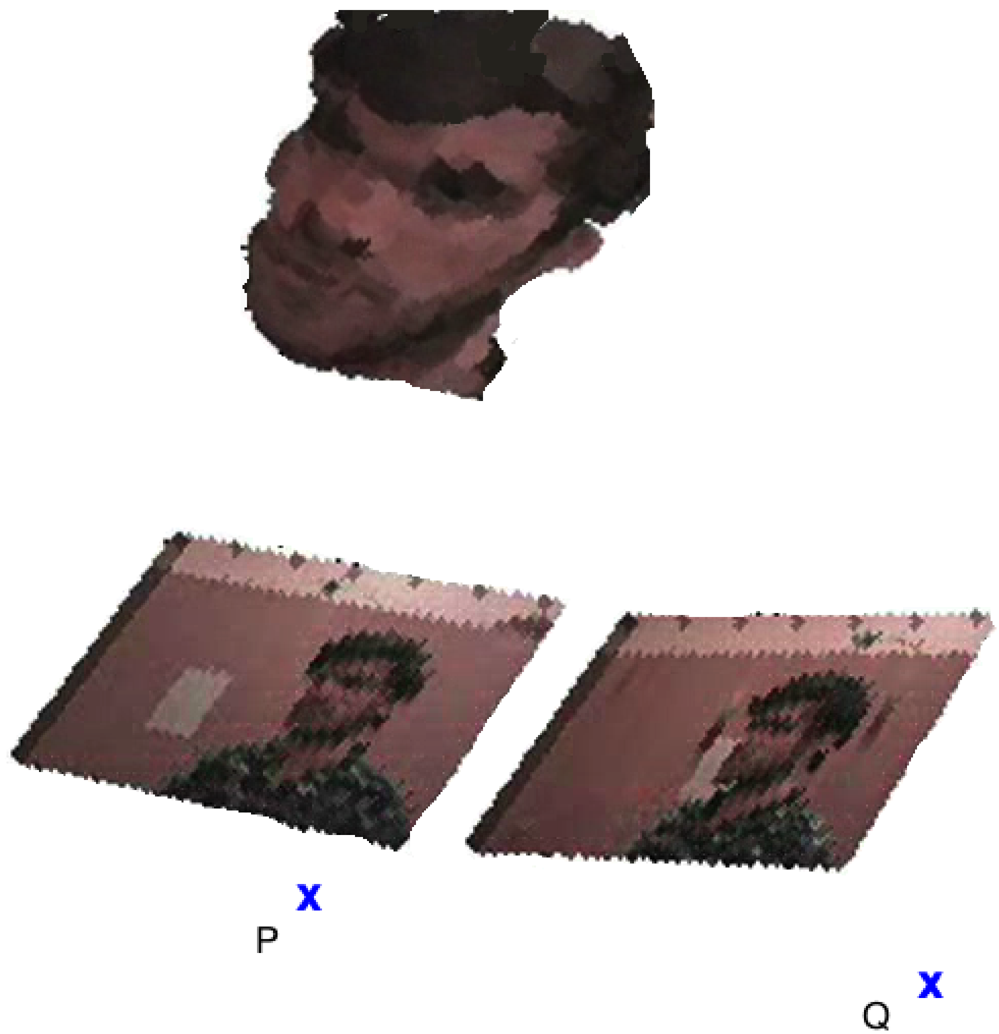
Epipolar Geometry and its application for the construction of state-of-the-art sensors.

Karel Zimmermann

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics
Center for Machine Perception
http://cmp.felk.cvut.cz/~zimmerk, zimmerk@fel.cvut.cz

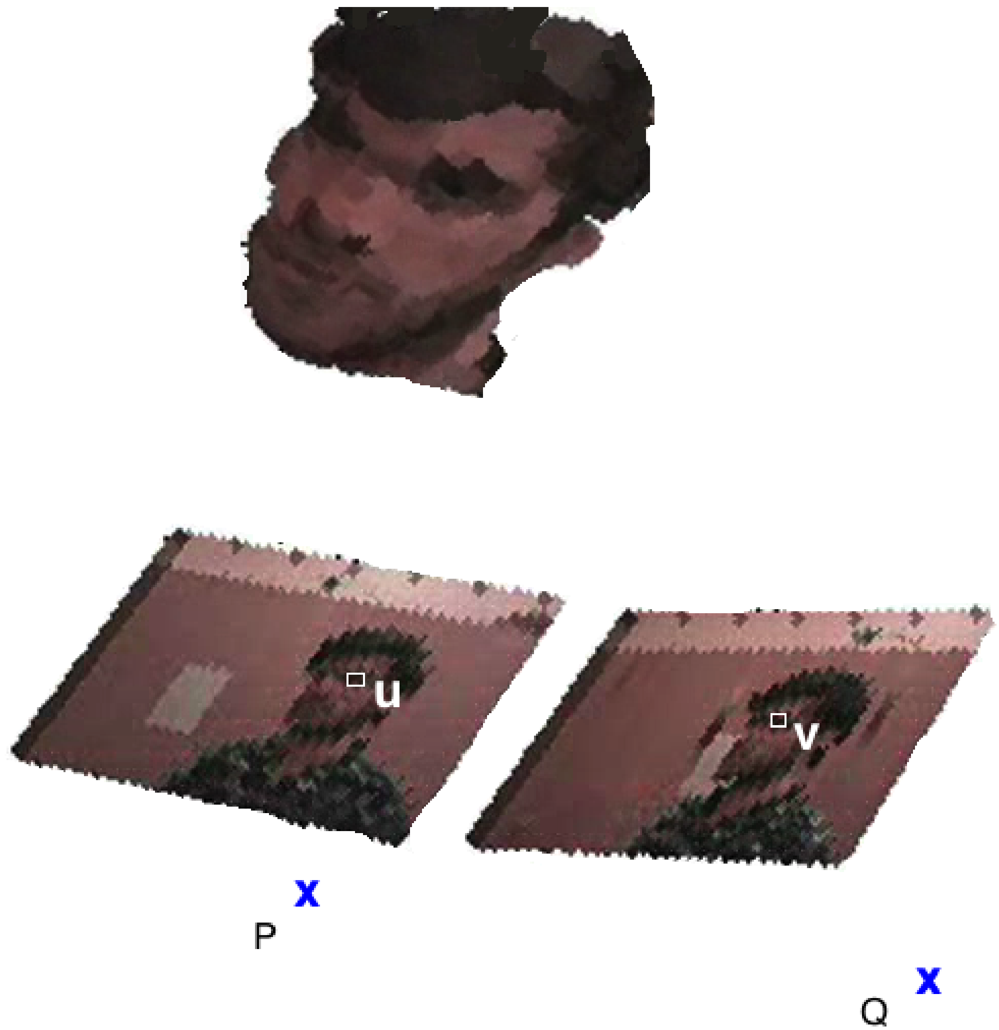
Motivation

- You are given two images of an object captured by two cameras P and Q from different view-points.



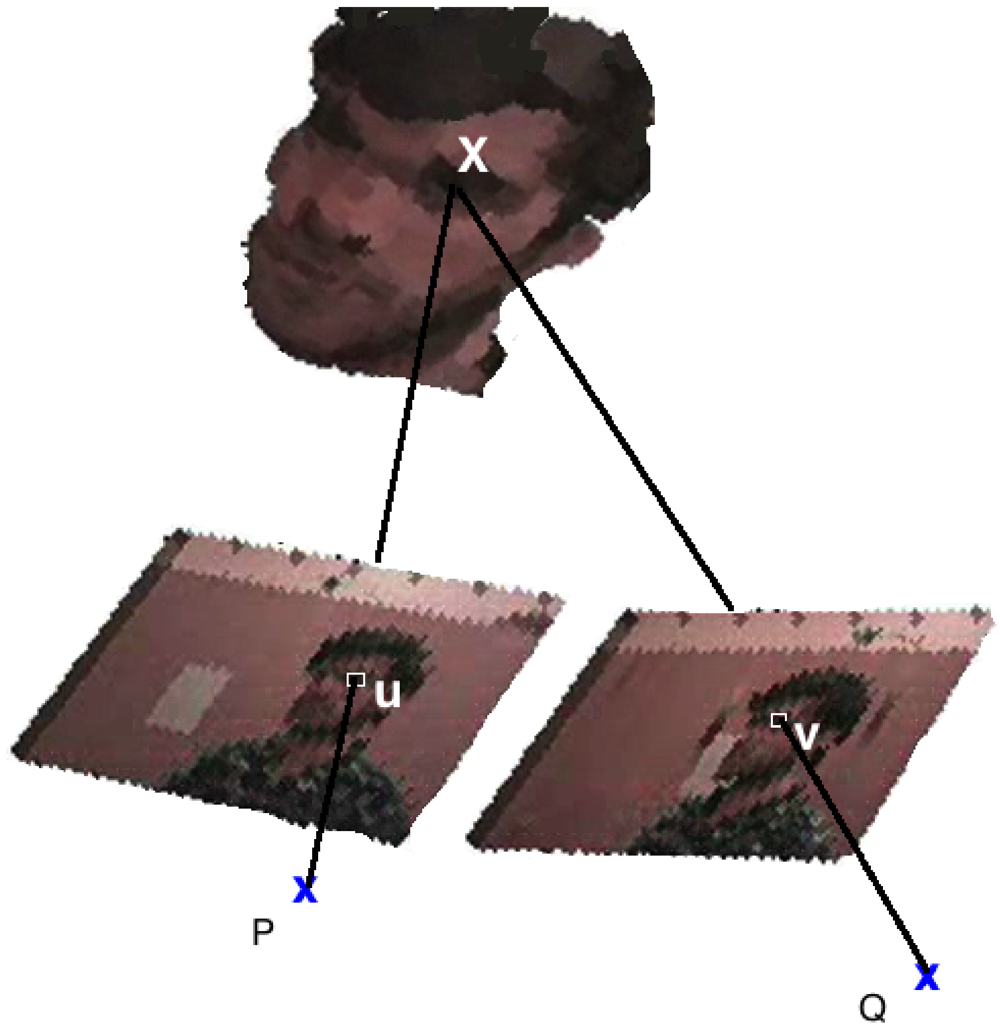
Motivation

- Given pair of corresponding pixels $(\mathbf{u}, \mathbf{v})$ (i.e. pixels corresponding to the same unknown 3D point $\mathbf{X}$ on the object), you can easily compute $\mathbf{X}$.



