Stereo Vision: Algorithms and Applications

Stefano Mattoccia

DEIS
University of Bologna

stefano.mattoccia@unibo.it
www.vision.deis.unibo.it/smatt

Updates

- Added details about our stereo camera with FPGA processing
- November 21, 2011: added experimental results for "Linear stereo matching" (ICCV2011), Min et al's algorithm (ICCV2011), description of "Fast Segmentation driven (FSD)" (IC3D) algorithm and description of SGM
- May 19, 2011: added experimental results of FBS on the GPU [71] and the VisionSt stereo camera
- July 25, 2010: Linux and Windows implementations of the Fast Bilateral Stereo algorithm available at: www.vision.deis.unibo.it/smatt/fast_bilateral_stereo.htm
- April 20th, 2010: included descriptions and experimental results for papers [67], [68], [69]

The latest version of this document is available here:

http://www.vision.deis.unibo.it/smatt/Seminars/StereoVision.pdf

Outline

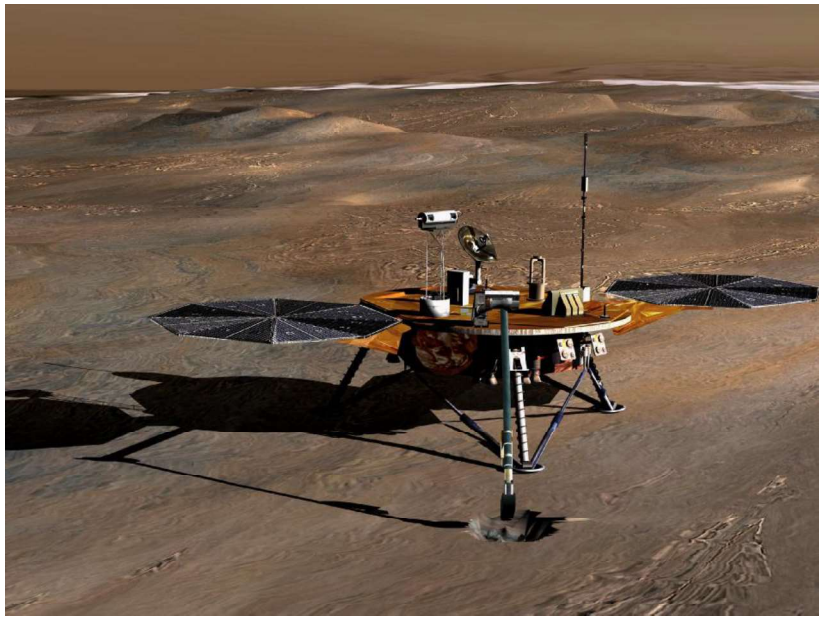
- Introduction to stereo vision
- Overview of a stereo vision system
- Algorithms for visual correspondence
- Computational optimizations
- Hardware implementation
- Applications

What is stereo vision ?

- Is a technique aimed at inferring depth from two or more cameras
- Wide research topic in computer vision
- This seminar is concerned with
  - binocular stereo vision systems
  - dense stereo algorithms
  - stereo vision applications
- Emphasis is on approaches that are (or might be hopefully soon) feasible for real-time/hardware implementation

Applications

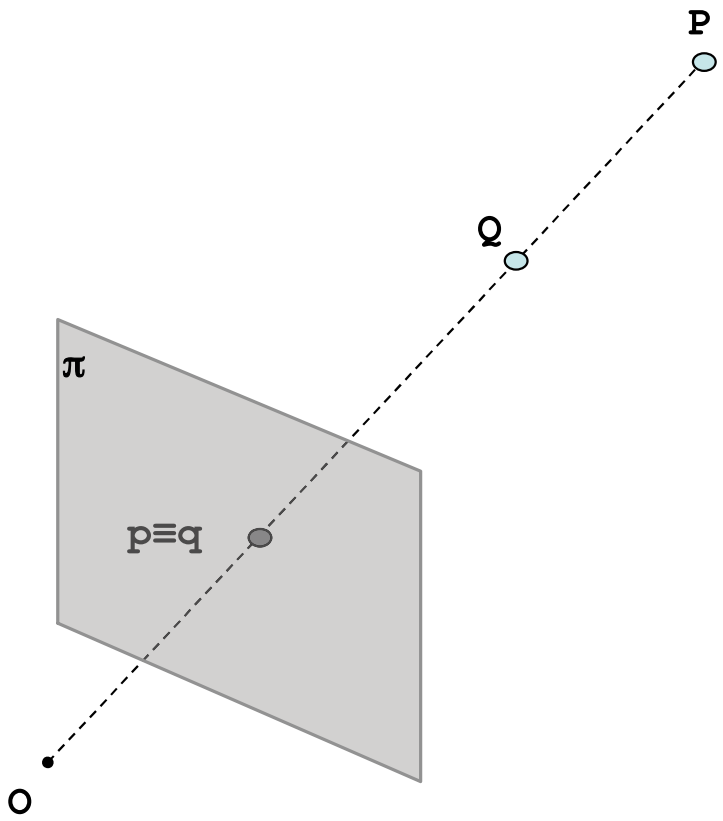
www.nasa.gov



www.nasa.gov

www.vision.deis.unibo.it/smatt/stereo

Single camera



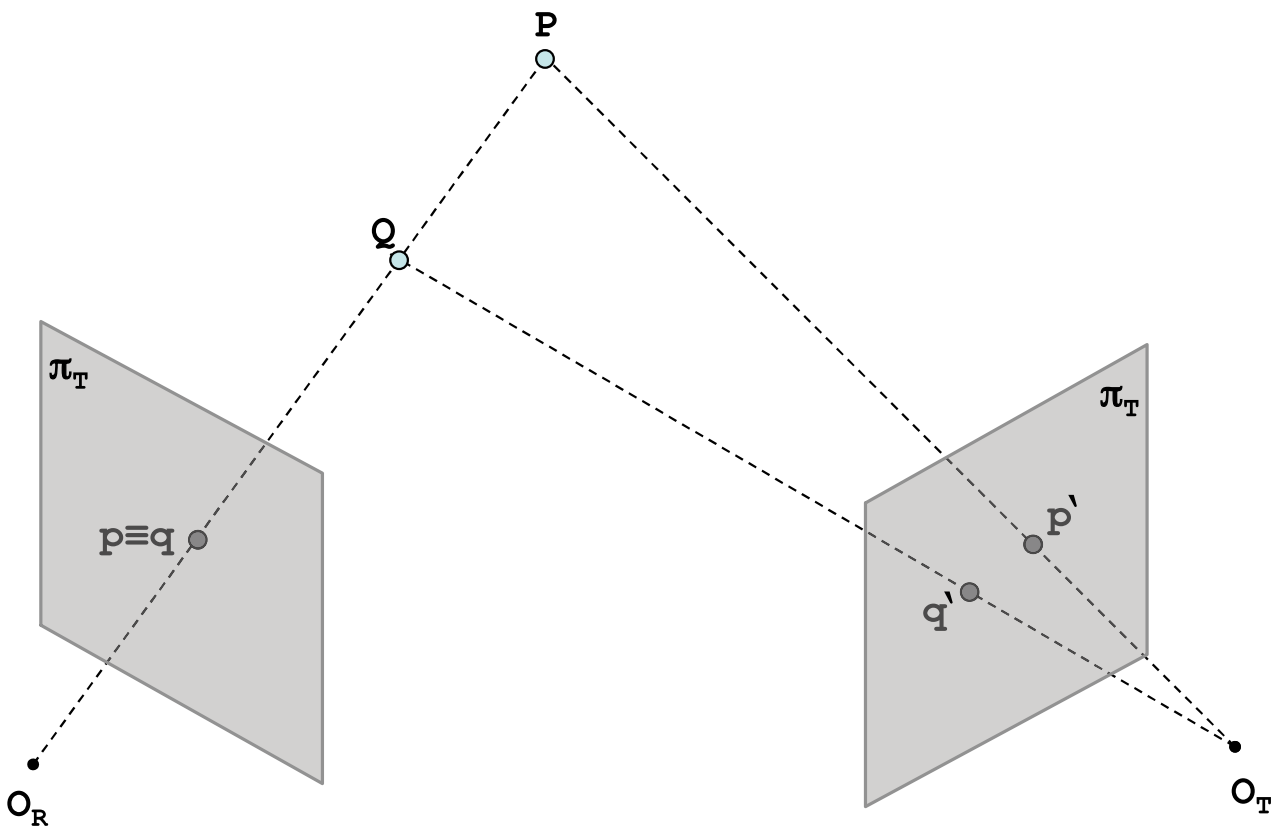
- Both (real) points (P and Q) project into the same image point (p ≡ q)
- This occurs for each point along the same line of sight
- Useful for optical illusions…

Courtesy of http://www.coolopticalillusions.com/

π: image plane

O: optical center

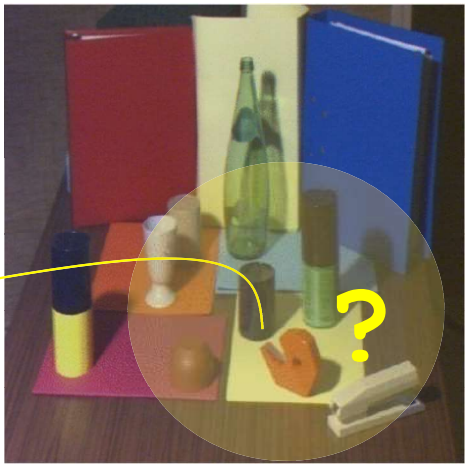
Stereo camera



With two (or more) cameras we can infer depth, by means of triangulation, if we are able to find corresponding (homologous) points in the two images

How to solve the correspondence problem ?

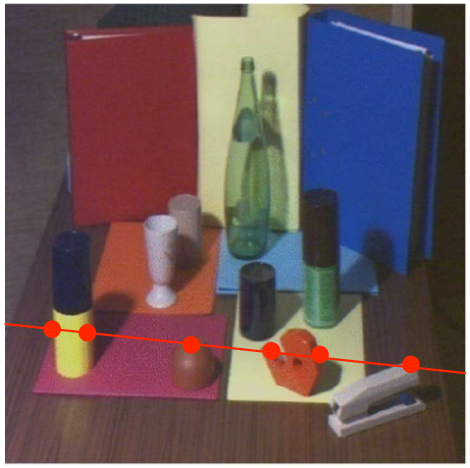
Reference (R)



Target (T)

2D search domain ?

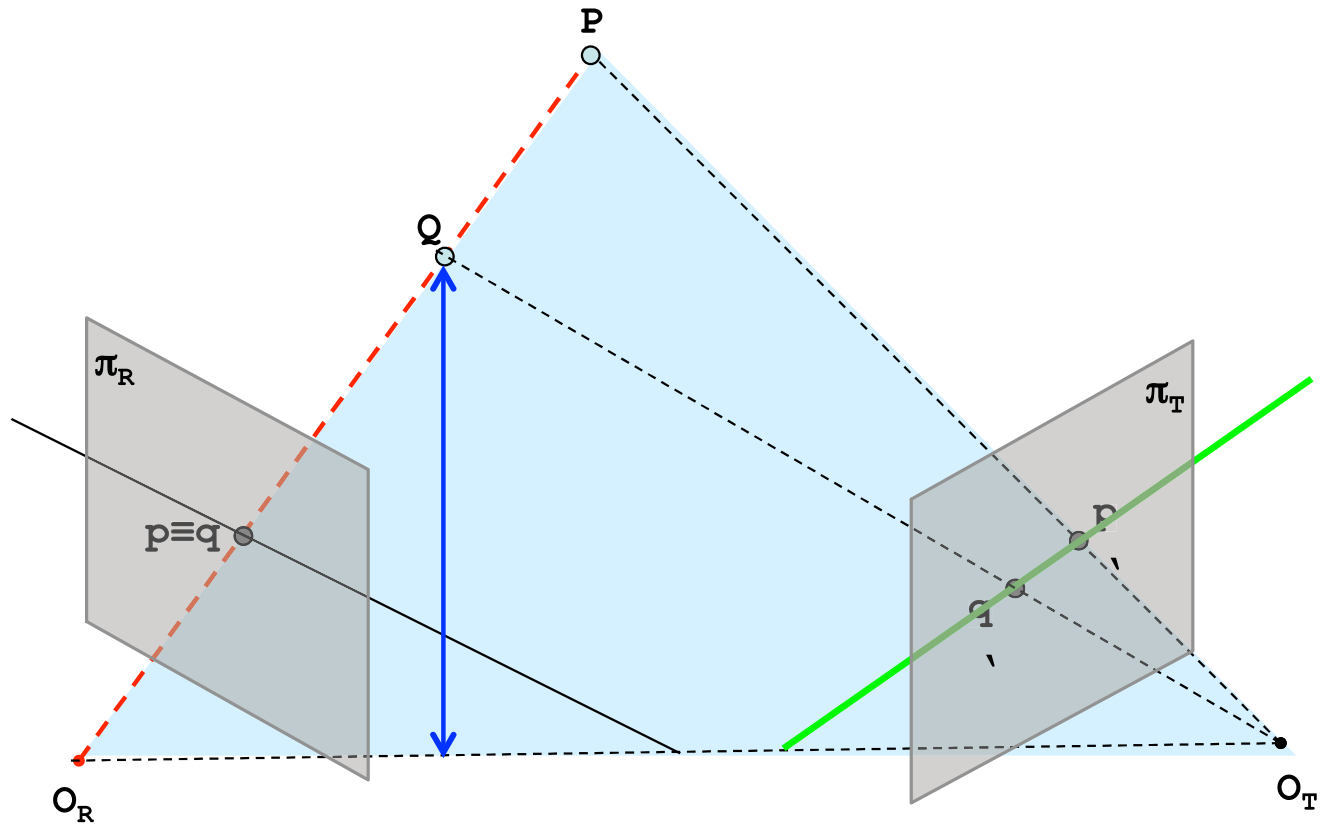
Reference (R)



Target (T)

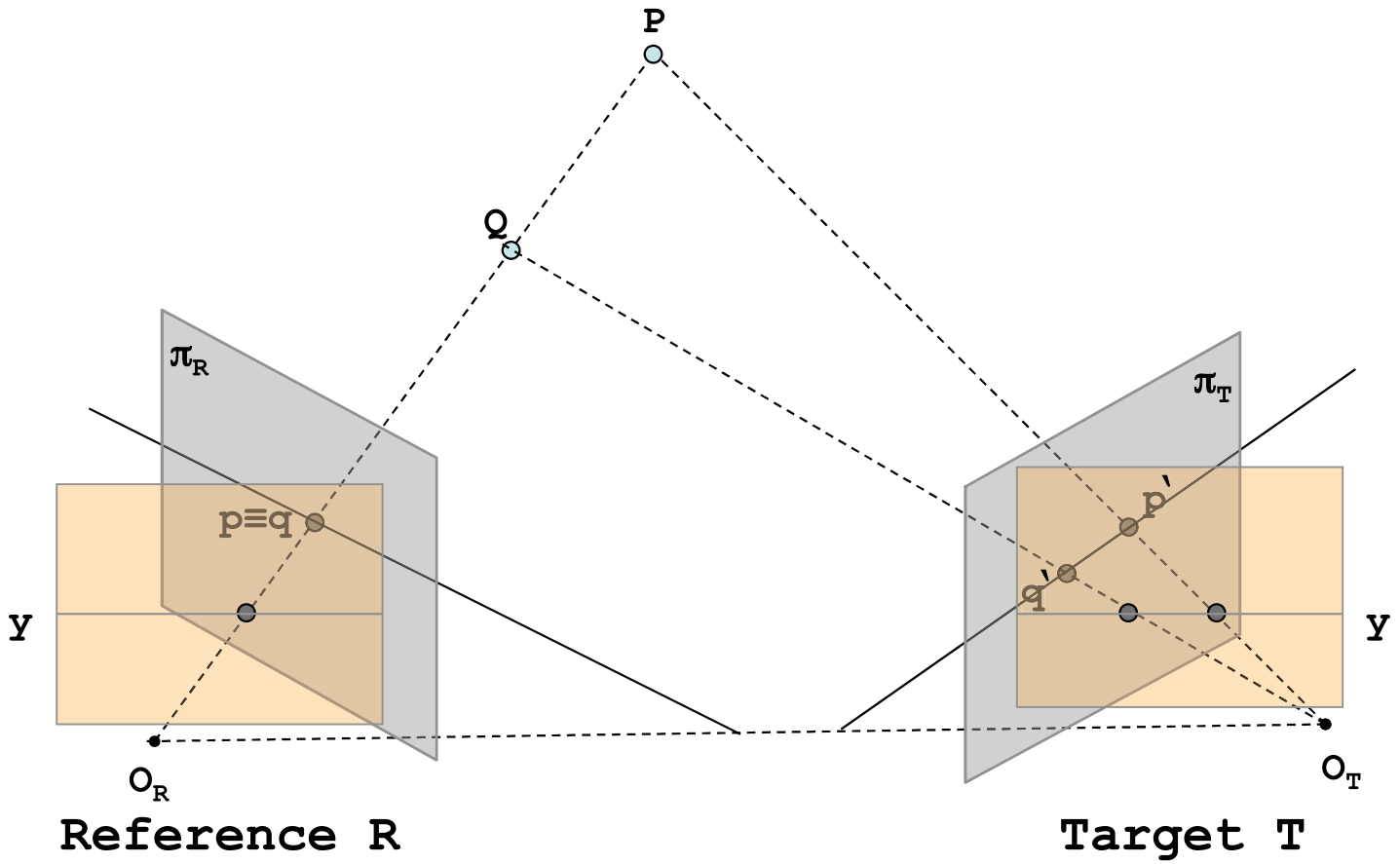
No!! Thanks to the epipolar constraint

Epipolar constraint

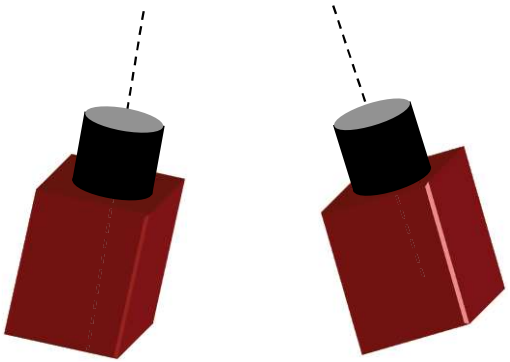


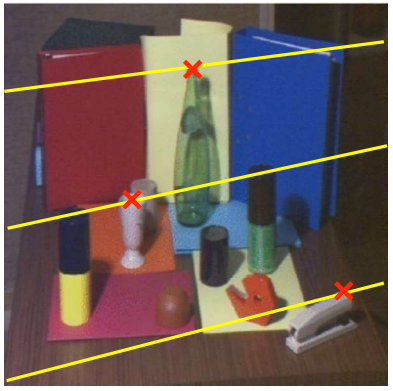
- Consider two points P and Q on the same line of sight of the reference image R (both points project into the same image point p≡q on image plane π_R of the reference image)
- The epipolar constraint states that the correspondence for a point belonging to the (red) line of sight lies on the green line on image plane π_T of target image

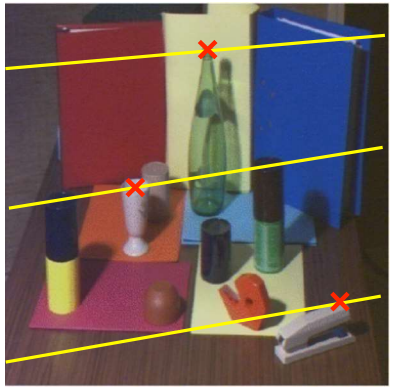
Stereo camera in standard form



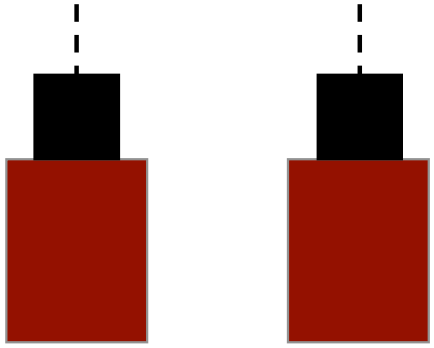
Once we know that the search space for corresponding points can be narrowed from 2D to 1D, we can put (virtually) the stereo rig in a more convenient configuration (standard form) - corresponding points are constrained on the same image scanline

