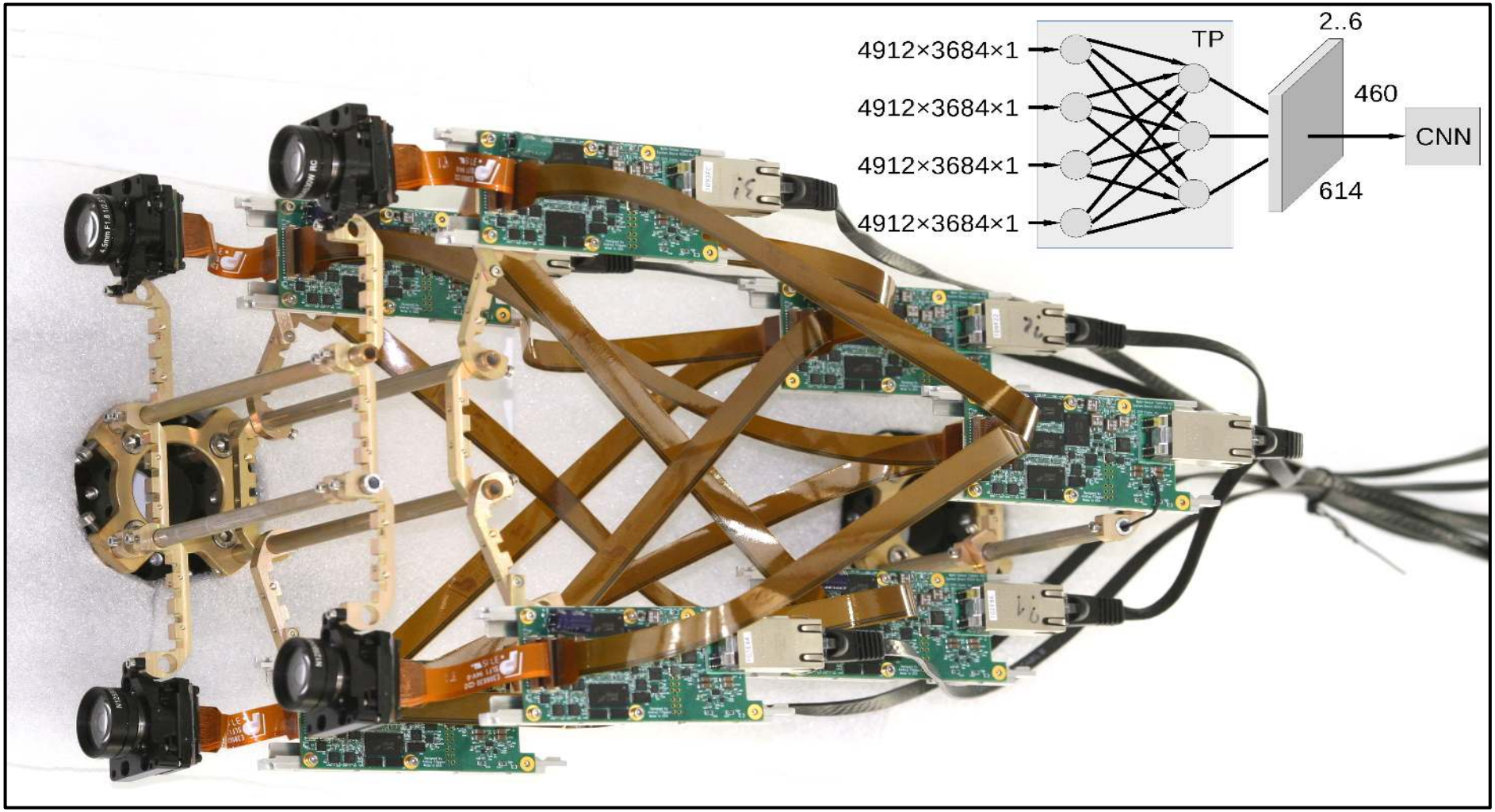
TP network with four 18 MPix image channels for multi-veiw stereo (MVS) applications

imaging solutions with free software and open hardware

elphel

Technical Approach: Multi-Board Prototype for the TP+CNN

Partitioning of the system into multiple existing Elphel 10393 camera system modules allows to fit the required TP functionality into smaller FPGAs



imaging solutions with free software and open hardware

elphel

Technical Approach: Quad Sensor Cameras and the Image Sets

High Resolution Image Sets

More research is needed for the CNN part of the system. Available image sets, such as KITTI have insufficient resolution (1.4 MPix) and they use different spatial arrangement of the cameras. We plan to **capture high resolution quad camera image sets** using available NC393-based cameras.