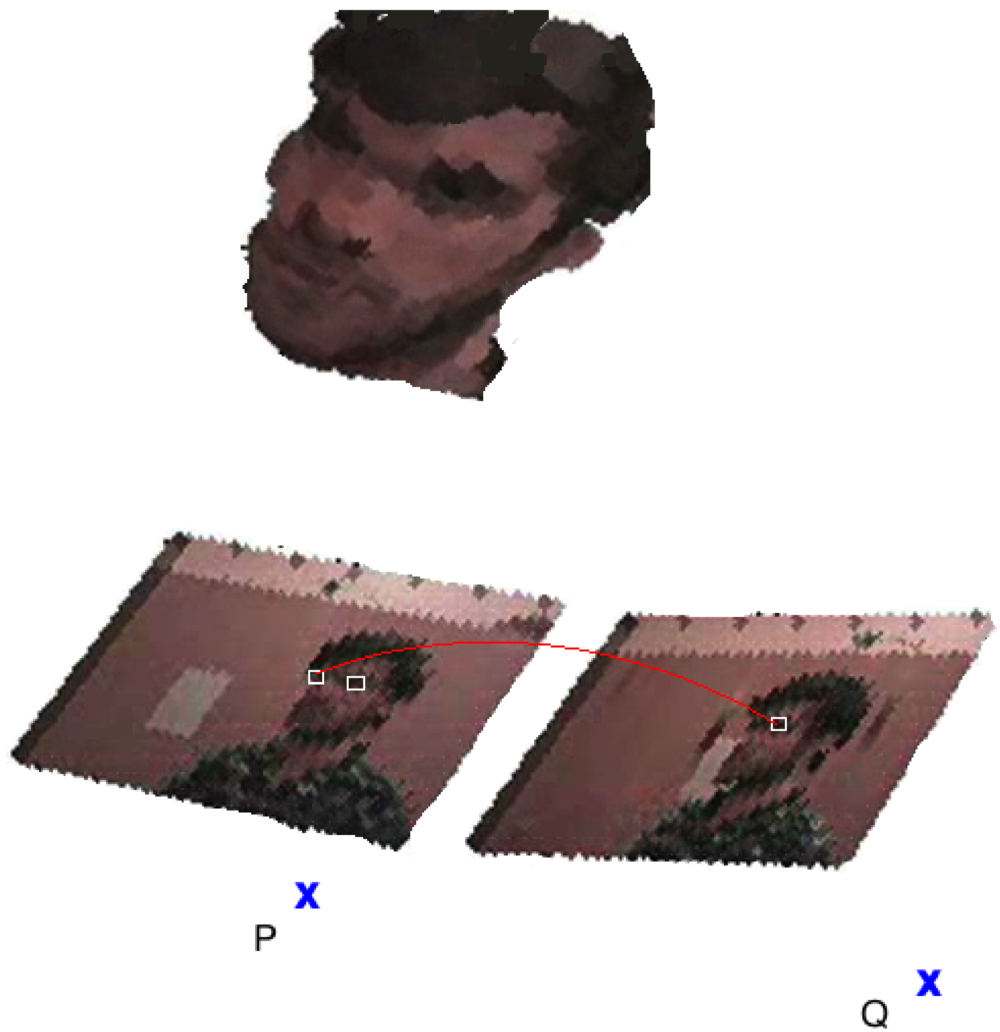
Motivation

- Given pair of corresponding pixels $(\mathbf{u}, \mathbf{v})$ (i.e. pixels corresponding to the same unknown 3D point $\mathbf{X}$ on the object), you can easily compute $\mathbf{X}$.



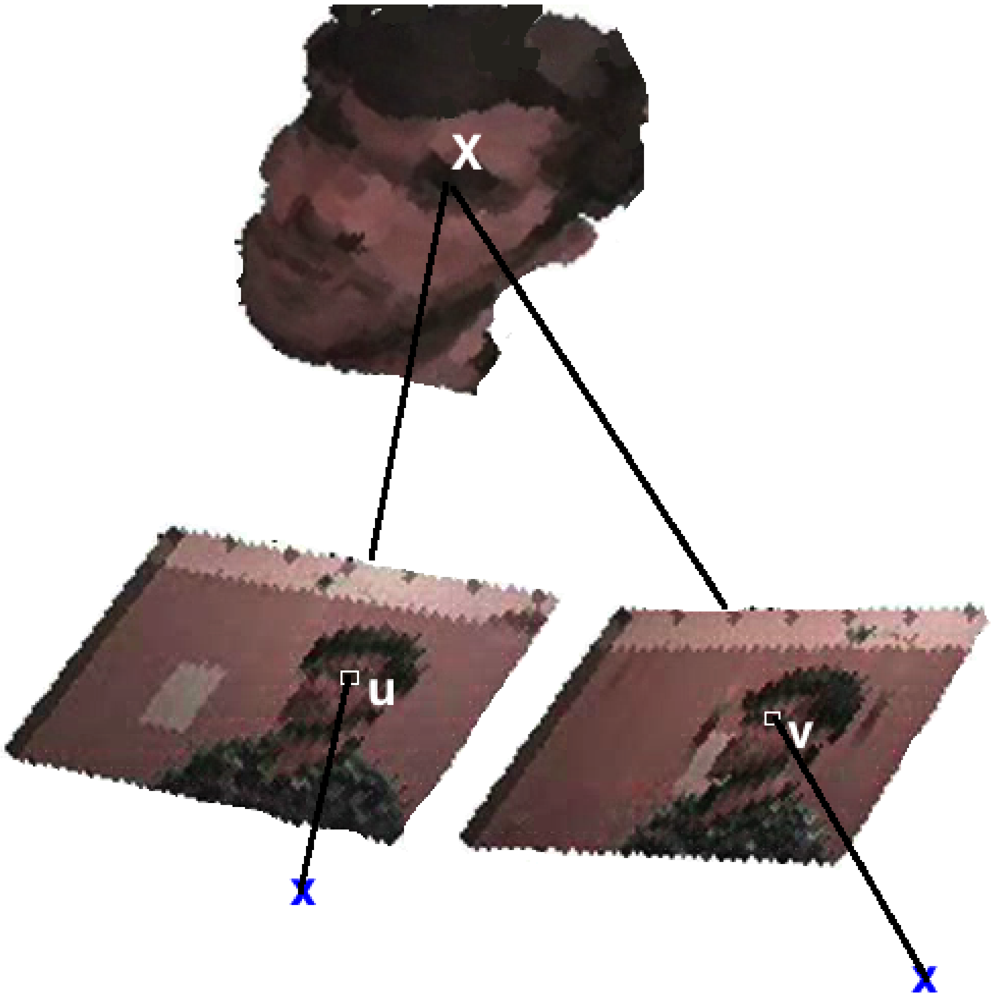
Motivation

- The only problem is, that you do not have the correspondence $(\mathbf{u}, \mathbf{v})$ and naïve matching of pixel neighbourhoods does not work.



Motivation

- This lecture is about
  - how to get 3D points from images captured by known cameras and
  - how to use this knowledge to built state-of-the-art depth sensors.



Outline

- Epipolar geometry
  - Epipolar line, essential and fundamental matrix
  - L_2 estimation of the essential matrix
- Depth sensors: Stereo, Kinect. RealSense, Lidar
- Depth from a single camera and the robust estimation of the essential matrix (RANSAC).

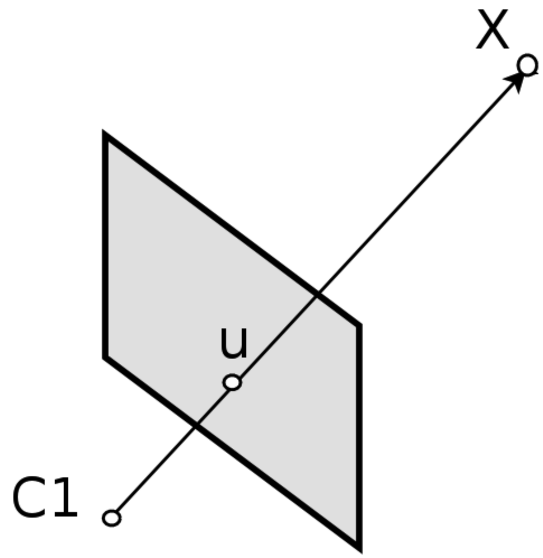
Projection of the 3D point to a single camera

- You are given 3×4 camera matrix $\mathtt{P} = \begin{bmatrix} \mathbf{p}_1^\top \\ \mathbf{p}_2^\top \\ \mathbf{p}_3^\top \end{bmatrix}$
- 3D point with homogeneous coordinates $\mathbf{X}$ projects on pixel $\mathbf{u}$

Projection of the 3D point to a single camera

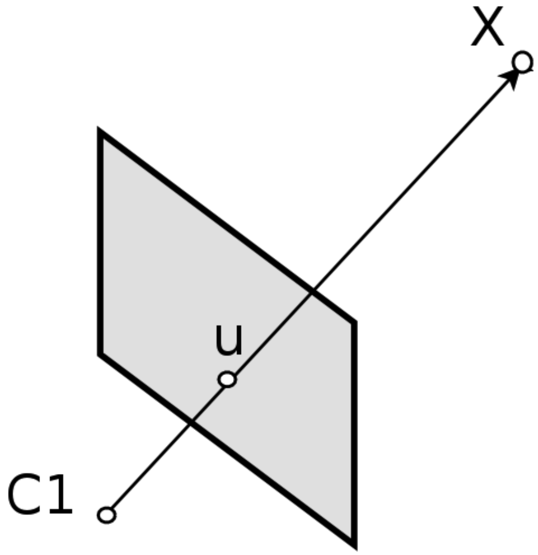
- You are given 3×4 camera matrix $\mathtt{P} = \begin{bmatrix} \mathbf{p}_1^\top \\ \mathbf{p}_2^\top \\ \mathbf{p}_3^\top \end{bmatrix}$
- 3D point with homogeneous coordinates $\mathbf{X}$ projects on pixel $\mathbf{u}$

$$u_1 = \frac{\mathbf{p}_1^\top \mathbf{X}}{\mathbf{p}_3^\top \mathbf{X}}, \qquad u_2 = \frac{\mathbf{p}_2^\top \mathbf{X}}{\mathbf{p}_3^\top \mathbf{X}}$$



Projection of the 3D point to a single camera

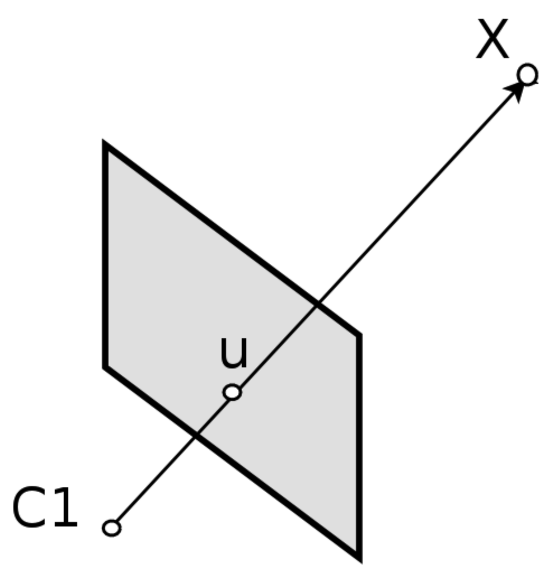
- What if $\mathbf{u}$ is known? Which $\mathbf{X}$ correspond to $\mathbf{u}$?



