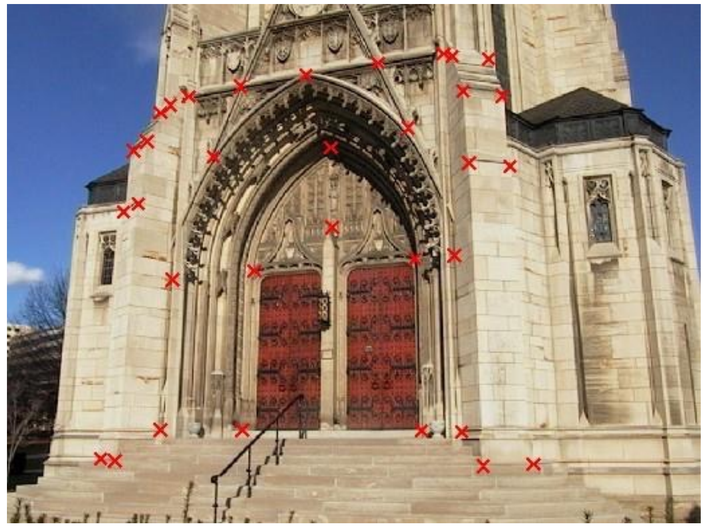
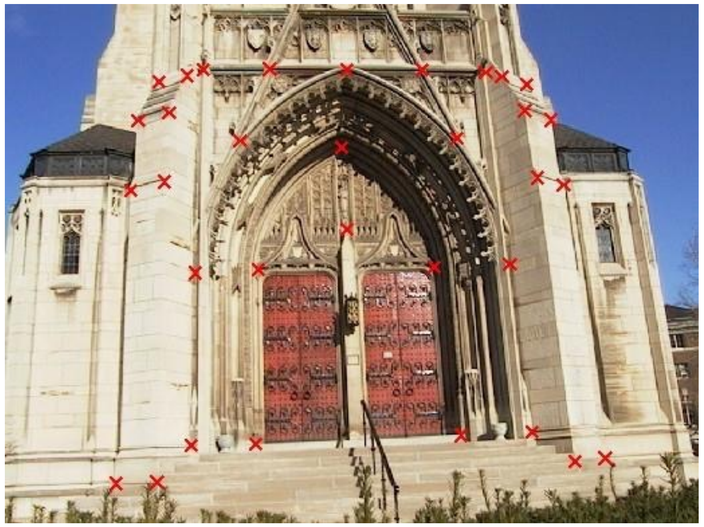
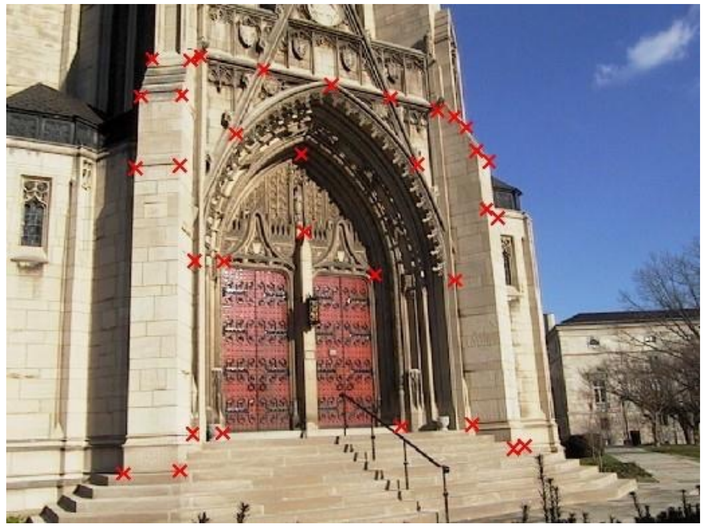
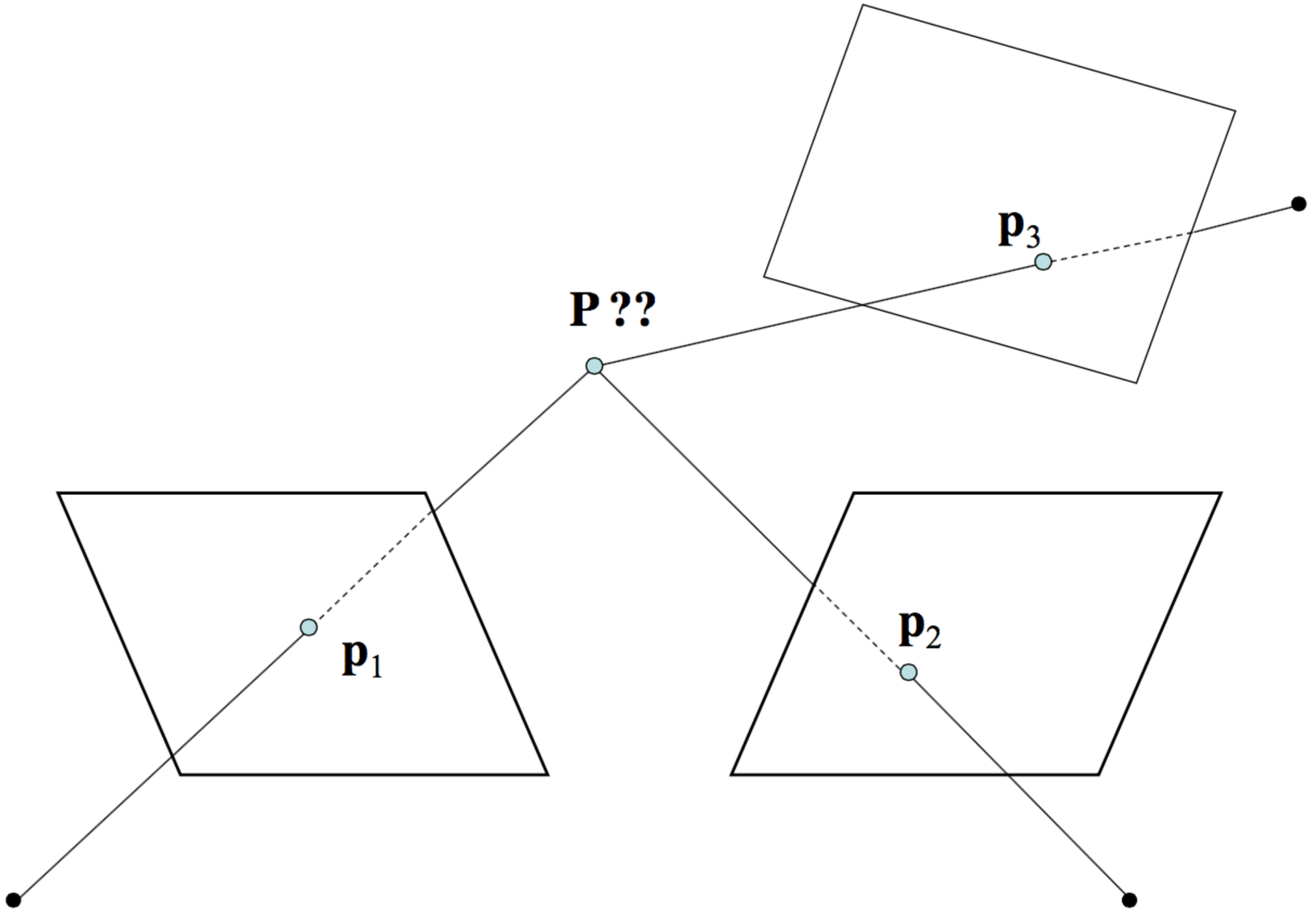
Structure from Motion

Gary Overett (Slides adapted from CMU 16-720 2014)

