3d Reconstruction from 2d Images

Antoine Manzanera

ENSTA Paris

ROB313 - Robotic Vision
December 2022

3d Reconstruction from Videos

Reconstructing the scene geometry from videos is useful in many applications: Robot navigation (obstacle detection), Metrology, 3d Cartography, Medicine...

+ It is a cheap and flexible approach: One single passive camera, Adaptive baseline,...

– It strongly relies on scene structure (texture) and precise camera positioning.

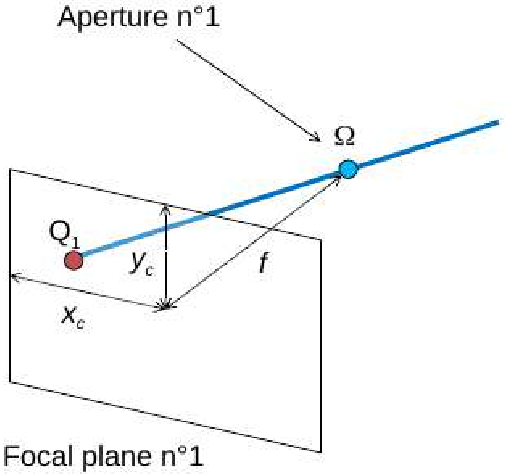
Presentation Outline

1. Introduction to analytical methods
2. Projective Geometry and Camera Matrices
3. Epipolar Geometry and the Fundamental Matrix
4. Depth Estimation and Epipolar Flow
5. Learning based depth prediction

Presentation Outline

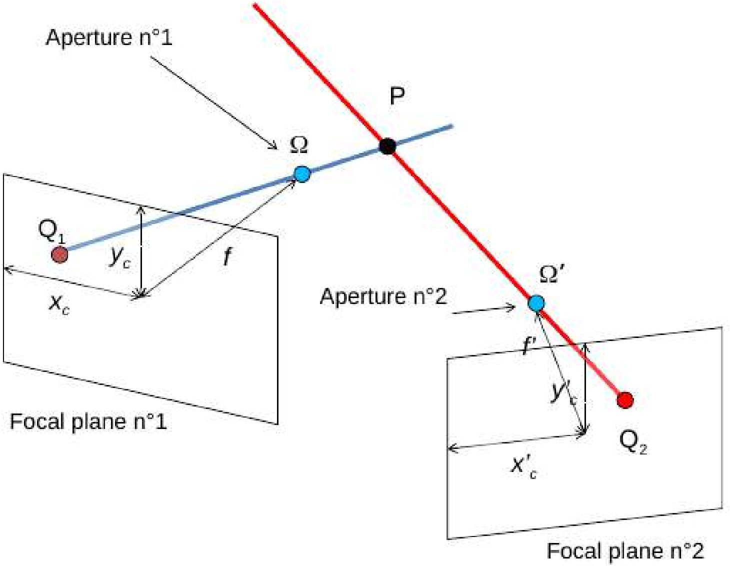
1. Introduction to analytical methods
2. Projective Geometry and Camera Matrices
3. Epipolar Geometry and the Fundamental Matrix
4. Depth Estimation and Epipolar Flow
5. Learning based depth prediction

Principles of Analytical Methods



The geometry of the camera (intrinsic parameters) identifies the projection line of any point in the focal plane.

Principles of Analytical Methods

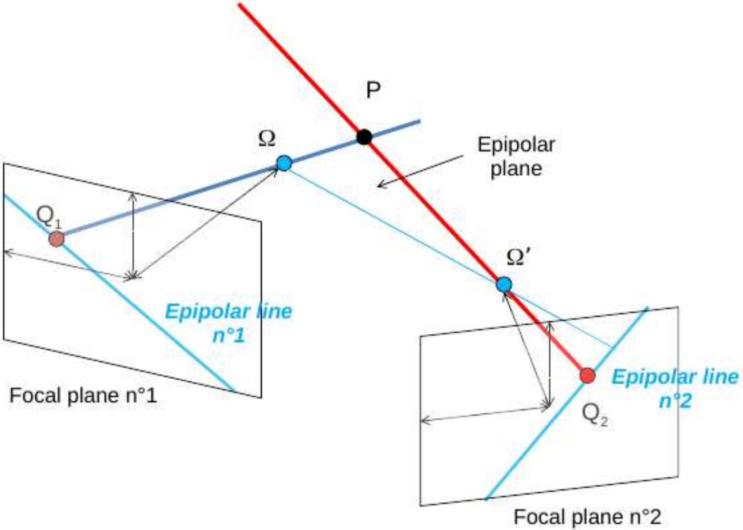


Another position of the camera (extrinsic parameters) allows to recover the 3d position of a point projected on the two focal planes:

$$\Omega P = \Omega\Omega' \frac{\sin \hat{\Omega}'}{\sin \hat{P}}$$

$$\Omega' P = \Omega\Omega' \frac{\sin \hat{\Omega}}{\sin \hat{P}}$$

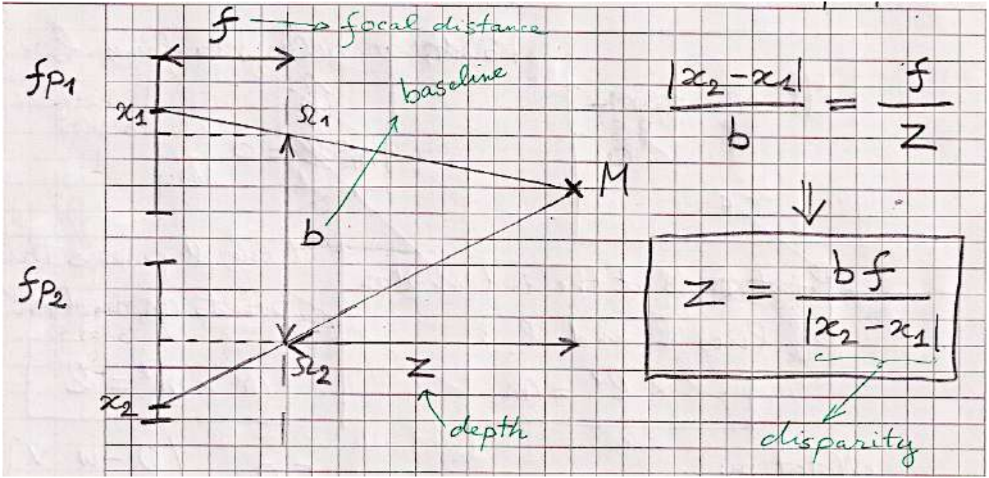
Principles of Analytical Methods



The epipolar constraints may reduce the search area for matching points. It is expressed by the fundamental matrix **F** in the projective geometry framework: $Q_2^t\mathbf{F}Q_1 = 0$.

- $\mathbf{F}Q_1$: epipolar line n.2.
- $Q_2^t\mathbf{F}$: epipolar line n.1.

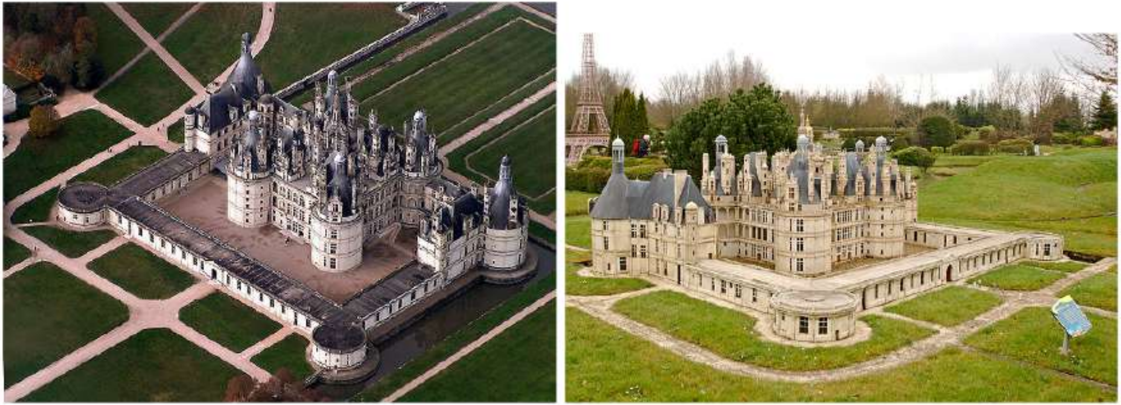
In-plane Ideal Stereovision



Ideal or Rectified or Plenoptic (Single-Lens) Stereovision

Scale Ambiguity

Without knowledge of focal and baseline, depth can at best be estimated up to scale factor! (But look at the contextual clues...):



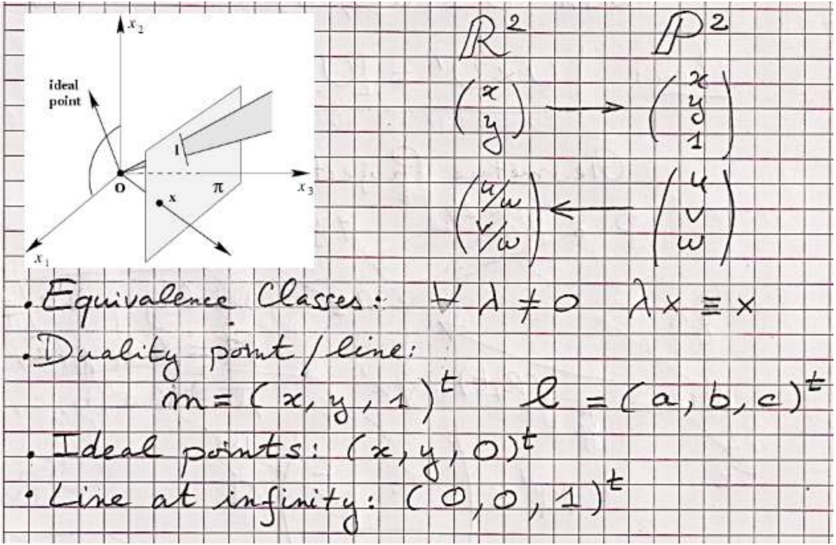
Aerial view of Chambord Castle and 1/30-scale model miniature model in La France Miniature

From **[PhD C. Pinard 2019]**

Presentation Outline

1. Introduction to analytical methods
2. Projective Geometry and Camera Matrices
3. Epipolar Geometry and the Fundamental Matrix
4. Depth Estimation and Epipolar Flow
5. Learning based depth prediction