Projection of the 3D point to a single camera

- What if $\mathbf{u}$ is known? Which $\mathbf{X}$ correspond to $\mathbf{u}$?
- All 3D points corresponding to pixel $\mathbf{u}$ lies in 1D linear subspace (ray) of 3D space (2 linear equations with 3 unknowns):

$$\begin{matrix} u_1 \mathbf{p}_3^\top \mathbf{X} = \mathbf{p}_1^\top \mathbf{X}, \\ u_2 \mathbf{p}_3^\top \mathbf{X} = \mathbf{p}_2^\top \mathbf{X} \end{matrix} \Rightarrow \begin{bmatrix} u_1 \mathbf{p}_3^\top - \mathbf{p}_1^\top \\ u_2 \mathbf{p}_3^\top - \mathbf{p}_2^\top \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \mathbf{0}$$

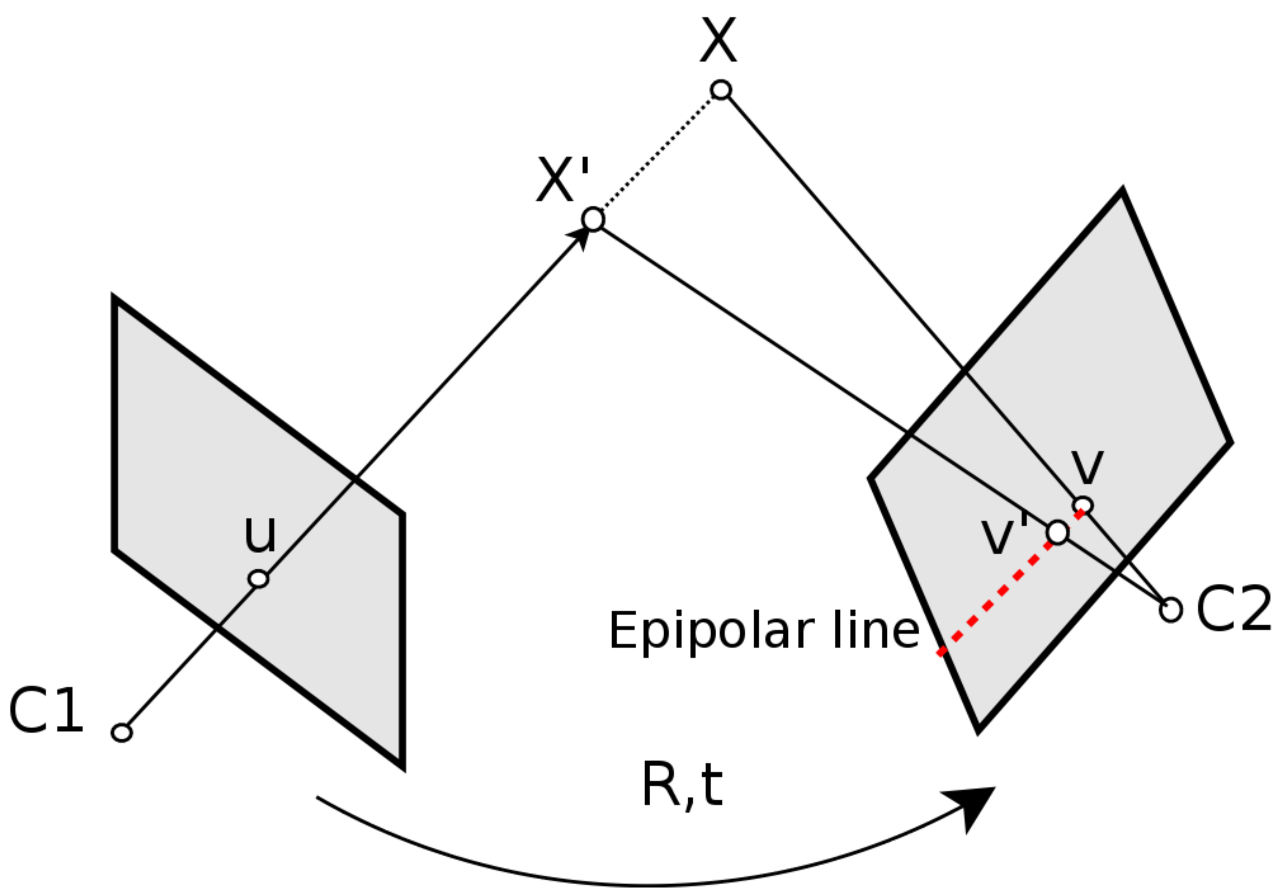


Fundamental matrix

- Projection of the ray from $\mathbf{u}$ into a second camera is called epipolar line

$$\{\mathbf{v} \mid \mathbf{u}^\top \mathrm{F}\,\mathbf{v} = 0\},$$

- where matrix $\mathtt{F} = \mathtt{K}^{-\top}(\mathtt{R} \times \mathbf{t})\mathtt{K}^{-1}$ is called fundamental matrix.



Essential matrix

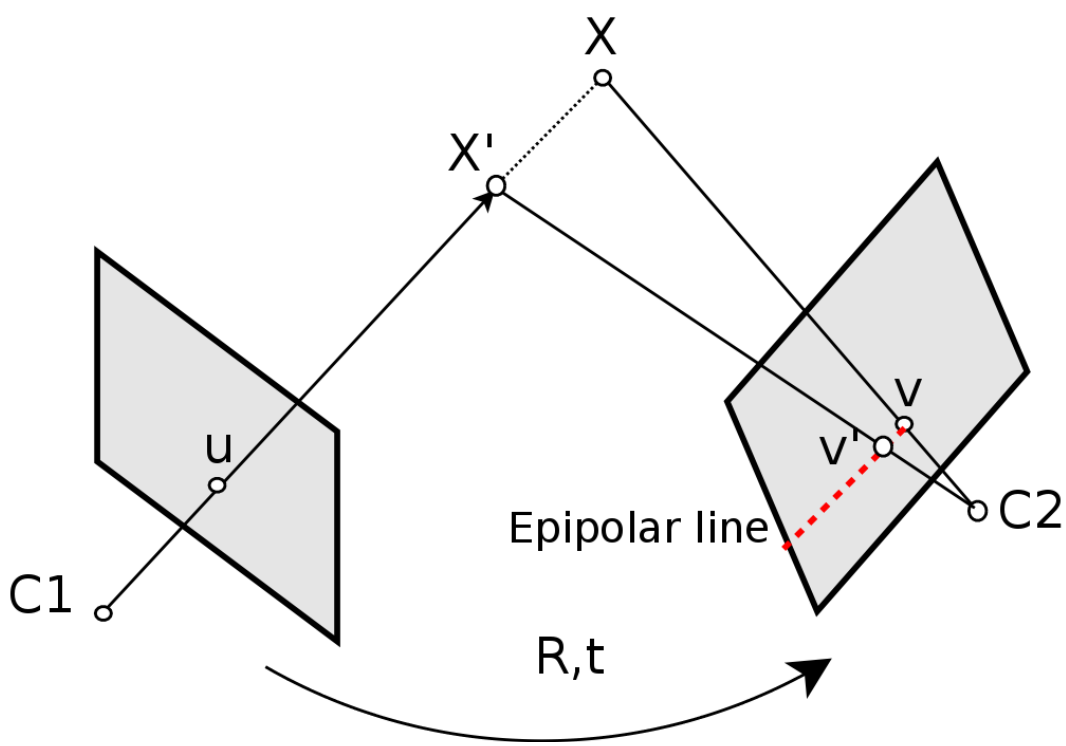
- We assume that $\mathtt{K}$ is known (i.e. the camera is calibrated).

Essential matrix

- We assume that $\mathtt{K}$ is known (i.e. the camera is calibrated).
- We normalize coordinates $\mathbf{u_n} = \mathtt{K}^{-1}\mathbf{u}$, $\mathbf{v_n} = \mathtt{K}^{-1}\mathbf{v}$ and pretend that $\mathtt{K}$ is identity.

Essential matrix

- We assume that $\mathtt{K}$ is known (i.e. the camera is calibrated).
- We normalize coordinates $\mathbf{u_n} = \mathtt{K}^{-1}\mathbf{u}$, $\mathbf{v_n} = \mathtt{K}^{-1}\mathbf{v}$ and pretend that $\mathtt{K}$ is identity.
- Epipolar line wrt normalized coordinates is $\{\mathbf{v_n} \mid \mathbf{u_n}^\top \mathtt{E}\, \mathbf{v_n} = 0\}$, where matrix $\mathtt{E} = \mathtt{R} \times \mathbf{t}$ is called essential matrix.



Derivation: `https://www.robots.ox.ac.uk/~vgg/hzbook/hzbook2/HZepipolar.pdf`

What is the essential matrix good for?

- **Important result 1:**
  - If camera motion is **known** (e.g. stereo), then
  - all possible correspondences of point $\mathbf{u}$ lie on the epipolar line (i.e. either $\{\mathbf{v} \mid \mathbf{u}^\top \mathbf{F}\,\mathbf{v} = 0\}$ or $\{\mathbf{v_n} \mid \mathbf{u_n}^\top \mathbf{E}\,\mathbf{v_n} = 0\}$).

What is the essential matrix good for?

- **Important result 1:**
  - If camera motion is **known** (e.g. stereo), then
  - all possible correspondences of point $\mathbf{u}$ lie on the epipolar line (i.e. either $\{\mathbf{v} \mid \mathbf{u}^\top \mathbf{F}\,\mathbf{v} = 0\}$ or $\{\mathbf{v_n} \mid \mathbf{u_n}^\top \mathbf{E}\,\mathbf{v_n} = 0\}$).

- **Important result 2:**
  - If camera motion is **unknown** (e.g. motion of a single camera), then
  - the essential matrix determines relative position of cameras (i.e. motion), since there exist unique decomposition $\mathtt{E} = \mathtt{R} \times \mathbf{t}$.

What is the essential matrix good for?

- **Important result 1:**
  - If camera motion is **known** (e.g. stereo), then
  - all possible correspondences of point $\mathbf{u}$ lie on the epipolar line (i.e. either $\{\mathbf{v} \mid \mathbf{u}^\top \mathtt{F}\,\mathbf{v} = 0\}$ or $\{\mathbf{v_n} \mid \mathbf{u_n}^\top \mathtt{E}\,\mathbf{v_n} = 0\}$).

- **Important result 2:**
  - If camera motion is **unknown** (e.g. motion of a single camera), then
  - the essential matrix determines relative position of cameras (i.e. motion), since there exist unique decomposition $\mathtt{E} = \mathtt{R} \times \mathtt{t}$.

- From now on, we drop the index n in normalized coordinates.

- How do we obtain the essential/fundamental matrix?

Compute essential matrix by minimizing L2-norm

- Let us assume that we have several correct correspondences.

Compute essential matrix by minimizing L2-norm

- Let us assume that we have several correct correspondences.
- Essential matrix $\mathtt{E}$ is just a solution of (overdetermined) homogeneous system of linear equations.