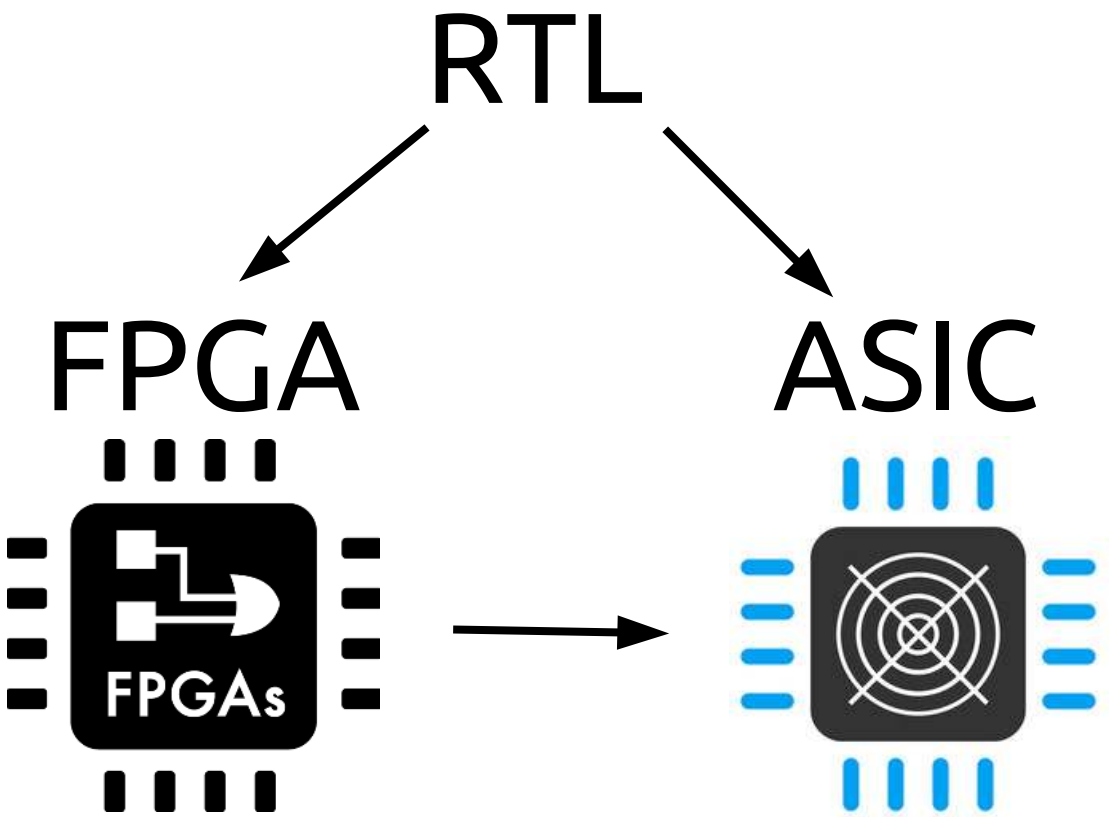
Ground Truth Data

We are primarily interested in **long distance ranging** (few hundreds to thousands meters), use of the LIDARs to capture **ground truth data** is not practical. Instead we use a pair of he similar quad cameras mounted on a car top, pointed in the same direction and fuse the 3-D measurements' results. This combined data has higher precision than that of each individual one and can be used instead of the ground truth data for them.