Forsyth&Ponce: Chap. 12 and 13 Szeliski: Chap. 7

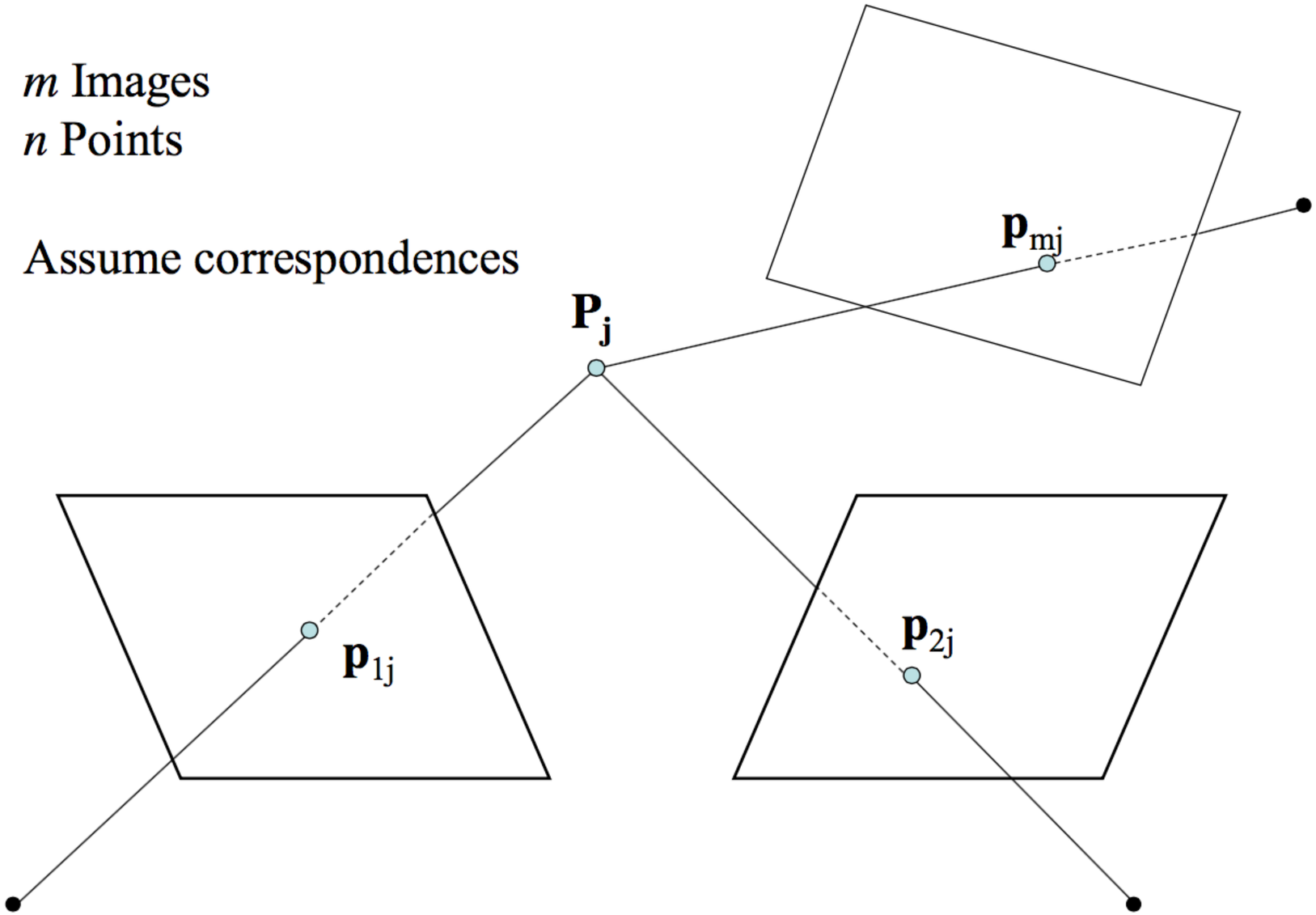
The Reconstruction Problem



The Reconstruction Problem

m Images
n Points

Assume correspondences



The Reconstruction Problem

m Images
n Points

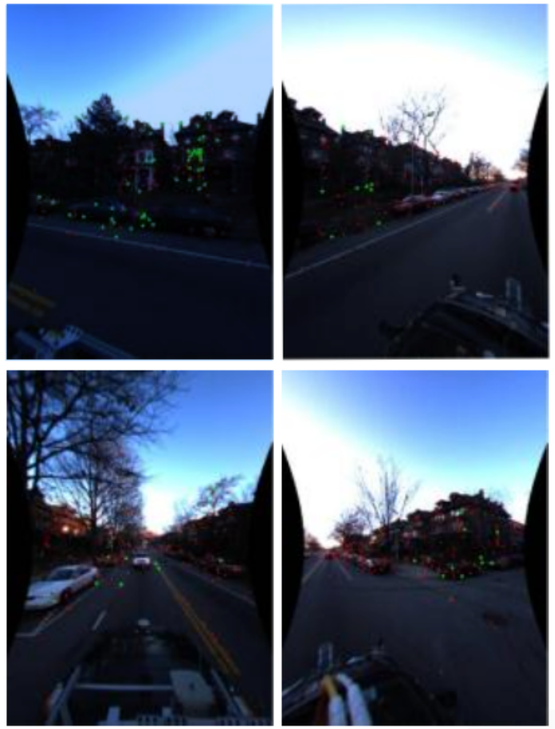
As

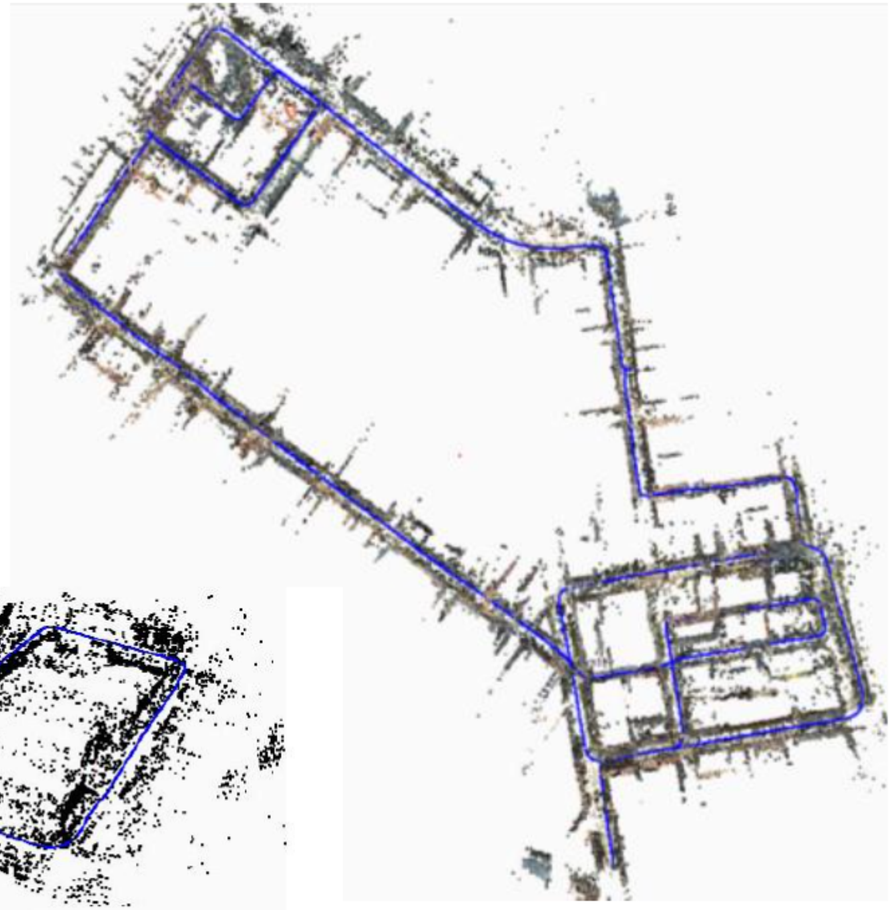
∇

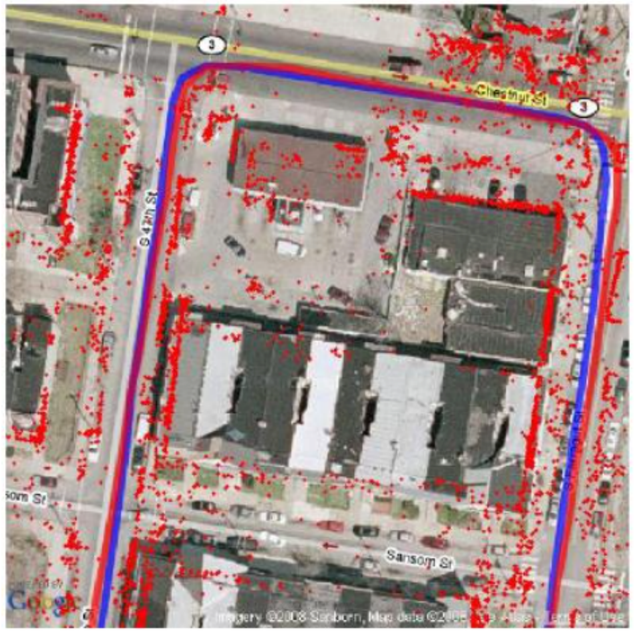
"Reconstruction" =
Scene points $\mathbf{P}_j$, j=1,..n
Camera projection matrices $\mathbf{M}_i$, i=1,..,m
Such that:

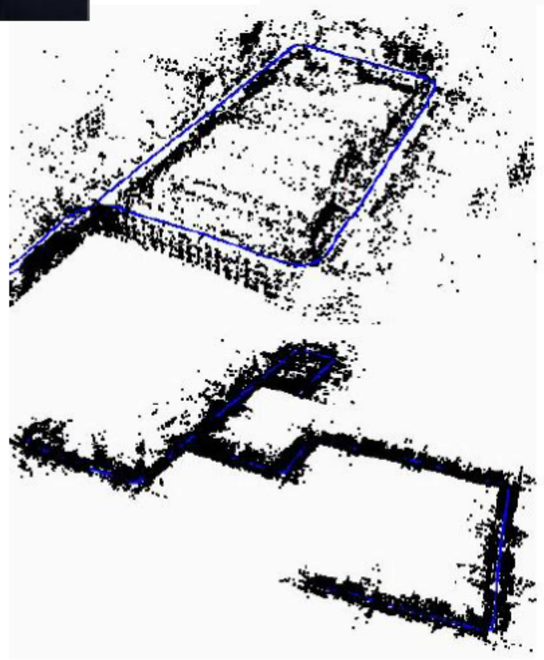
$$\boldsymbol{p}_{ij} \equiv \boldsymbol{M}_i \boldsymbol{P}_j \ \forall \boldsymbol{i}, \boldsymbol{j}$$