Compute essential matrix by minimizing L2-norm

- Let us assume that we have several correct correspondences.
- Essential matrix $\mathtt{E}$ is just a solution of (overdetermined) homogeneous system of linear equations.
- For each correspondence pair $\mathbf{u}, \mathbf{v}$, the following holds:

$$\mathbf{u}^\top \mathbf{E}\,\mathbf{v} = \mathbf{u}^\top \begin{bmatrix} \mathbf{e}_1^\top \\ \mathbf{e}_2^\top \\ \mathbf{e}_3^\top \end{bmatrix} \mathbf{v} = \mathbf{u}^\top \begin{bmatrix} \mathbf{e}_1^\top \mathbf{v} \\ \mathbf{e}_2^\top \mathbf{v} \\ \mathbf{e}_3^\top \mathbf{v} \end{bmatrix} = [u_1 \mathbf{e}_1^\top \mathbf{v} + u_2 \mathbf{e}_2^\top \mathbf{v} + u_3 \mathbf{e}_3^\top \mathbf{v}] =$$

Compute essential matrix by minimizing L2-norm

- Let us assume that we have several correct correspondences.
- Essential matrix $\mathtt{E}$ is just a solution of (overdetermined) homogeneous system of linear equations.
- For each correspondence pair $\mathbf{u}, \mathbf{v}$, the following holds:

$$\mathbf{u}^\top \mathbf{E}\,\mathbf{v} = \mathbf{u}^\top \begin{bmatrix} \mathbf{e}_1^\top \\ \mathbf{e}_2^\top \\ \mathbf{e}_3^\top \end{bmatrix} \mathbf{v} = \mathbf{u}^\top \begin{bmatrix} \mathbf{e}_1^\top \mathbf{v} \\ \mathbf{e}_2^\top \mathbf{v} \\ \mathbf{e}_3^\top \mathbf{v} \end{bmatrix} = [u_1 \mathbf{e}_1^\top \mathbf{v} + u_2 \mathbf{e}_2^\top \mathbf{v} + u_3 \mathbf{e}_3^\top \mathbf{v}] =$$

$$= [u_1 \mathbf{v}^\top \; u_2 \mathbf{v}^\top \; u_3 \mathbf{v}^\top] \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{bmatrix} = 0$$

- It must hold for all correspondece pairs $\mathbf{u}_i$, $\mathbf{v}_i$, therefore:

$$\begin{bmatrix} u_{11}\mathbf{v}_1^\top & u_{12}\mathbf{v}_1^\top & u_{13}\mathbf{v}_1^\top \\ u_{21}\mathbf{v}_2^\top & u_{22}\mathbf{v}_2^\top & u_{23}\mathbf{v}_2^\top \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{bmatrix} = \mathbf{0}$$

Compute essential matrix by minimizing L2-norm

- It is just homogeneous set of linear equations:

$$\underbrace{\begin{bmatrix} u_{11}\mathbf{v}_1^\top & u_{12}\mathbf{v}_1^\top & u_{13}\mathbf{v}_1^\top \\ u_{21}\mathbf{v}_2^\top & u_{22}\mathbf{v}_2^\top & u_{23}\mathbf{v}_2^\top \\ & \vdots\vdots\vdots & \end{bmatrix}}_{\mathtt{A}} \underbrace{\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{bmatrix}}_{\mathbf{e}} = \mathbf{0}$$

- We want to avoid trivial solution $\mathbf{e}_1 = \mathbf{e}_2 = \mathbf{e}_3 = \mathbf{0}$,
- therefore the following optimization task (constrained LSQ) is solved:

$$\arg\min_{\mathbf{e}} \|\mathtt{A}\mathbf{e}\| \;\; \text{subject to} \;\; \|\mathbf{e}\| = 1$$

- the solution is singular vector of matrix $\mathtt{A}$ corresponding to the smallest singular value (can be found via SVD or eigenvectors/eigenvalues of $\mathtt{A}\mathtt{A}^\top$)

Compute essential matrix by minimizing L2-norm

- The same is valid for the estimation of the fundamental matrix from not normalized coordinates.

Compute essential matrix by minimizing L2-norm

- The same is valid for the estimation of the fundamental matrix from not normalized coordinates.
- L_2-norm works only in a controlled environment (e.g. offline stereo calibration).

Compute essential matrix by minimizing L2-norm

- The same is valid for the estimation of the fundamental matrix from not normalized coordinates.
- L_2-norm works only in a controlled environment (e.g. offline stereo calibration).
- I will show how essential/fundamental matrix allows to estimate correspondences in state-of-the-art depth (3D) sensors.

Stereo

- Pair of cameras mounted on a rigid body, which provides depth (3D points) of the scene (simulates human binocular vision).
- Relative position of cameras fixed

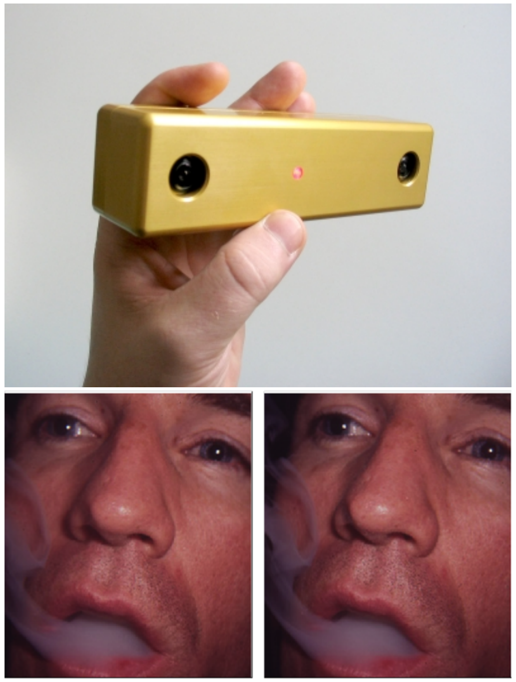
[0] Courtesy of prof.Boris Flach for original stereo images and depth images

Stereo

- Pair of cameras mounted on a rigid body, which provides depth (3D points) of the scene (simulates human binocular vision).
- Relative position of cameras fixed
- **offline**: fundamental matrix estimated from known correspondences.

[0] Courtesy of prof.Boris Flach for original stereo images and depth images

Stereo

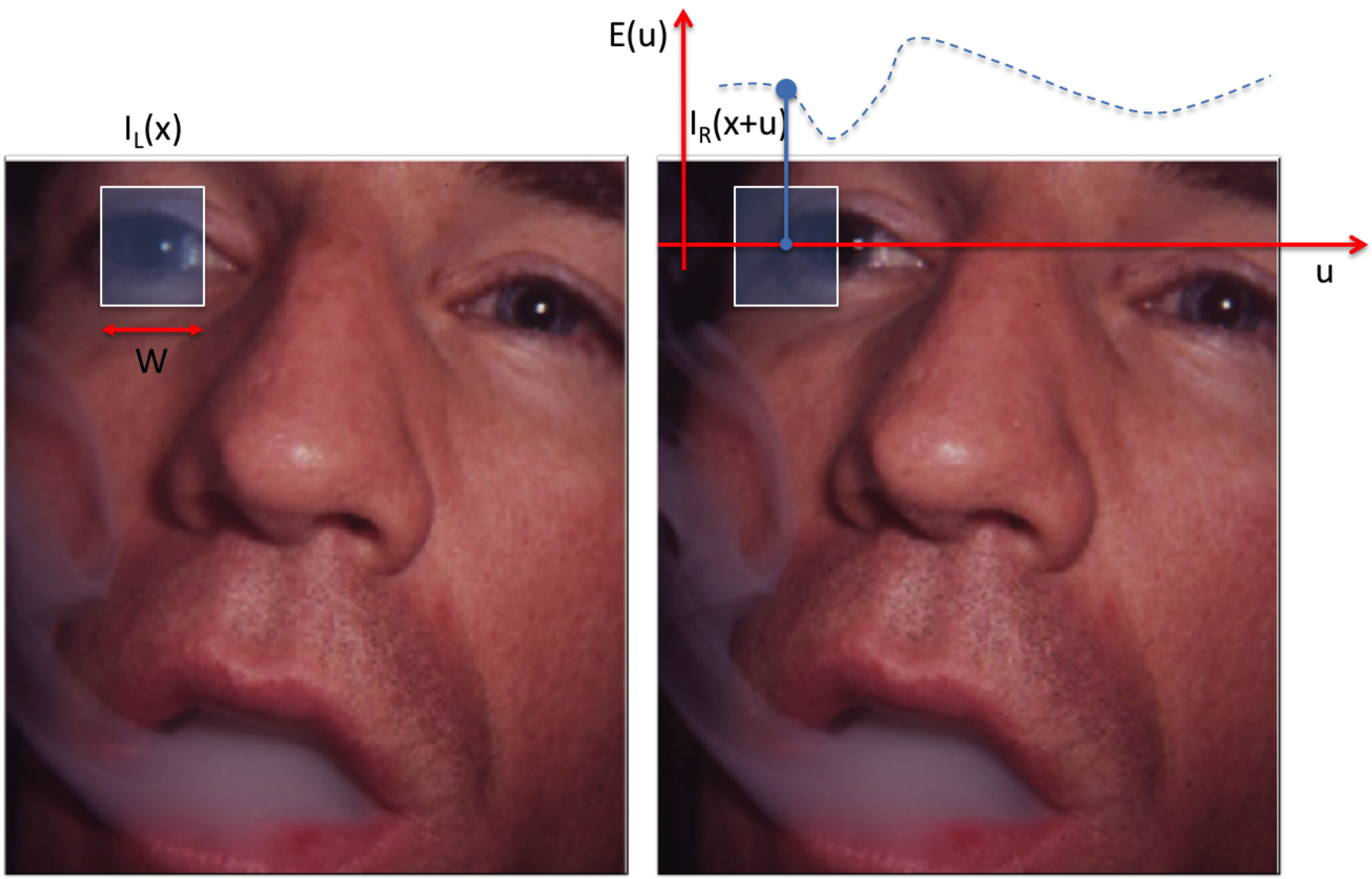


- Pair of cameras mounted on a rigid body, which provides depth (3D points) of the scene (simulates human binocular vision).
- Relative position of cameras fixed
- **offline**: fundamental matrix estimated from known correspondences.
- **online**: correspondences searched along epipolar lines.