Projective Geometry in $\mathbb{P}^2$: Reminder


x_2
ideal point
l
O
π
x
x_3
x_1

$$\mathbb{R}^2 \qquad \mathbb{P}^2$$

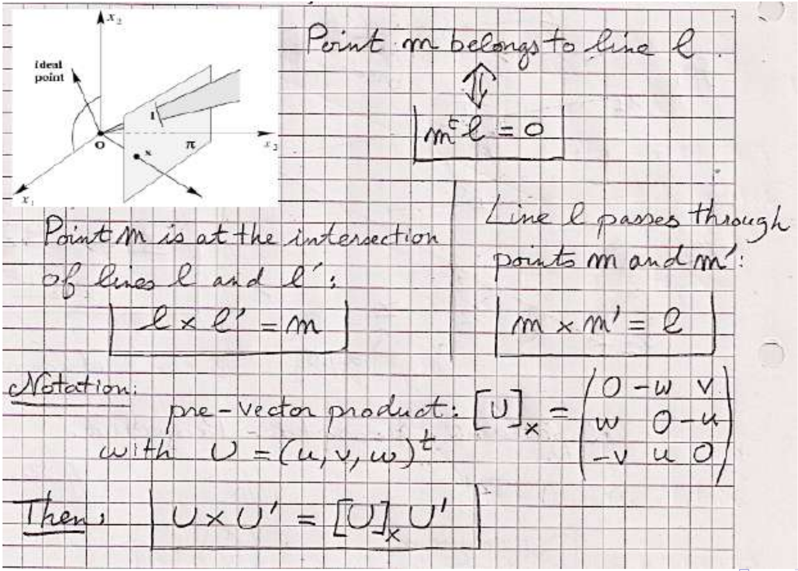
$$\begin{pmatrix} x \\ y \end{pmatrix} \longrightarrow \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} u/w \\ v/w \end{pmatrix} \longleftarrow \begin{pmatrix} u \\ v \\ w \end{pmatrix}$$

- Equivalence Classes: $\forall \lambda \neq 0 \quad \lambda x \equiv x$
- Duality point/line:

  $m = (x, y, 1)^t \qquad l = (a, b, c)^t$
- Ideal points: $(x, y, 0)^t$
- Line at infinity: $(0, 0, 1)^t$

Projective Geometry in $\mathbb{P}^2$: Reminder


x_2
ideal point
l
O
π
x
x_3
x_1

Point m belongs to line ℓ.

$$\Updownarrow$$

$$\boxed{m^t \ell = 0}$$

Point m is at the intersection of lines ℓ and ℓ':

$$\boxed{\ell \times \ell' = m}$$

Line ℓ passes through points m and m':

$$\boxed{m \times m' = \ell}$$

Notation:

pre-vector product: $[U]_\times = \begin{pmatrix} 0 & -w & v \\ w & 0 & -u \\ -v & u & 0 \end{pmatrix}$

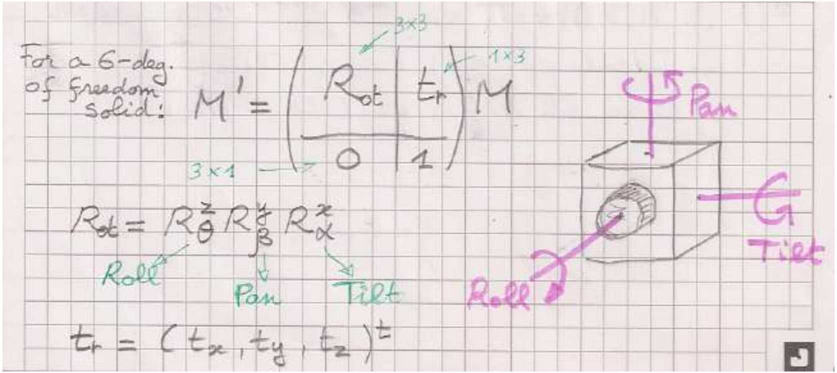
with $U = (u, v, w)^t$

Then, $\boxed{U \times U' = [U]_\times U'}$

Projective Geometry in $\mathbb{P}^3$

- $\mathbb{R}^3 \leftrightarrow \mathbb{P}^3$: $(X, Y, Z) \rightarrow (X, Y, Z, 1)$; $(u/h, v/h, w/h) \leftarrow (u, v, w, h)$
- Duality point / plane: $M = (X, Y, Z, 1)^t$ / $\Pi = (a, b, c, d)$.
- Lines are defined from 2 points or from 2 planes!

$\mathbb{P}^3$ allows to express linearly affine transformations:



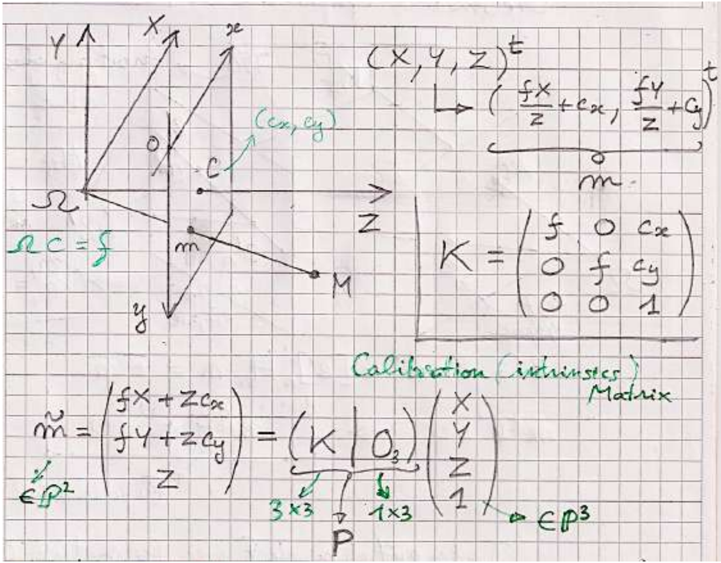
For a 6-deg. of freedom solid:

$$M' = \begin{pmatrix} \overset{3\times3}{R_{ot}} & \overset{1\times3}{t_r} \\ \underset{3\times1}{0} & 1 \end{pmatrix} M$$

$$R_{ot} = \underbrace{R^z_\theta}_{\text{Roll}} \underbrace{R^y_\beta}_{\text{Pan}} \underbrace{R^x_\alpha}_{\text{Tilt}}$$

$$t_r = (t_x, t_y, t_z)^t$$

Camera (Calibration) Matrix: Intrinsics


Y
X
x
O
C
(cx, cy)
Ω
Z
m
M
y
ΩC = f

$$(X,Y,Z)^t \mapsto \underbrace{\left(\frac{fX}{Z}+c_x,\ \frac{fY}{Z}+c_y\right)^t}_{m}$$

$$K = \begin{pmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{pmatrix}$$

Calibration (intrinsics) Matrix

$$\underbrace{\tilde{m}}_{\in \mathbb{P}^2} = \begin{pmatrix} fX + Zc_x \\ fY + Zc_y \\ Z \end{pmatrix} = \underbrace{(\underbrace{K}_{3\times 3} \mid \underbrace{0_3}_{1\times 3})}_{P} \underbrace{\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}}_{\in \mathbb{P}^3}$$

Projection and Back-Projection Matrices

$M = (X, Y, Z)^t \in \mathbb{R}^3$
$m = (x, y)^t \in \mathbb{R}^2$, and $\tilde{m} = (x, y, 1)^t \in \mathbb{P}^2$

Camera (Projection) Matrix

$$m = \pi(M) = \left(f\frac{X}{Z} + c_x, f\frac{X}{Z} + c_x\right)$$

Equivalent to:

$$\tilde{m} = KM$$

with: $K = \begin{pmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{pmatrix}$

Back-Projection Matrix

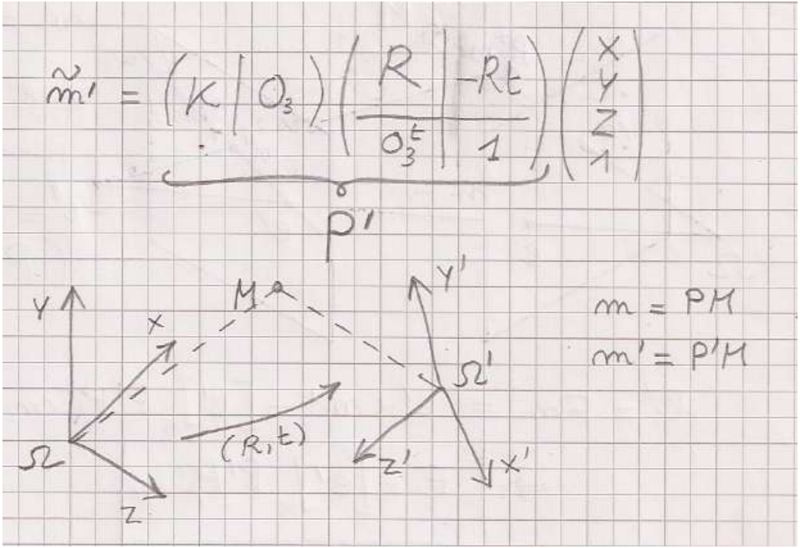
$$M = \pi^{-1}(m, Z) = \left(Z\frac{x - c_x}{f}, Z\frac{y - c_y}{f}, Z\right)$$

Equivalent to:

$$M = \underbrace{Z}_{\text{Depth}}\underbrace{K^{-1}\tilde{m}}_{\text{Direction}}$$

with: $K^{-1} = \begin{pmatrix} \frac{1}{f} & 0 & -\frac{c_x}{f} \\ 0 & \frac{1}{f} & -\frac{c_y}{f} \\ 0 & 0 & 1 \end{pmatrix}$

Displacement Matrix: Extrinsics



$$\tilde{m}' = \underbrace{\left(K \mid 0_3\right)\left(\begin{array}{c|c} R & -Rt \\ \hline 0_3^t & 1 \end{array}\right)}_{P'} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

$m = PM$

$m' = P'M$

Presentation Outline

1. Introduction to analytical methods

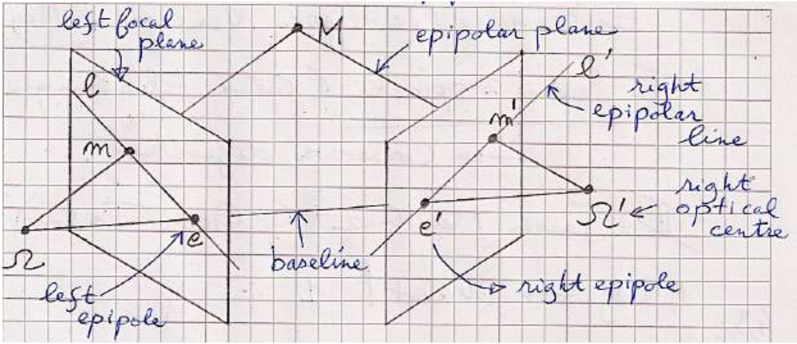
2. Projective Geometry and Camera Matrices

3. Epipolar Geometry and the Fundamental Matrix

4. Depth Estimation and Epipolar Flow

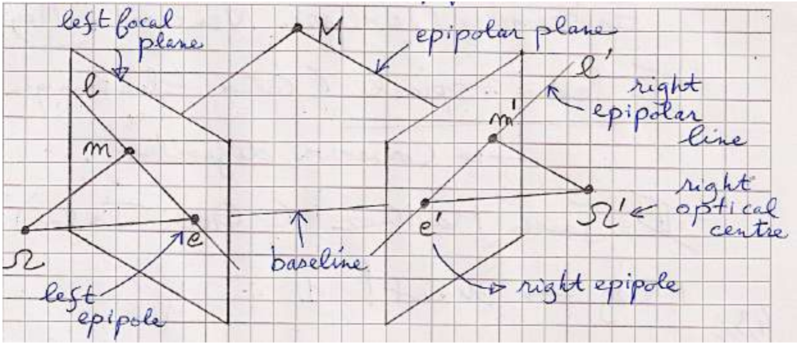
5. Learning based depth prediction

Epipolar Geometry



- Ω, m, M, m' and Ω' are coplanar.
- The epipolar plane cuts each focal plane through the epipolar line.
- Each point M has its own epipolar plane.
- All epipolar planes (epipolar pencil) intersect at the baseline $(\Omega\Omega')$

Epipolar Geometry



- The right (resp. left) epipole is the projection of the left (resp. right) optical centre on the right (resp. left) focal plane.
- All epipolar lines intersect at the epipole.