Example from UPenn's group

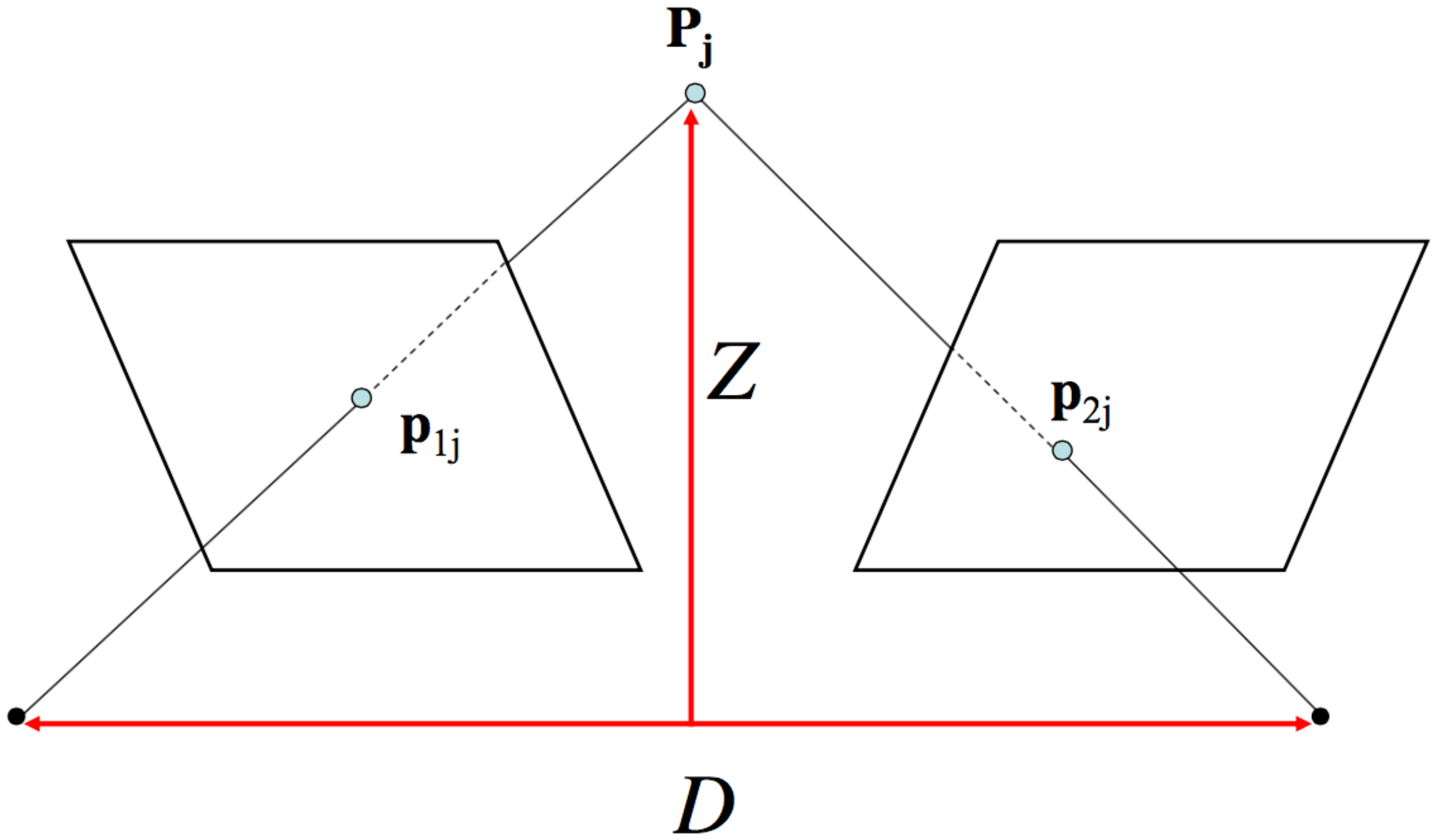






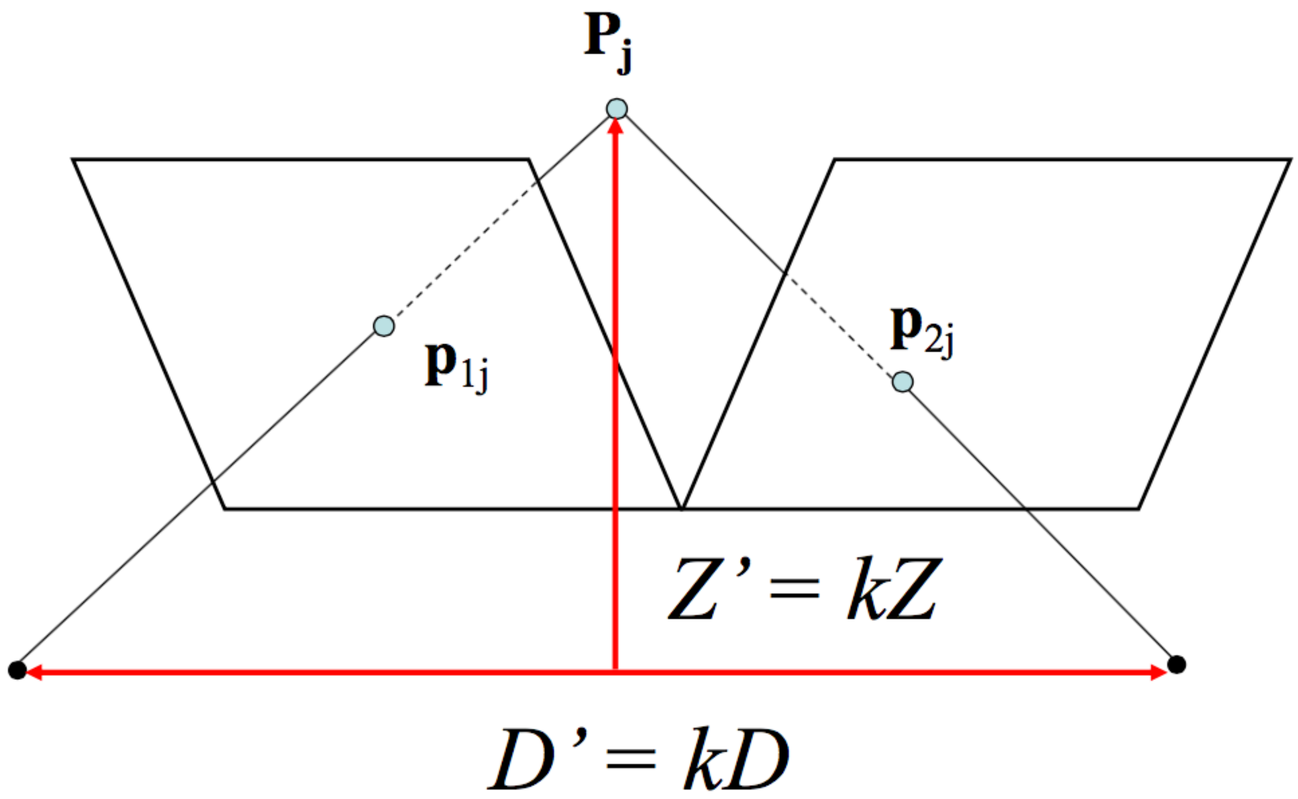


Scale Ambiguity



Scale Ambiguity

If we scale the entire scene by some factor k and, at the same time, we scale the distance between the cameras by the same scale factor, the projections of the scene points in the image remain exactly the same!
→ It is impossible to recover the absolute scale of the scene.



If $\tilde{P}_j (j = 1,..,n)$ is the "true" reconstruction of all the points and all the projection matrices $\tilde{M}_i (i = 1,..,m)$

Then, given a 4x4 matrix Q, the other reconstruction defined as:

$$P_j = Q^{-1}\tilde{P}_j$$
$$M_i = \tilde{M}_i Q$$

is also a solution because:

$$p_{ij} \equiv M_i P_j = M_i Q Q^{-1} P_j = \tilde{M}_i \tilde{P}_j$$

$$\forall i, j$$

Therefore: *The reconstruction is defined only up to a global transformation. We consider 3 types of transformations: metric, affine, projective.*

If $\tilde{P}_j\ (j = 1,..,n)$ is the "true" reconstruction of all the points and all the projection matrices

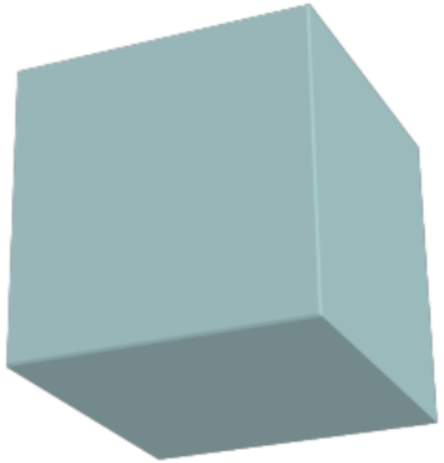
In plain English: If we transform the scene and we transform the cameras in the opposite way, then the images *do not* change!

$$p_{ij} \equiv M_i P_j = M_i Q Q^{-1} P_j = M_i P_j$$

$$\forall i, j$$

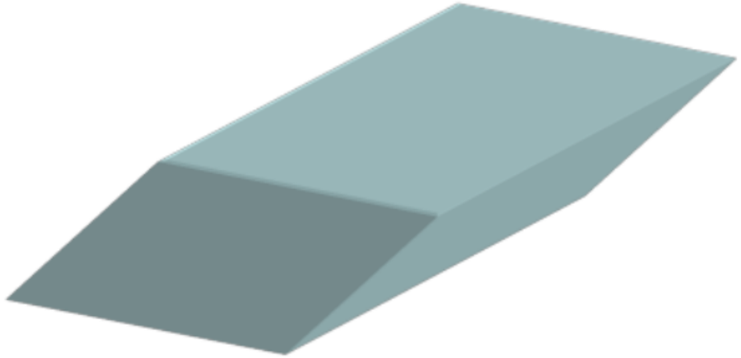
Therefore: *The reconstruction is defined only up to a global transformation. We consider 3 types of transformations: metric, affine, projective.*

Q = similarity
(scale+rotation+translation)



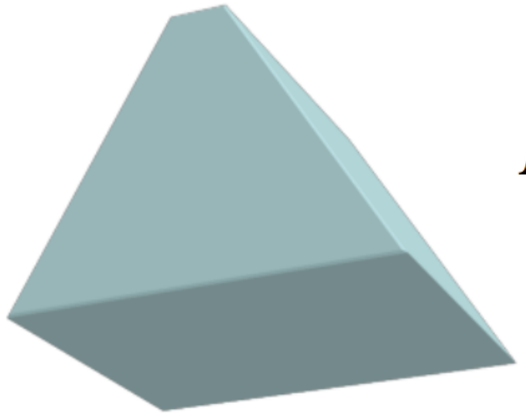
Metric

Q = affine
(last row is [0 0 0 1])