[0] Courtesy of prof.Boris Flach for original stereo images and depth images

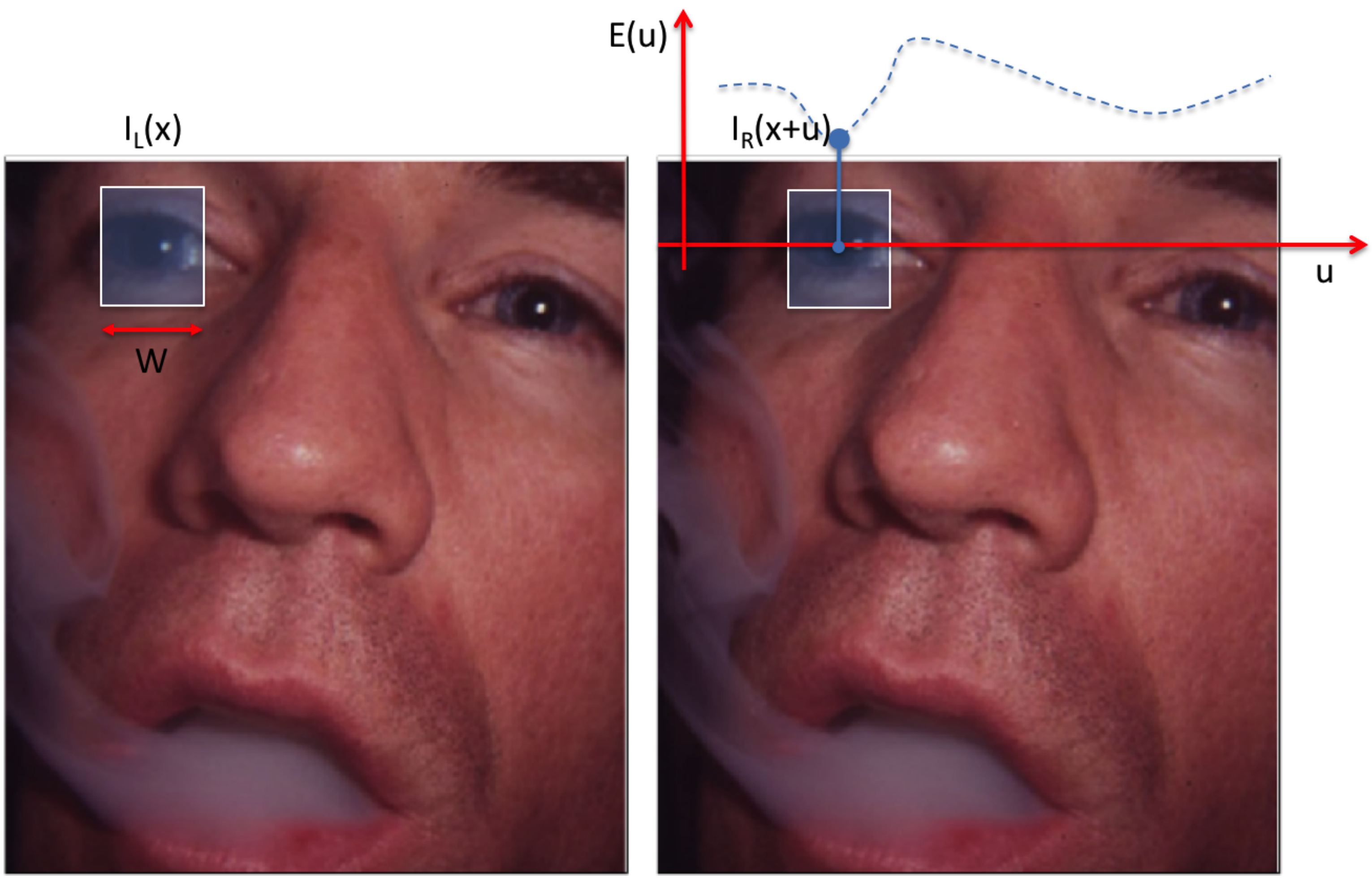
Stereo

Block-matching energy function: $E(u) = \sum_{x \in W}(I_L(x) - I_R(x+u))^2$



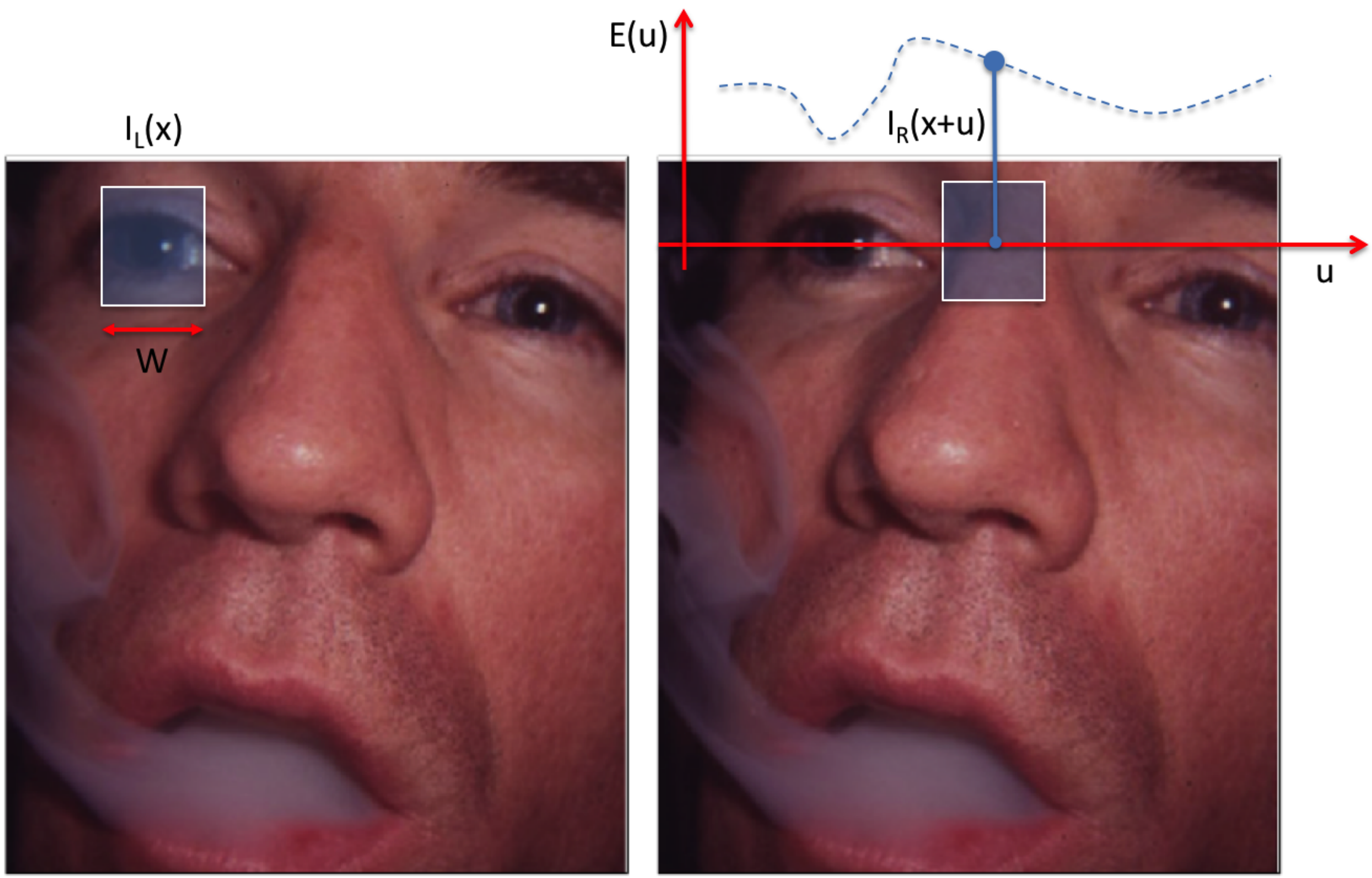
Stereo

Block-matching energy function: $E(u) = \sum_{x \in W} (I_L(x) - I_R(x+u))^2$



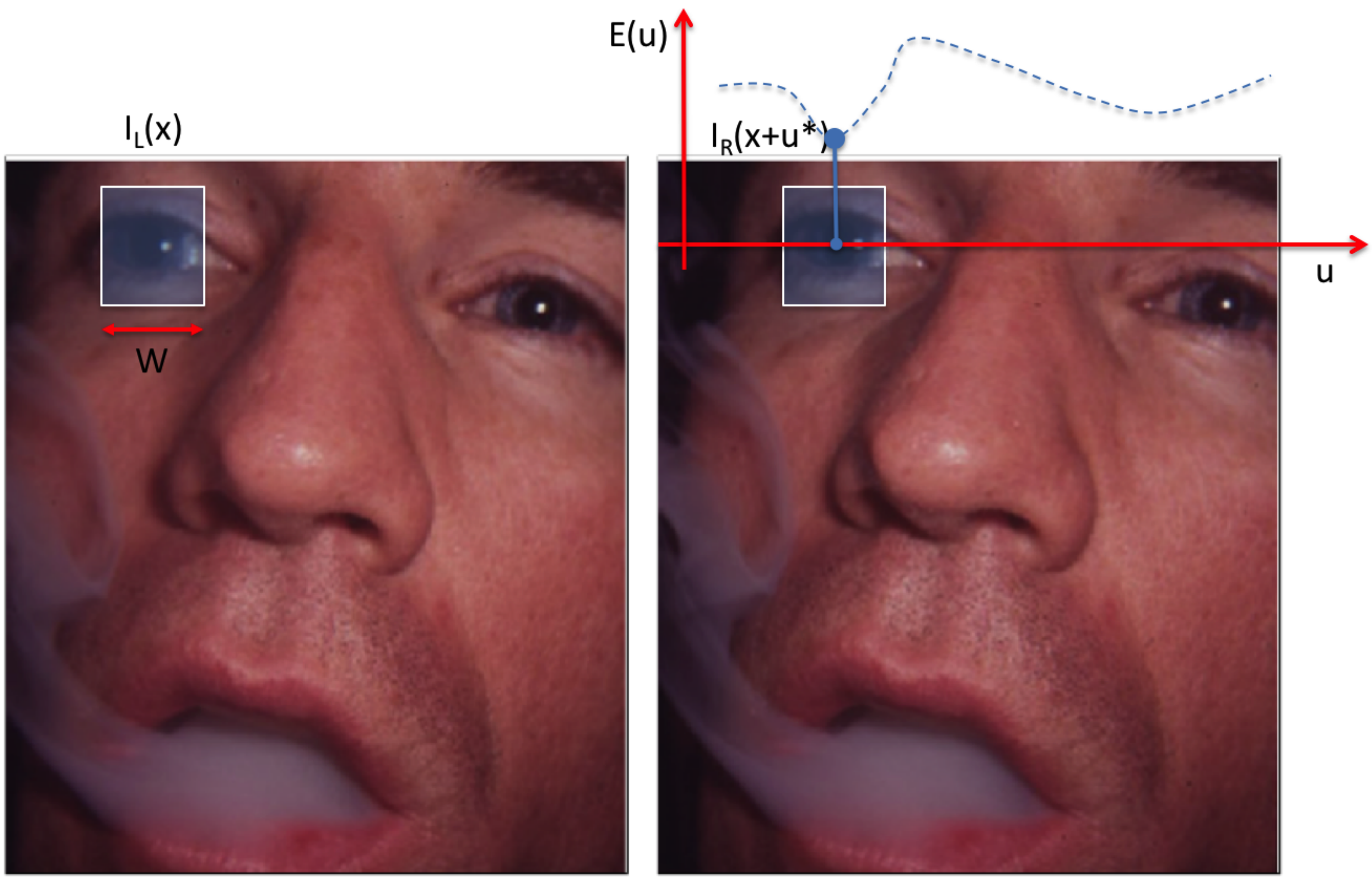
Stereo

Block-matching energy function: $E(u) = \sum_{x \in W} (I_L(x) - I_R(x+u))^2$



Stereo

Correspondence for each pixel estimated separately: $u^* = \arg\min_u E(u)$



Stereo

Correspondence for each pixel estimated separately:
$u_1^* = \arg\min_u E_1(u)$,

Stereo

Correspondence for each pixel estimated separately:

$u_1^* = \arg\min_u E_1(u), \quad u_2^* = \arg\min_u E_2(u)$

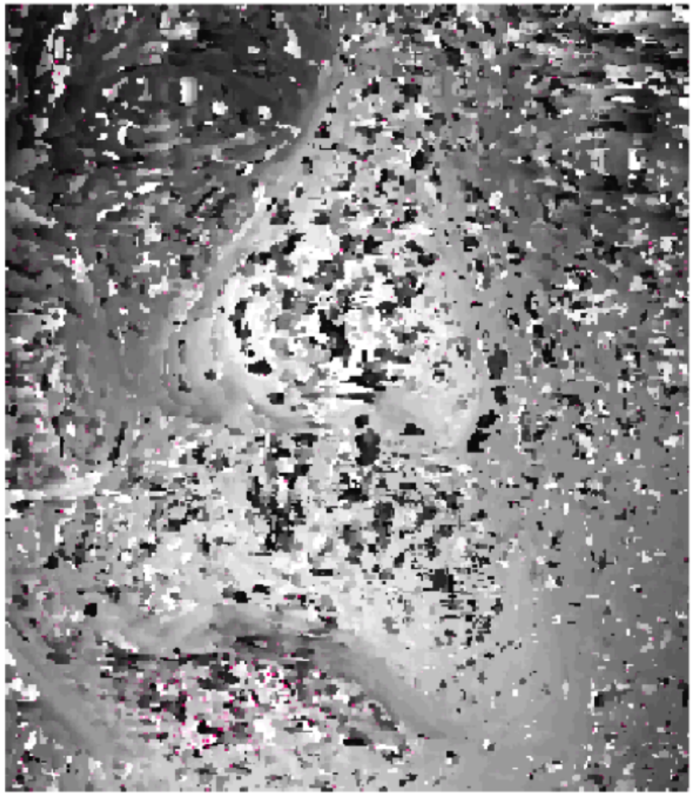
Stereo

Correspondence for each pixel estimated separately:

$$u_1^* = \arg\min_u E_1(u), \quad u_2^* = \arg\min_u E_2(u) \quad \dots \quad u_N^* = \arg\min_u E_N(u)$$

Stereo

How can we improve the result?

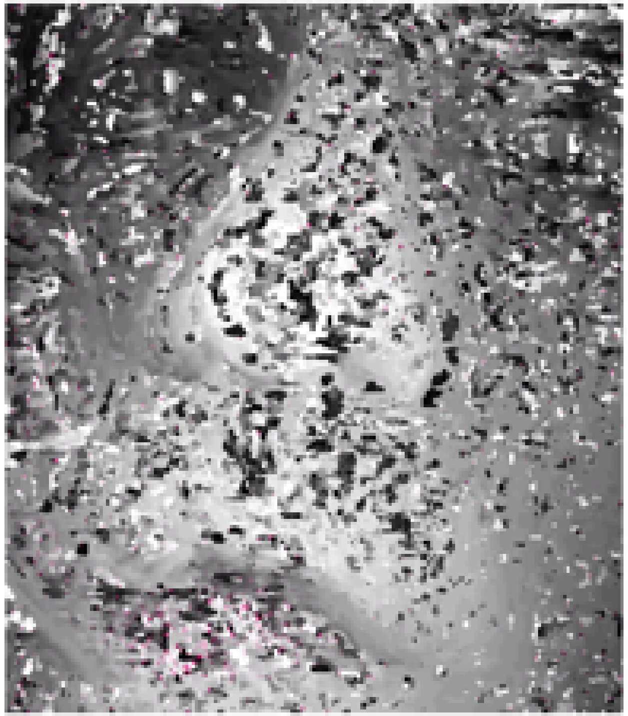


Stereo

Energy with horizontal smoothness term:

$$E_1(u_1) + C(u_2 - u_1)^2 + E_2(u_2) + C(u_3 - u_2)^2 + E_3(u_3) + \cdots + E_N(u_N)$$

Image



Block matching

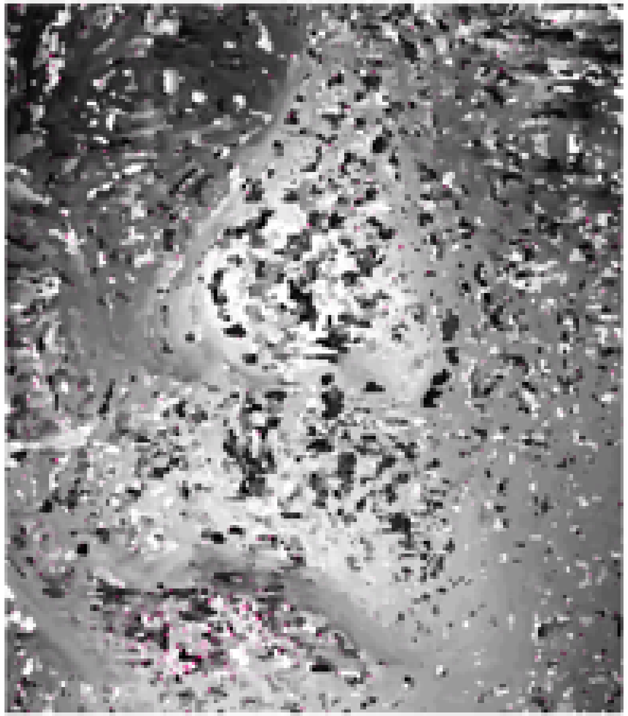
Dynamic programming

Stereo

Dynamic programming solves each line of N pixels separately:

$$u_1^* \dots u_N^* = \arg \min_{u_1 \dots u_N} \sum_{i=1}^{N-1} E_i(u_i, u_{i+1})$$

Image



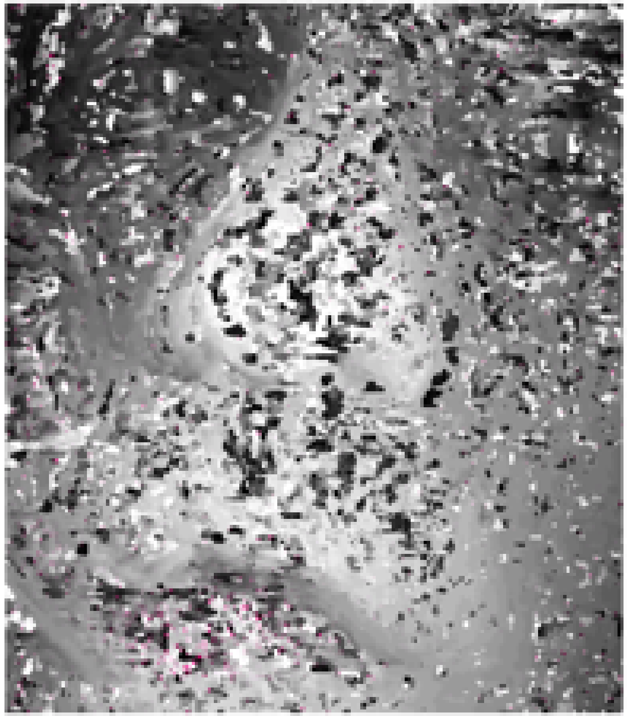
Block matching

Dynamic programming

Stereo

What else can we do?

Image

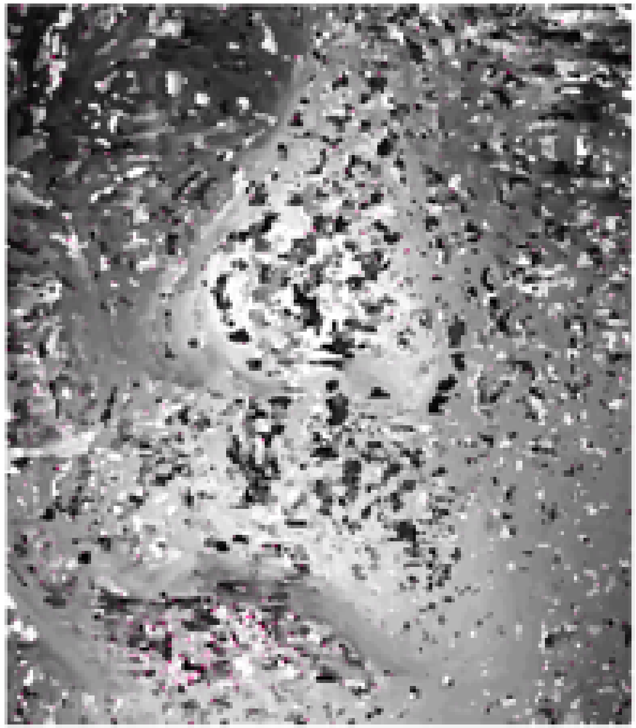


Block matching

Dynamic programming

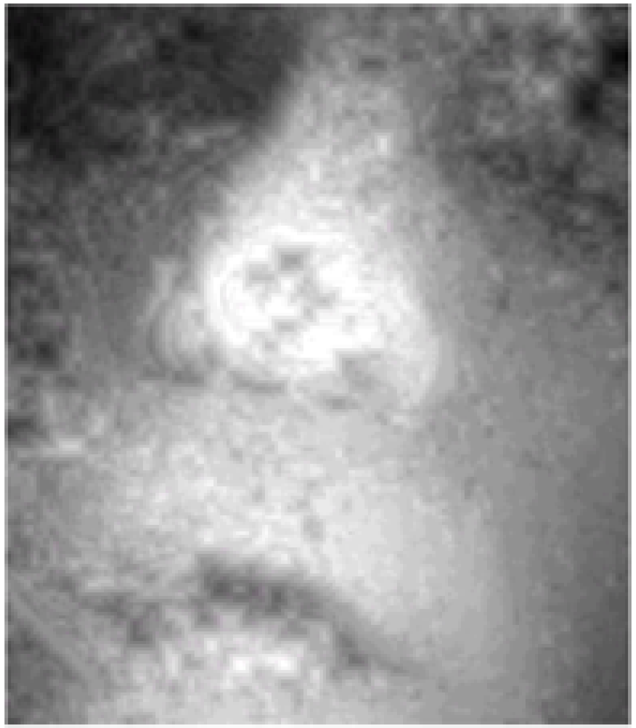
Stereo

Enforce also vertical smoothness $\Rightarrow$ graph energy minimization (computationally demanding optimization solved on specialized chips).