Example 1: Converging Cameras

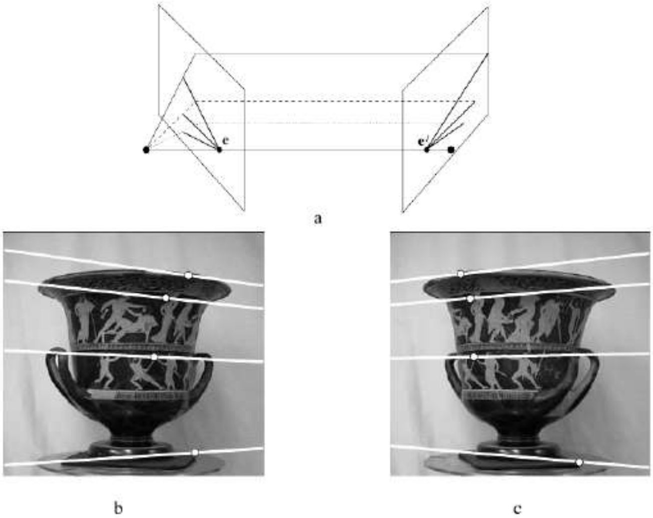


Figure from **[Hartley and Zissermann 2003]**

Example 2: In-Focal-Plane Moving Camera

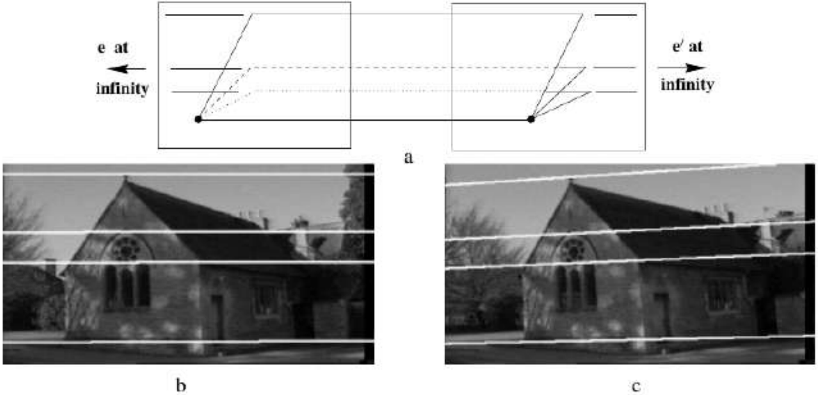


Figure from [Hartley and Zissermann 2003]

Example 3: Radially Moving Camera

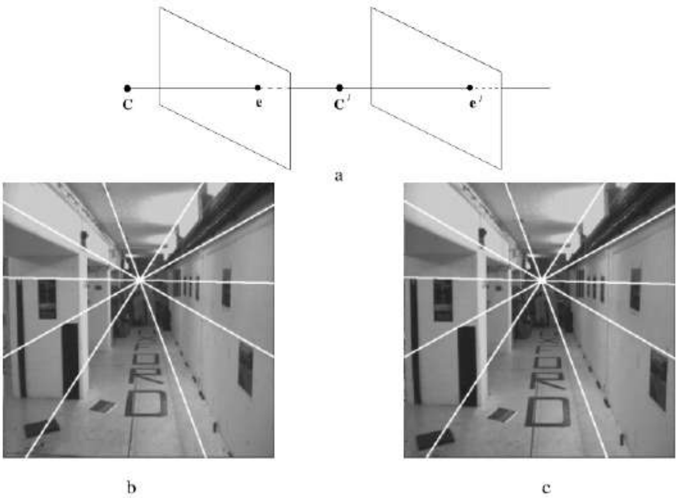
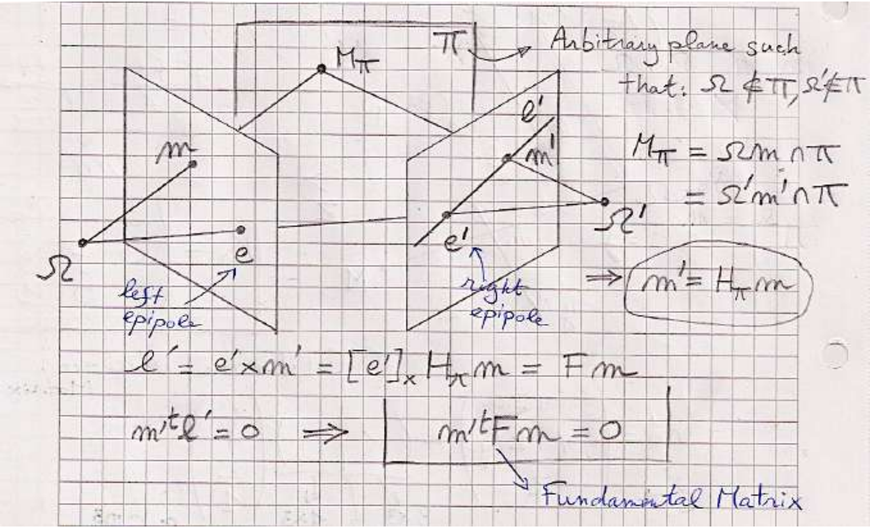


Figure from [Hartley and Zissermann 2003]

Fundamental matrix derived from a plane



π → Arbitrary plane such that: $\Omega \notin \pi, \Omega' \notin \pi$

$$M_\pi = \Omega m \cap \pi$$
$$= \Omega' m' \cap \pi$$

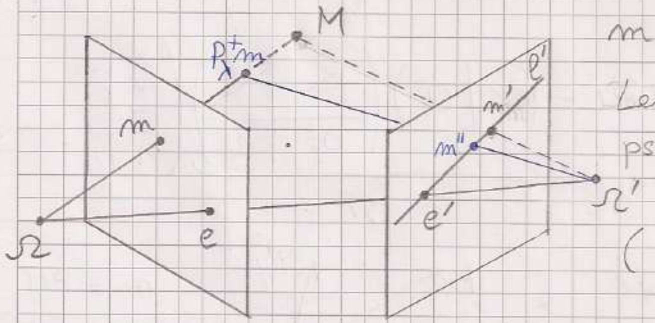
$$\Rightarrow \boxed{m' = H_\pi m}$$

$$\ell' = e' \times m' = [e']_\times H_\pi m = F m$$

$$m'^t \ell' = 0 \Rightarrow \boxed{m'^t F m = 0}$$

Fundamental Matrix

Fundamental matrix derived from the camera matrices



$m = PM$

Let P_λ^+ the pseudo-inverse of P

$(P_\lambda^+ P = I)$

$$\ell' = Fm = e' \times m' = [e']_\times P' P_\lambda^+ m$$

$$\Rightarrow \boxed{F = [e']_\times P' P_\lambda^+}$$

Fundamental matrix from the camera matrices - Essential matrix

Starting from the equation $F = [e']_\times P' P^+_\lambda$, if we consider one single moving camera with projection matrix K, and right pose given by displacement matrix R, we use $e' = KR\Omega$, $P' = KR$, and $P^+_\lambda = K^{-1}m$, and then:

$$\begin{aligned} l' &= [KR\Omega]_\times KRK^{-1}m \\ &= (KR)^{-t}[\Omega]_\times K^{-1}m \end{aligned}$$

And so:

$$\boxed{F = (KR)^{-t}[\Omega]_\times K^{-1}}$$

In the calibrated case (i.e. when K is known beforehand), we can use the *essential* matrix, which only depends on the displacement of the camera, and is defined as:

$$\boxed{E = K^t F K = R^{-t}[\Omega]_\times}$$

Fundamental Matrix Summary

For 2 images captured by cameras with distinct optical centres, the fundamental matrix is the unique 3×3 rank 2 matrix F that satisfies $m'^t F m = 0$, for all corresponding pairs of points (m, m').

- **Epipolar lines**: $l' = Fm$ and $l = m'^t F$ are the right and left epipolar lines respectively.
- **Epipoles**: Since $e' \in l'$, we have $\forall m$, $e'^t F m = 0$. Then $e'^t F = 0$. Similarly, $Fe = 0$.
- **Rank**: F is an homogeneous (8 DoF) 3×3 matrix, and has rank 2 ($\det F = 0$), so it actually has 7 DoF.

Estimation of Fundamental Matrix F

Each correspondence $m \leftrightarrow m'$ provides one scalar equation:

$$m'^t F m = 0$$

The developed equation writes:

$$xx'f_{11} + x'yf_{12} + x'f_{13} + y'xf_{21} + yy'f_{22} + y'f_{23} + xf_{31} + yf_{32} + f_{33} = 0$$

Or, by separating data and unknowns:

$$\underbrace{\begin{pmatrix} x'x & x'y & x' & y'x & y'y & y' & x & y & 1 \end{pmatrix}^t}_{\mathbf{d}} \underbrace{\begin{pmatrix} f_{11} & f_{12} & f_{13} & f_{21} & f_{22} & f_{23} & f_{31} & f_{32} & f_{33} \end{pmatrix}}_{\mathbf{f}} = 0$$

And, by using N correspondence pairs $\{m_i \leftrightarrow m'_i\}_{1 \leq i \leq N}$:

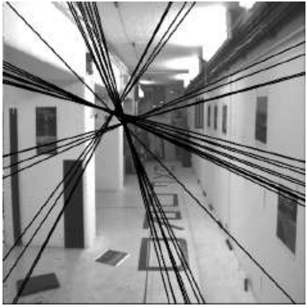
$$\mathbf{Df} = \begin{pmatrix} \mathbf{d}_1 \\ \vdots \\ \mathbf{d}_N \end{pmatrix} \mathbf{f} = \mathbf{O}_N$$

Estimation of Fundamental Matrix F

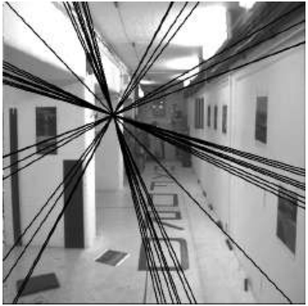
- The system $\mathbf{Df} = \mathbf{O}_N$ is solved using SVD.
- Since the columns of $\mathbf{D}$ range over several order of magnitudes, it is better to normalise the data, for numerical stability purposes.
- Once F is estimated, it is usually imposed that: $e'^{t}F = 0$, $Fe = 0$, and $\text{rank}(F) = 2$.
  - This is done by finding F' such that $F' = \arg \min_{G;\text{rank}(G)=2} ||F - G||_{\mathrm{F}}$
- RANSAC is used to minimise the number of outliers in the N correspondences $\{m_i \leftrightarrow m'_i\}_{1 \leq i \leq N}$.