Application Areas: 3D-Aware ML systems

1. **Long distance passive ranging and 3D scene reconstruction**
   - Autonomous vehicle navigation
   - Noise-resilient 3D object detection, classification and localization
   - Motion detection and object tracking
2. **High resolution high frame rate photogrammetry**
   - Real-time optical aberration and distortion correction
   - Image enhancement, including LWIR and multi-spectral
3. **New application areas of the real-time machine learning systems**
   - 3D reconstruction of the complex scenes
   - 3D multiple view and optical flow objects classification and localization
   - Merging with other ML systems as an efficient pre-processor

FPGA – RTL - ASIC

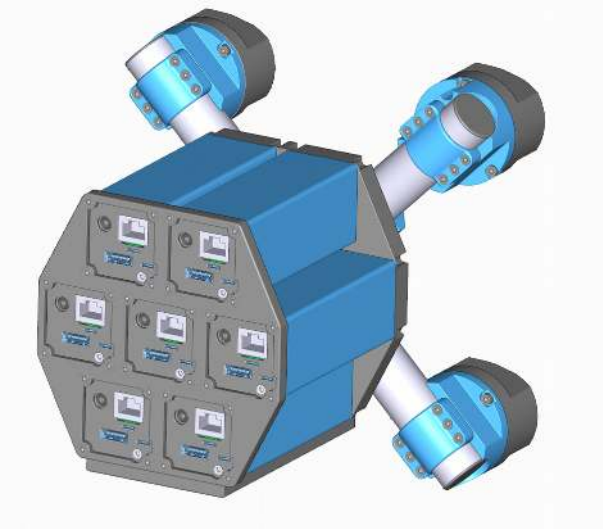


FPGA-based TP implementation is designed for the future conversion to ASIC.
It may be combined with the CNN on the same chip or communicate to an external ML system.

imaging solutions with free software and open hardware

elphel

Directions of the Project and Timeline



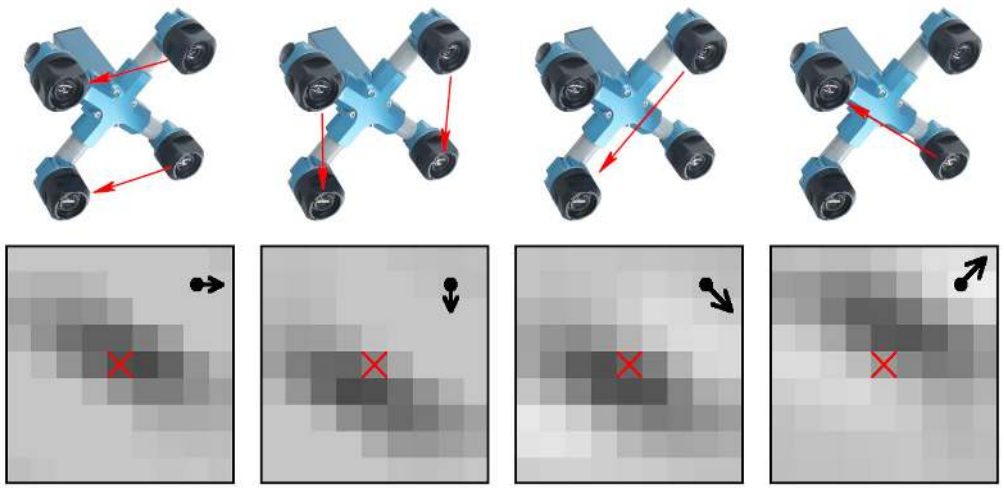
1	Create calibrated image sets for the TP software emulation and CNN training/testing (5 MPix images)
2	Upgrade a pair of quad cameras to 18 MPix sensors, create image sets
3	Interface software TP implementation with the CNN, evaluate disparity-space images from the captured sets with several CNN architectures
4	Build an MVS camera with the seven FPGA-based system boards capable of at least 10 Hz TP pre-processing for 5MPix, 2.5 Hz for 18 MPix quad images
5	Create and test RTL code for the TP and related functionality on top of existing camera FPGA code
6	Develop and test MVS camera interface with external GPU-based TensorFlow implementation

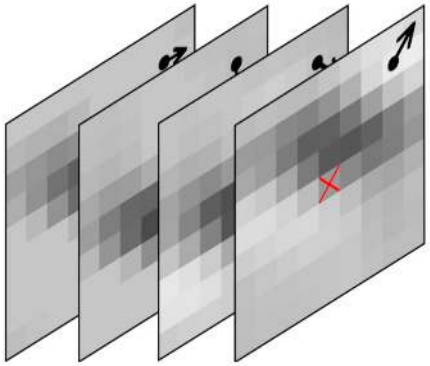
	1	2	3	4	5	6	7	8	9	10	11	12
1												
2												
3												
4												
5												
6												

Elphel is seeking collaboration with machine learning research teams working in the area of ML applications for 3-D object classification, localization and tracking.