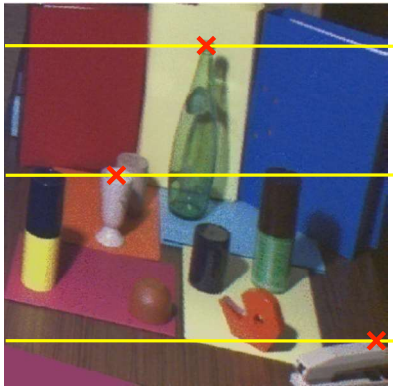


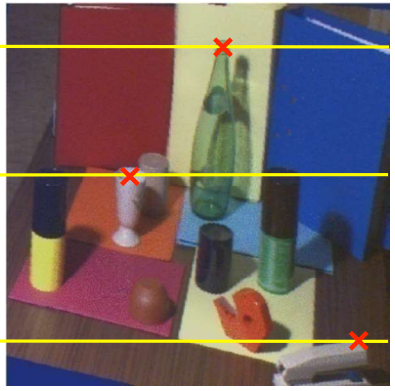




Original stereo pair



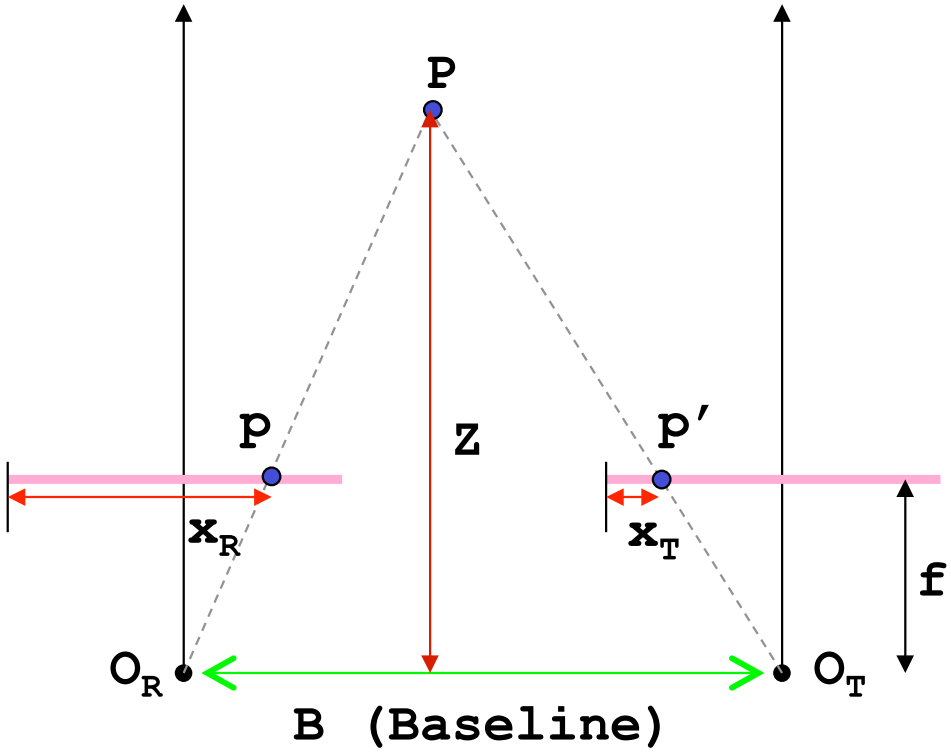




Stereo pair in standard form

Cameras are "*perfectly*" aligned
and with the same *focal length*

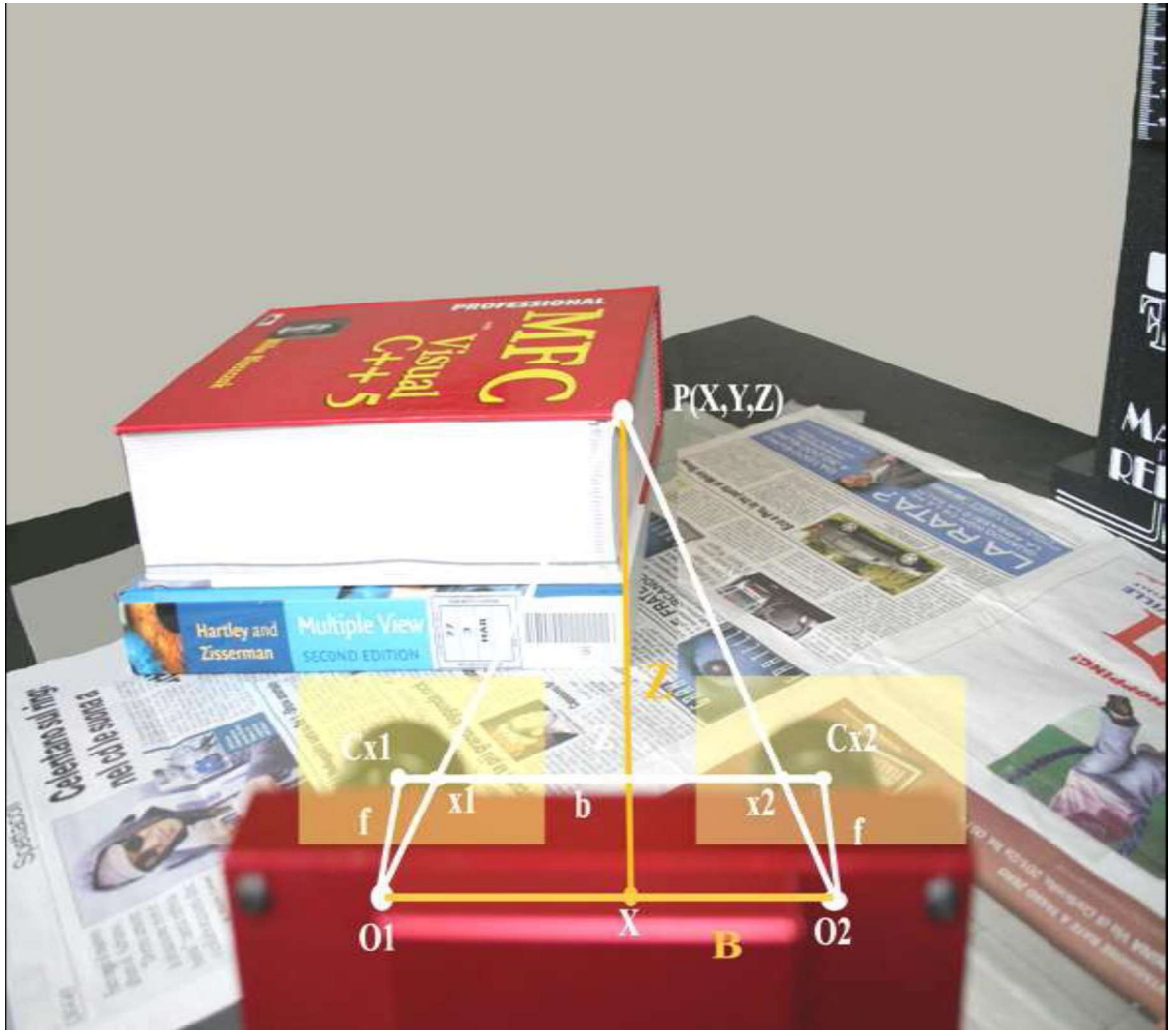
Disparity and depth



With the stereo rig in standard form and by considering similar triangles (PO_RO_T and Ppp'):

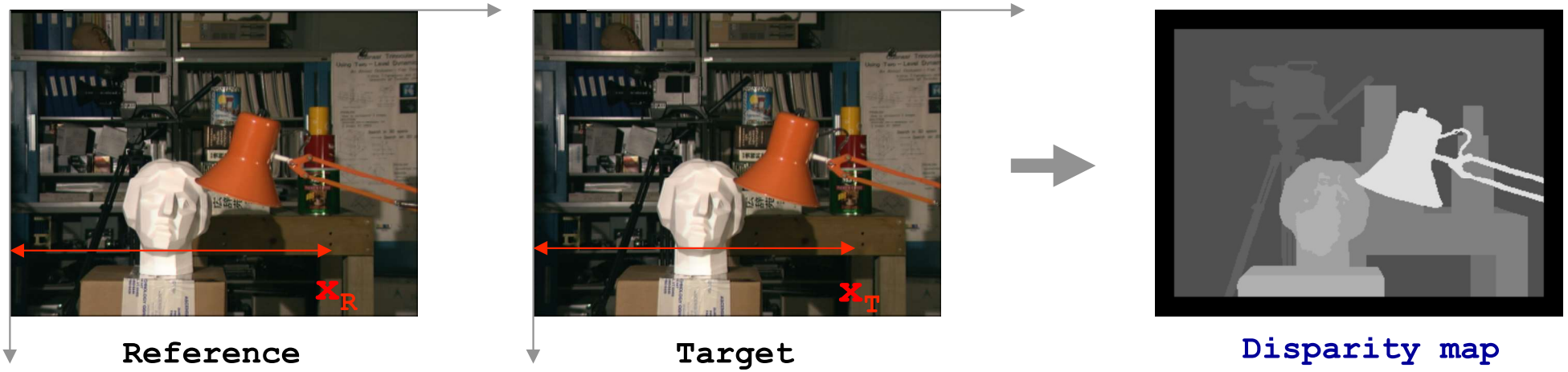
$$\frac{b}{Z} = \frac{(b + x_T) - x_R}{Z - f} \rightarrow Z = \frac{b \cdot f}{x_R - x_T} = \frac{b \cdot f}{d}$$

$x_R - x_T$ is the **disparity**

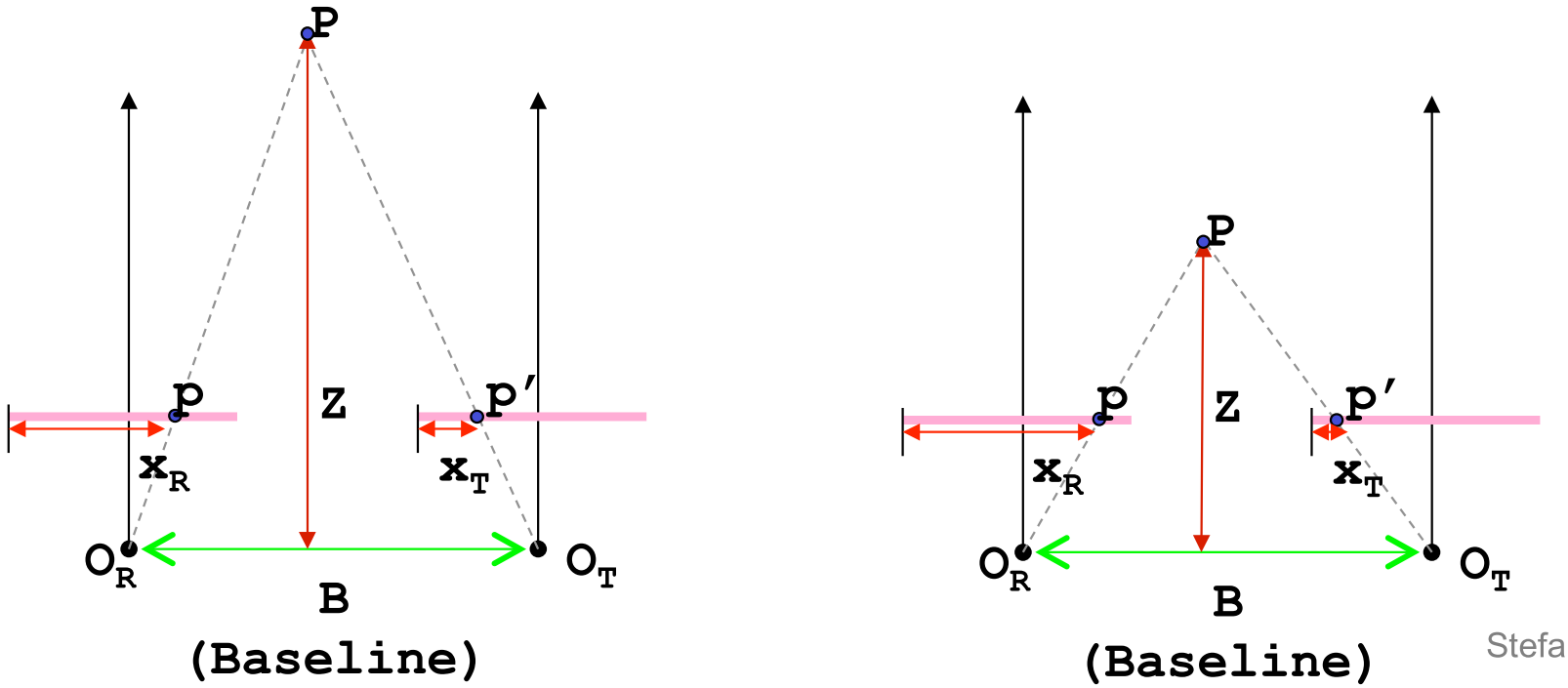


Disparity and depth

The disparity is the difference between the x coordinate of two corresponding points; it is typically encoded with greyscale image (closer points are brighter).

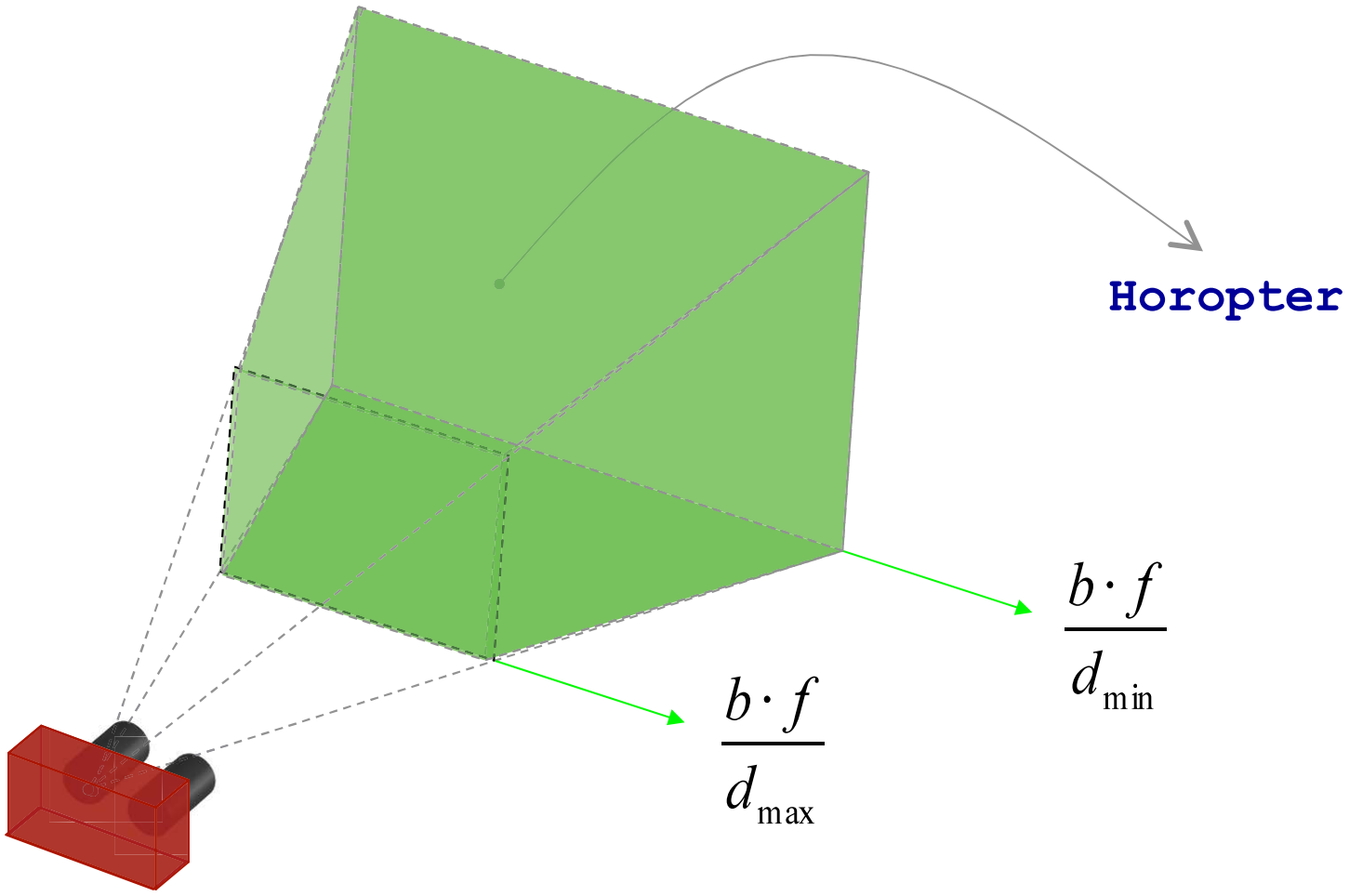


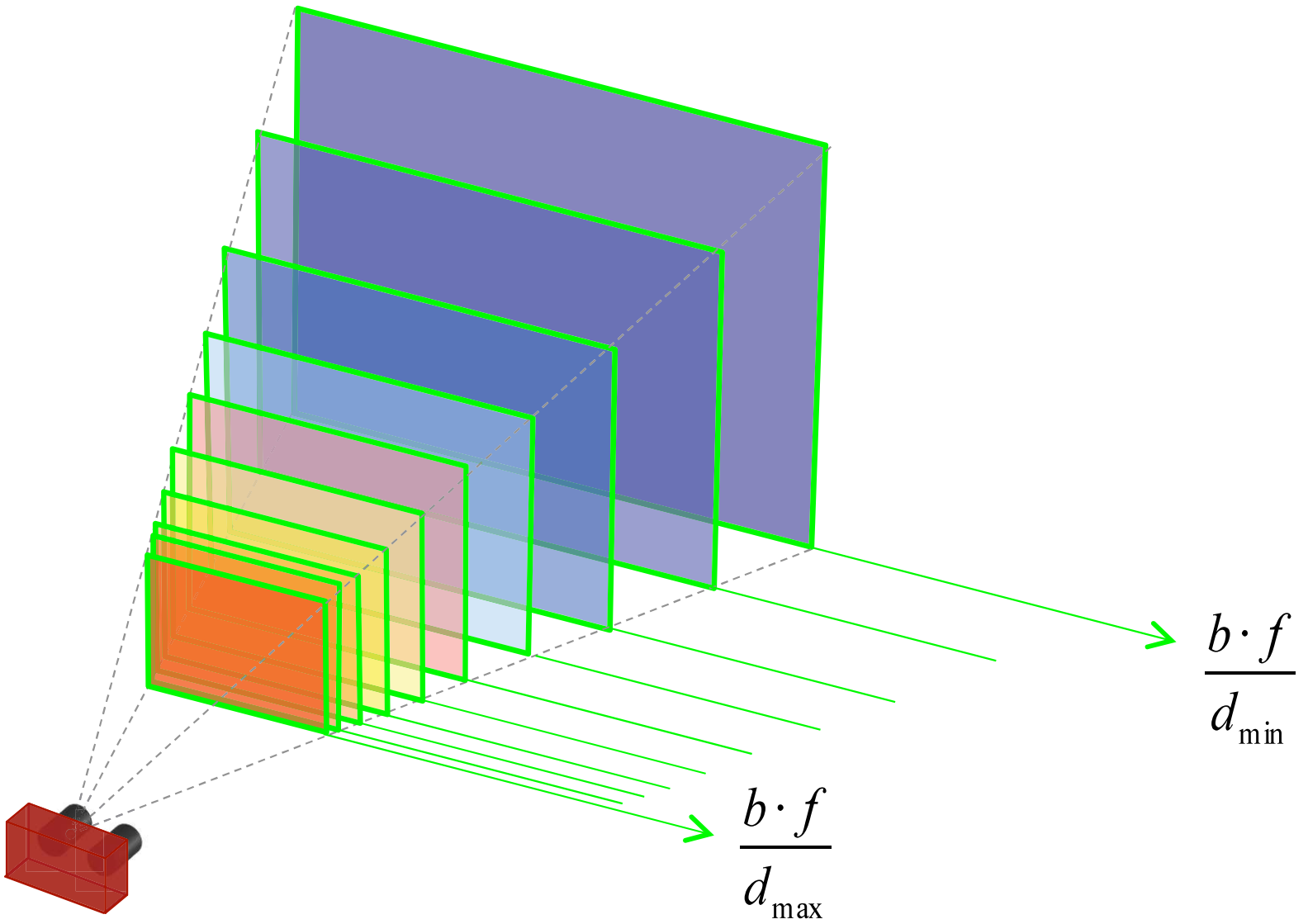
Disparity is higher for points closer to the camera