Estimation of Fundamental Matrix F - Rank Constraint

- Once F is estimated, it is usually imposed that: $e'^t F = 0$, $Fe = 0$, and $\text{rank}(F) = 2$.
  - This is done by finding F' such that $F' = \arg \min_{G;\text{rank}(G)=2} ||F - G||_{\mathrm{F}}$



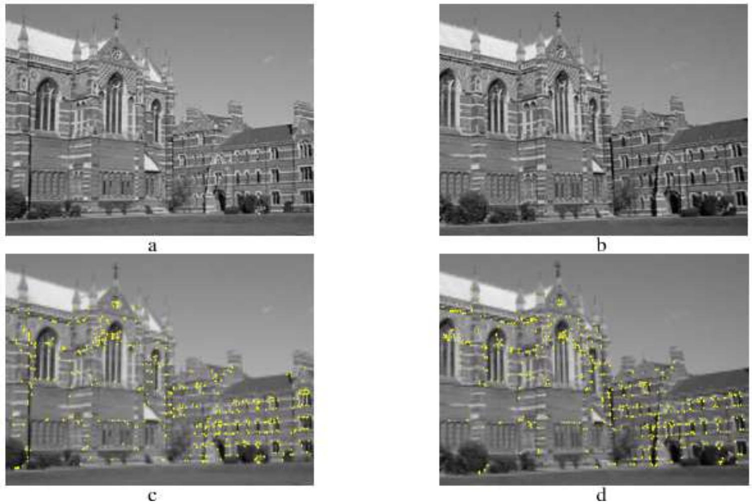
Rank 3



Rank 2

Figure from [Hartley and Zissermann 2003]

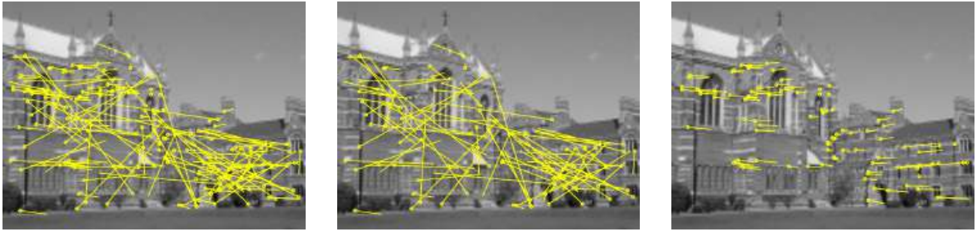
Estimation of Fundamental Matrix F - RANSAC



From **[Hartley and Zissermann 2003]** - There are ≈ 500 keypoints on each image.

Estimation of Fundamental Matrix F - RANSAC

- RANSAC is used to minimise the number of outliers in the N correspondences $\{m_i \leftrightarrow m'_i\}_{1 \leq i \leq N}$.



188 Matches (<< 500!)

89 Outliers

99 Inliers

Figure from **[Hartley and Zissermann 2003]**

Presentation Outline

1. Introduction to analytical methods

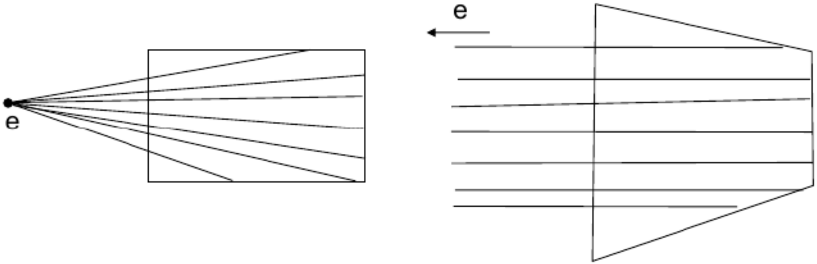
2. Projective Geometry and Camera Matrices

3. Epipolar Geometry and the Fundamental Matrix

4. **Depth Estimation and Epipolar Flow**

5. Learning based depth prediction

Stereo Rectification

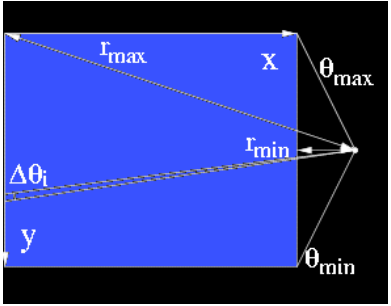


From [Pollefeys 2002]

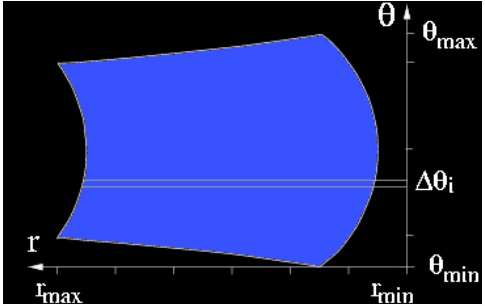
- Objective: come back to the ideal stereo case.
- Find the homography H that makes epipolar lines parallel.
- H transfers the epipole to infinity: $He = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^t$.
- Numerical problems when e is close to (or inside!) the image.

Polar Rectification (Pollefeys et al 1999)

Solution: Polar re-parameterization of the two images around their epipoles.



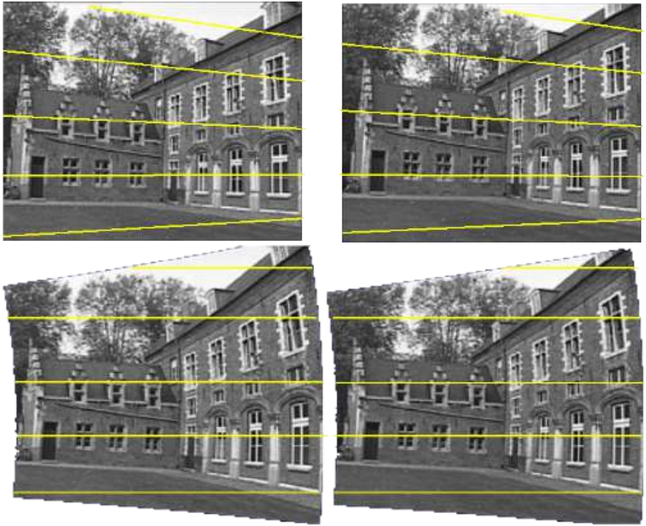
Original



Rectified

From [Pollefeys 2002]

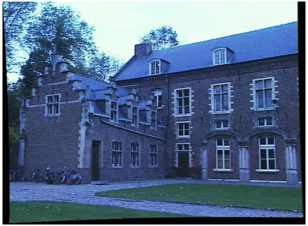
Polar Rectification (Pollefeys et al 1999)



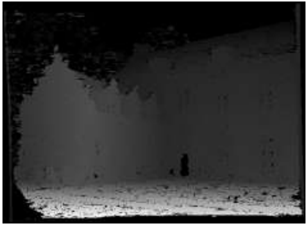
From **[Pollefeys 2002]**

Disparity and Depth estimation

- Rectify the two images.
- Compute the dense correspondence between the two images along each epipolar line.
- The horizontal shift between the two images is the disparity.
- The depth is inversely proportional to the disparity.

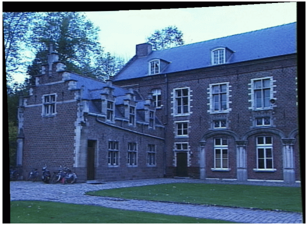


Left



Disparity

From [Pollefeys 2004]



Right

Epipolar Flow Estimation

Input: Image pair

Keypoint-based sparse flow estimation

- Detector: Blockwise FAST
- Descriptor: 11x11 pixel patch
- Error filtering based on local coherence

Fundamental matrix estimation

- With the 8-points algorithm + RANSAC

Dense optical flow estimation

- Search domain reduced to the epipolar lines
- Propagation of the seed flow vectors coming from the sparse flow estimation

Error filtering

- Make erroneous pixels diverge from epipolar lines
- Filter them according to the epipolar line distance and to local coherence

Small holes filling

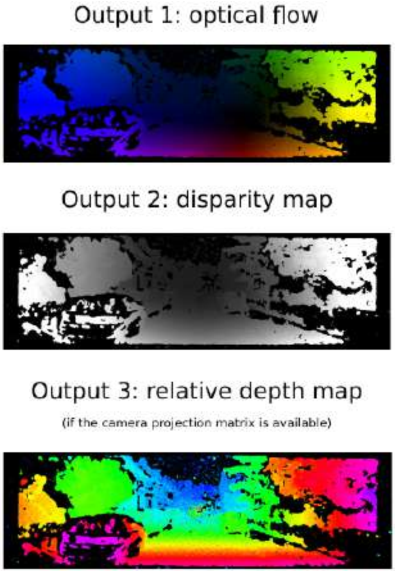
- Simple linear interpolation of the disparity to fill small holes caused by error filtering

[Garrigues 17]

Epipolar Flow Estimation

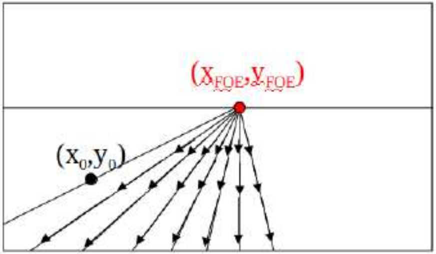
[Garrigues 17]:

- Real-Time semi-dense optical flow and relative depth estimation.
- Was ranked #1 on Kitti 2012 Optical Flow dataset (on sparse optical flow category).



Conclusion: Limitations of analytical methods

- Estimation strongly relies on local structure (texture), then depth estimation on textureless areas depends on complicated regularization methods.
- Depth calculation depends on the apparent displacement (speed) of a point with respect to the epipole (i.e. the Focus of Expansion FoE, that indicates the translation direction of the camera). Such calculation turns undetermined when the point gets close to the FoE.



Presentation Outline

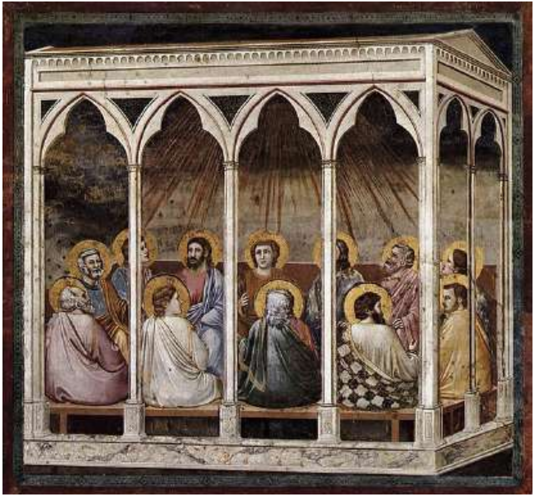
1. Introduction to analytical methods
2. Projective Geometry and Camera Matrices
3. Epipolar Geometry and the Fundamental Matrix
4. Depth Estimation and Epipolar Flow
5. Learning based depth prediction

DNN for 3d reconstruction

- Like Optical Flow, Depth can benefit from Deep Networks dense prediction capabilities.
- Training can be easily done on *synthetic* or *real RGB-d* data, and loss function is also relatively straightforward.
- One determining benefit of DNN is their ability to exploit potentially *all the depth indices:* parallax, perspective, size and texture gradients, shading,...