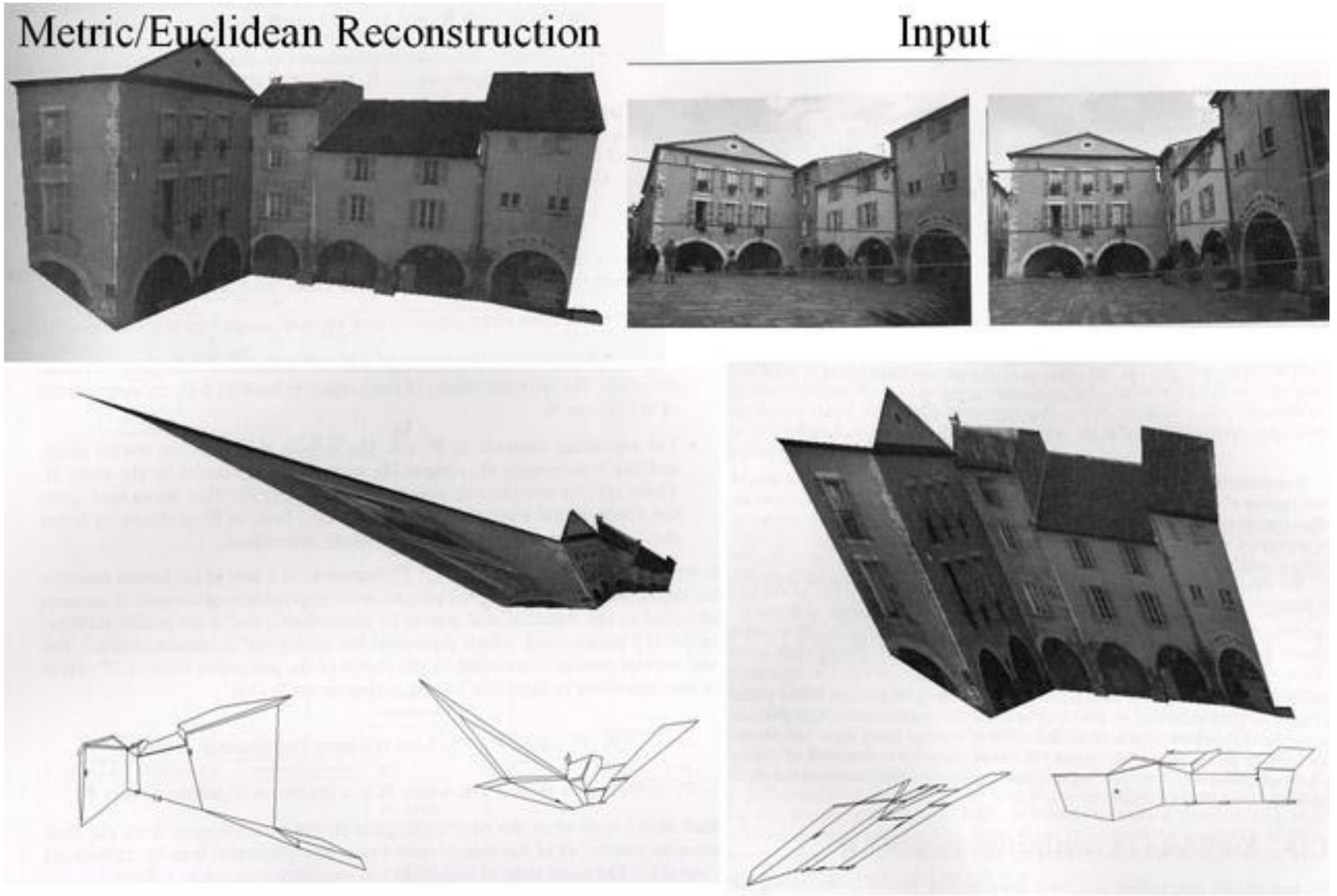
Affine

Q = general projective
(arbitrary 4x4 matrix)



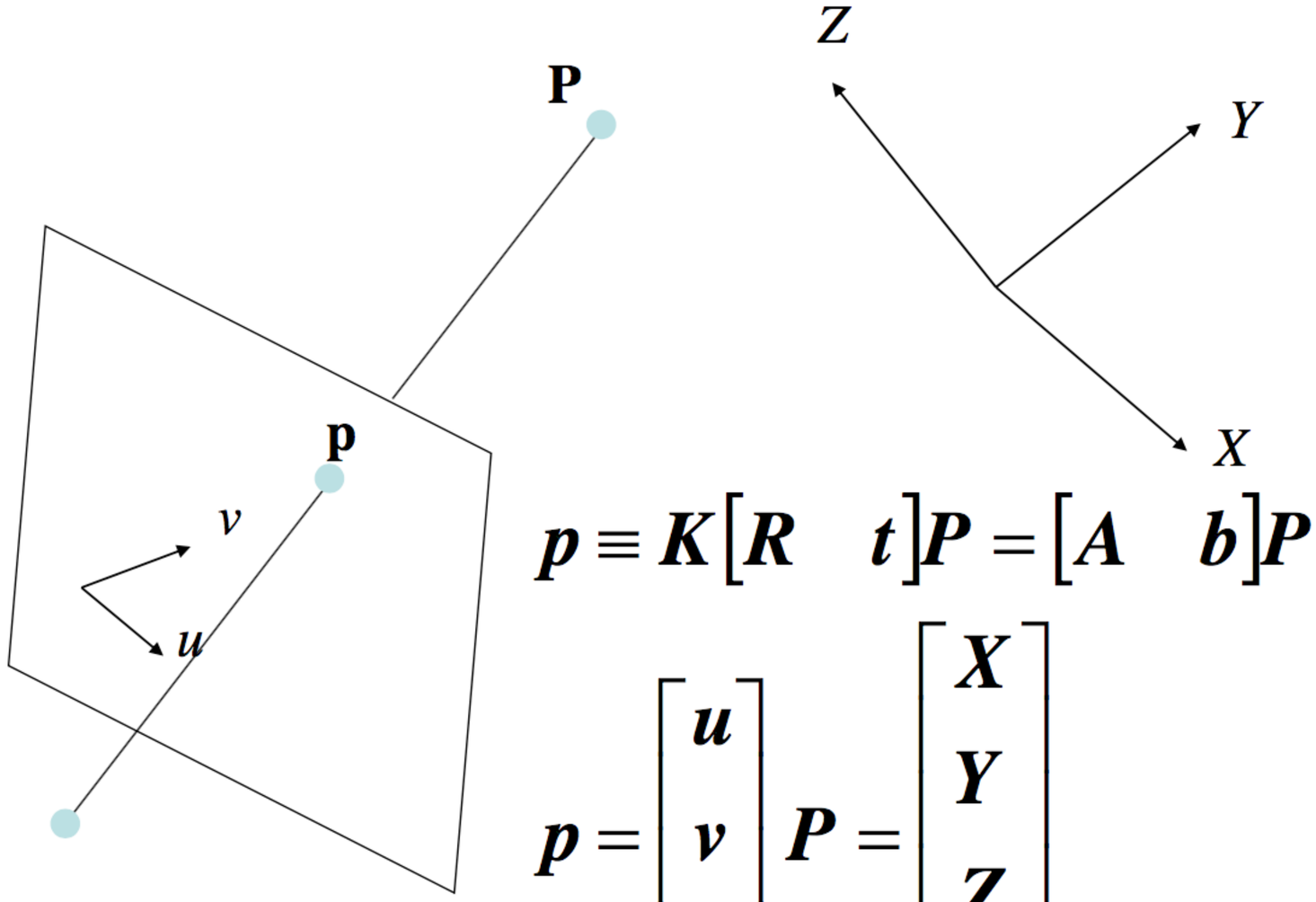
Projective

Example Transforms



General Case : Projective Reconstruction

General Projection Model

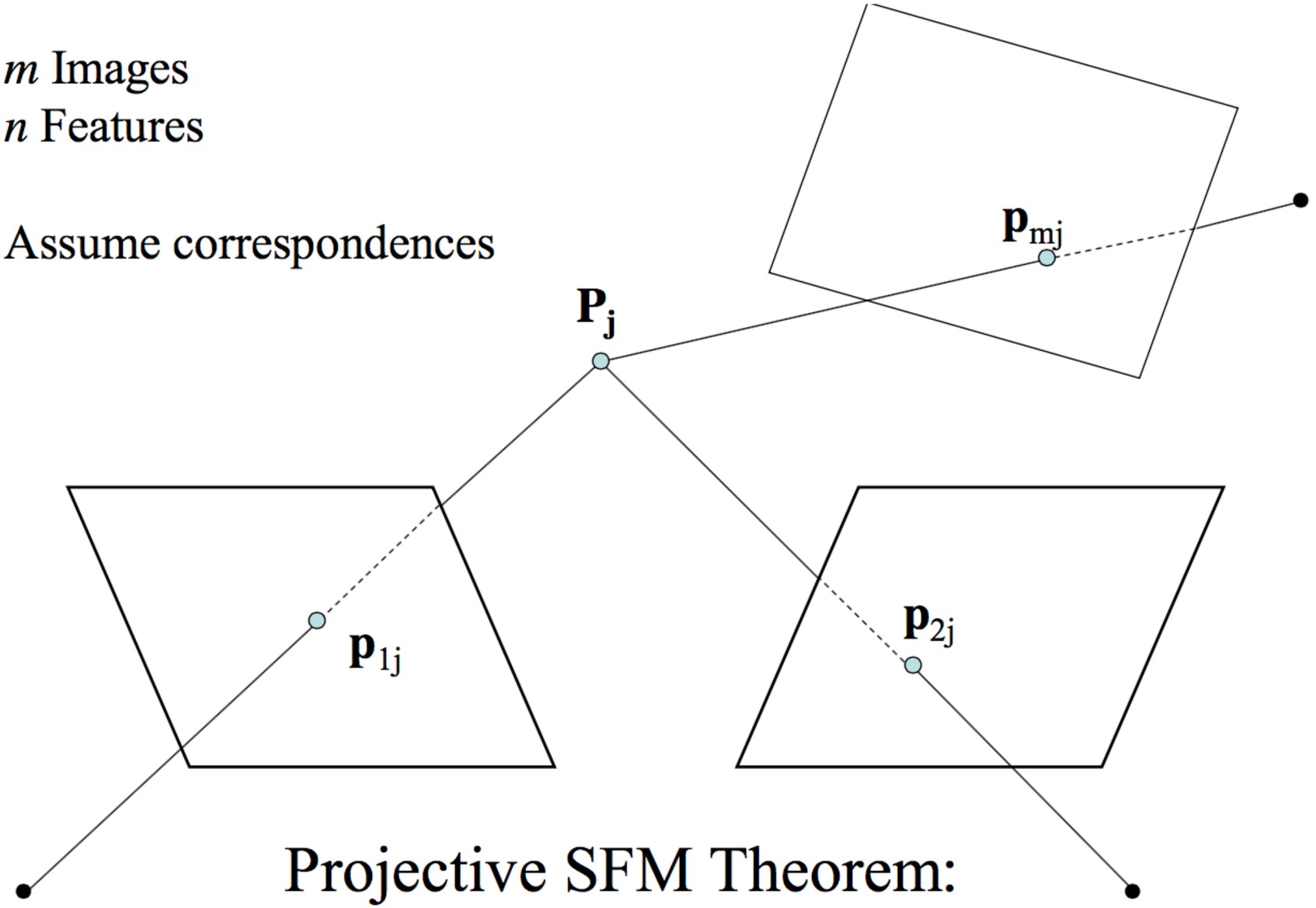


$$p \equiv K\begin{bmatrix} R & t \end{bmatrix} P = \begin{bmatrix} A & b \end{bmatrix} P$$

$$p = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad P = \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