We will demonstrate results of our current research during CVPR 2018 in Salt Lake City, Utah (June 18-22, 2018).

Preparing Space-Invariant Input Data





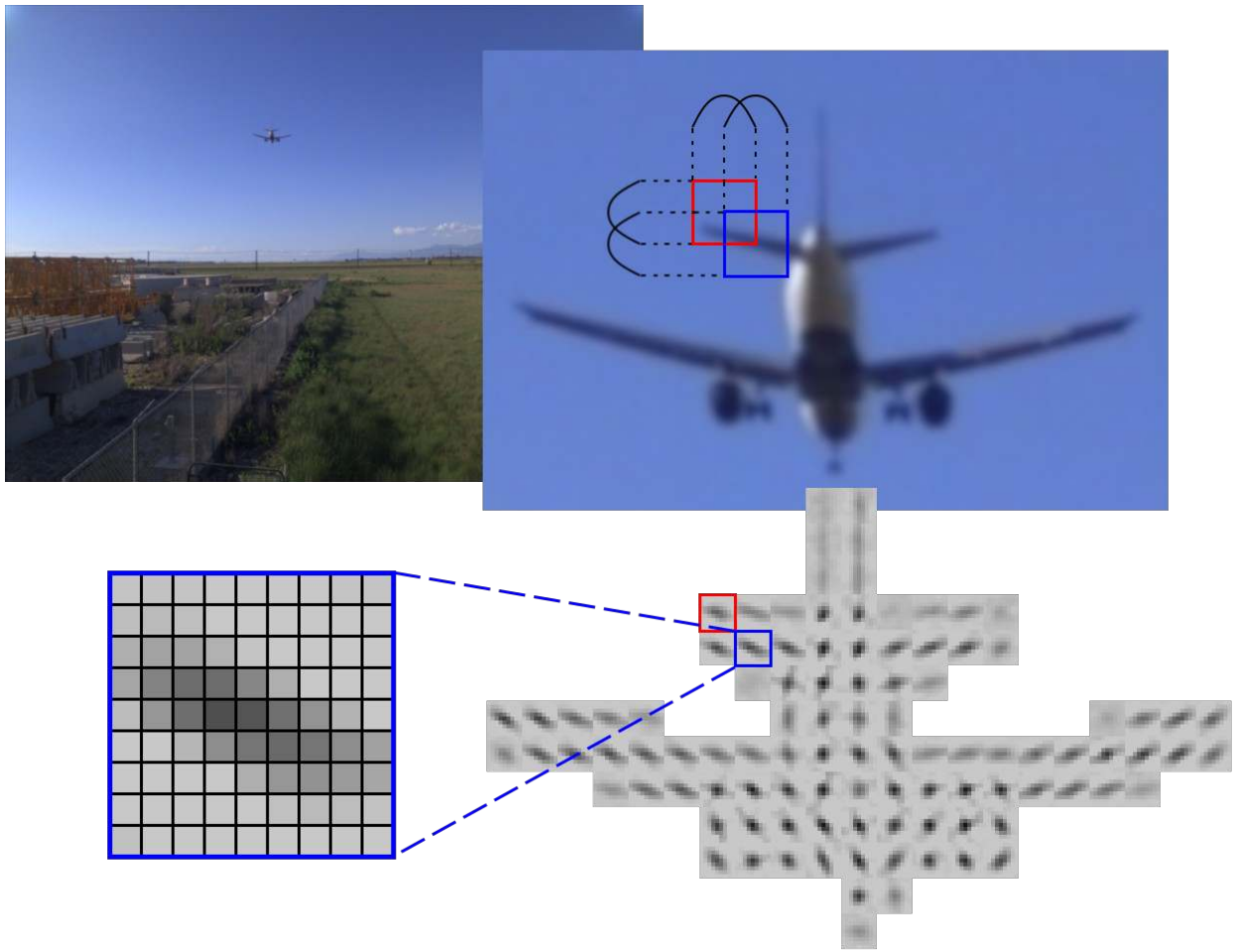
Each camera image is split into 16×16 pixel tiles (stride 8), converted to Frequency Domain (FD) and subject to space-variant deconvolution for aberration correction.

Remaining processing is *space-invariant*.

Six image pairs result in four 2D phase correlation outputs.