Range field (Horopter)

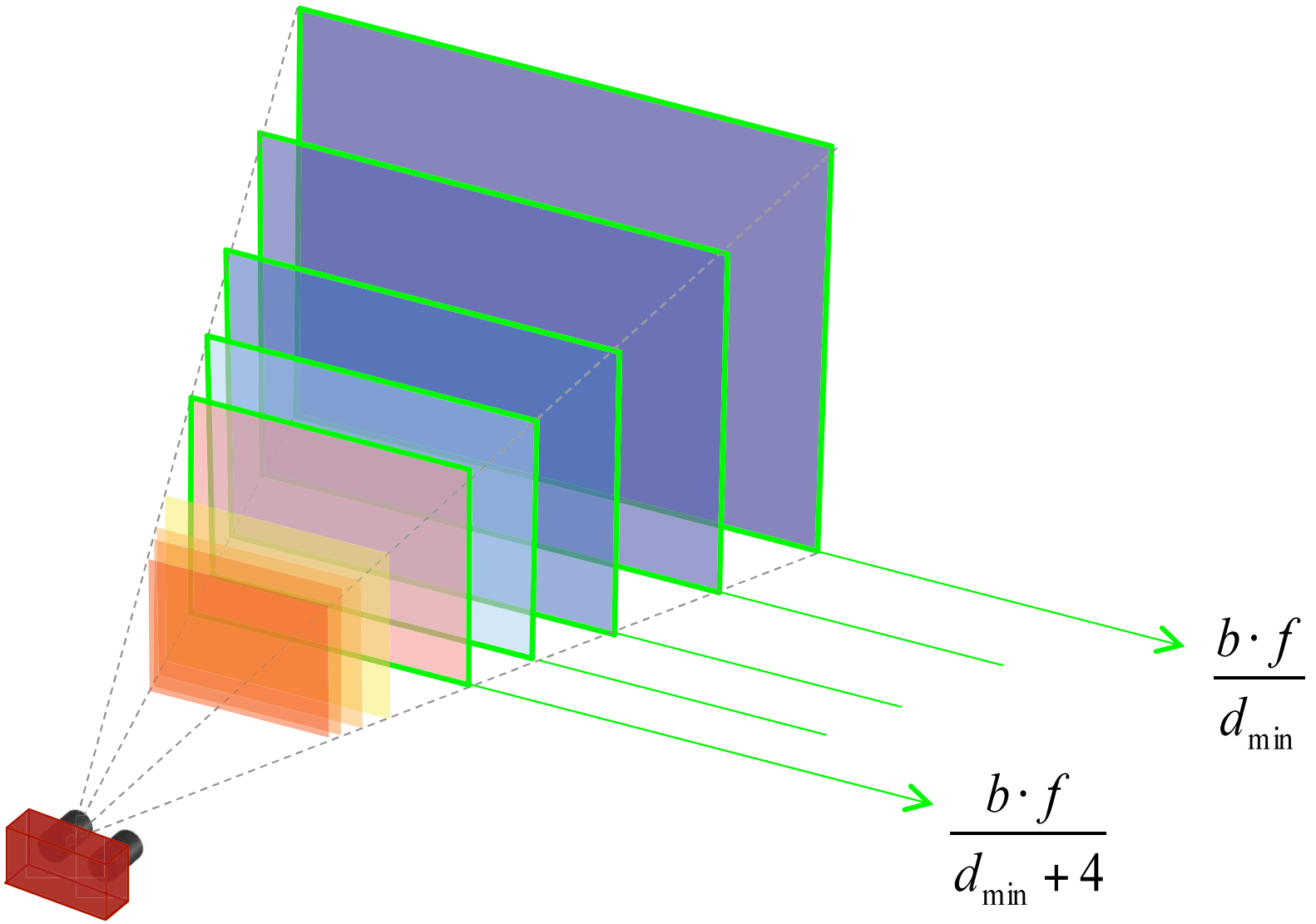
Given a stereo rig with baseline b and focal length f, the range field of the system is constrained by the disparity range $[d_{min}, d_{max}]$.



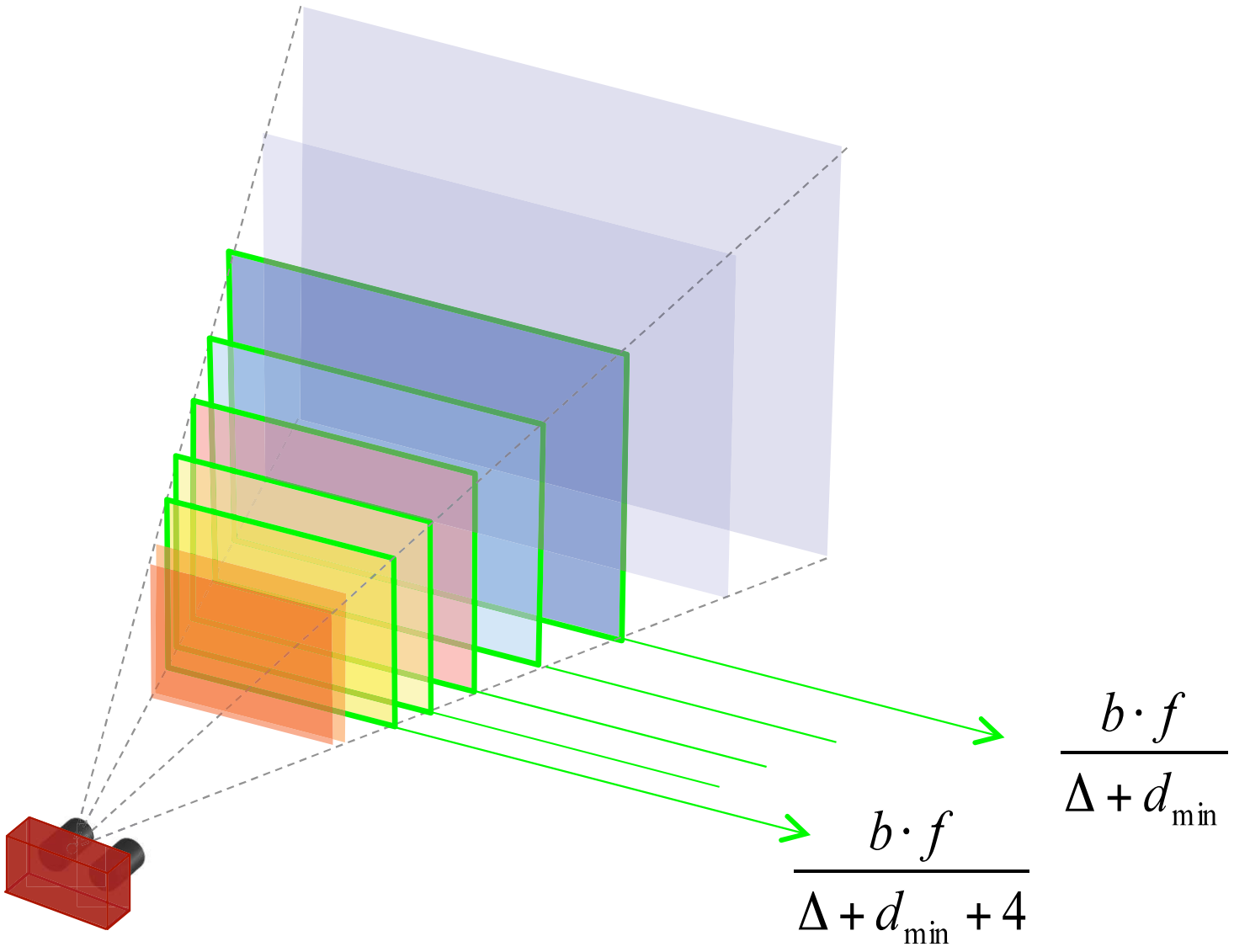


- **Depth measured by a stereo vision system is discretized into parallel planes (one for each disparity value)**

- **A better (virtual) discretization can be achieved with subpixel techniques (see Disparity Refinements)**



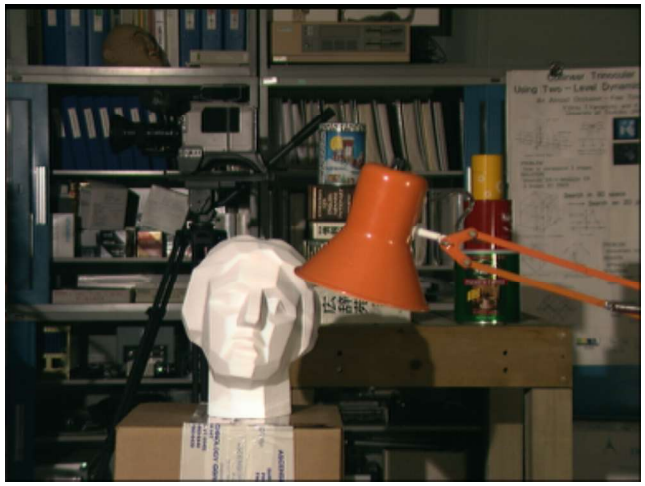
- **The range field (horopter) using 5 disparity values [d_{min}, d_{min}+4]**



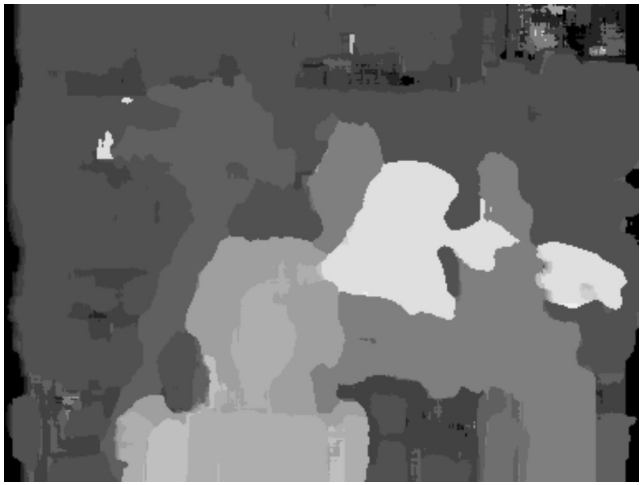
- **Using 5 disparity values [$\Delta+d_{min}$,$\Delta+d_{min}+4$]**
- **With $\Delta>0$, *horopter* gets closer and shrinks (depth and obviously area/volume)** horopter đến gần hơn và co lại

Key module in stereo vision?

The algorithm is crucial in this technology

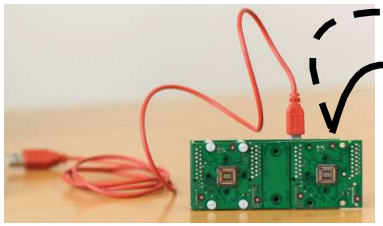


Traditional algorithm



State of the art (ICCV 2011)

Overview of a stereo vision system



Calibration (offline)

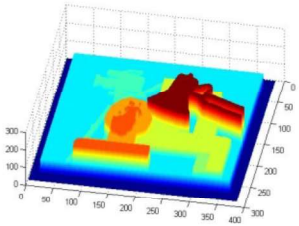
Intrinsic and extrinsic parameters

Stereo pair

Rectified stereo pair

Disparity map

Depth map



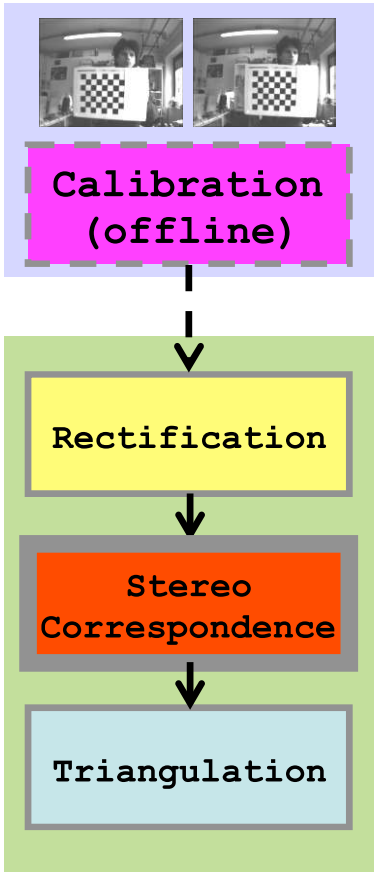
Rectification

Stereo Correspondence

Triangulation

PC, FPGA

Calibration (offline)



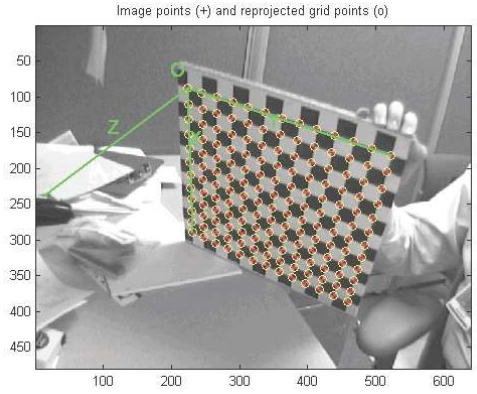
Offline procedure aimed at finding:

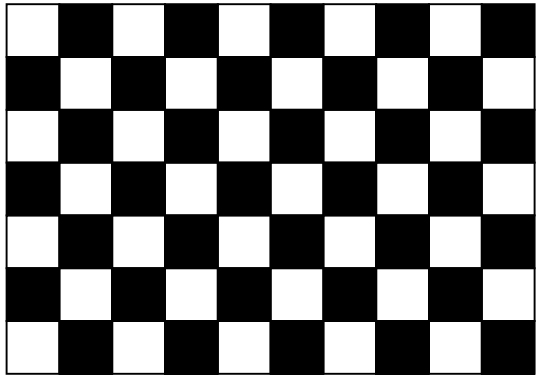
- **Intrinsic parameters of the two cameras (focal length, image center, parameters of lenses distortion, etc)**

  Thông số nội tại của 2 cameras (focal length, image center, parameters of lenses distortion, etc)

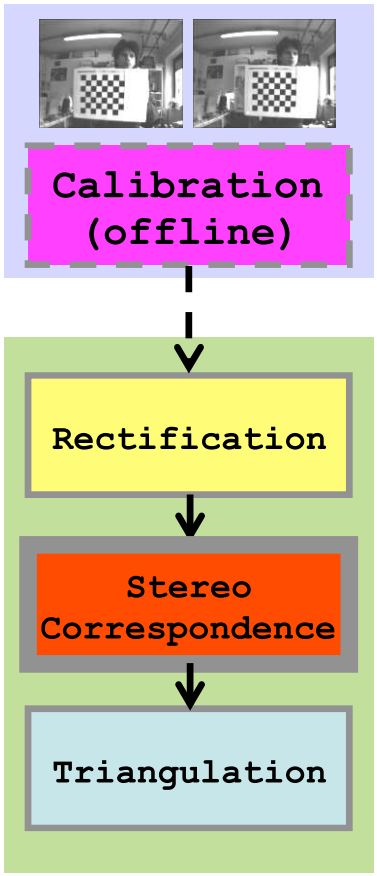
- **Extrinsic parameters (R and T that aligns the two cameras)**

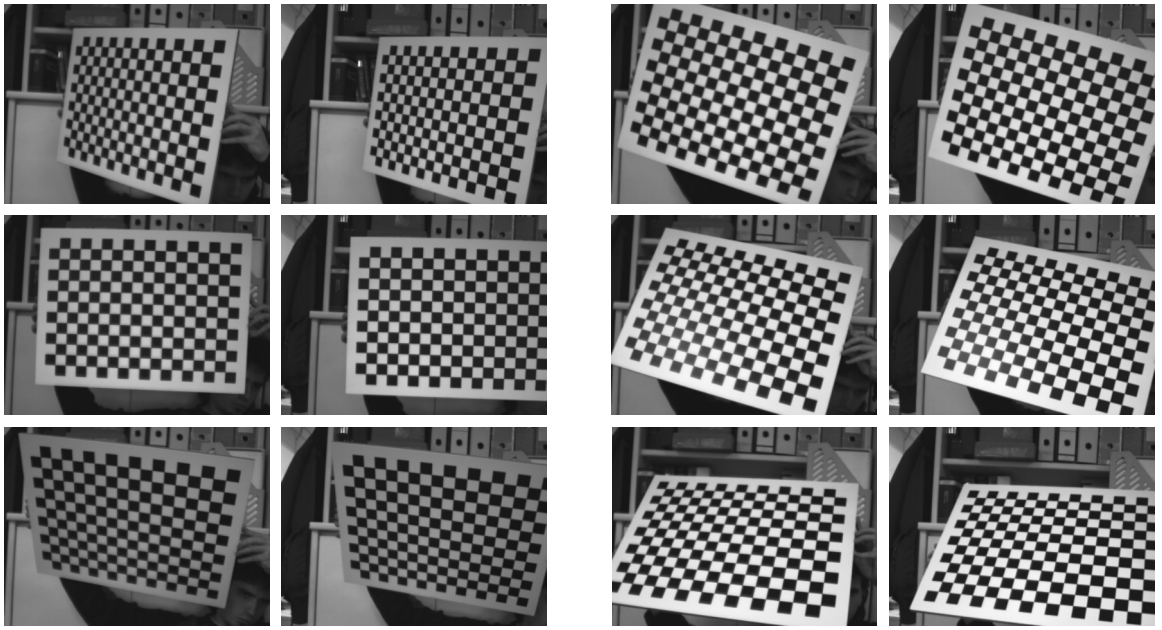
  Thông số ngoại tại (R và T)





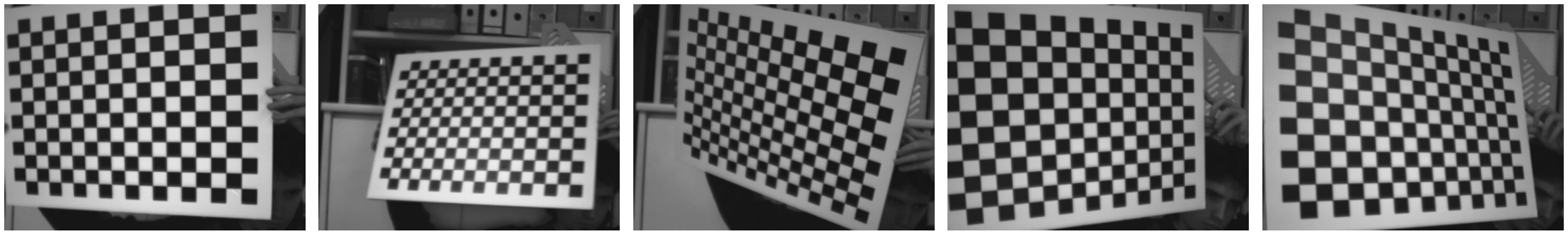
Calibration is carried out acquiring and processing 10+ stereo pairs of a known pattern (typically a checkerboard)