Block matching

Dynamic programming



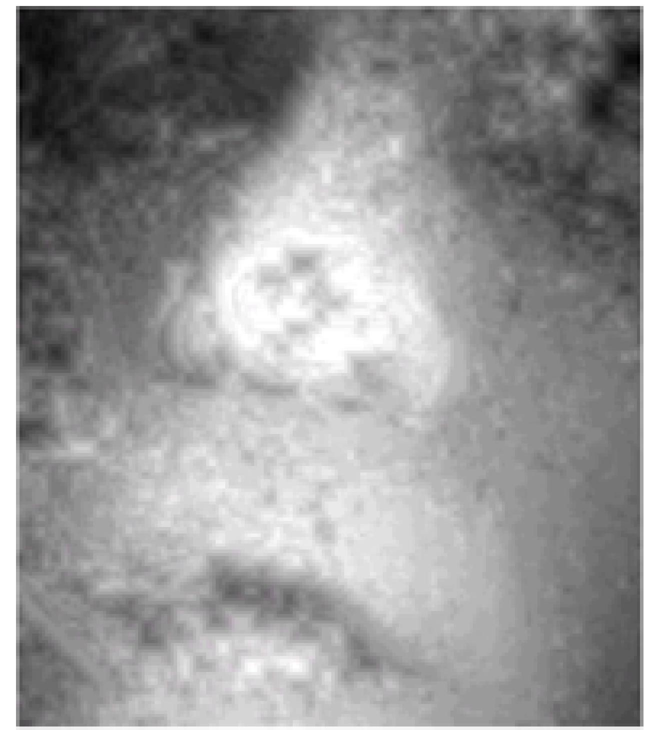
(Min,+) solution

Stereo

Enforce also vertical smoothness $\Rightarrow$ graph energy minimization (computationally demanding optimization solved on specialized chips).

Block matching

Dynamic programming



(Min,+) solution

- **Limitation:** usually works only on sufficiently rich patterns and sufficiently smooth depths.

Stereo competition

- Do you have your own idea how to estimate the depth from stereo images?
- http://vision/middlebury.edu/stereo/data/2014/

Stereo conclusion

- What makes stereo depth estimation complicated?

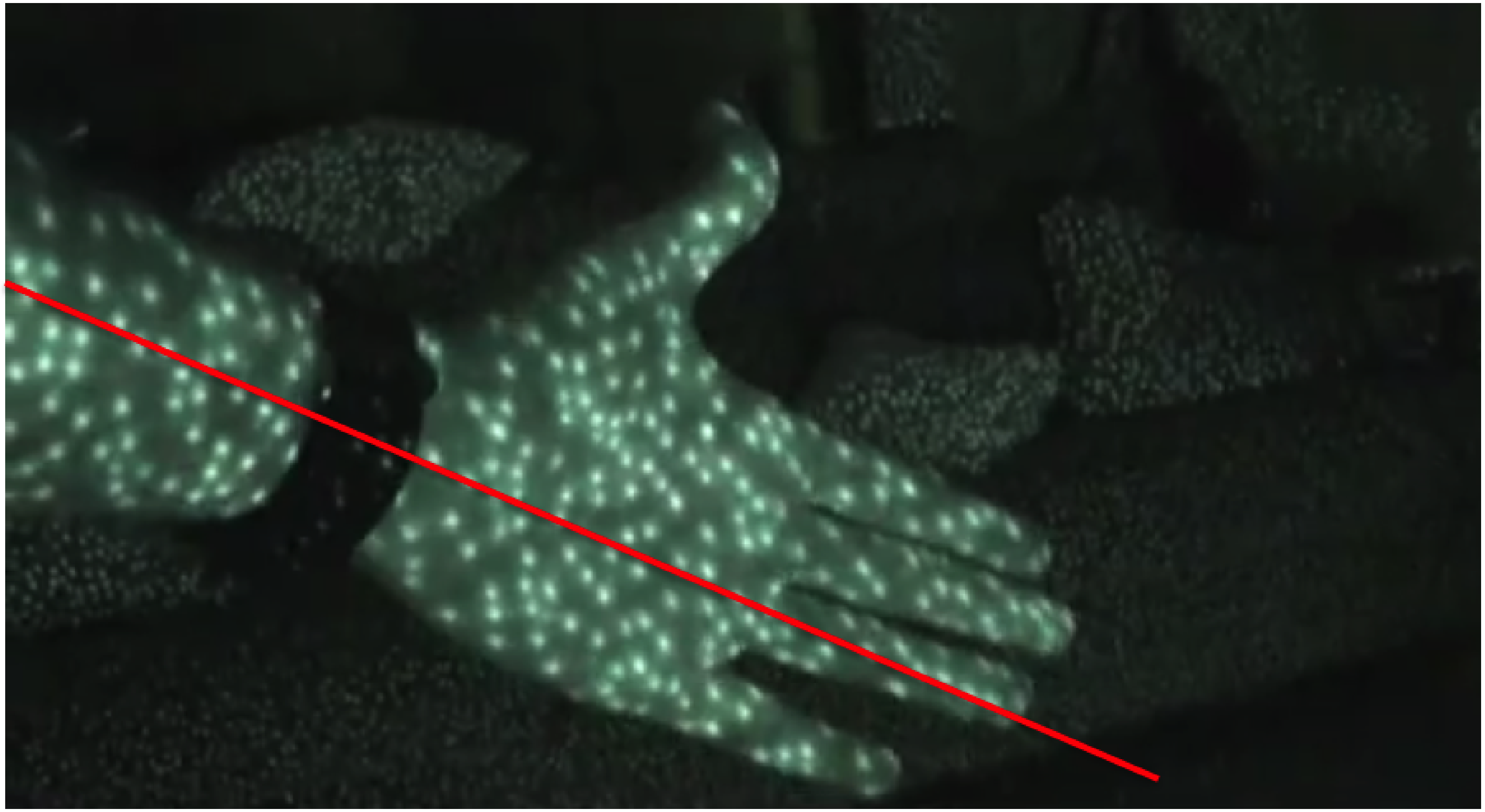
Stereo conclusion

- What makes stereo depth estimation complicated?
- Can you get rid of it?

Kinect (structured-light approach)

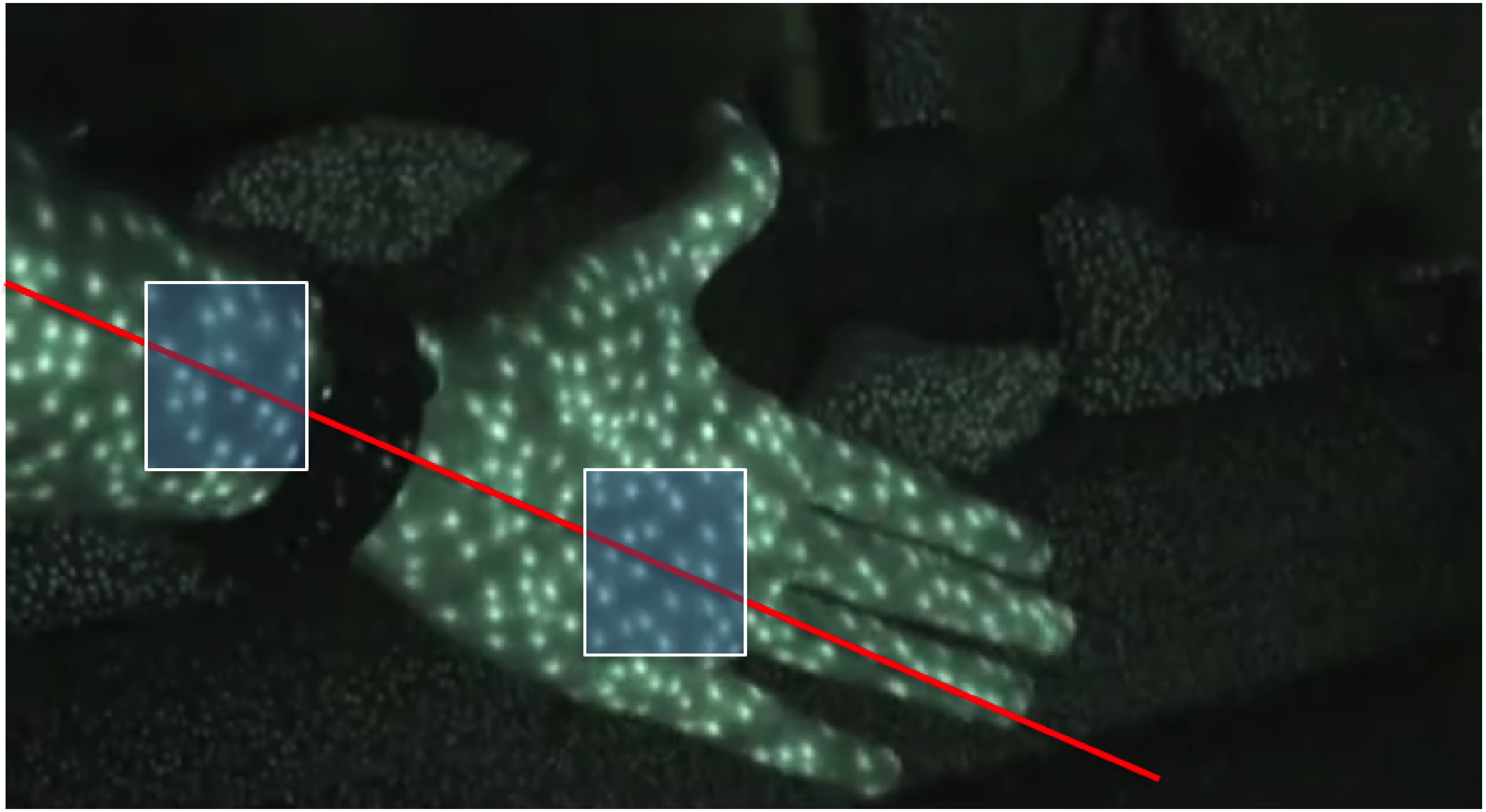
- **Stereo** looks at the same object two-times and estimates the correspondence from two passive RGB images.
- **Kinect** avoids ambiguity by actively projecting a unique IR pattern on the surface and search for its known appearance in the IR camera.