Each tile is pre-shifted to the estimated disparity with FD phase rotation to subpixel accuracy, similar to eye convergence in human/animal binocular vision, only the residual disparity has to be determined.

Anisotropy of the 2D Correlation



Fine textured areas result in sharp maximums, linear features (object edges) produce elongated peaks.

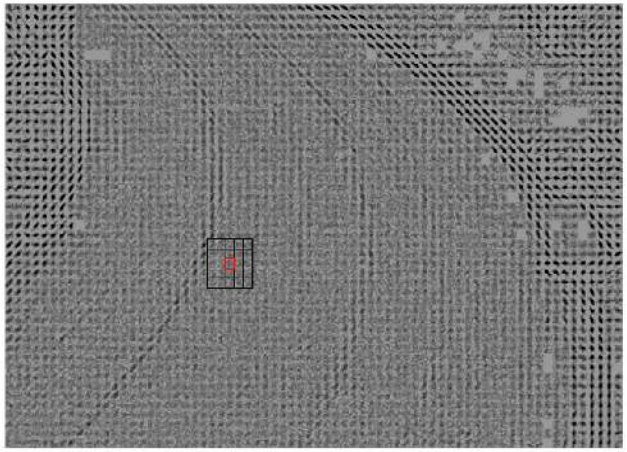
Pairs with the baseline close to orthogonal to the feature direction provide most accurate disparity (and so distance) measurements.

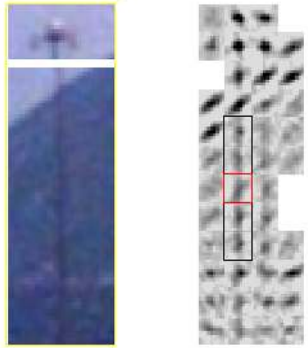
16×16 2D correlation results in 15×15 output, but most data is contained in the center area, we keep 9×9 = 81 features.