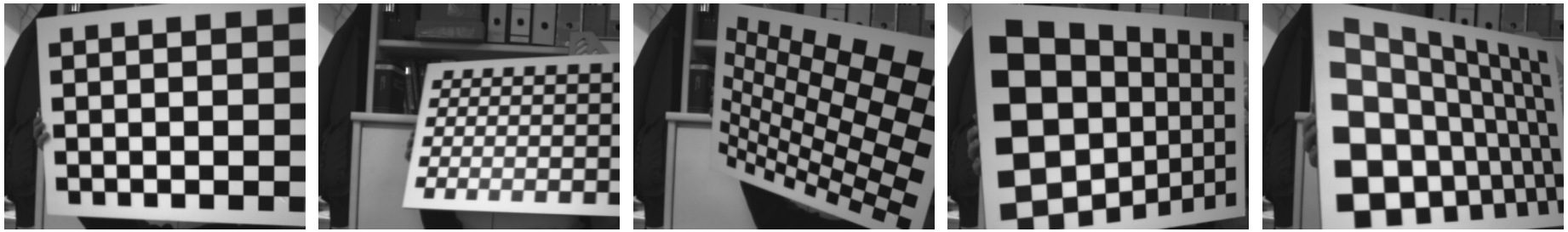


- Calibration is available in OpenCV [39] and Matlab [40]
- A detailed description of calibration can be found in [20,21,22]
- Next slides show 20 stereo pairs used for calibrating a stereo camera

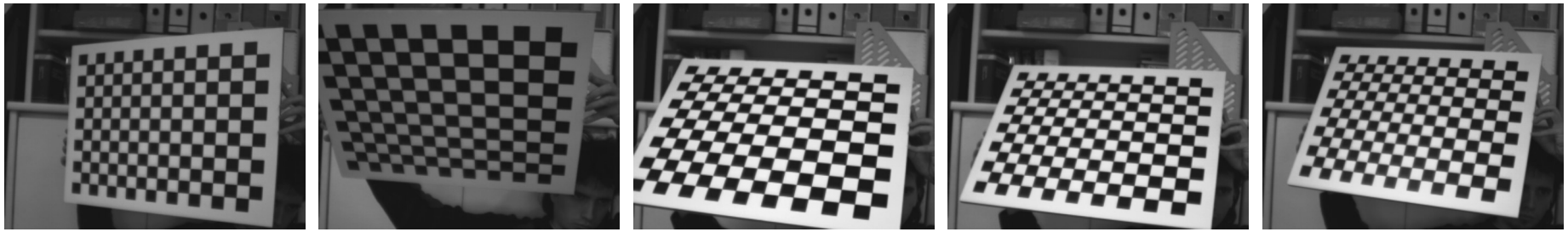
R



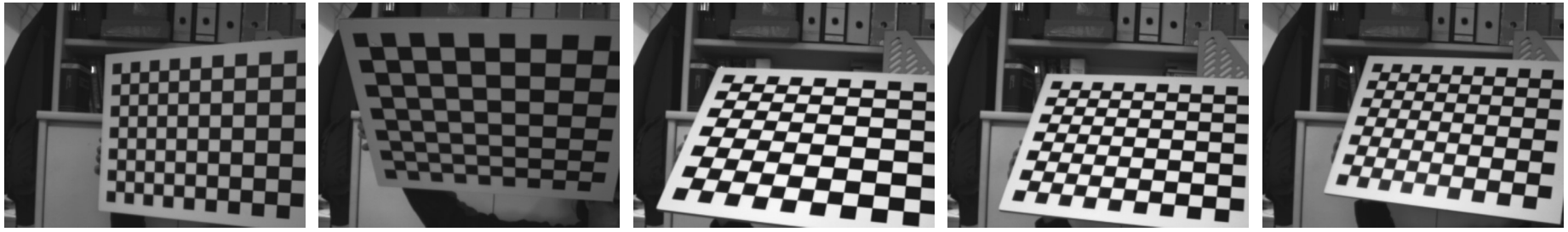
T



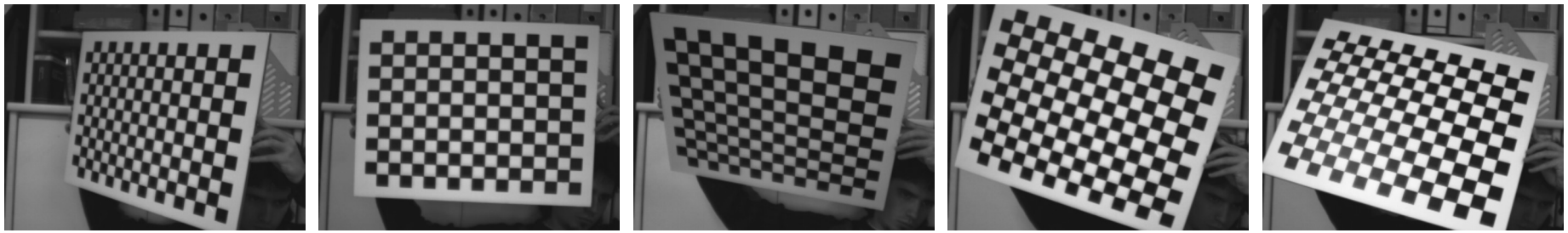
R



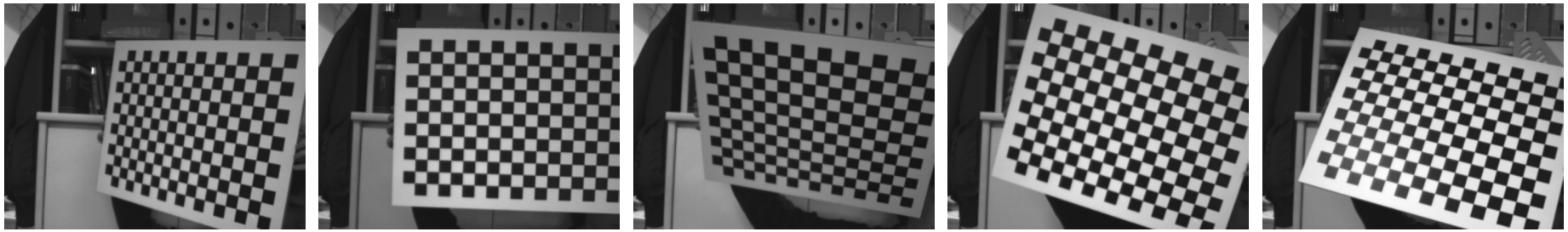
T



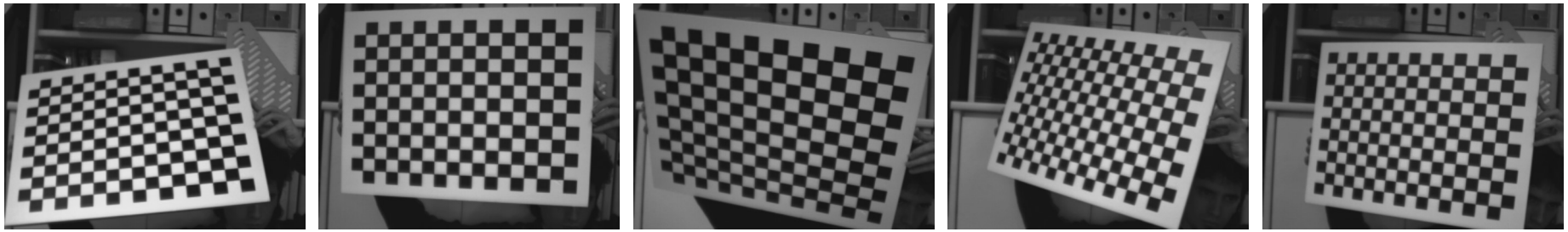
R



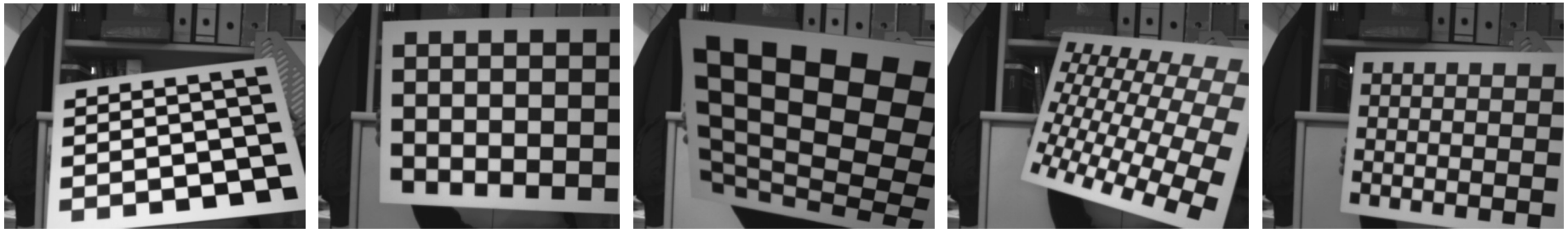
T



R



T



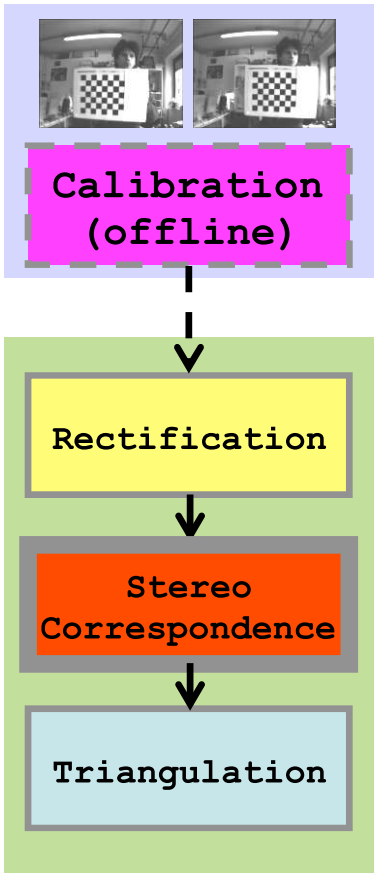
Rectification

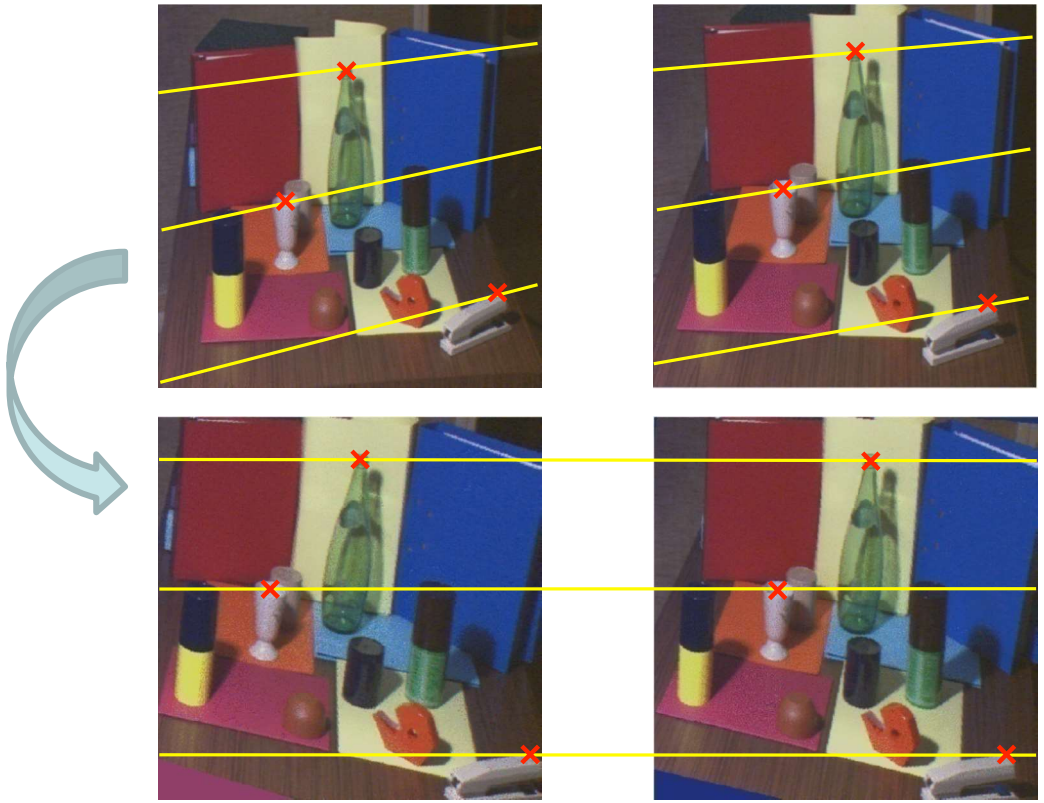
Sự cải chỉnh

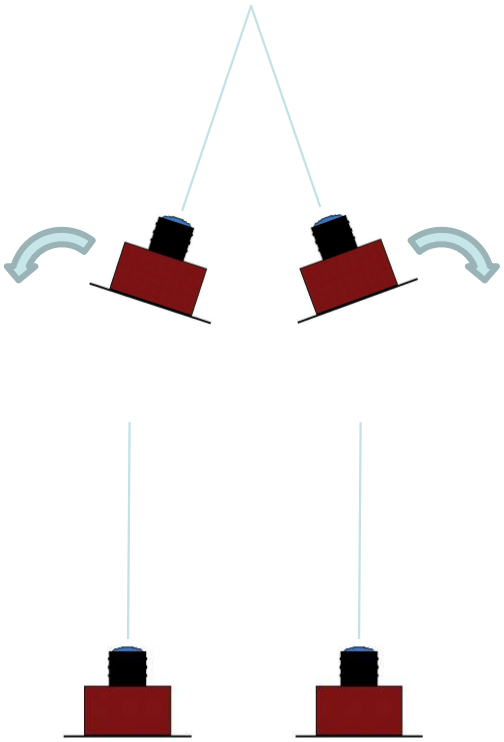
Using the information from the calibration step:

a) removes lens distortions loại bỏ sự biến dạng ống kính

b) turns the stereo pair in standard form

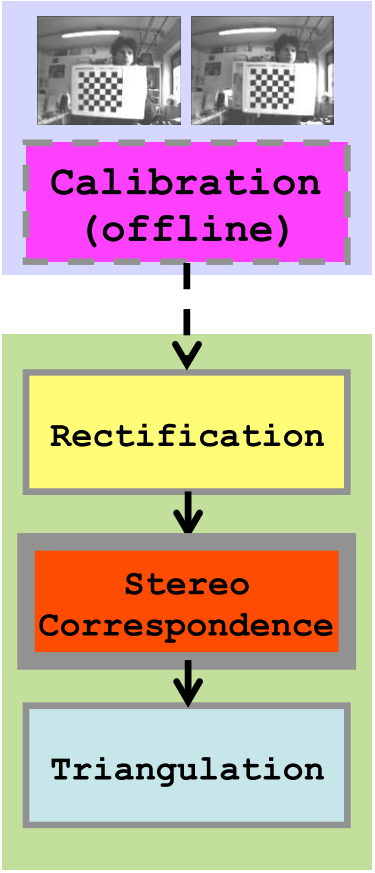




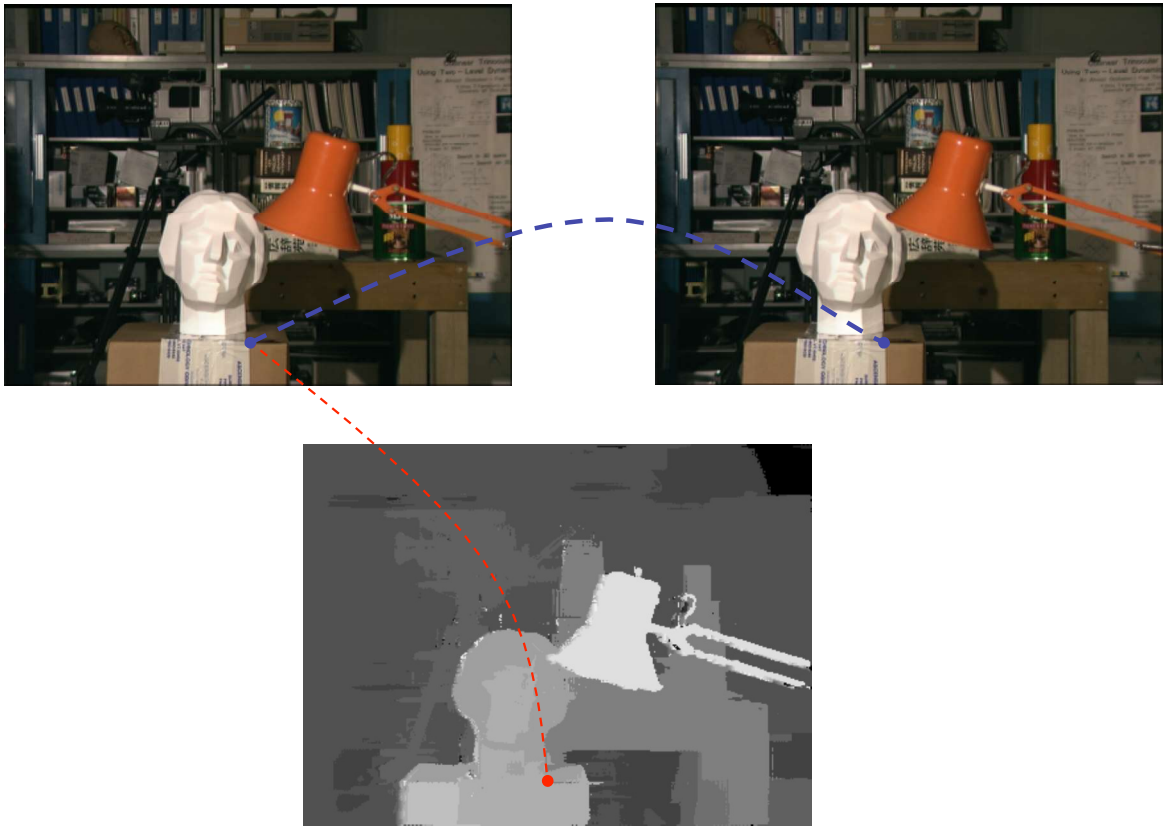


Stereo camera in standard form

Stereo correspondence



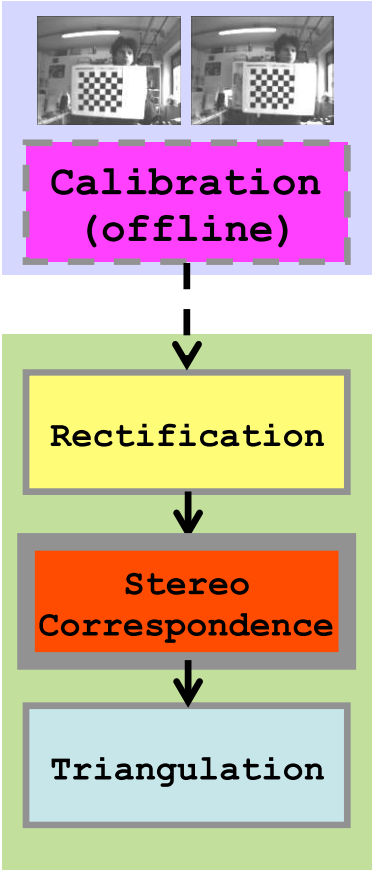
Aims at finding homologous points in the stereo pair. homologous points: điểm tương đồng



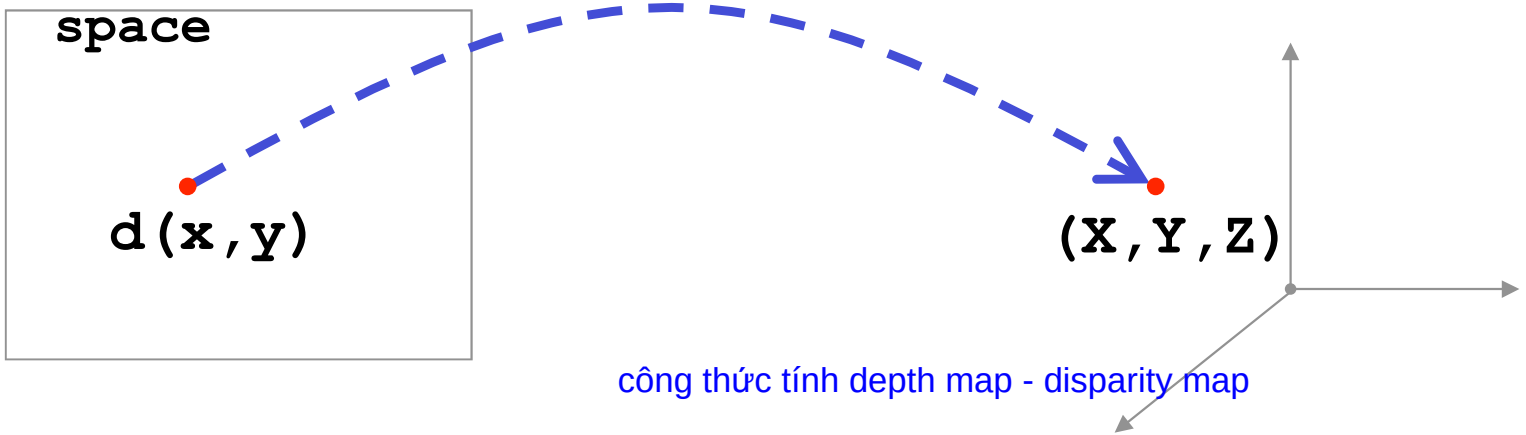
disparity map bản đồ chênh lệch

This topic will be extensively analyzed in the next slides...