Kinect



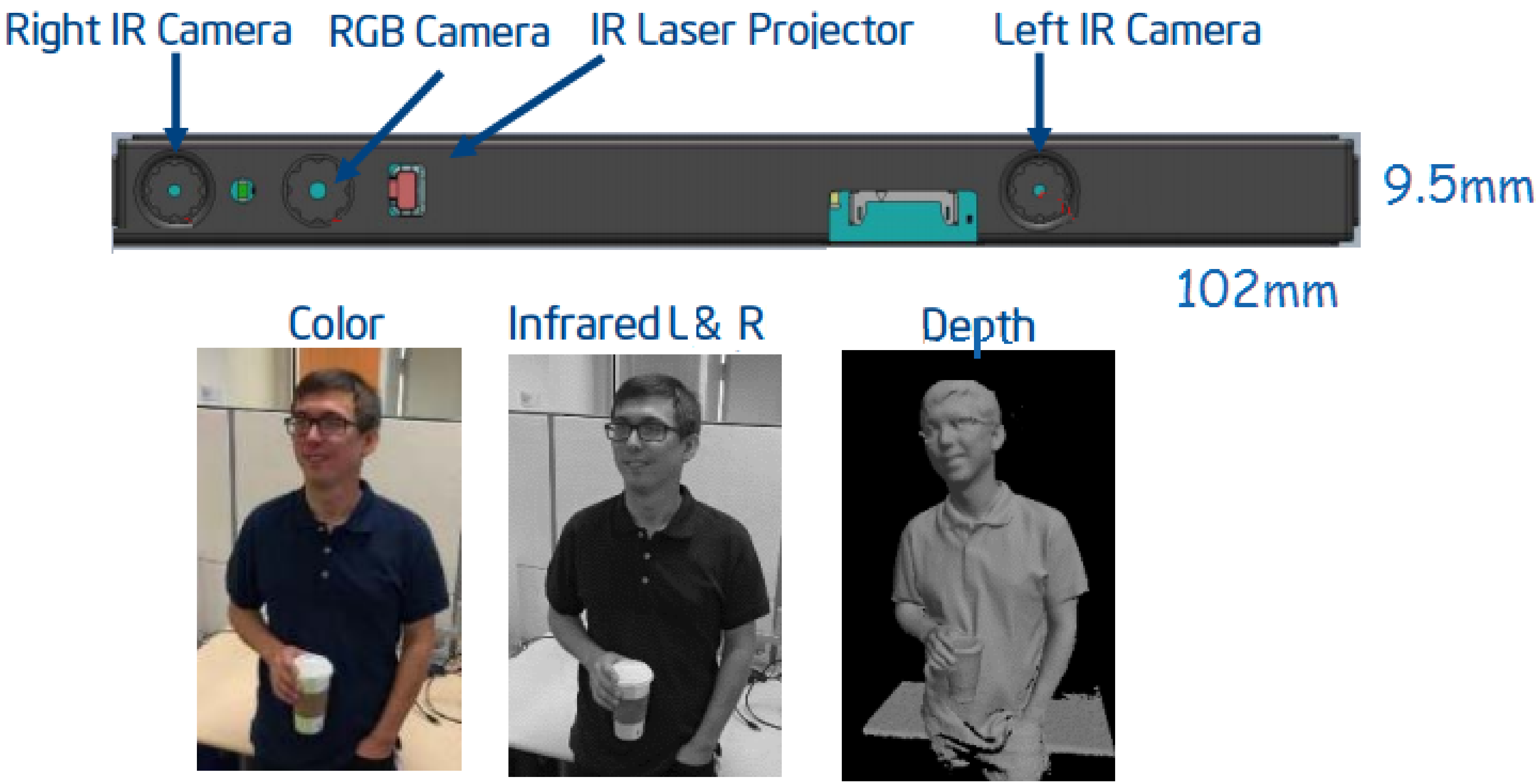
- Since camera-projector relative position is known, correspondence between projected pixel and observed pixel lies again on epipolar lines.

Kinect



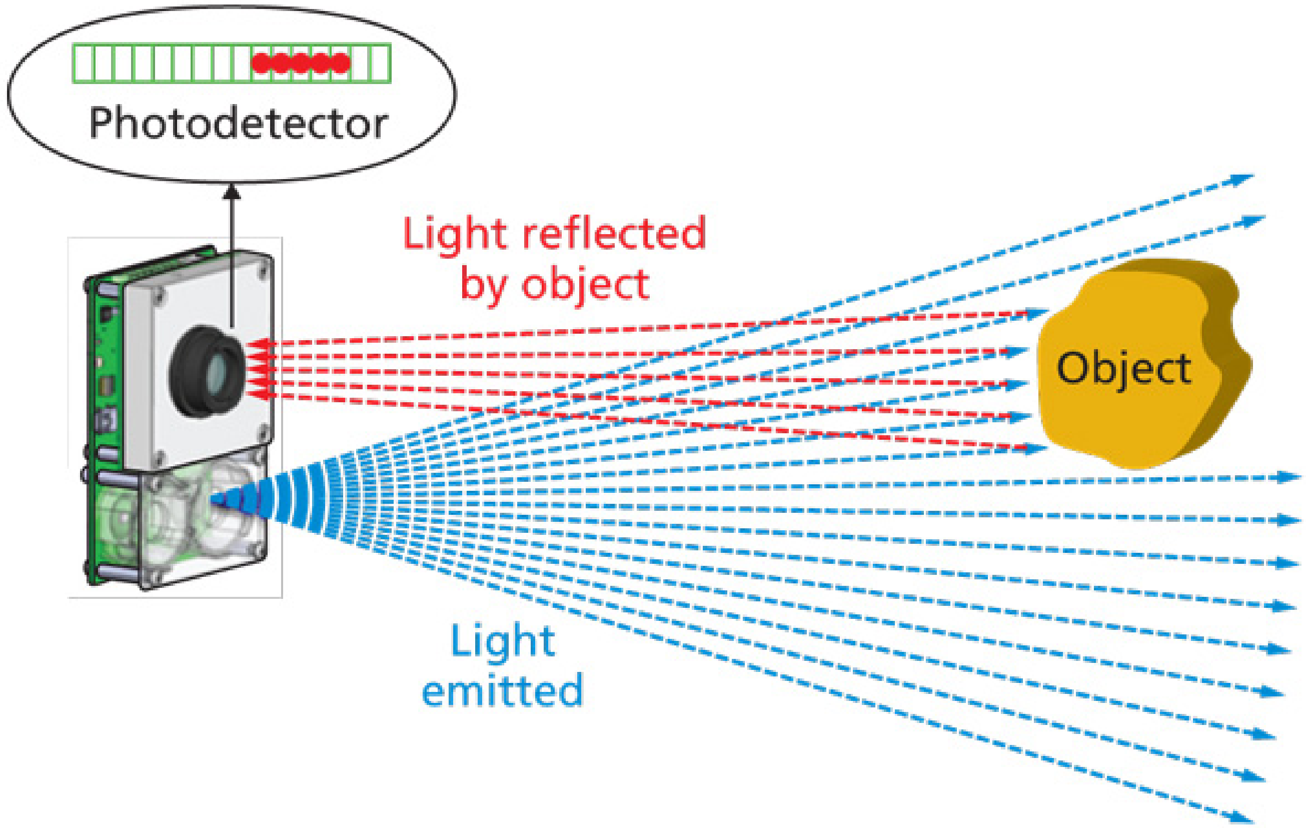
- Unique IR speckle-pattern: no two sub-windows with the same pattern
- Energy along epipolar line has only one strong minimum.
- **Limitation:** works only indoor.

RealSense



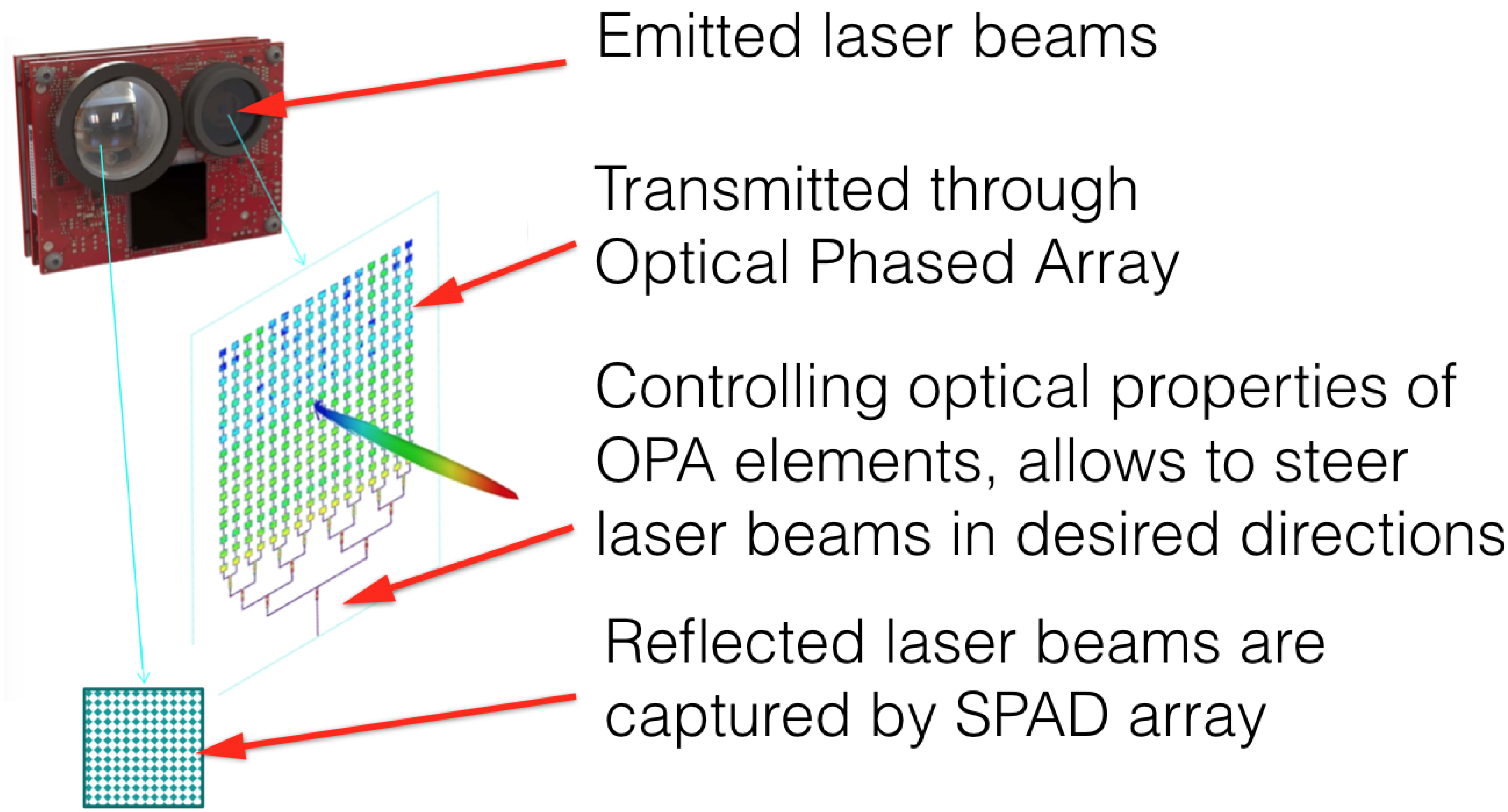
- Hybrid approach one IR projector and two IR cameras.
- Combines advantages of stereo and structured light approach. So far best solution for robotics.

Lidar (Time-of-Flight sensor)



- Light emitted from laser projector is reflected by the object and then captured by photodetector.
- Delay between the light emission and detection determines the depth.
- Usually expensive, low resolution (sweeping plane rotation), heavy, prone to mechanical wear.

Solid State Lidar (Steerable Time-of-Flight sensor)



Images of S3 Lidar redistributed with permission of Quanergy Systems (http://quanergy.com)

- Active ray steering allows to focus measurements on the parts of the scene relevant for the scenario.
- Not yet commercially available.

Conclusions

- Stereo is passive sensor, which works on well only on sufficiently rich patterns
- Structured-light works inside
- Time-of-flight is heavy and prone to mechanical wear
- Active sensors (those which projects something) might interfere!