imaging solutions with free software and open hardware

elphel

Pooling Multiple Tiles Correlation Outputs



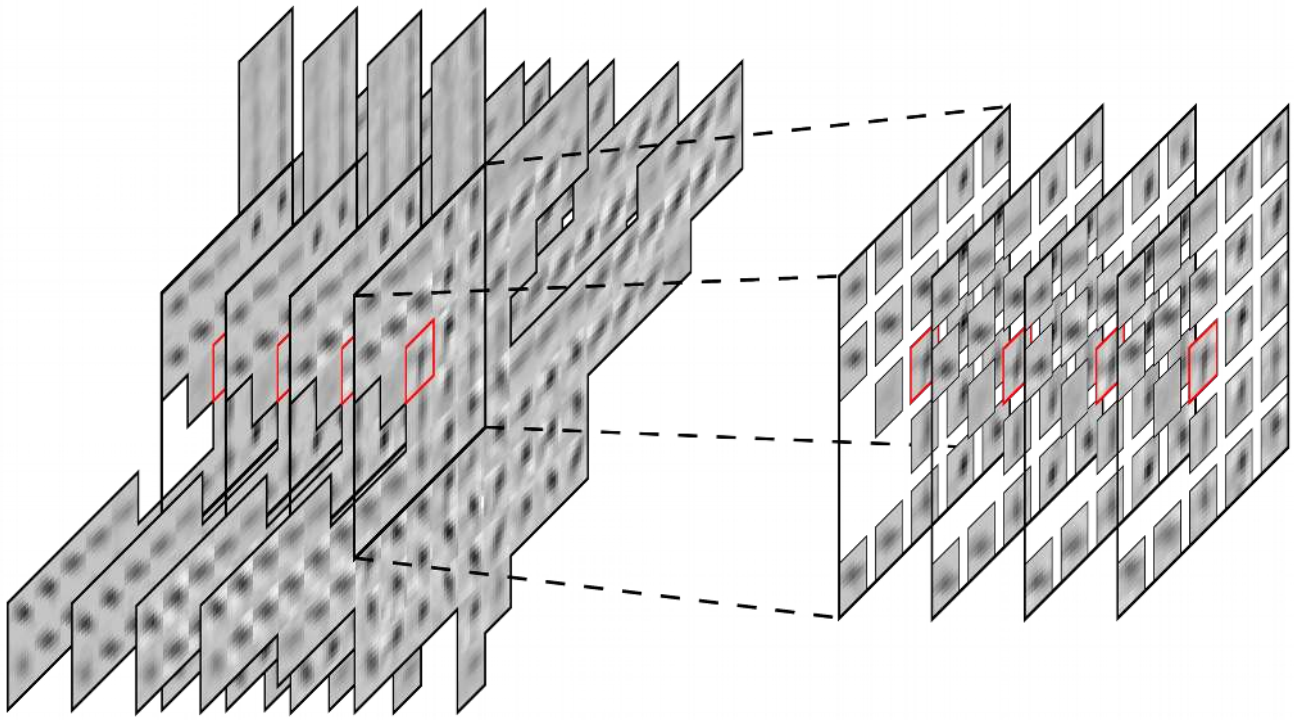


While in most image areas each tile outputs sufficient data for the disparity calculation, there are cases when neighbor tiles have to be considered too: following object edges (a) and increasing S/N for poorly textured areas (b).

a) Small disparity difference (~1 pix above) merge foreground and background peaks

b) Poorly textured areas are challenging to match. But lack of the sharp features usually means smooth in 3D surfaces so pooling of adjacent tiles is justified.

Input Features for the 5×5 Tiles Group

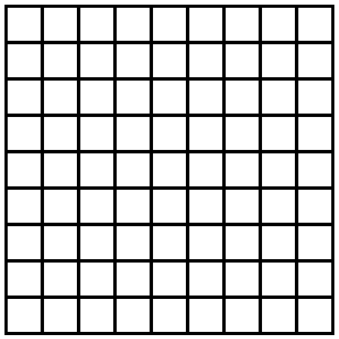


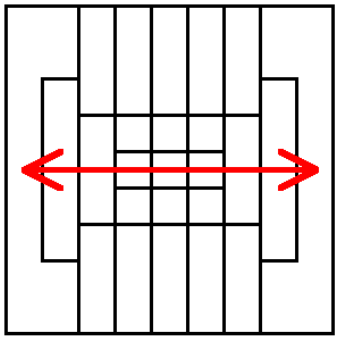
Feeding network directly with data from the tile correlation output and its reasonable (5×5) vicinity would involve 5×5×4×9x9=8100 features (plus 1 - “eye convergence”).

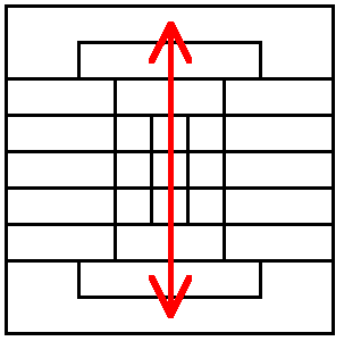
Alternative would be to process each tile in a separate subnet and then combine the outputs to exploit tiles locality.

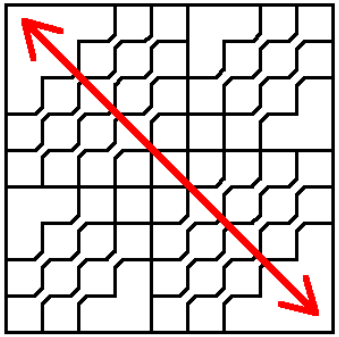
We propose a combined approach: reduce number of the input features by application-specific pooling, provide each tile subnet with the linear combination of the neighbor outputs (in addition to the local tile data) and connect tiles outputs with a convolutional network to deal with the higher level semantic 3D correspondence.