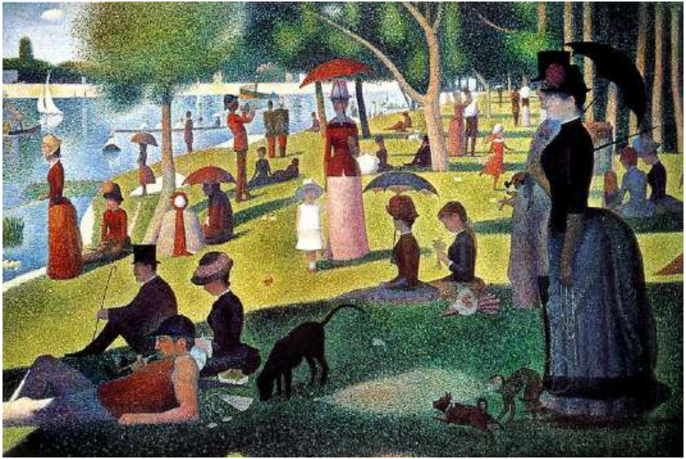
Monocular Depth Cues? Occlusions!

Giotto - Pentecoste
(*circa* 1305)



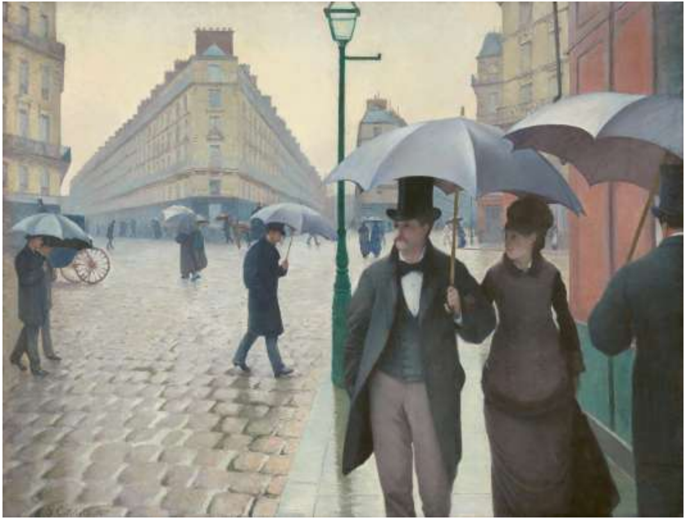
Monocular Depth Cues? Object sizes!

Georges Seurat - Un après-midi à l'île de la Grande Jatte (1884-1886)



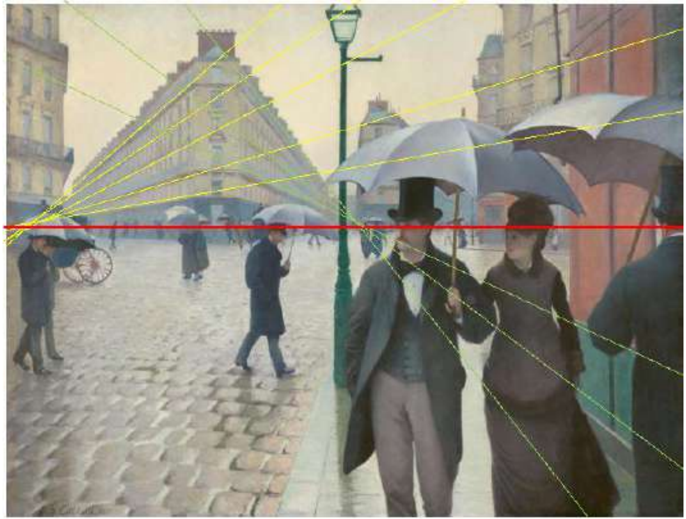
Monocular Depth Cues? Object sizes, Perspective, and Texture Gradients!

Gustave Caillebotte - Rue de Paris, temps de pluie (1877)

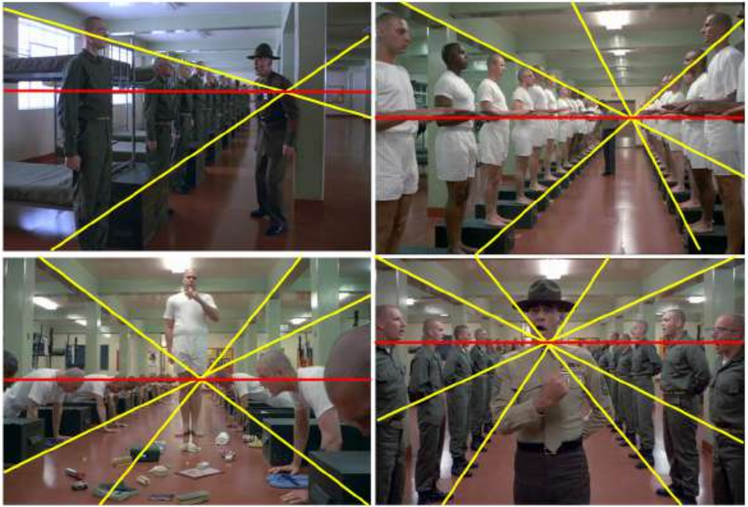


Monocular Depth Cues? Perspective, Horizon and Vanishing Points!

Gustave Caillebotte - Rue de Paris, temps de pluie (1877)

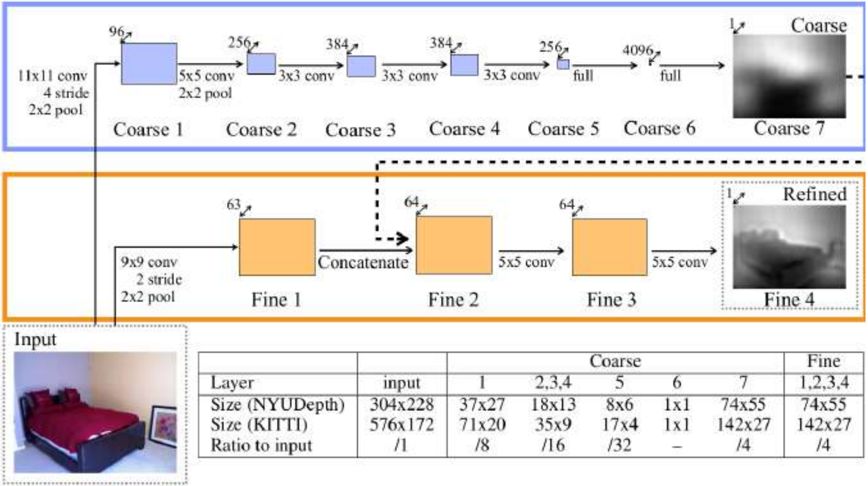


Monocular Depth Cues? Horizon and Camera Pose!



Stanley Kubrick – Full Metal Jacket (1987)

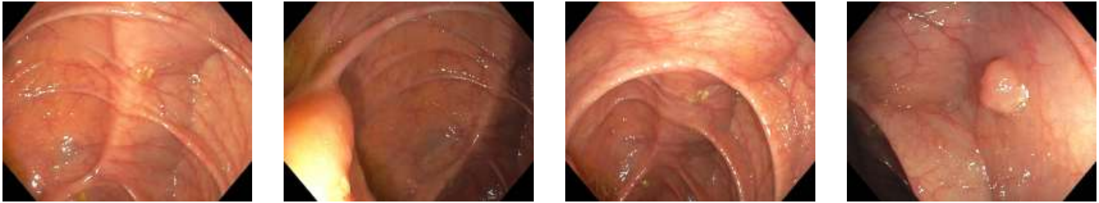
Depth inference from single view!



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

CNN based Depth estimation from single view **[Eigen 14]** works well on a particular context!

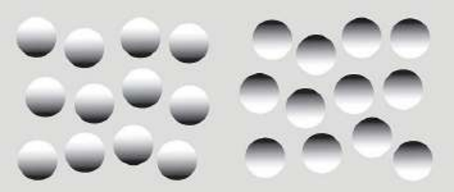
One very particular context...

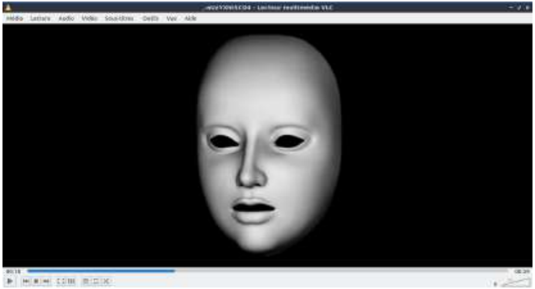


Colonoscopy images **[Ruano 19]**

Monocular Depth Cues? Shading!

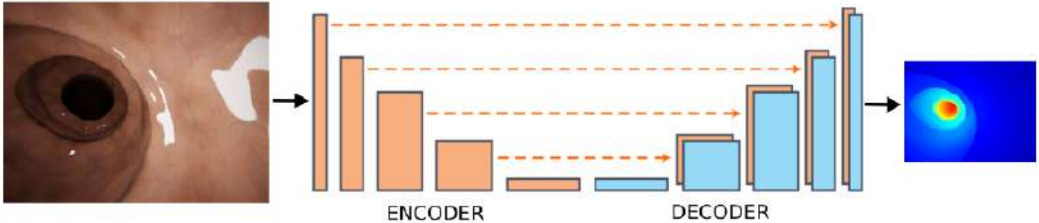
Self shadowing is a strong but ambiguous depth cue (light source position *vs* concavity). Without shape prior, the concavity is determined by a prior of top lighting (right image).





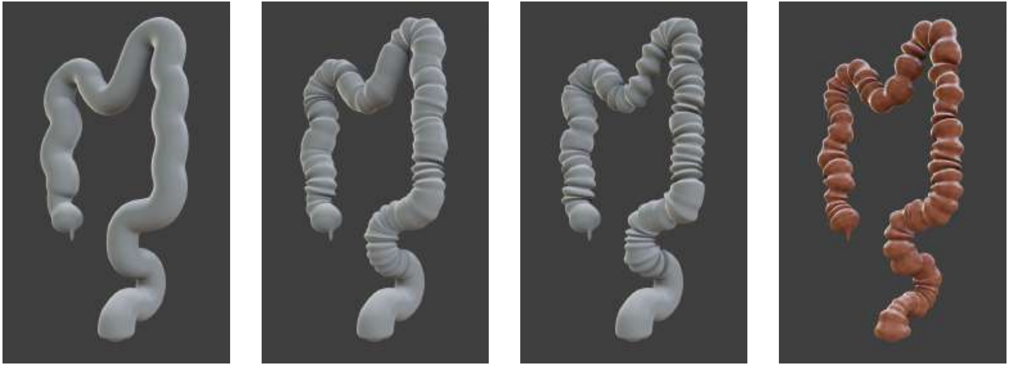
When the shape prior is strong (face then convex), the concavity prior dominates the lighting prior (top-down effect, animation on the left).