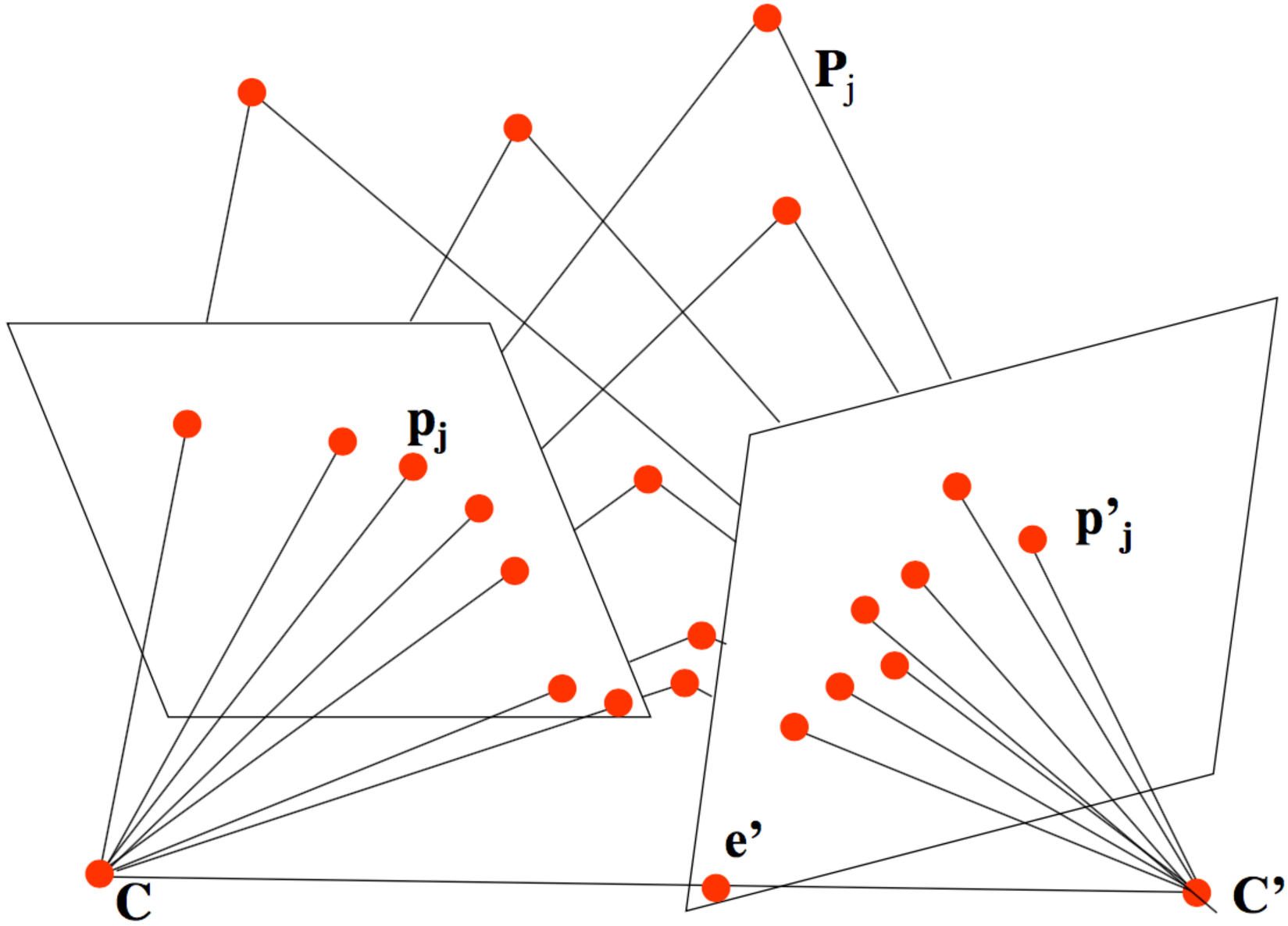
m Images
n Features

Assume correspondences



Projective SFM Theorem:
Reconstruction is possible as long as

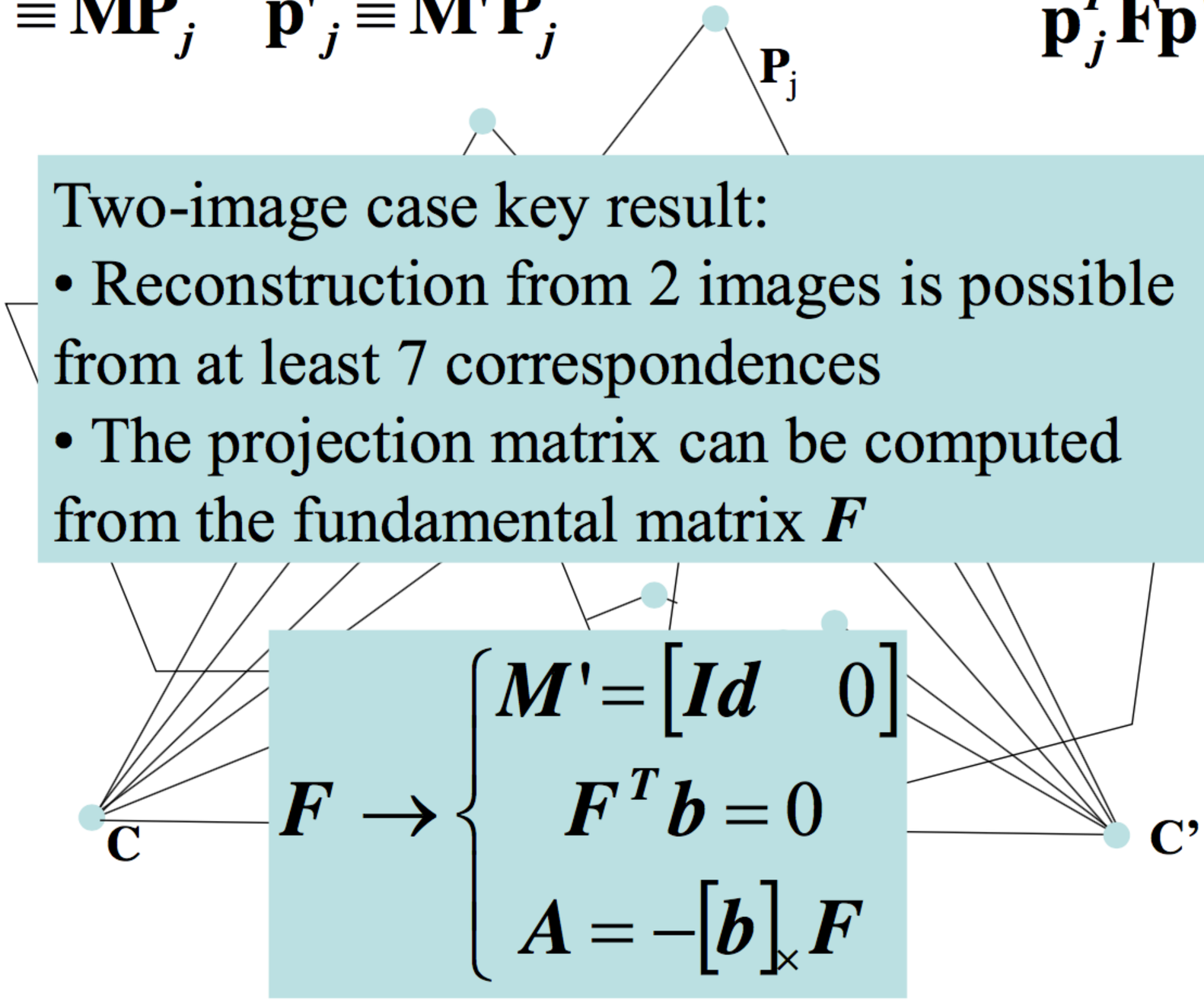
$$2mn >= 11m + 3n - 15$$



$$\mathbf{p}_j \equiv \mathbf{M}\mathbf{P}_j \quad \mathbf{p}'_j \equiv \mathbf{M}'\mathbf{P}_j$$

$$\mathbf{p}_j^T \mathbf{F} \mathbf{p}'_j = 0$$

$$\mathbf{p}_j \equiv \mathbf{M}\mathbf{P}_j \quad \mathbf{p}'_j \equiv \mathbf{M}'\mathbf{P}_j \qquad \mathbf{p}_j^T \mathbf{F} \mathbf{p}'_j = 0$$



Two-image case key result:

- Reconstruction from 2 images is possible from at least 7 correspondences
- The projection matrix can be computed from the fundamental matrix $\boldsymbol{F}$

$$\boldsymbol{F} \rightarrow \begin{cases} \boldsymbol{M}' = [\boldsymbol{Id} \quad 0] \\ \boldsymbol{F}^T \boldsymbol{b} = 0 \\ \boldsymbol{A} = -[\boldsymbol{b}]_\times \boldsymbol{F} \end{cases}$$