Triangulation



Given the disparity map, the baseline and the Focal length (calibration): triangulation computes the position of the correspondence in the 3D space

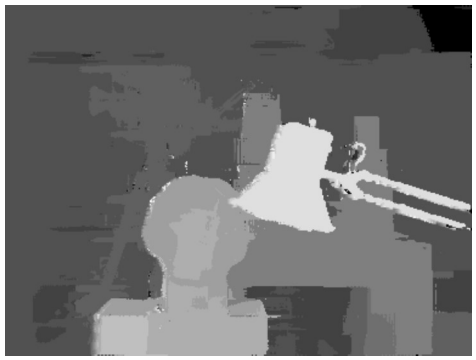


công thức tính depth map - disparity map

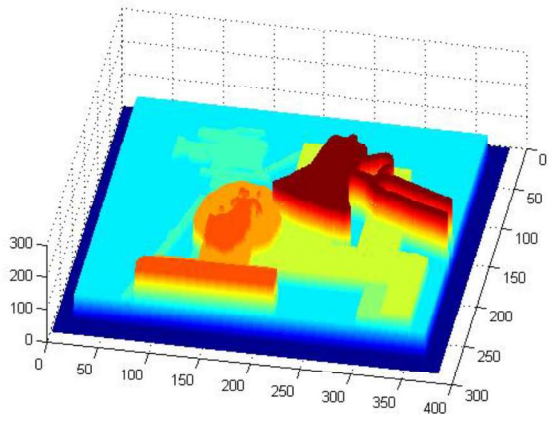
$$Z = \frac{b \cdot f}{d}$$

$$X = Z\frac{x_R}{f}$$

$$Y = Z\frac{y_R}{f}$$



disparity map



depth map

Datasets: stereo sequences

Sequences acquired with stereo cameras are available at:

http://www.vision.deis.unibo.it/smatt/stereo.htm

The datasets include:

- **calibration parameters**
- **original sequences**
- **rectified sequences**
- **disparity maps**

Architectures

- Microprocessors
  - Floating Point (FP) units + SIMD
  - C/C++ (+ assembly)
  - power,cost and size are the main drawbacks

- Low power & low cost processor
  - C/c++
  - no FP
  - no SIMD (often)

- GPUs (Graphic Processing Units)
  - raw power
  - high power dissipation and cost
  - programming is difficult (CUDA and OpenCL help)

- FPGA (Field Programmable Gate Array)
  - efficient, low power (<1 W), low cost
  - programming language: VHDL
  - coding is difficult and tailored for specific devices

Commercial (binocular) stereo cameras

www.videredesign.com

www.visionst.com

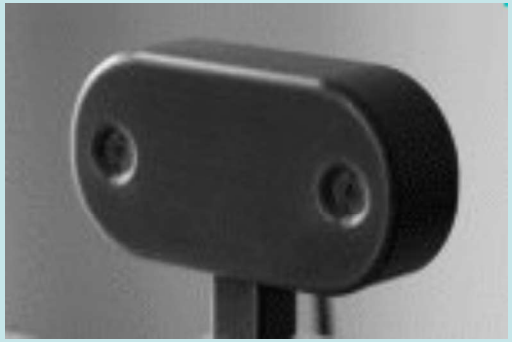
www.valdesystems.com

www.focusrobotics.com

FPGA/ASIC
DSP

www.tyzx.com

www.ptgrey.com

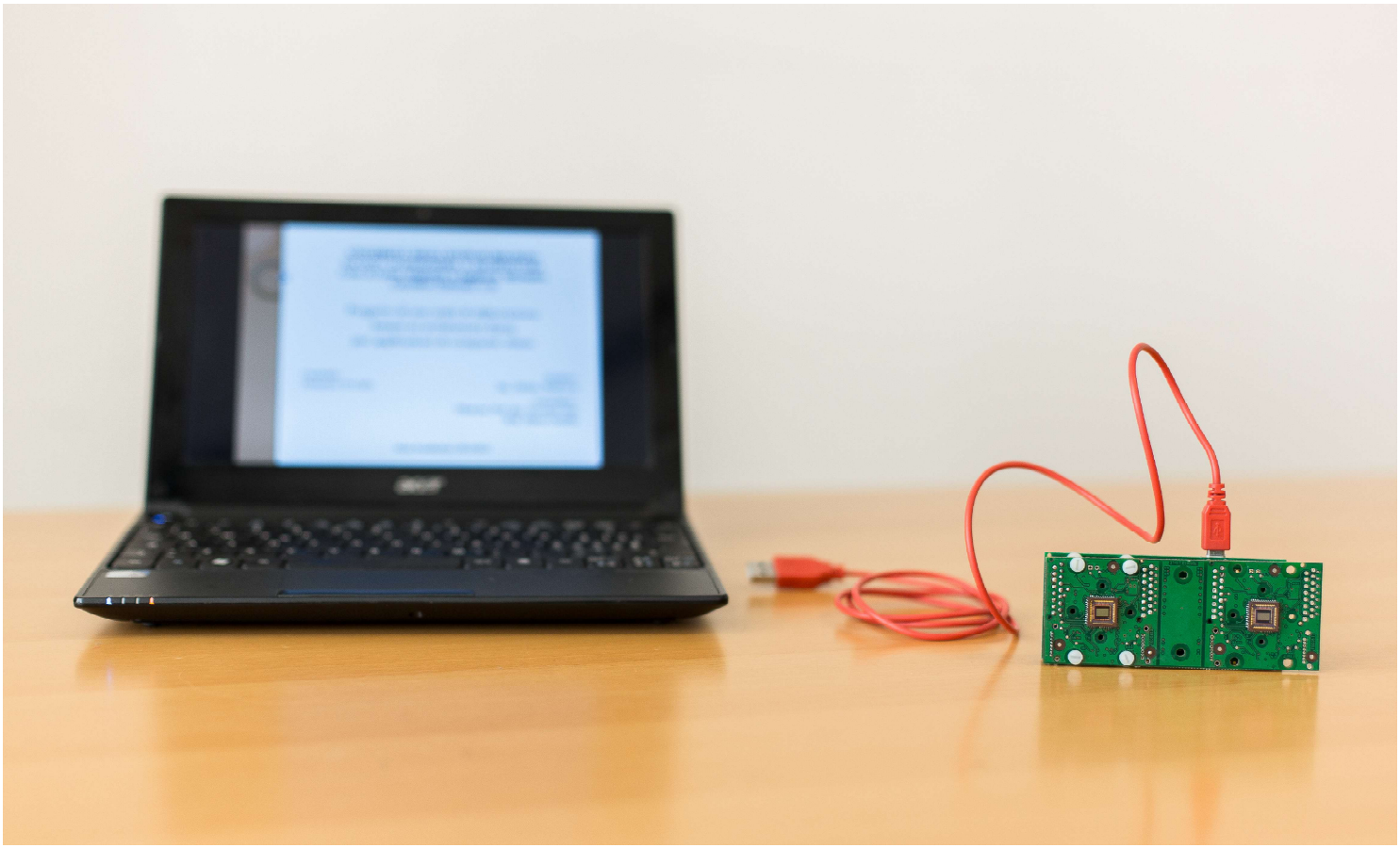


www.nvela.com

*

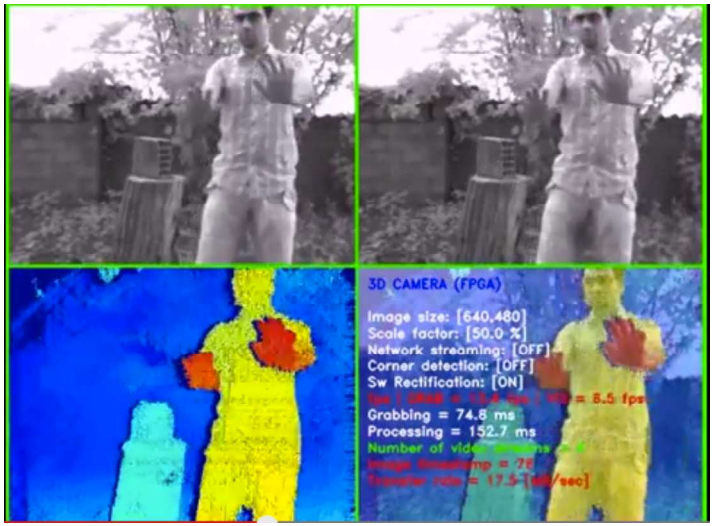
www.minoru3dwebcam.com

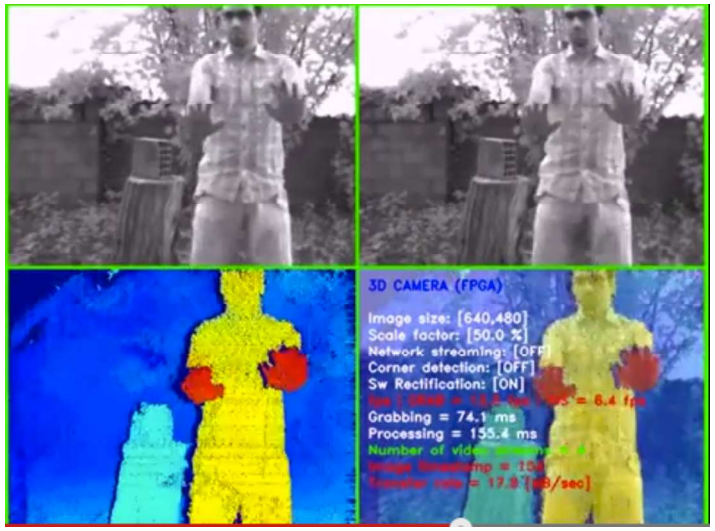
Custom FPGA-based stereo camera 1/2



- Real-time depth maps computed on FPGA with state of the art algorithms

Custom FPGA-based stereo camera 2/2

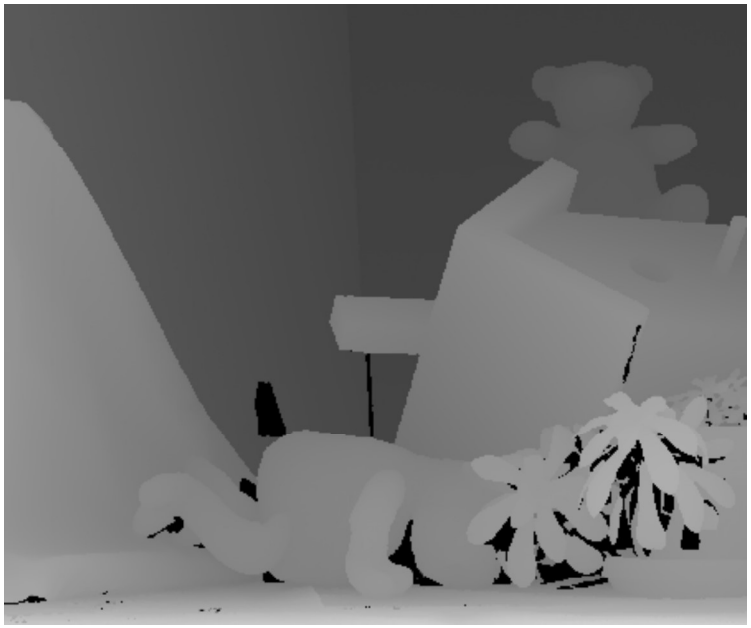




http://www.youtube.com/watch?v=STHZyvHVptw

http://www.youtube.com/watch?v=VV9cfJTs9OE

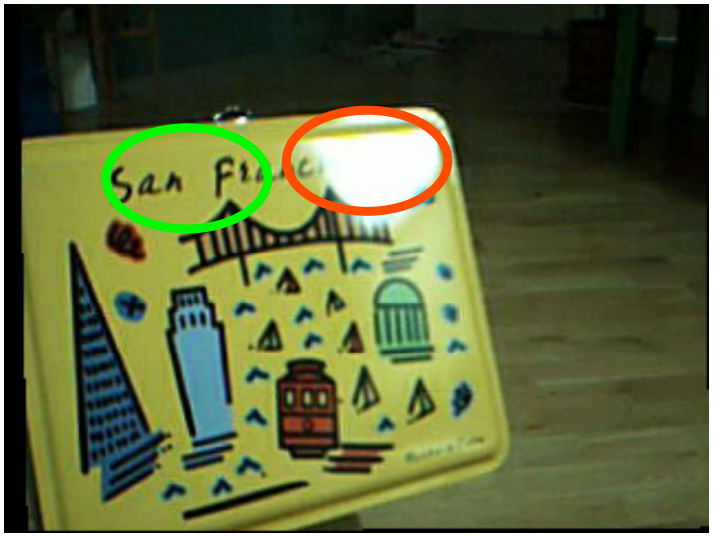
Why is stereo correspondence so challenging ?