Learning Shape from Shading for Automated Colonoscopy



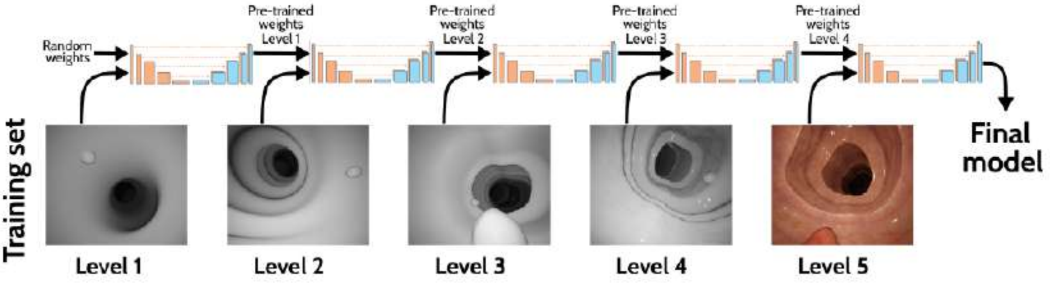
Images from synthetic videos are used to train a CNN using a loss function based on the ground truth depthmap **[Ruano 19]**

Curriculum Learning Shape from Shading for Automated Colonoscopy



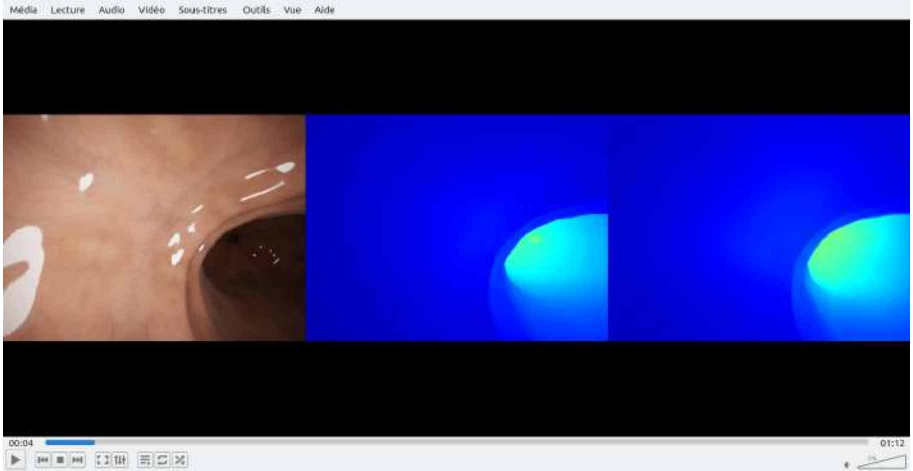
Synthetic exploration videos are created from a hierarchy of synthetic colons of increasing complexity **[Ruano 19]**

Curriculum Learning Shape from Shading for Automated Colonoscopy


Random weights
Pre-trained weights Level 1
Pre-trained weights Level 2
Pre-trained weights Level 3
Pre-trained weights Level 4
Final model
Training set
Level 1
Level 2
Level 3
Level 4
Level 5

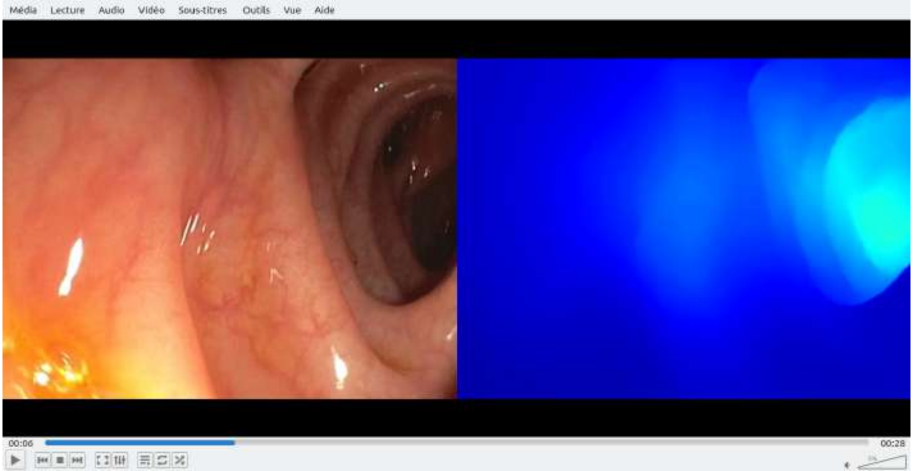
The training is performed with progressive complexity **[Ruano 19]**

SfSNet on Synthetic Videos



ShapeFromShadingNet on Synthetic Test Videos [Ruano 19]

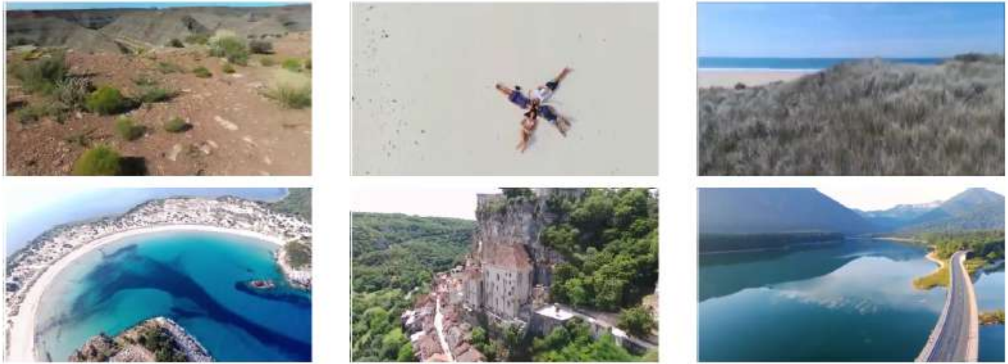
SfSNet on Real Videos



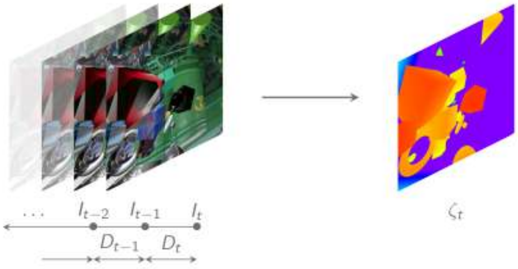
ShapeFromShadingNet on Real Videos **[Ruano 19]**. Single images seem to be sufficient in such particular context!

What about UAV's context?

These scenes are all taken from the same drone !



Non photorealistic synthesis for learning SfM



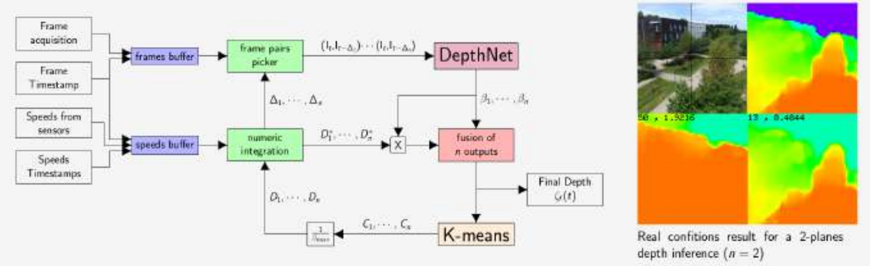
Supervised learning of depth from synthetic sequences

[Pinard 17a]

- Network is based on FlowNet_S
- Unrealistic scenes ↔ Abstraction of the context
- Focus on geometry / motion, not on appearance /context
- Trained on rotationless movement, at a constant speed

Baseline adaptation using multiple image pairs

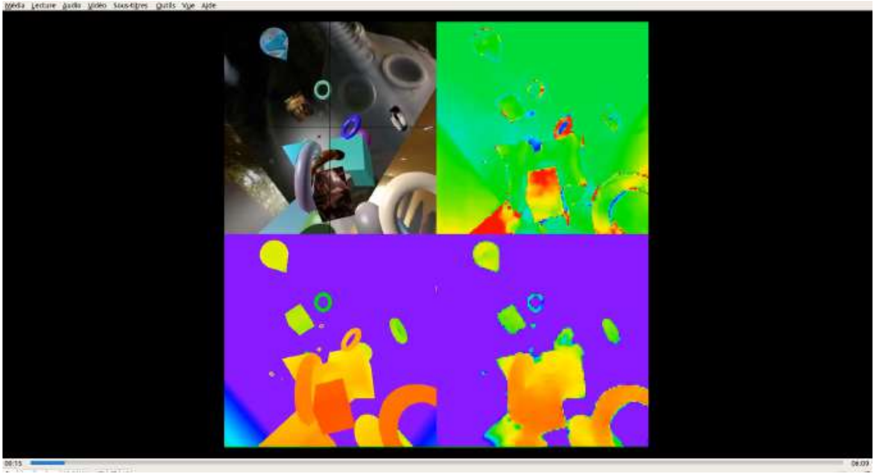
- At the inference time, the depth which is relative to the trained speed, is scaled with respect to the actual velocity.
- Adaptable precision is achieved by dynamically adapting the image pairs (baselines) to the depth distribution.



Real confitions result for a 2-planes depth inference ($n = 2$)

Adaptation of the baselines to the depth distribution [Pinard 17b]

Supervised DepthNet



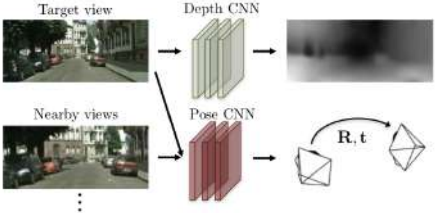
Supervised DepthNet results [Pinard 17a]: See

```
https://perso.ensta-paris.fr/~manzaner/Download/ECMR2017/DepthNetResults.mp4
```

Unsupervised depth estimation CNN

- Re-training on real/operative context is still essential.
- But data are rarely annotated.
- Self-supervised learning is then necessary.
- *Photometric loss function* can be used, that compares a pair of registered images, knowing the depth and the camera pose.
- Camera pose then needs to be known, or predicted!

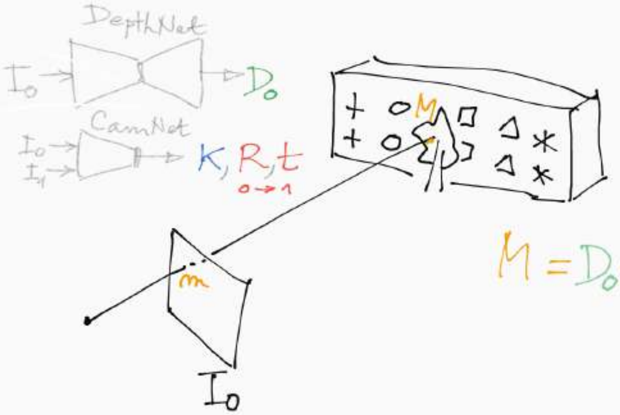
(a) Training: unlabeled video clips.



(b) Testing: single-view depth and multi-view pose estimation.