Issues

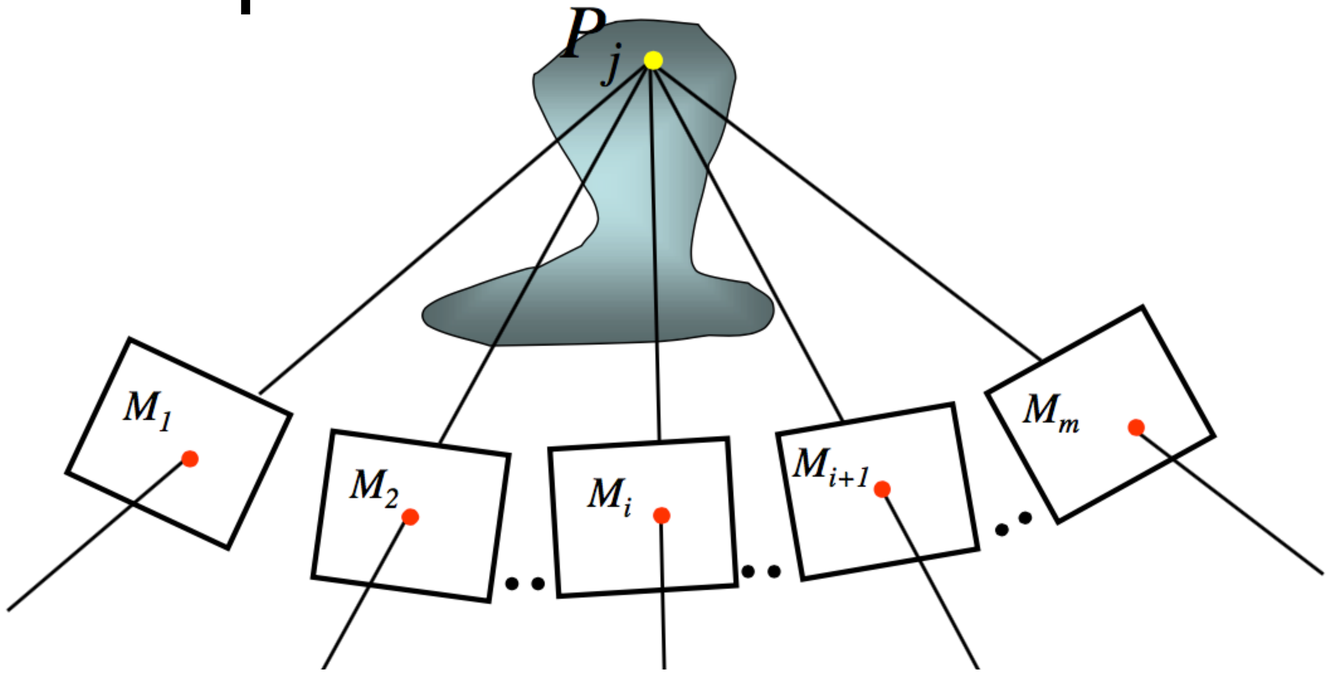
- We know how to compute the camera geometry from 2 images*
- Remaining issues:
  1. What about m > 2 images?
  2. Projective reconstruction is not unique: How can we find the metrically correct reconstruction?
  3. How can we find the “optimal” reconstruction?

*: (or 3, not shown in class)

Issue : Arbitrary number of images

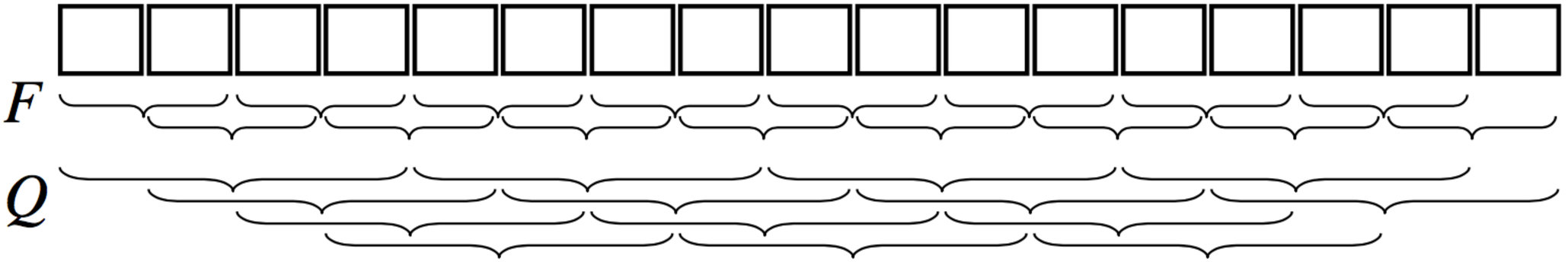
- Factorization (e.g., assuming affine first)
- Sequential methods
- Hierarchical methods
- …

Sequential Methods



- Start reconstruction from two (or three) images by using the fundamental matrix (or the trifocal tensor).
- Given projection matrix M_i of image i, find the projection matrix of image i+1 such that:
- $M_iP_j = M_{i+1}P_j$ for the points P_j in common between the two images
- M_{i+1} can be determined from 6 points → RANSAC if correspondences not known

Hierarchical Methods



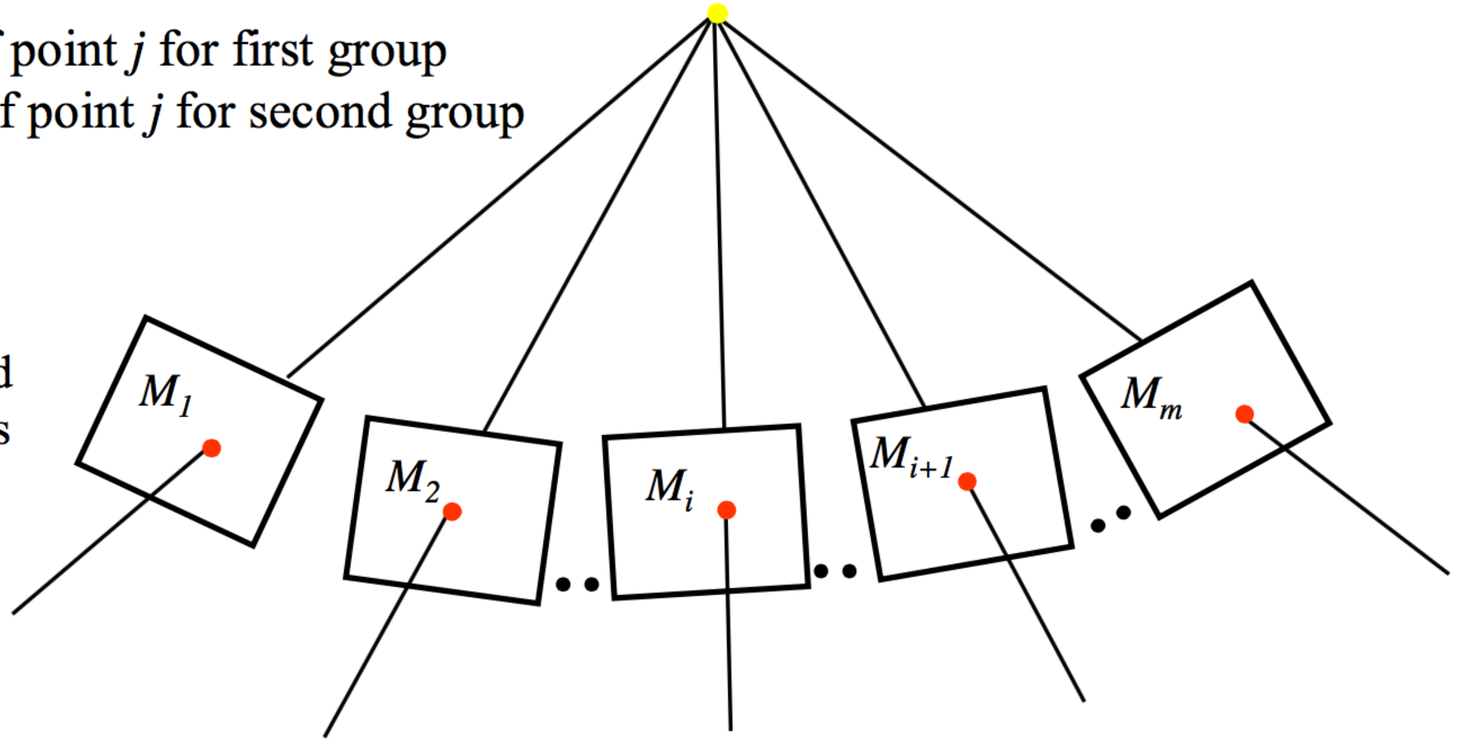
Example from Pollefeys

- Compute F from groups of 2 or 3 images
- Iteratively “stitch” the images into larger groups
- Stitching = Find a matrix Q that aligns the reconstruction from one group with the reconstruction from another group → Need some features in common between the groups

Stitching

P_j=coordinates of point j for first group
P'_j=coordinates of point j for second group

Find Q that transforms the point in the second group to the points in the first group



Distance in space

$$\min \sum d(P_j, QP'_j)$$

Distance in camera geometry

$$\min \sum d(M_i, M'_i Q^{-1})$$

Reprojection error

$$\min \sum d\left(M_i Q P'_j, p_{ij}\right) + \sum d\left(M'_i Q^{-1} P_j, p_{ij}\right)$$

Think about it…

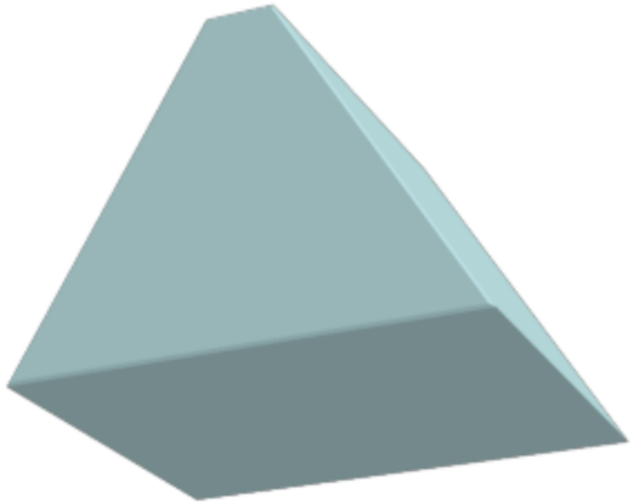
- In what situations would you prefer to use sequential or hierarchical methods?
- “Situations” might be:
  - Reconstruct a city from a database of Flicker images
  - Reconstruct a city from a drive-through using a vehicle or drone

Issues

- We know how to compute the camera geometry from 2 images*
- Remaining issues:
  1. What about m > 2 images?
  2. Projective reconstruction is not unique: How can we find the metrically correct reconstruction?
  3. How can we find the “optimal” reconstruction?