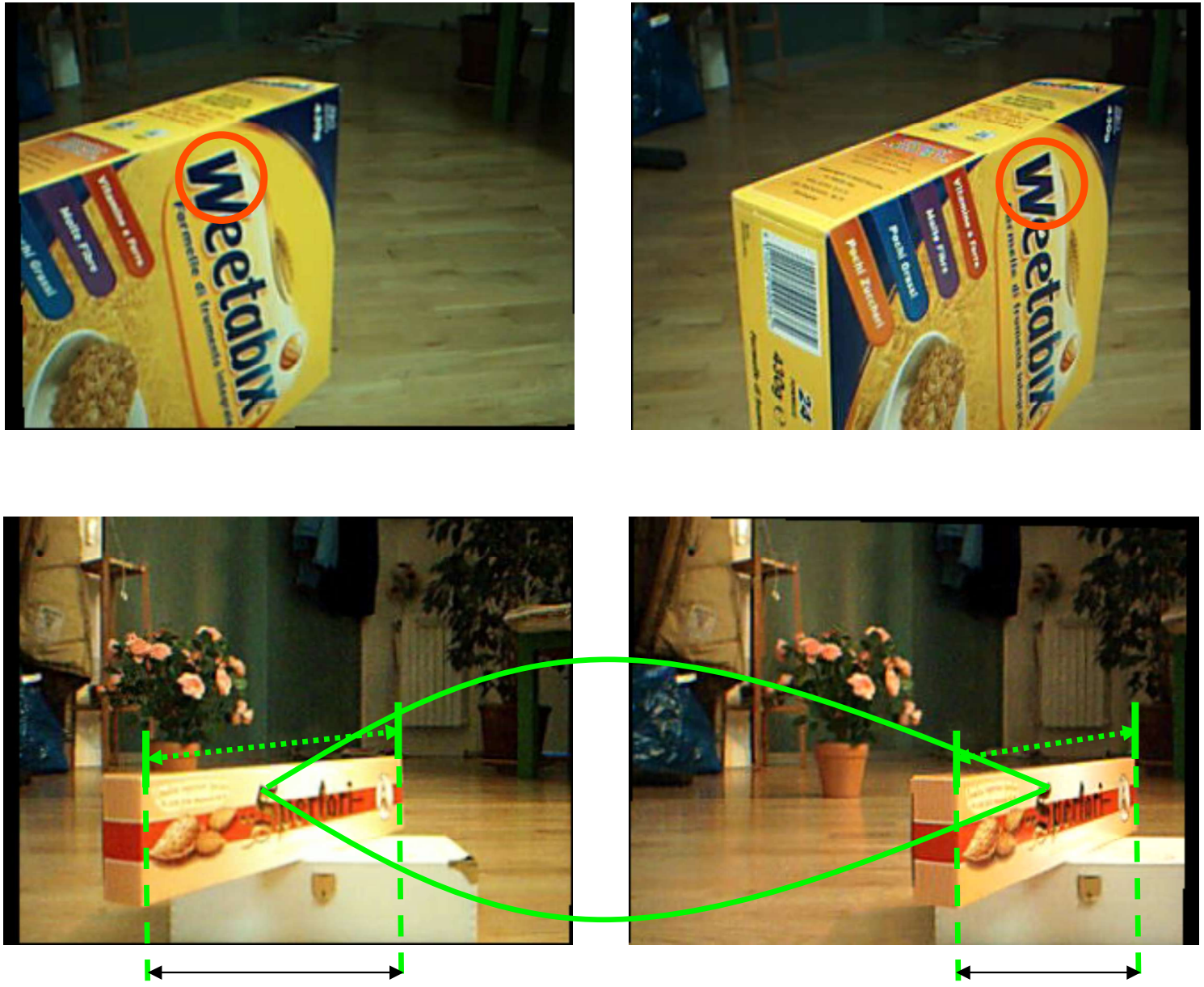
Next slides show common pitfalls…

Photometric distortions and noise

Specular surfaces



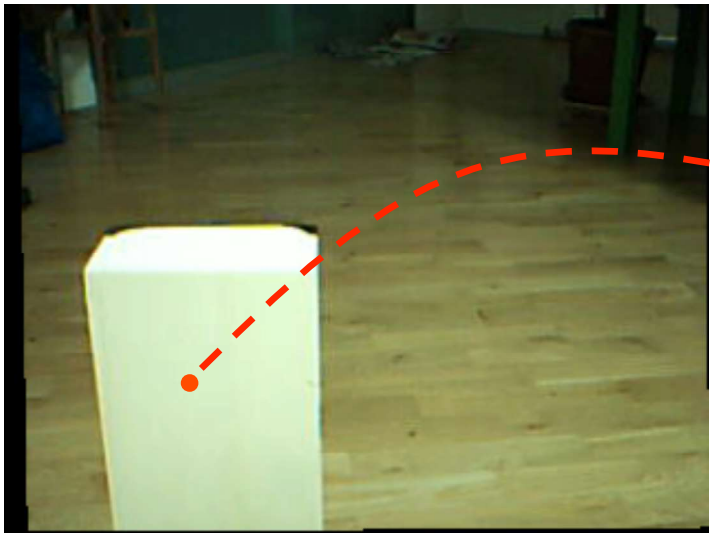
Foreshortening

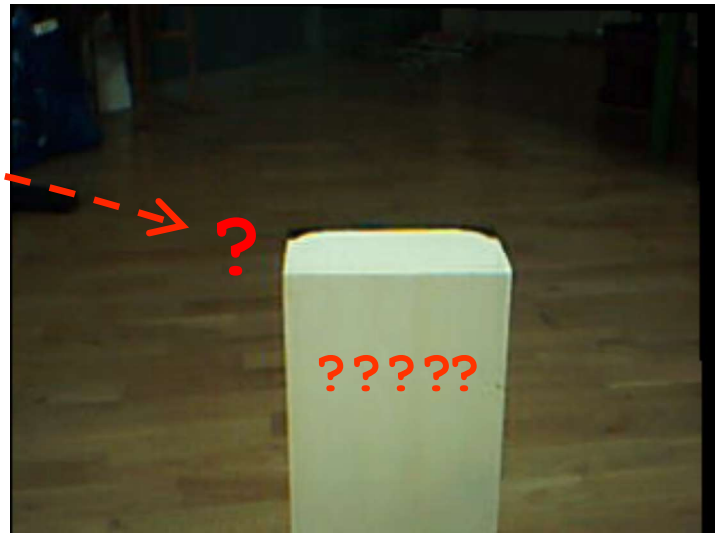


Uniqueness constraint ? :-(

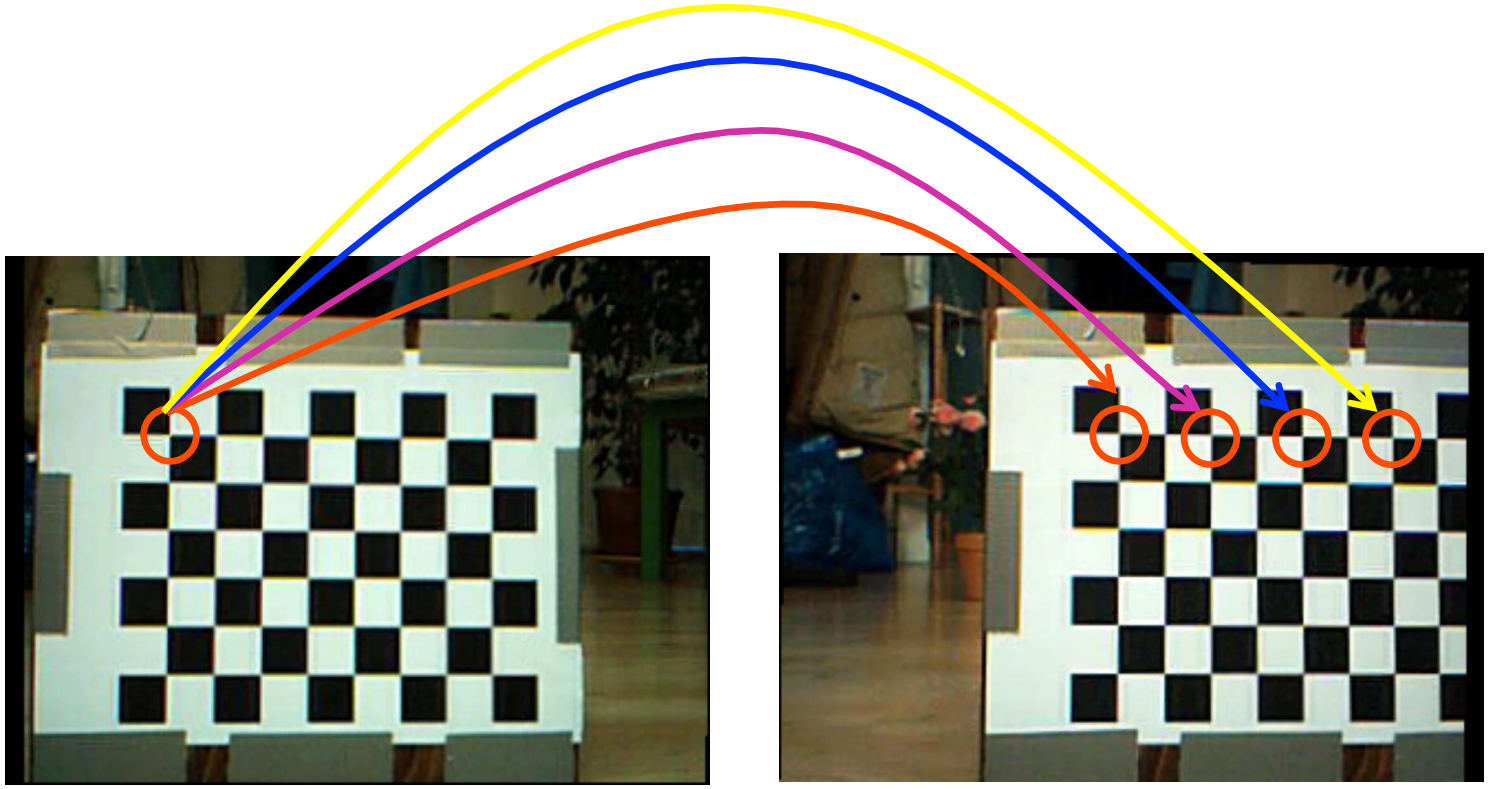
Perspective distortions

Uniform/ambiguous regions





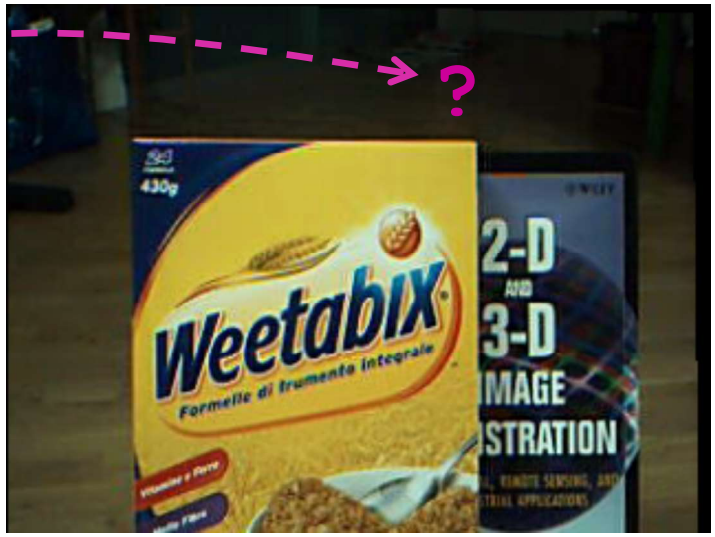
Repetitive/ambiguous patterns



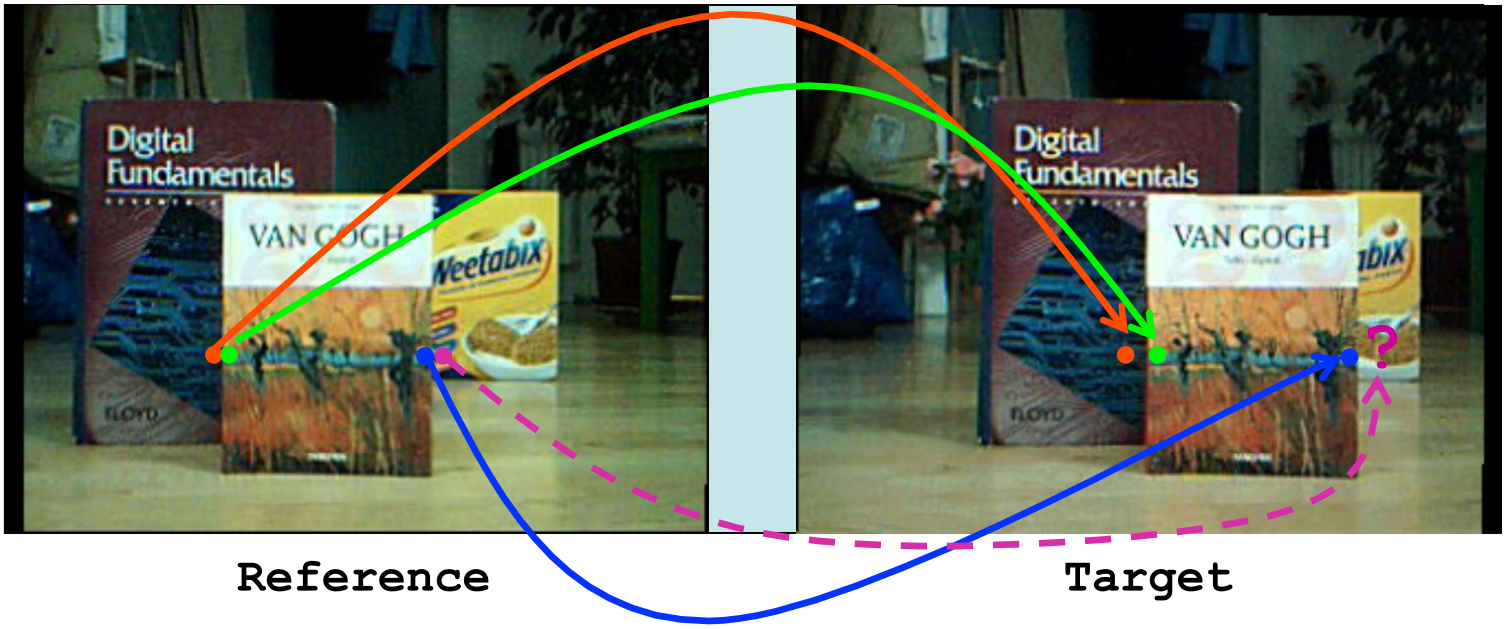
How to reduce ambiguity... ?

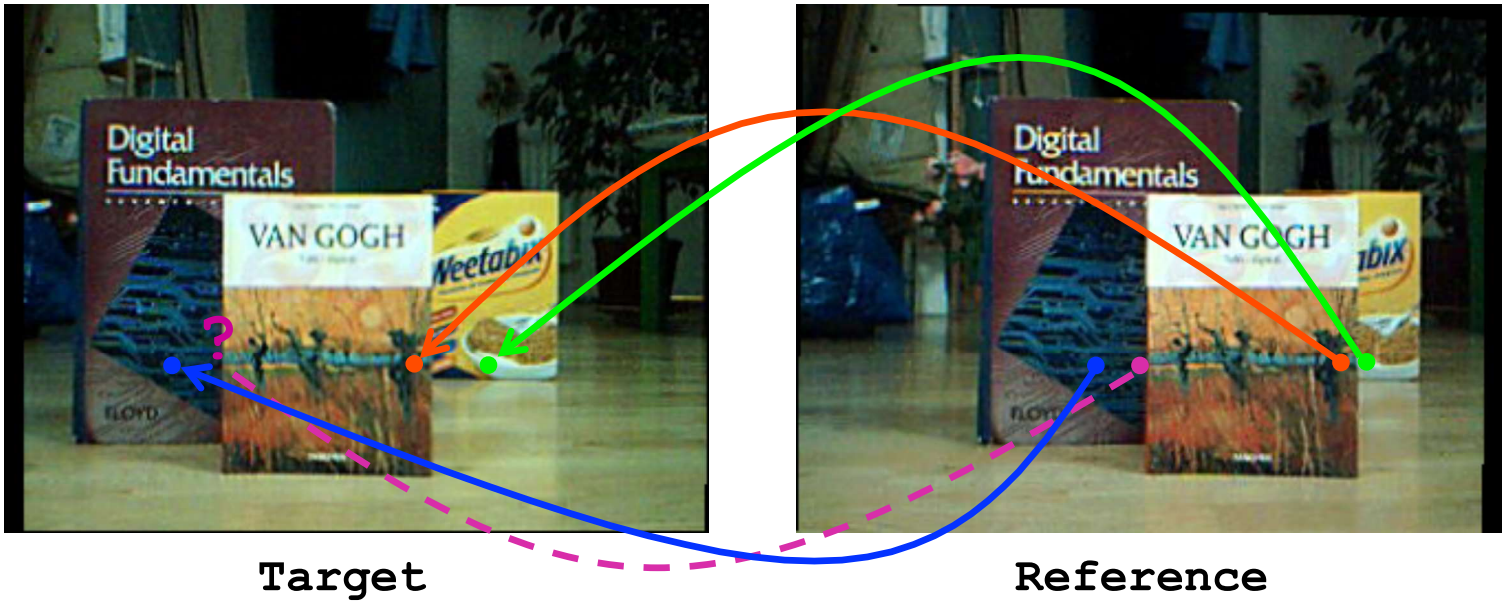
Transparent objects

Occlusions and discontinuities 1/2



Occlusions and discontinuities 2/2





Middlebury stereo evaluation

The Middlebury stereo evaluation site [15] provides a framework and a dataset (showed in the next slide) for benchmarking novel algorithms.

Scharstein and Szeliski provide:

- **a methodology for the evaluation of (binocular) stereo vision algorithms [11]**
- **datasets with groundtruth [11,15,17,18,19]**
- **online evaluation procedure and ranking [15]**

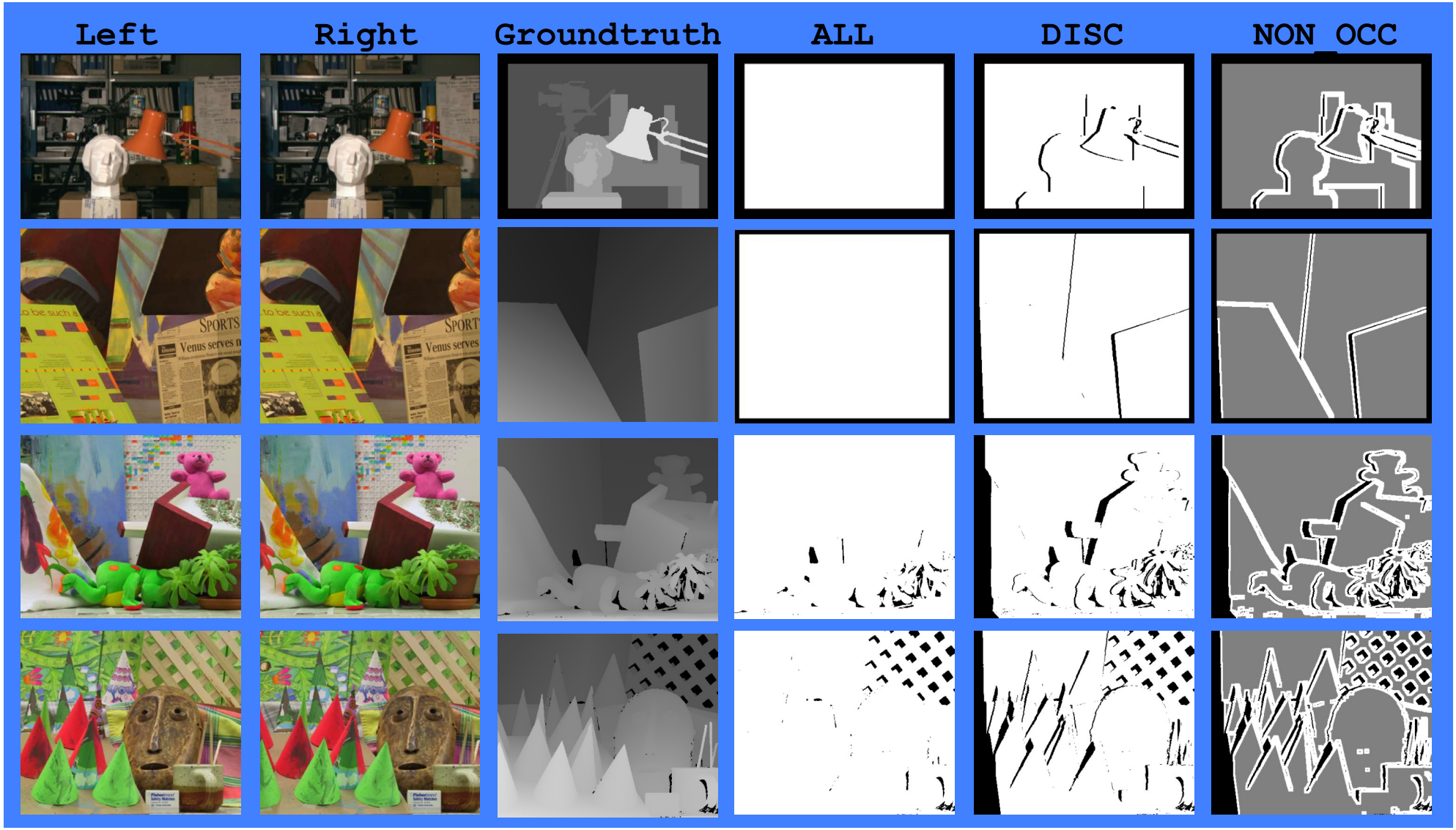
Datasets (with groundtruth) of stereo pairs affected by photometric distortions are also available in [15].

[15] D. Scharstein and R. Szeliski, http://vision.middlebury.edu/stereo/eval/

[11] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms" Int. Jour. Computer Vision, 47(1/2/3):7–42, 2002

Middlebury dataset (2003) [15]

Tsukuba, Venus, Teddy and Cones stereo pairs



D. Scharstein and R. Szeliski, http://vision.middlebury.edu/stereo/eval/

The correspondence problem

According to the taxonomy proposed in [11] most stereo algorithms perform (subset of) these steps:

1) Matching cost computation

2) Cost aggregation

3) Disparity computation/optimization

4) Disparity refinement

Local algorithms perform:
1 ⇒ 2 ⇒ 3 (with a simple Winner Takes All (WTA) strategy)

Global Algorithms perform:
1 (⇒ 2) ⇒ 3 (with global or semi-global reasoning)

Pre-processing (0)

Sometime is deployed a pre-processing stage mainly to compensate for photometric distortions.

Typical operations include:

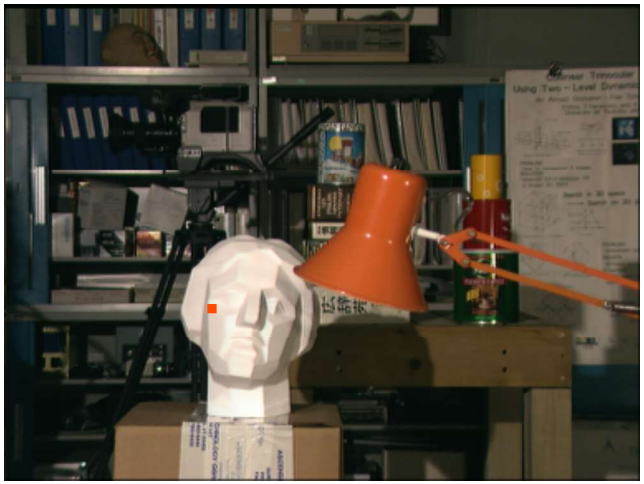
- **Laplacian of Gaussian (LoG) filtering [41]**
- **Subtraction of mean values computed in nearby pixels [42]**
- **Bilateral filtering [16]**
- **Census transform**

[41] T. Kanade, H. Kato, S. Kimura, A. Yoshida, and K. Oda, Development of a Video-Rate Stereo Machine International Robotics and Systems Conference (IROS '95), Human Robot Interaction and Cooperative Robots, 1995

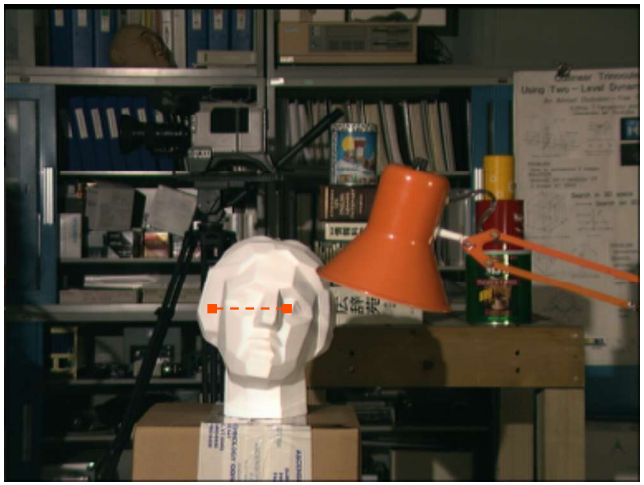
[42] O. Faugeras, B. Hotz, H. Mathieu, T. Viville, Z. Zhang, P. Fua, E. Thron, L. Moll, G. Berry, Real-time correlation-based stereo: Algorithm. Implementation and Applications, INRIA TR n. 2013, 1993

[16] A. Ansar, A. Castano, L. Matthies, Enhanced real-time stereo using bilateral filtering IEEE Conference on Computer Vision and Pattern Recognition 2004