[Zhou 17]

Photometric Loss (1): Back-projection from first image

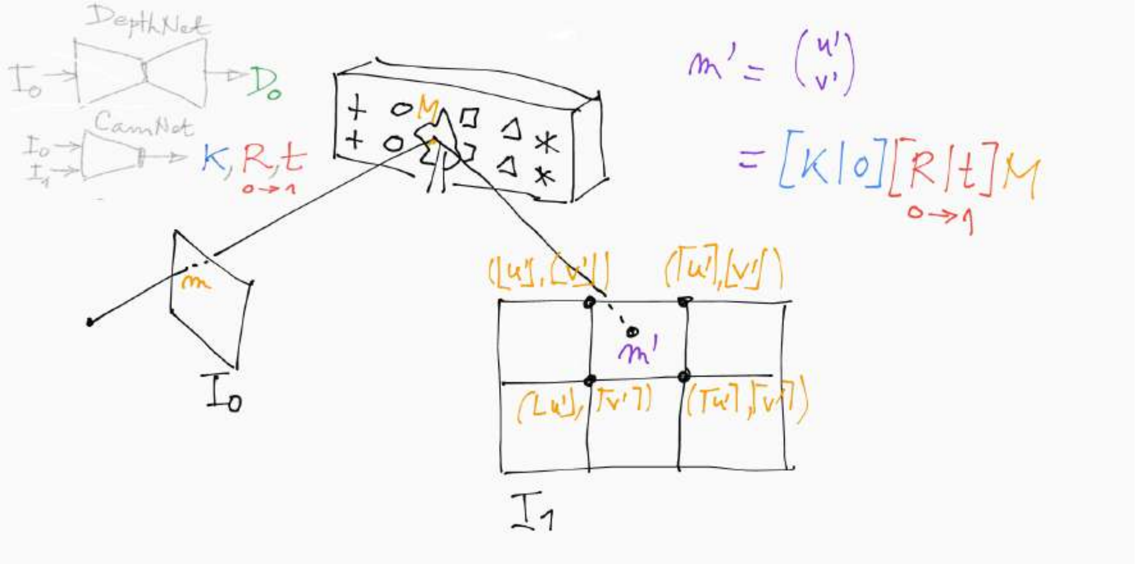


$$m = \begin{pmatrix} u \\ v \end{pmatrix}$$

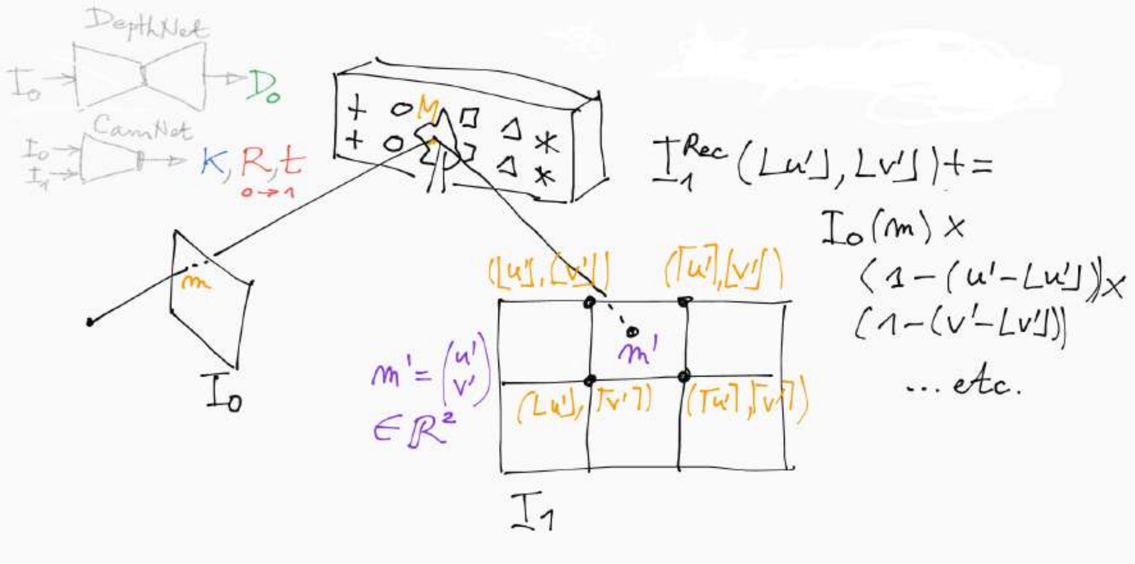
$$M = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

$$M = D_0(m) \times K^{-1} m$$

Photometric Loss (2): Re-projection onto second image


DepthNet
I_0
D_0
CamNet
I_0
I_1
K, R, t
$0 \to 1$
M
m
I_0
$m' = \begin{pmatrix} u' \\ v' \end{pmatrix}$
$= [K|0][R|t]_{0 \to 1} M$
$(\lfloor u' \rfloor, \lfloor v' \rfloor)$
$(\lceil u' \rceil, \lfloor v' \rfloor)$
m'
$(\lfloor u' \rfloor, \lceil v' \rceil)$
$(\lceil u' \rceil, \lceil v' \rceil)$
I_1

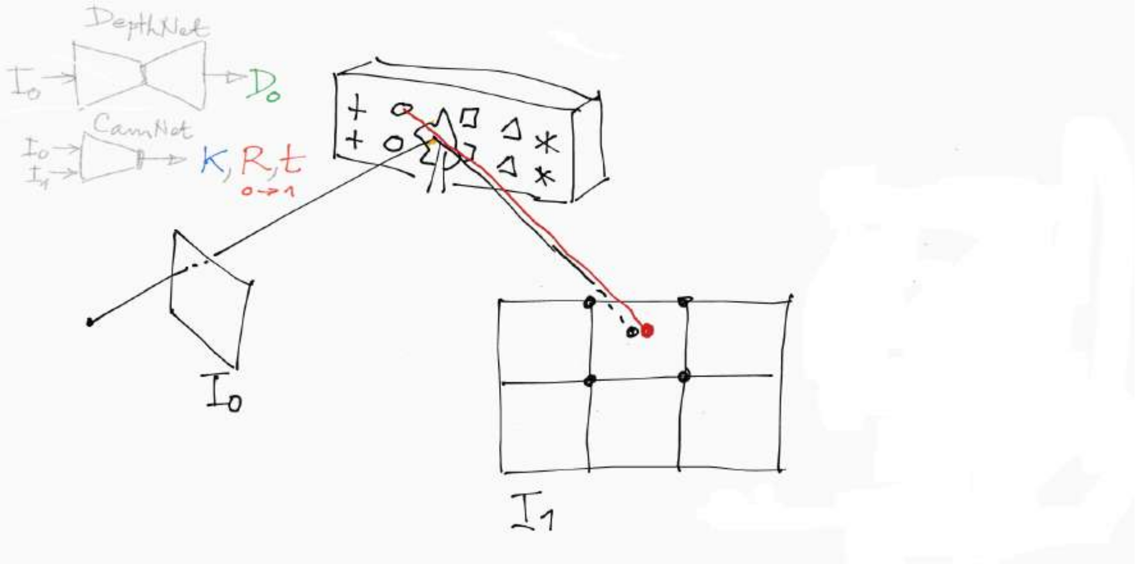
Photometric Loss (3): Interpolation within second image


DepthNet
I_0
D_0
CamNet
I_0
I_1
$K, R, t_{0 \to 1}$
M
m
I_0
$(\lfloor u' \rfloor, \lfloor v' \rfloor)$
$(\lceil u' \rceil, \lfloor v' \rfloor)$
m'
$(\lfloor u' \rfloor, \lceil v' \rceil)$
$(\lceil u' \rceil, \lceil v' \rceil)$
I_1
$m' = \begin{pmatrix} u' \\ v' \end{pmatrix} \in \mathbb{R}^2$

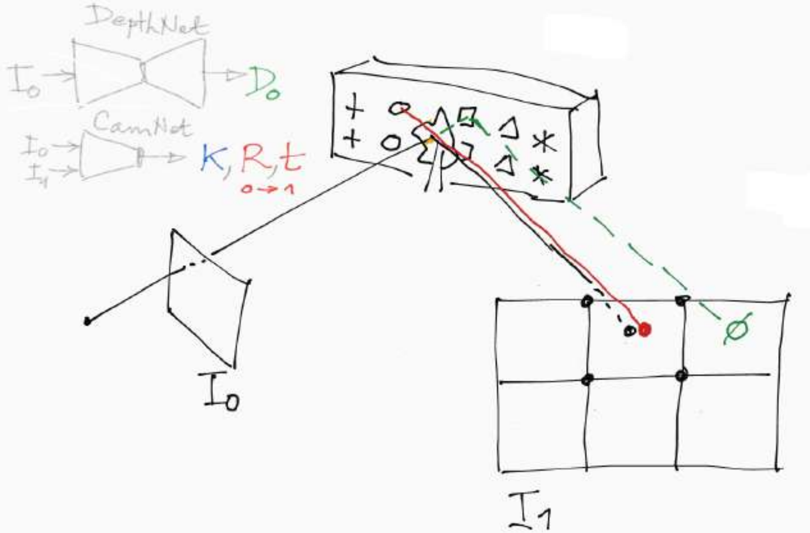
$$I_1^{Rec}(\lfloor u' \rfloor, \lfloor v' \rfloor) += I_0(m) \times (1 - (u' - \lfloor u' \rfloor)) \times (1 - (v' - \lfloor v' \rfloor))$$

... etc.

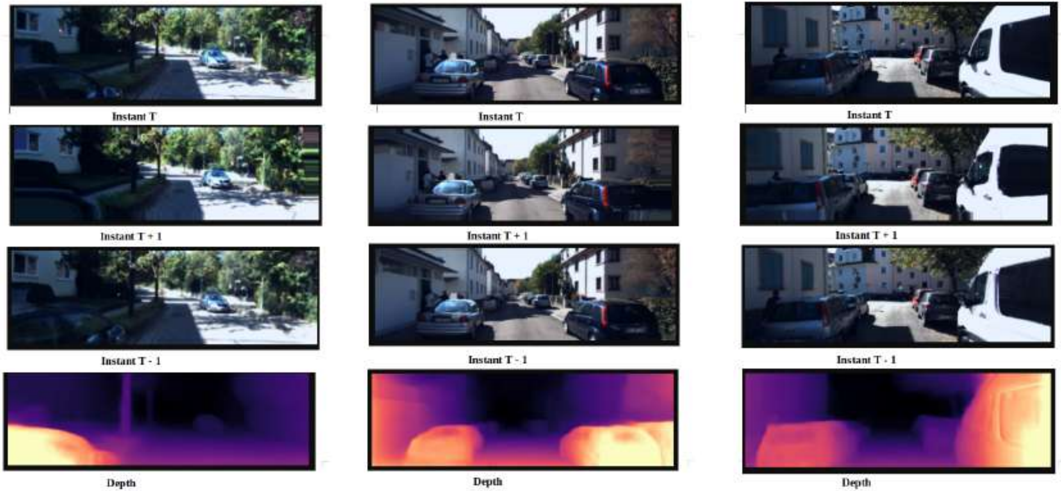
Photometric Loss: Occlusion issue



Photometric Loss: Un-occlusion issue

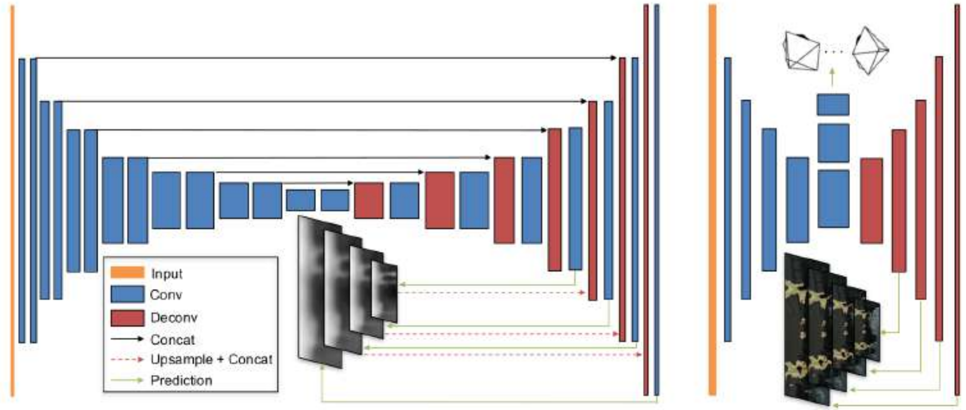


Examples of reprojected images



[PhD Marwane Hariat]