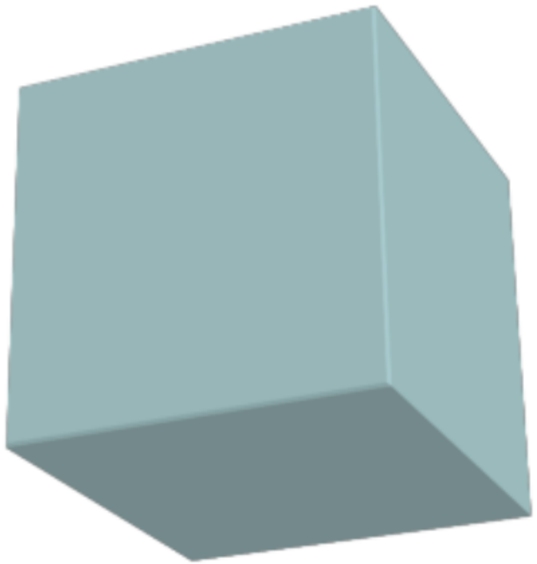
*: (or 3, not shown in class)

From Projective to Metric Reconstruction Auto Calibration



$$\mathbf{M}_i \leftarrow \mathbf{M}_i\mathbf{Q}$$

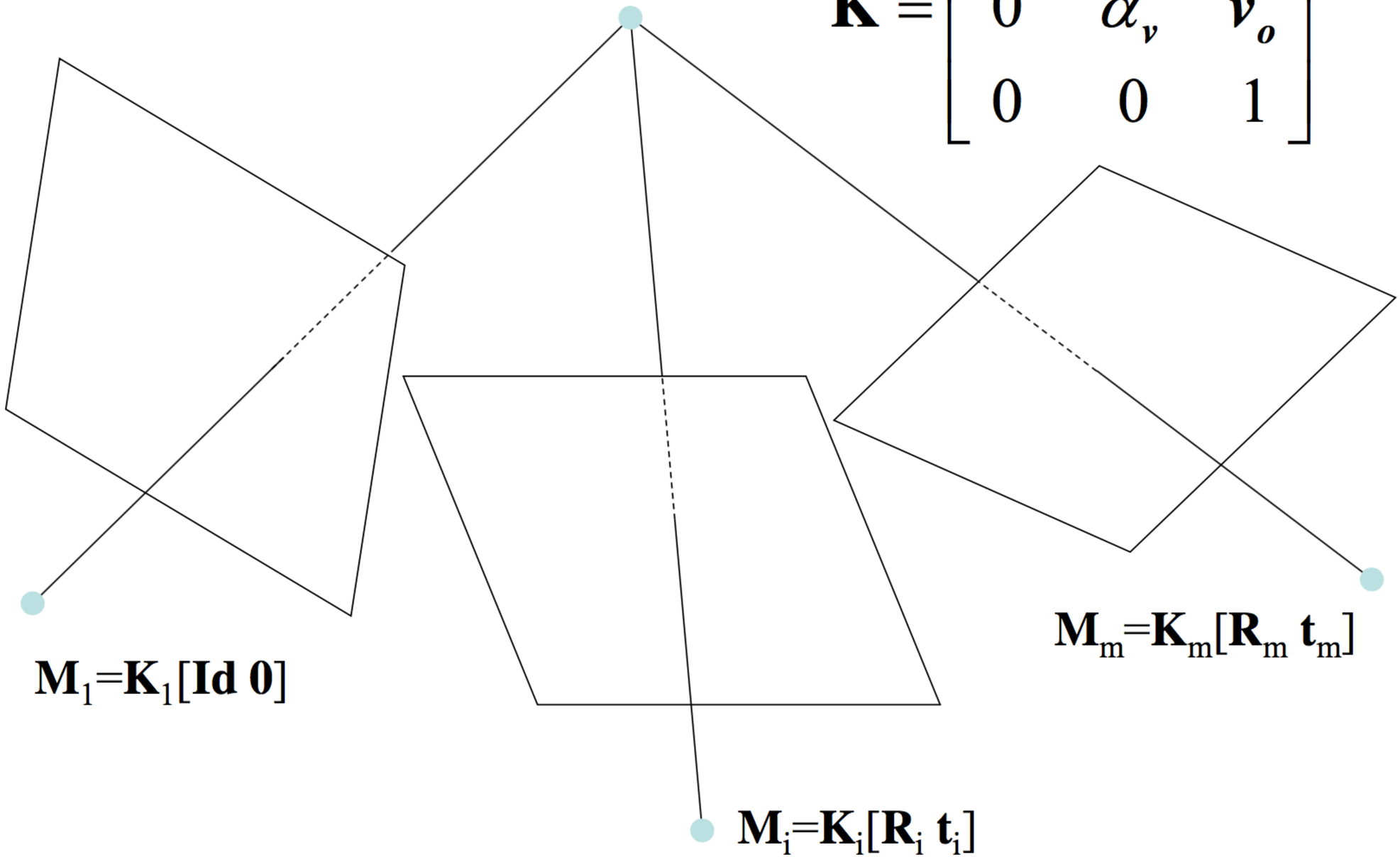
$$\mathbf{P}_j \leftarrow \mathbf{Q}^{-1}\mathbf{P}_j$$



Projective

Metric

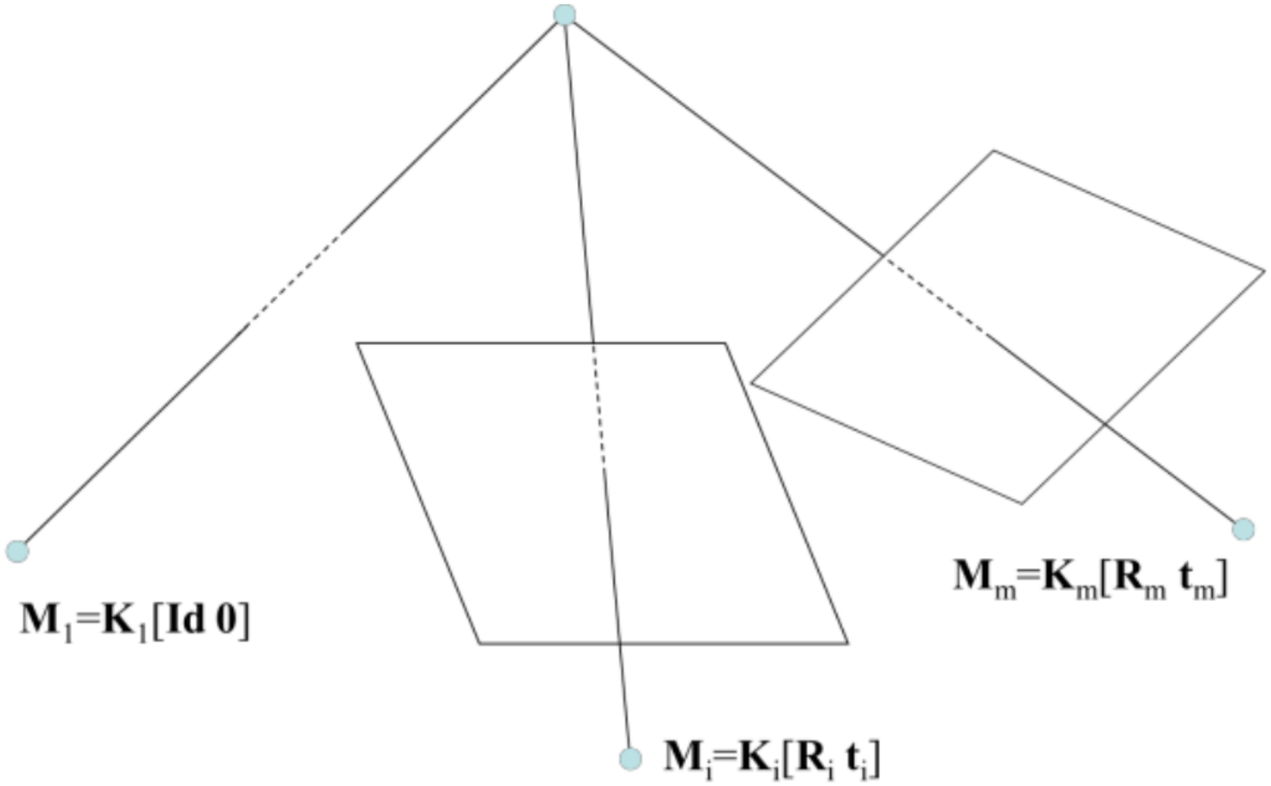
$$\mathbf{K} = \begin{bmatrix} \alpha_u & s & u_o \\ 0 & \alpha_v & v_o \\ 0 & 0 & 1 \end{bmatrix}$$