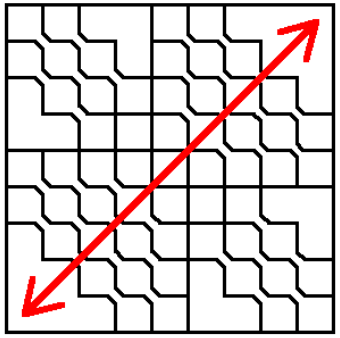
Pooling Tile Correlation Outputs









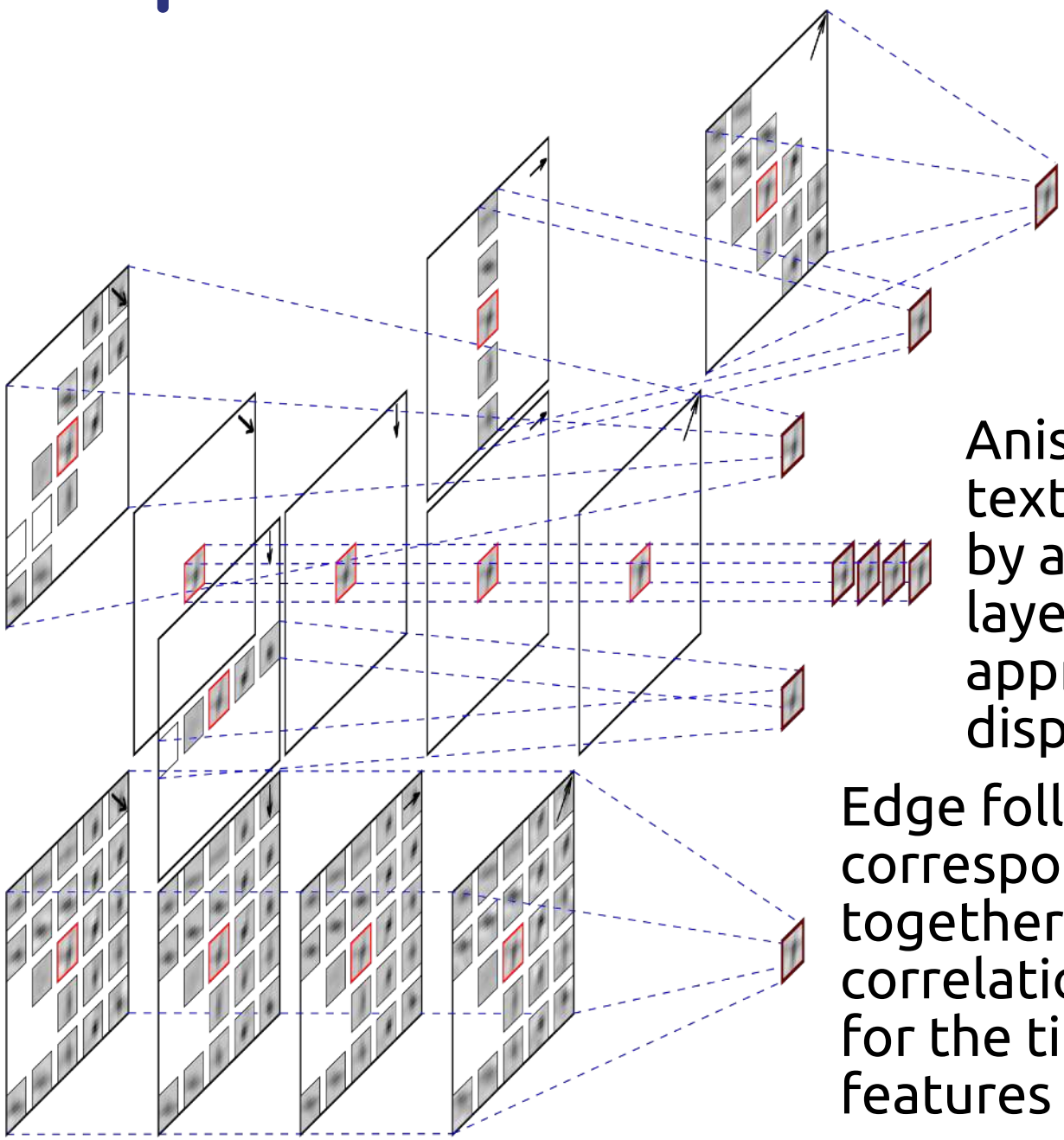


2D output from the phase correlation of the pre-shifted windowed tiles differs from the generic 2D image:

- Value of the center tiles is higher than that of the peripheral ones
- Resolution in the disparity direction is more important than in an orthogonal one

The four layers of 81 features each can be replaced with the linear combinations of them resulting in 104 features instead of the 81×4=324 of them.

Consolidating Data from Neighbor Tiles



Another significant reduction of the number of input features is achieved by providing only consolidated information from the neighbor tiles that adds to the tile itself full data.

Anisotropic data for the low-textured areas can be produced by averaging all 4 directional layers of all 25 tiles (with appropriate rotation to match disparity direction).

Edge following requires only the corresponding directional layer, so all together neighbors add 5 averaged correlation layers to the 4 individual for the tile itself with less than 250 features total (instead of the 8100).