Unsupervised depth estimation CNN

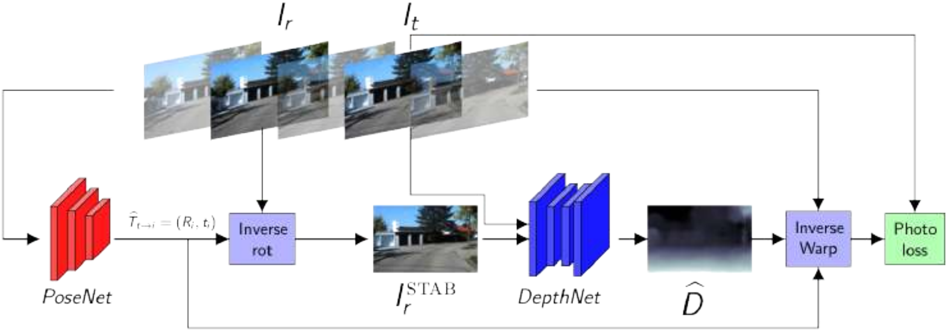


(a) Single-view depth network

(b) Pose/explainability network

[Zhou 17]

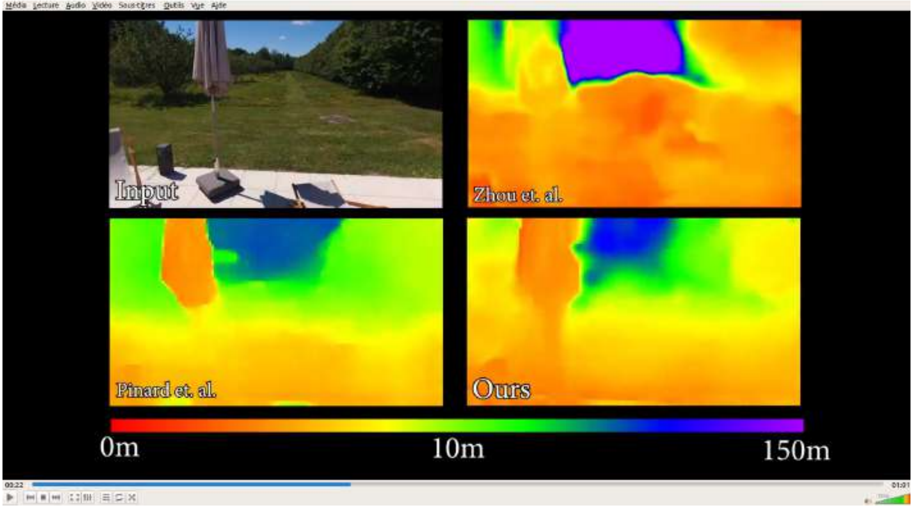
Unsupervised DepthNet



$$\forall i,\ t_i^{\mathrm{NORM}} = t_i \frac{T_0}{\epsilon + \|t_r\|}$$

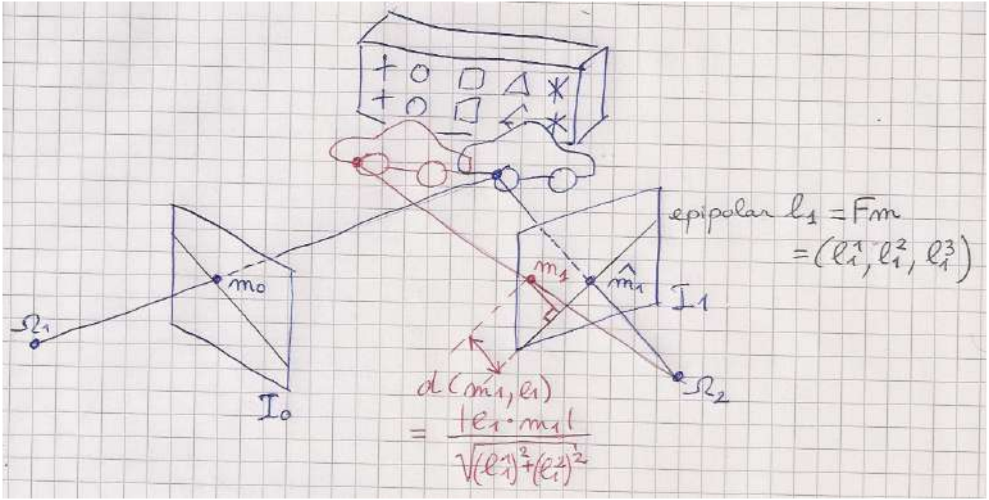
Unsupervised re-learning of Structure from Motion with adaptive baseline **[Pinard 18]**

Unsupervised DepthNet

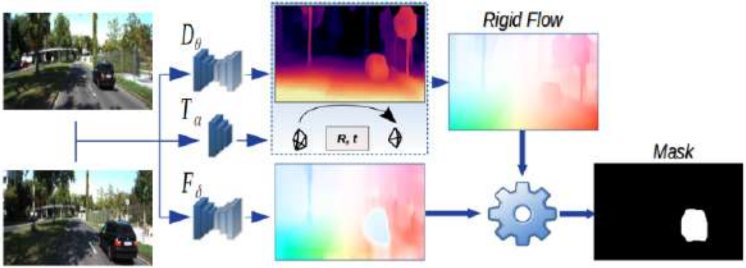


Unsupervised DepthNet real fly demo **[Pinard 18]**: See `https://www.youtube.com/watch?v=ZDgWAWTwU7U`

Photometric Loss: Moving objects issue



CoopNet: Joint training of Optical Flow, Odometry and Depth

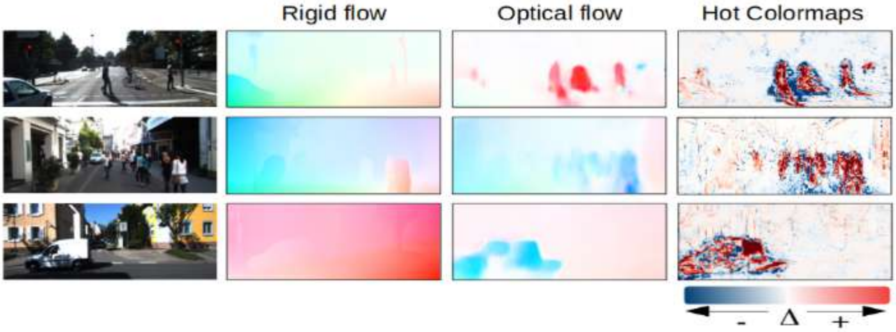


CoopNet [Hariat 23]

By estimating (or predicting) the optical flow, moving objects can also be predicted by comparing the optical flow with the *rigid flow*, which is the apparent velocity field under rigid assumption scene (i.e. only due to camera motion), defined as:

$$[\mathbf{K}|\mathbf{O}_4]\,[\mathbf{R}|\mathbf{t}]\,D_0(\mathbf{m}) \times \mathbf{K}^{-1}\mathbf{m} - \mathbf{m}$$

CoopNet: Joint training of Optical Flow, Odometry and Depth



CoopNet [Hariat 23]

The CoopNet network is trained based on the difference between the photometric losses from the optical flow and from the depth networks:

$$\Delta(\mathbf{m}) = \mathcal{L}_{\text{photo}}^{\text{depth,odometry}} - \mathcal{L}_{\text{photo}}^{\text{flow}}$$

Conclusion on Learning-based methods

- Learning optical flow and depth from videos has many advantages:
  - Globally addressing the context
  - Multi-cues depth inference
  - Natural regularization of ill-posed problem
- The main issues to adress are the hard dependence to the learned context, and the difficulties inherent to online learning. The current work perspectives are:
  - Domain adaptation: ground robotics, medical robotics,...
  - Incremental and online learning...
  - Explainability and Reliability...

Contributors for this lecture

- **Matthieu Garrigues**: PhD student 2012-2016
- **Clément Pinard**: PhD student (CIFRE ANRT Parrot) 2016-2019
- **Josué Ruano Balseca**: PhD student (w. UNAL Bogotá) 2018-
- **Marwane Hariat**: PhD student 2021-

References (1)

[Hartley and Zisserman 2003] R. Hartley and A. Zisserman
Multiple View Geometry in Computer Vision
Cambridge University Press, 2003

[Pollefeys 2002] M. Pollefeys
Visual 3D Modeling from Images (Tutorial and slides)
`https://www.cs.unc.edu/~marc/tutorial/`, 2002

[Pollefeys 99] M. Pollefeys, R. Koch and L. Van Gool
A simple and efficient rectification method for general motion
Proc. International Conference on Computer Vision, pp.496-501, 1999.

References (2)

[Garrigues 17] M. Garrigues and A. Manzanera
Fast Semi Dense Epipolar Flow Estimation
IEEE Winter Conf. on Applications of Computer Vision (WACV). Sta Rosa, CA, pp.1-8, 2017

[Eigen 14] D. Eigen and C. Puhrsch and R. Fergus
Depth map prediction from a single image using a multi-scale deep network
Advances in neural information processing systems (NIPS), pp.2366–2374, 2014

[Zhou 17] T. Zhou and M. Brown and N. Snavely and D.G. Lowe
Unsupervised learning of depth and ego-motion from video
Computer Vision and Pattern Recognition (CVPR), 2017.

[Ruano 19] J. Ruano Balseca and A. Manzanera and E. Romero Castro
Curriculum-based strategy for learning shape-from-shading from colonoscopy synthetic database
Research Report, 2019

References (3)

[Pinard 17a] C. Pinard and L. Chevalley and A. Manzanera and D. Filliat
End-to-end depth from motion with stabilized monocular videos
Int. Conf. on Unmanned Aerial Vehicles in Geomatics (UAV-g) Bonn, pp. 67-74, 2017

[Pinard 17b] C. Pinard and L. Chevalley and A. Manzanera and D. Filliat
Multi range Real-time depth inference from a monocular stabilized footage using a Fully Convolutional Neural Network
European Conference on Mobile Robotics (ECMR), Palaiseau, 2017

[Pinard 18] C. Pinard and L. Chevalley and A. Manzanera and D. Filliat
Learning structure-from-motion from motion
European Conf. on Computer Vision Workshops (ECCV-W), pp.363-376, 2018

[Hariat 23] M. Hariat and A. Manzanera and D.Filliat
Rebalancing gradient to improve self-supervised co-training of depth, odometry and optical flow predictions
IEEE Winter Conf. on Applications of Computer Vision (WACV). Waikoloa, 2023