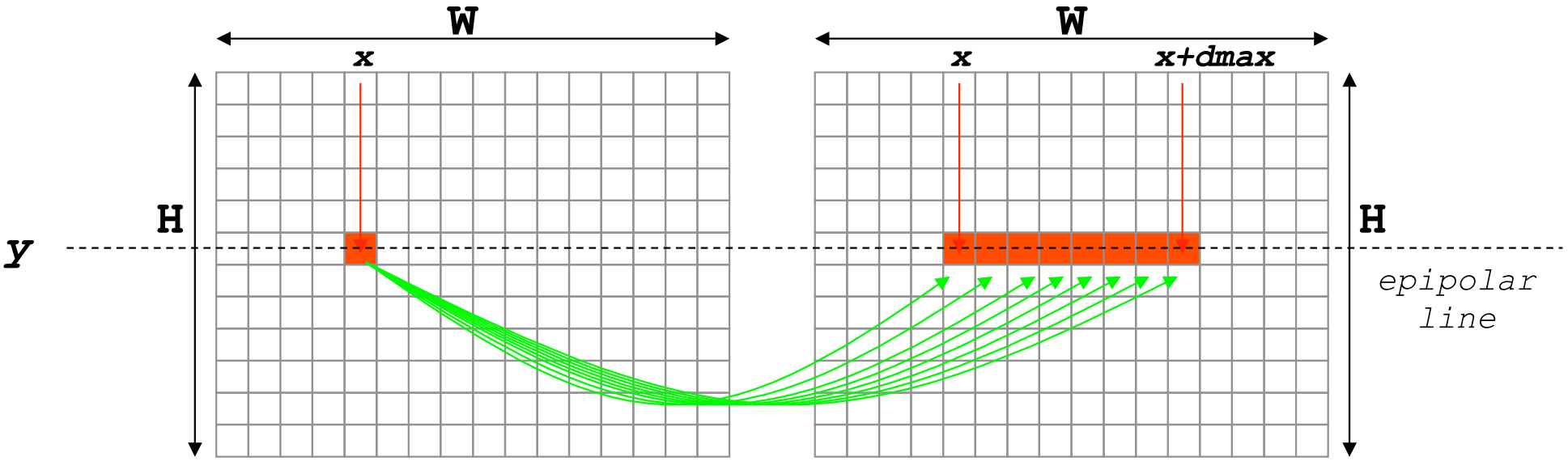
The simplest (naive and unused) local approach:



Reference (R)

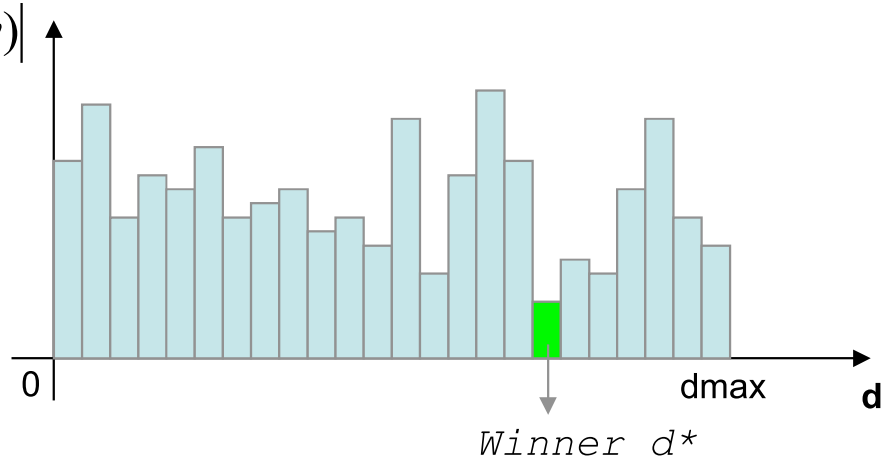


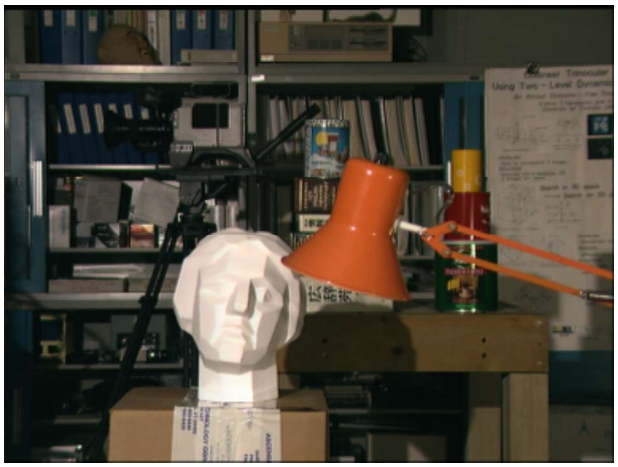
Target (T)



- **matching cost (1): pixel-based absolute difference between pixel intensities**
- **disparity computation (3): Winner Takes All (WTA)**

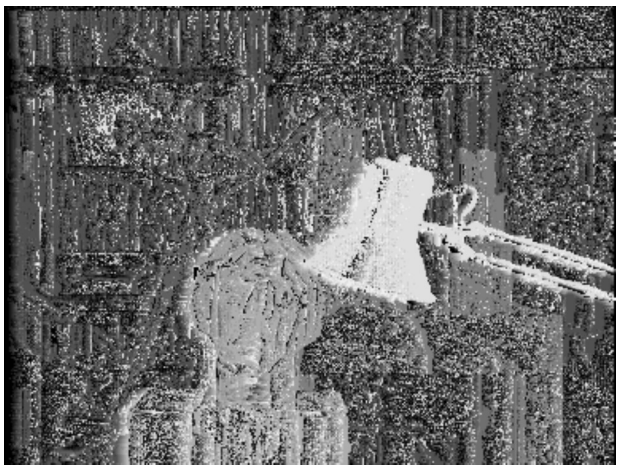
$$\left|I_R(x,y) - I_T(x+d,y)\right|$$





Reference

Groundtruth



Result (disappointing)

How to improve the results of the naive approach ?

Basically exist two different (not mutually exclusive) strategies:

- **Local algorithms** use the simple WTA disparity selection strategy but reduce ambiguity (increasing the signal to noise ratio (SNR)) by aggregating matching costs over a support window (aka kernel or correlation window). Sometime a smoothness term is adopted. **Steps 1+2 (+ WTA)**

- **Global (and semi-global*) algorithms** search for disparity assignments that minimize an energy function over the whole stereo pair using a pixel-based matching cost (sometime the matching cost is aggregated over a support). **Steps 1+3**

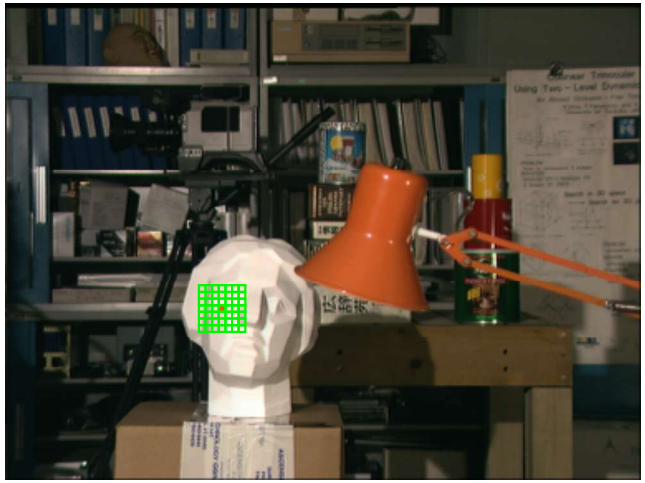
Both approaches assume that the scene is piecewise smooth. Sometime this assumption is violated...

This hypothesis is implicitly assumed by local approaches while it is explicitly modelled by global approaches

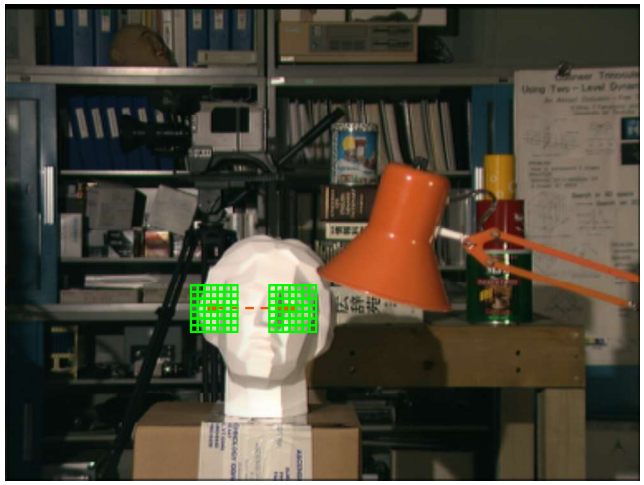
* subset of the stereo pair

Local approaches:

In order to increase the SNR (reduce ambiguity) the matching costs are aggregated over a support window



Reference (R)



Target (T)

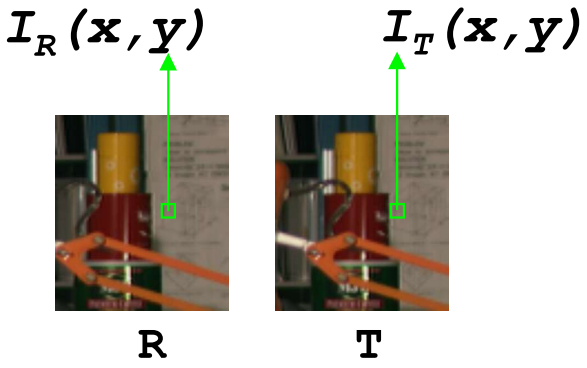
Global (and semi-global*) approaches:

Many algorithms search for the disparity assignment that minimize a certain cost function over the whole* stereo pair

$$E(d) = E_{data}(d) + E_{smooth}(d)$$

* subset of the stereo pair

Matching cost computation (1)



Pixel-based matching costs

- **Absolute differences**

$$e(x,y,d) = \left| I_R(x,y) - I_T(x+d,y) \right|$$

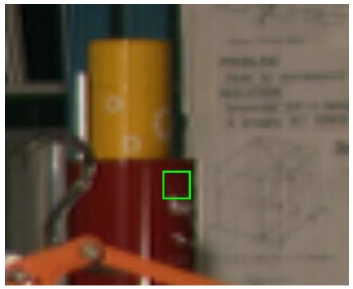
- **Squared differences**

$$e(x,y,d) = \left(I_R(x,y) - I_T(x+d,y) \right)^2$$

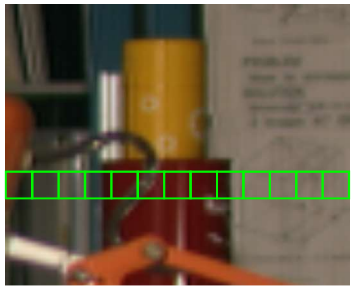
- **Robust matching measures (M-estimators)**
  - **Limit influence of outliers**
  - **Example: truncated absolute differences (TAD)**

$$e(x,y,d) = \min\left\{ \left| I_R(x,y) - I_T(x+d,y) \right|, T \right\}$$

- **Dissimilarity measure insensitive to image sampling (Birchfield and Tomasi [27])**

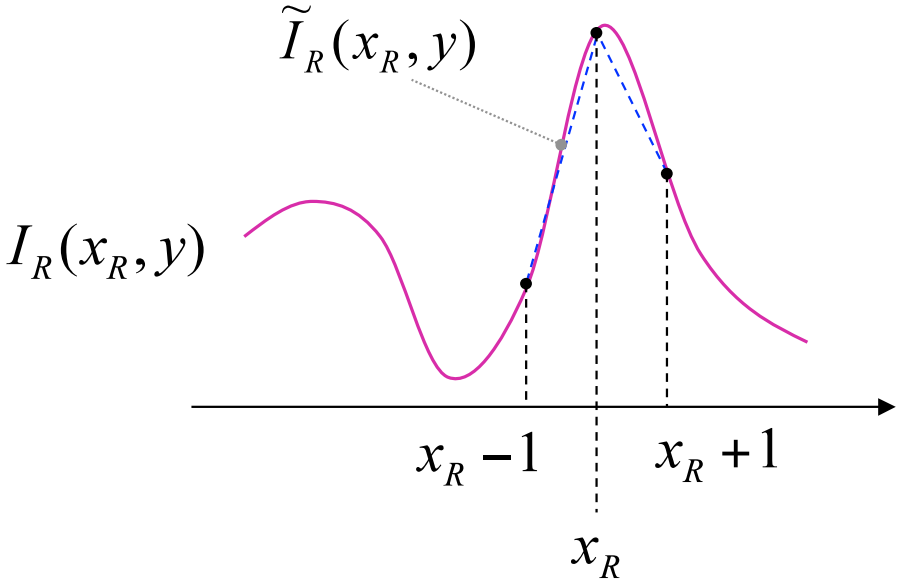


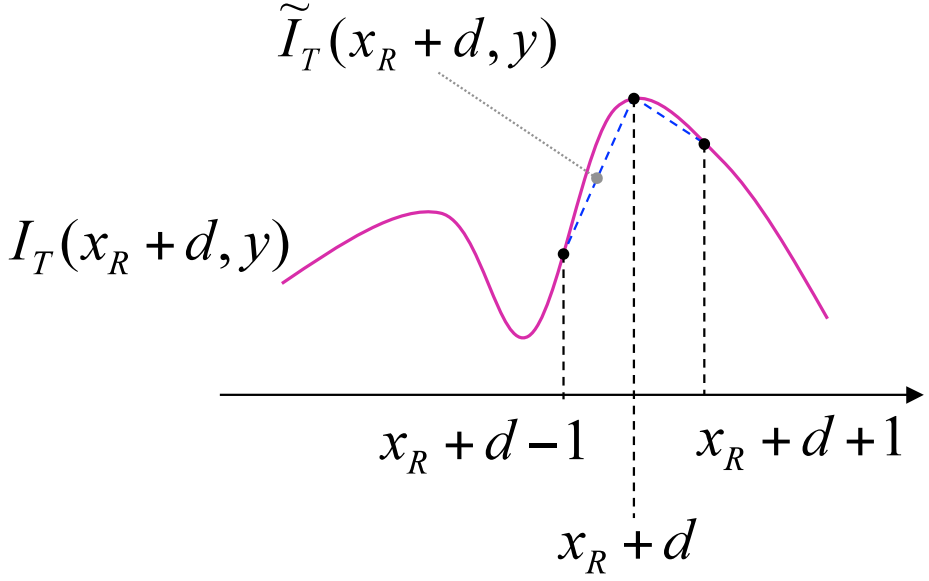
Reference (R)



Target (T)

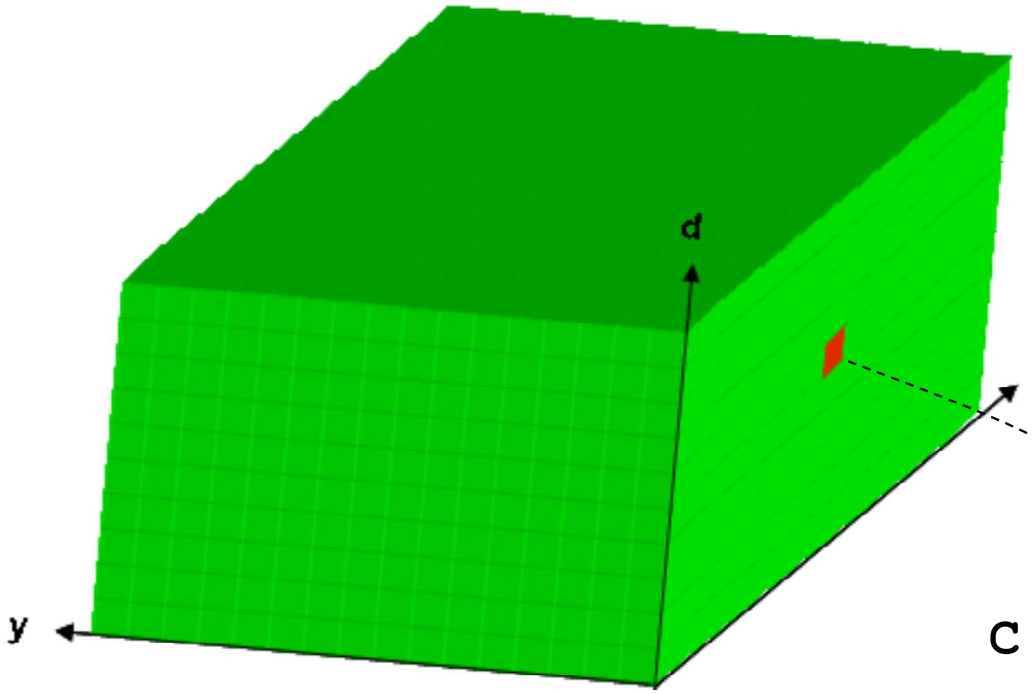
BT helps at depth and color discontinuities





$$e(x_R, y, d) = \min\left\{ \min_{x_R - \frac{1}{2} \le x \le x_R + \frac{1}{2}} \left| I_R(x_R, y) - \tilde{I}_T(x + d, y) \right|, \min_{x_R - \frac{1}{2} \le x \le x_R + \frac{1}{2}} \left| I_T(x_R + d, y) - \tilde{I}_R(x, y) \right| \right\}$$

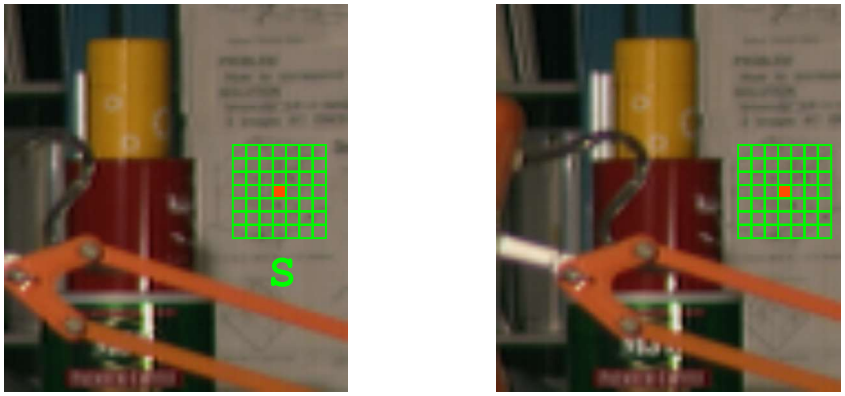
The Disparity Space Image (DSI) is a 3D matrix ($W \times H \times (d_{max} - d_{min})$)



likelihood/confidence
of each correspondence

Each element $C(x,y,d)$ of the DSI represents the cost of the correspondence between $I_R(x_R,y)$ and $I_T(x_R+d,y)$

Area-based matching costs:



- **Sum of Absolute differences (SAD)**

$$C(x,y,d) = \sum_{x \in S} \left| I_R(x,y) - I_T(x+d,y) \right|$$

- **Sum of Squared differences (SSD)**

$$C(x,y,d) = \sum_{x \in S} \left(I_R(x,y) - I_T(x+d,y) \right)^2$$

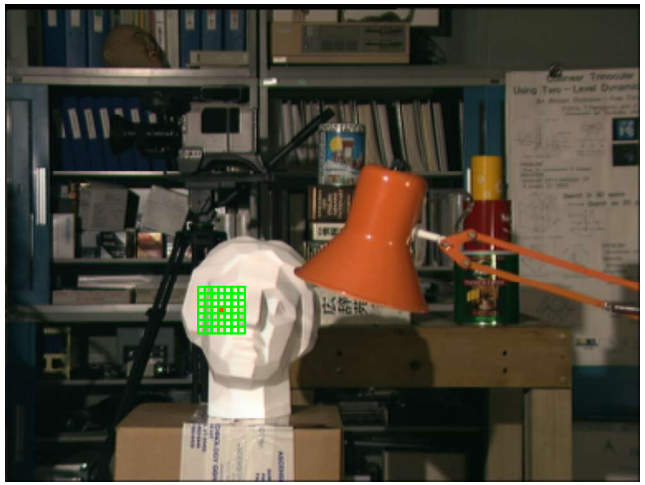
- **Sum of truncated absolute differences (STAD)**

$$C(x,y,d) = \sum_{x \in S} \min\left\{ \left| I_R(x,y) - I_T(x+d,y) \right|, T \right\}$$

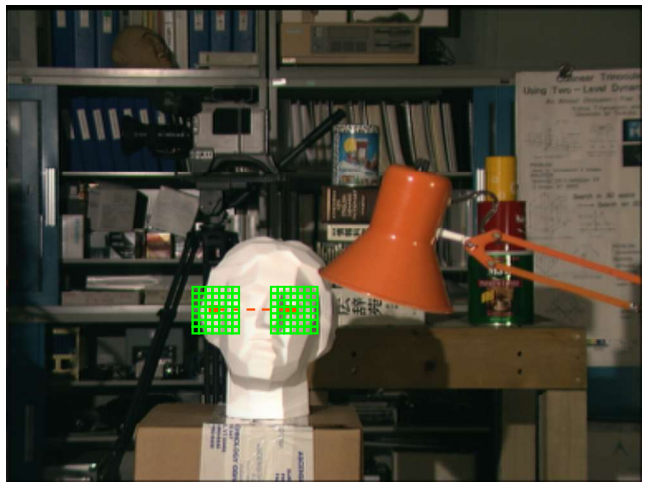
- **Normalized Cross Correlation [57]**
- **Zero mean Normalized Cross Correlation [58]**
- **Gradient based MF [59]**
- **Non parametric [60,61]**
- **Mutual Information [30]**

Cost aggregation (2)

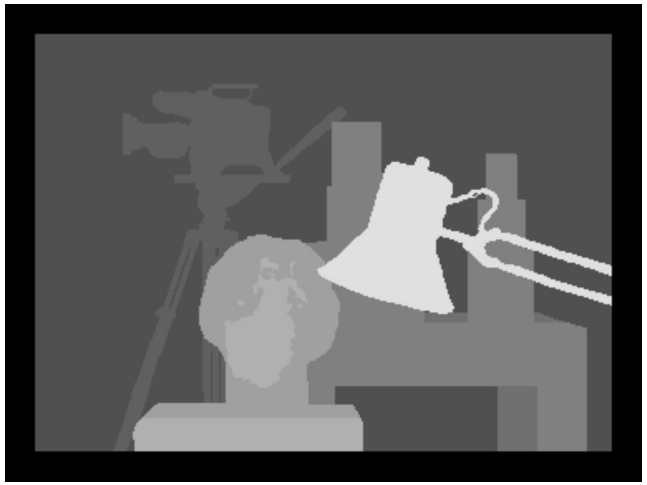
Let's start by examining the simplest Fixed Window (FW) cost aggregation strategy (TAD, disparity selection WTA)



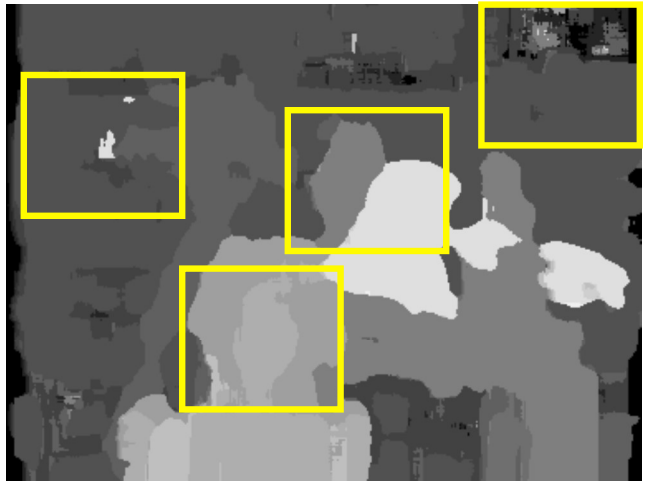
Reference (R)



Target (T)



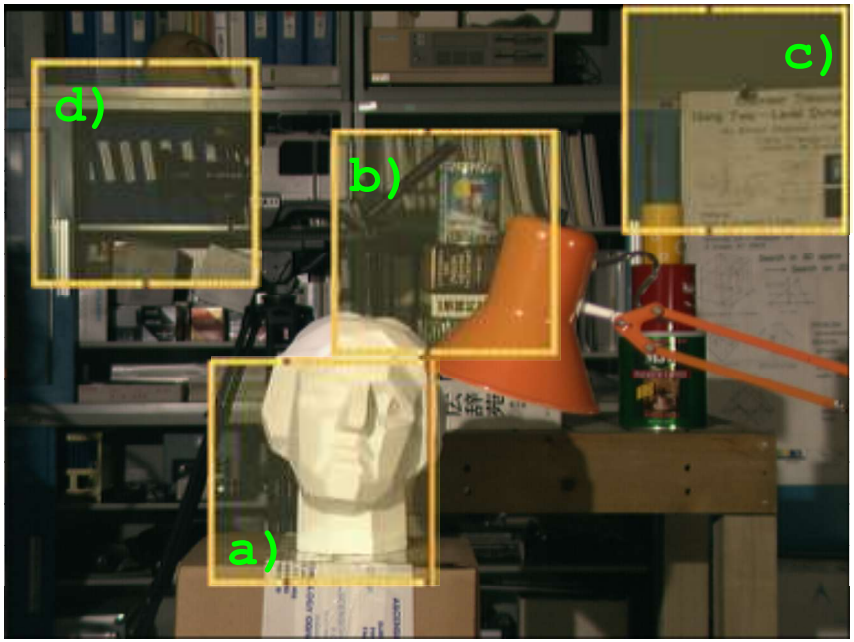
Groundtruth



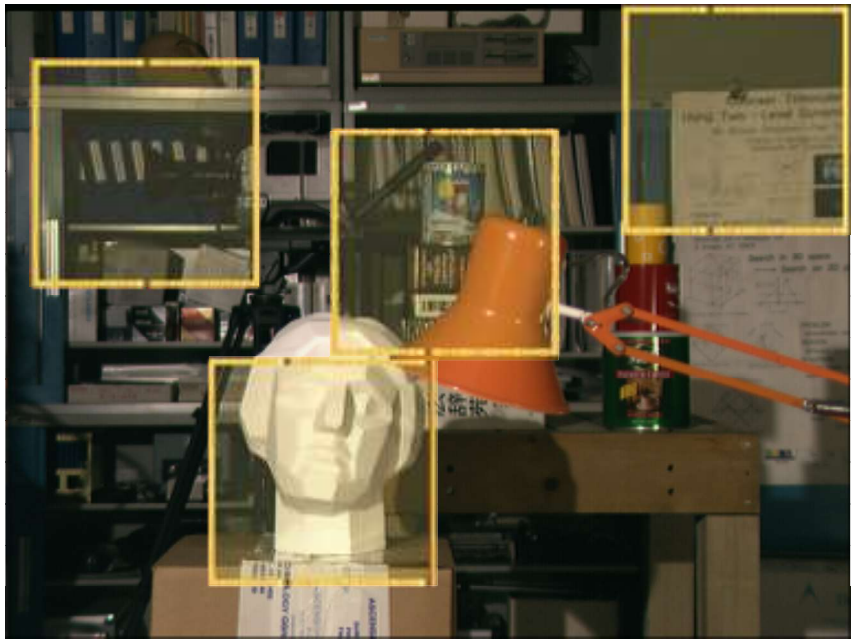
Fixed Window (FW)

What's wrong with FW ?

FW (with WTA reasoning) fails in most points for the following reasons:



Reference (R)



Target (T)

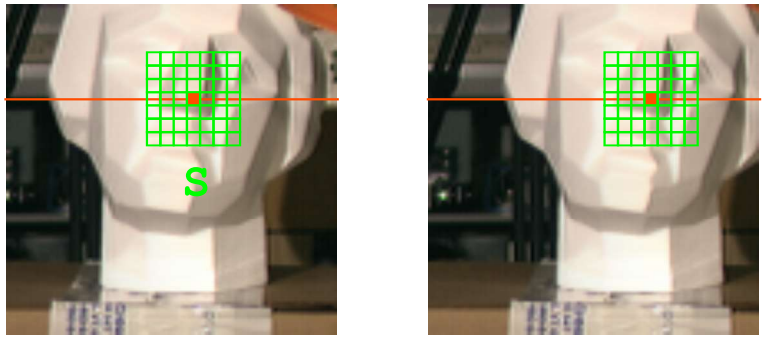
a) implicitly assumes frontal-parallel surfaces

b) ignores depth discontinuities

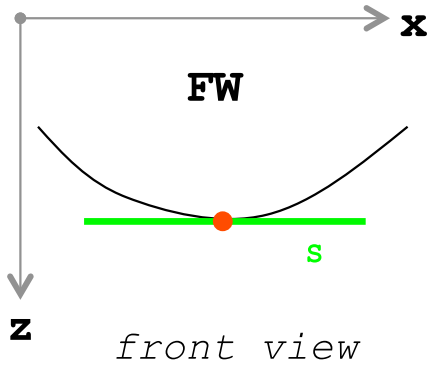
c) does not deal explicitly with uniform areas

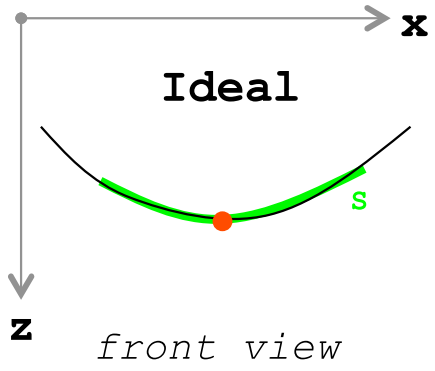
d) does not deal explicitly with repetitive patterns

a) FW implicitly assumes frontal-parallel surfaces

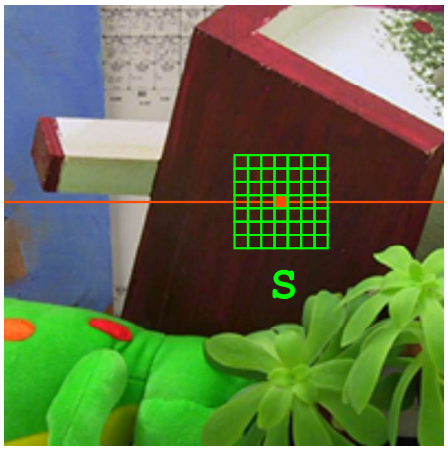


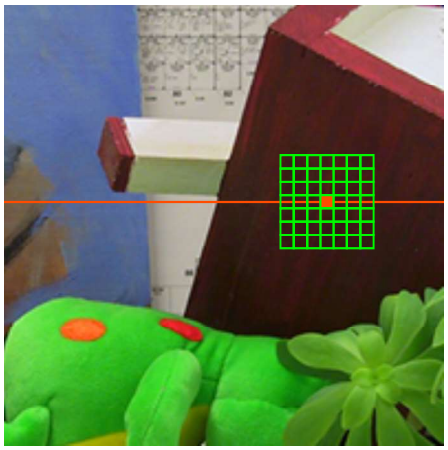
FW

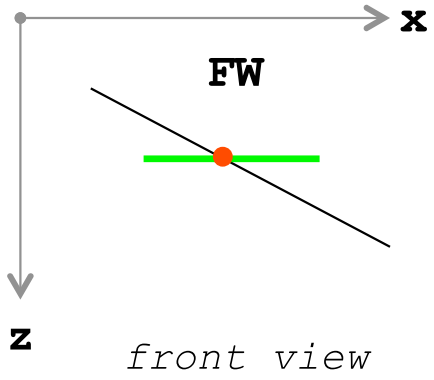


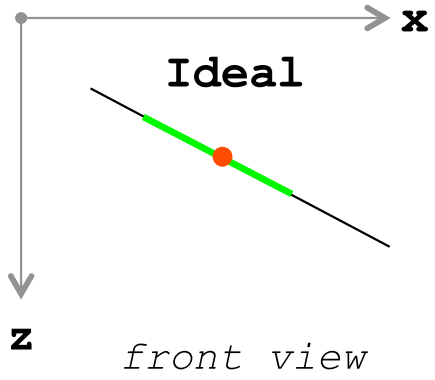


Often violated in practice: top figure, slanted surfaces (down), etc.







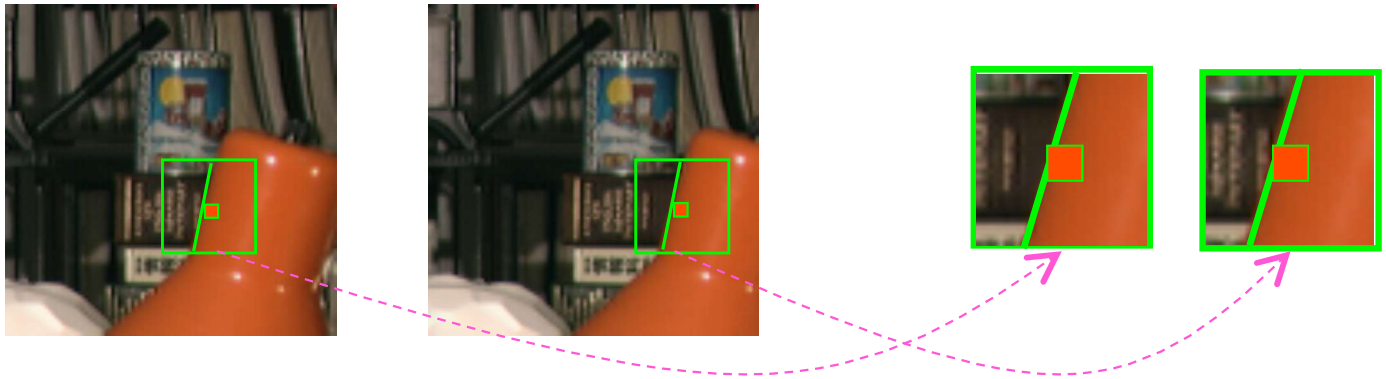


chiến lược tổng hợp

Nevertheless, almost all state-of-the-art cost aggregation strategies rely on the assumption that all the points belonging to the support share the same disparity (only few exceptions).

b) FW ignores depth discontinuities

Implicitly assuming frontal-parallel surface in the real scene is violated near depth discontinuities.



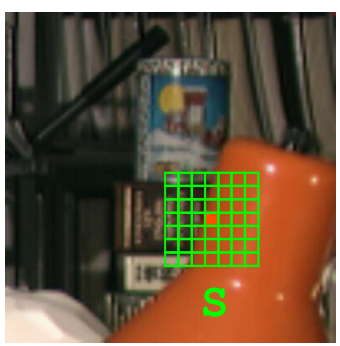
Background is misaligned !

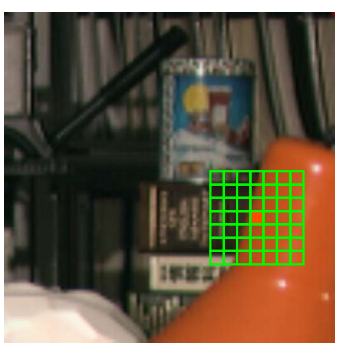
Aggregating the matching costs of two populations at different depth (aligned foreground and misaligned background (outliers)) results in the typical inaccurate localization of depth borders.

FW

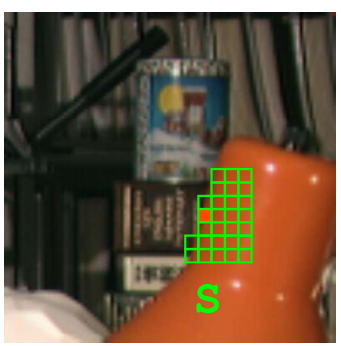
Robust matching measures (TAD) can partially reduce the influence of outliers

State-of-the-art cost aggregation strategies aim at shaping the support in order to include only points with the same (unknown) disparity.





FW



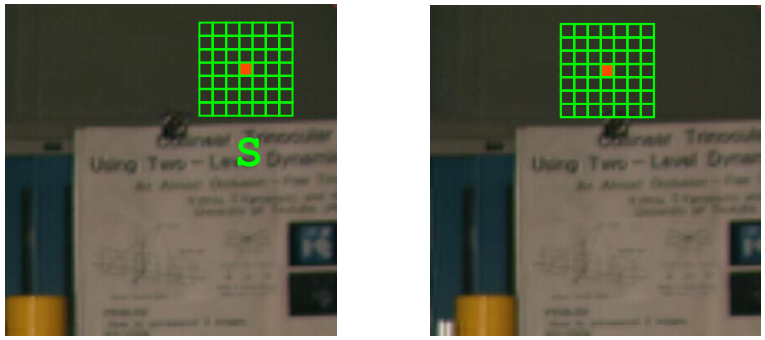
Ideal

For what concerns FW: decreasing the size of the support helps in reducing the border localization problem.

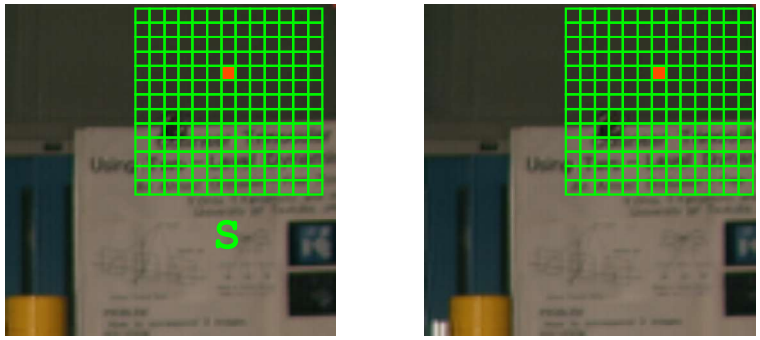
However, this choice renders the correspondence problem more ambiguous (especially when dealing with uniform regions and repetitive patterns, see the next slide).

In practice, for the FW approach the choice of the optimal size of the support is done empirically.

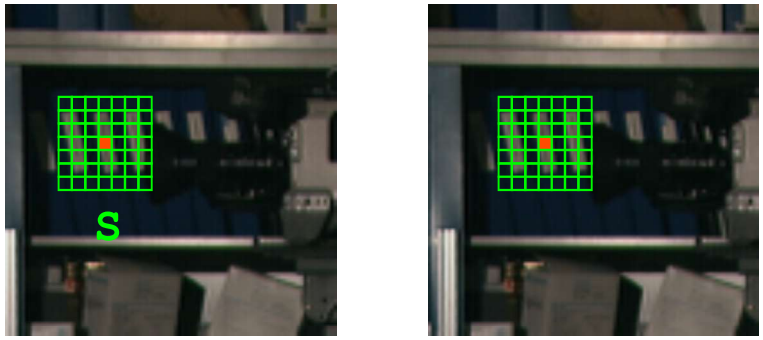
FW does not deal explicitly with ambiguous regions - uniform areas c) and repetitive patterns d)



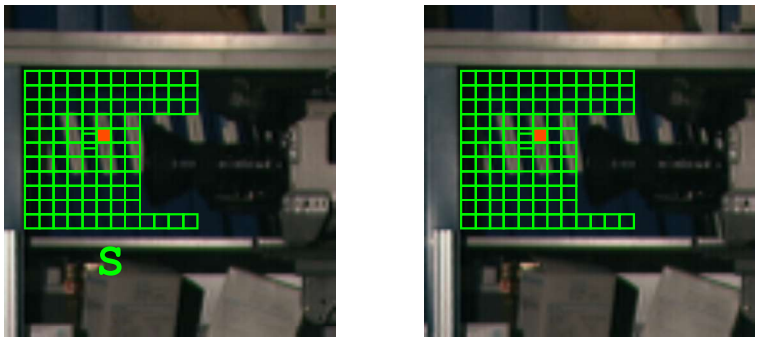
FW



Ideal



FW



Ideal

In both cases an ideal cost aggregation strategy should extend its support in order to include as much points at the same (unknown) depth as possible.

Quite surprisingly, in spite of its limitations, FW is widely adopted in practice (probably it is the most frequently used algorithm for real applications).

- **Easy to implement**
- **Fast, thanks to incremental calculation schemes**
- **Runs in real-time on standard processors (SIMD)**
- **Has limited memory requirements**
- **Hardware implementations (FPGA) run in real-time with limited power consumption (<1W)**

Before analyzing more sophisticated approaches let's consider two optimization techniques used by FW and other algorithms:

- **Integral Images (II)**
- **Box-Filtering (BF)**