$\mathbf{M}_1 = \mathbf{K}_1[\mathbf{Id}\ \mathbf{0}]$

$\mathbf{M}_m = \mathbf{K}_m[\mathbf{R}_m\ \mathbf{t}_m]$

$\mathbf{M}_i = \mathbf{K}_i[\mathbf{R}_i\ \mathbf{t}_i]$



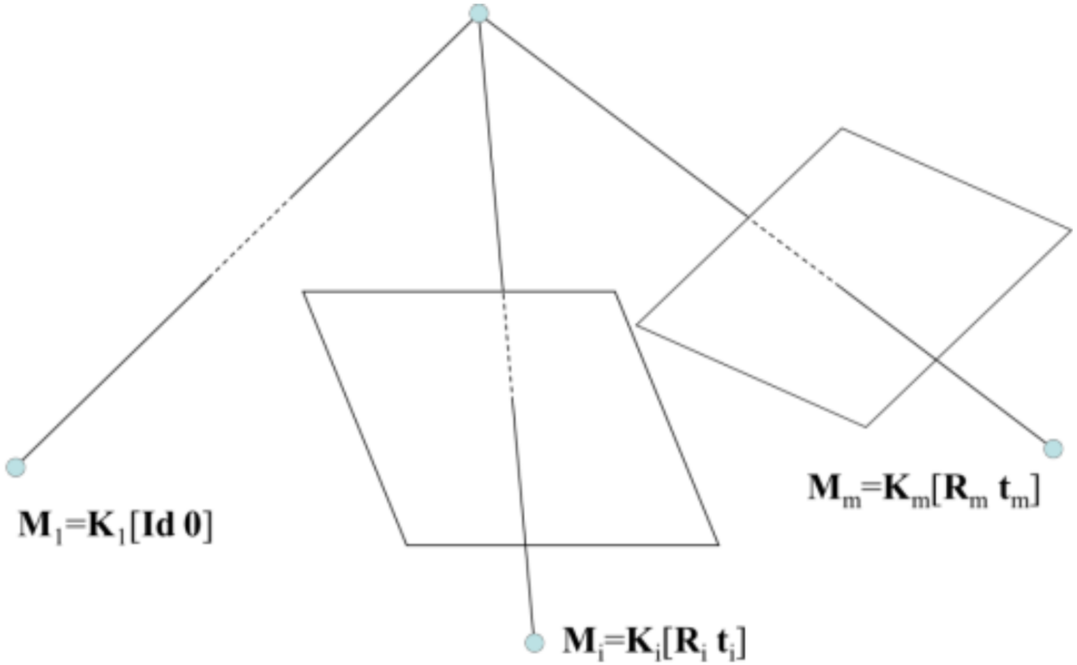
Metric Upgrade:

The projective reconstruction gives us a set of 3x4 projection matrices $\mathbf{M}_i$ for each camera i=1,..,m. The next problem is to convert this projective reconstruction to a metric reconstruction. Specifically, we want to find a 4x4 matrix $\mathbf{Q}$ such that:

$\mathbf{M}_i\mathbf{Q} \equiv \mathbf{K}_i[\mathbf{R}_i|\mathbf{t}_i]$

$\mathbf{R}_i$ and $\mathbf{t}_i$ are the rotation/translation between the coordinate system of camera i and an arbitrary coordinate system.

α_x and α_y are the scales in the x and y directions, x_o and y_o are the coordinates of the center, and s is the skew of the camera ($s = 0$ if the axes are orthogonal.)



Fundamental Transformation:

Our fundamental equation is:

$$\mathbf{M}_i\mathbf{Q} \equiv \mathbf{K}[\mathbf{R}_i\ \mathbf{t}_i]$$

Denoting the matrix formed by taking the first 3 columns of $\mathbf{Q}$ by $\mathbf{Q}_3$, such that $\mathbf{Q} = [\mathbf{Q}_3\ \mathbf{q}_4]$, we have: $\mathbf{M}_i\mathbf{Q}_3 \equiv \mathbf{K}_i\mathbf{R}_i$.

Taking the first 3 columns and observing that $\mathbf{R}_i$ is a rotation matrix:

$$\mathbf{M}_i\mathbf{Q}_3\mathbf{Q}_3^T\mathbf{M}_i^T \equiv \mathbf{K}_i\mathbf{K}_i^T$$

This is the key observation: By writing that the first three columns of the product of $\mathbf{M}$ by $\mathbf{Q}$ is a rotation, we are able to eliminate the rotation from the unknowns. All that is left are the matrices of internal parameters for each of the images. This process is sometimes called auto-calibration since it amounts to calibrating the internal parameters of the cameras directly from images.

It is important to understand the number of degrees of freedom in $\mathbf{Q}_3$. The total number of entries in $\mathbf{Q}_3$ is 4x3=12. The matrix is defined up to scale since all the equalities are homogeneous. Moreover, the matrix is defined up to a rotation since for any arbitrary rotation R: $\mathbf{Q}_3\mathbf{R}\mathbf{R}^T\mathbf{Q}_3^T = \mathbf{Q}_3\mathbf{Q}_3^T$ so that if $\mathbf{Q}_3$ is a solution, so is $\mathbf{Q}_3\mathbf{R}$. These additional degrees of freedom simply reflect the fact that one can choose the orientation and scale of the global coordinate system arbitrarily. Therefore, $\mathbf{Q}_3$ is characterized by $12 - 1 - 3 = 8$ unknowns.

Basic Trick

$$\boldsymbol{M}_i\boldsymbol{Q} \equiv \boldsymbol{K}_i\begin{bmatrix}\boldsymbol{R}_i & \boldsymbol{t}_i\end{bmatrix}$$

$$\Downarrow$$

Look only at the first 3 columns

$$\boldsymbol{M}_i\boldsymbol{Q}_3 \equiv \boldsymbol{K}_i\boldsymbol{R}_i$$

$$\Downarrow$$

Use the fact that $\boldsymbol{R}$ is a rotation

$$\boldsymbol{M}_i\boldsymbol{Q}_3\boldsymbol{Q}_3^T\boldsymbol{M}_i^T \equiv \boldsymbol{K}_i\boldsymbol{R}_i\boldsymbol{R}_i^T\boldsymbol{K}_i^T = \boldsymbol{K}_i\boldsymbol{K}_i^T$$

$$M_i Q \equiv K_i \begin{bmatrix} R_i & t_i \end{bmatrix}$$

$$\Downarrow$$

Look only at the first 3 columns

$$M_i Q_3 \equiv K_i R_i$$

Use the fact that R is a rotation

$$\Downarrow$$

$$M_i Q_3 Q_3^T M_i^T \equiv K_i R_i R_i^T K_i^T = K_i K_i^T$$

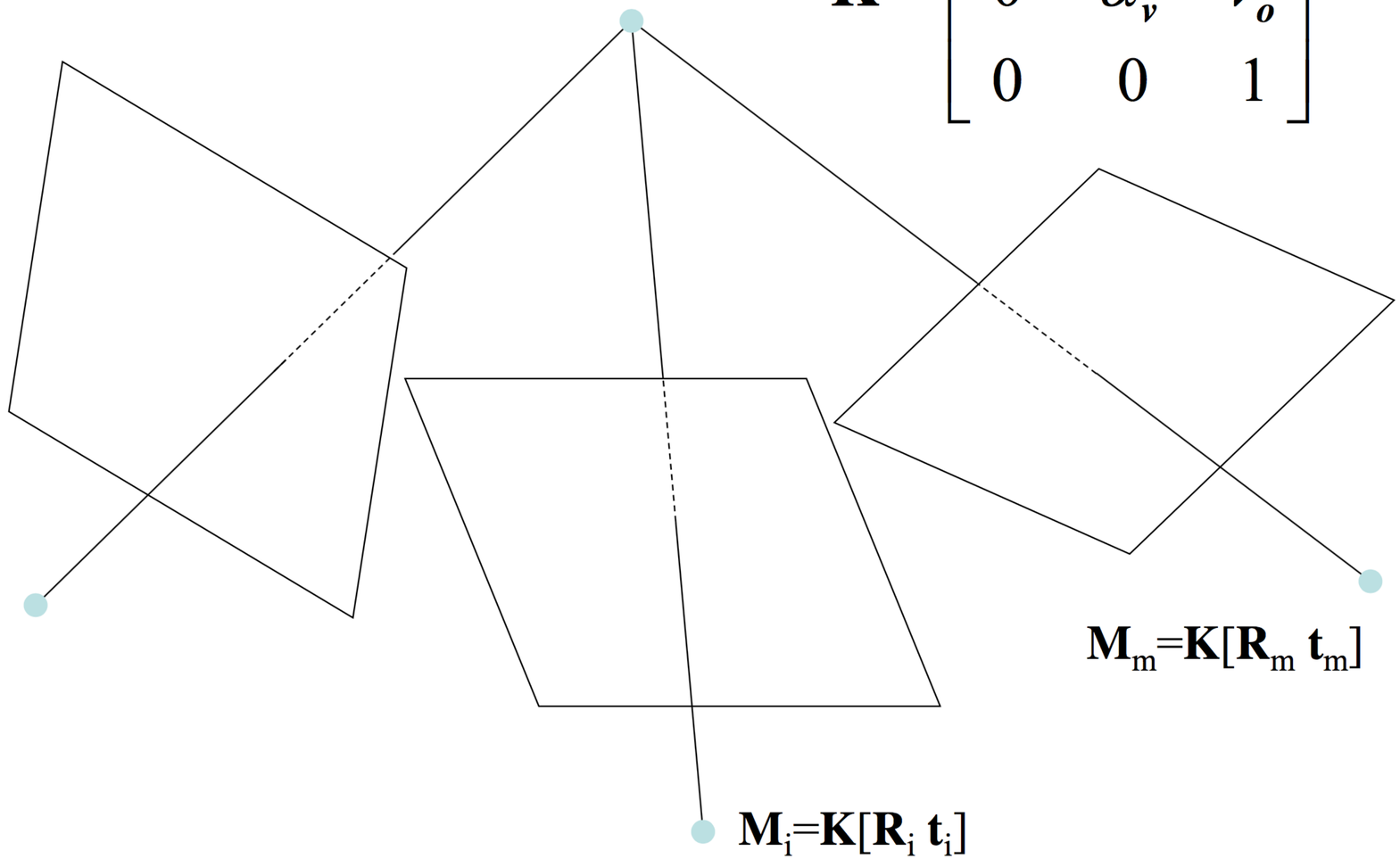
For convenience, we denote the matrix the matrix $\mathbf{Q}_3\mathbf{Q}_3^T$ by $\mathbf{L}$ (a 4x4 matrix) and $\mathbf{M}_i\mathbf{L}\mathbf{M}_i^T$ by $\boldsymbol{\omega}_i$. The set of equations to solve is:

$$\mathbf{M}_i\mathbf{L}\mathbf{M}_i^T \equiv \mathbf{K}_i\mathbf{K}_i^T \quad i=1,..,m$$

Each image generates 5 independent equations (the left hand side is a 3x3 symmetric matrix, but the equality is up to scale). The total number of unknowns is 8 ($\mathbf{Q}_3$) + 5m ($\mathbf{K}_i\mathbf{K}_i^T$). Therefore, the number of equations (5m) is *always* lower than the number of unknowns (8 + 5m) and we can never solve this system of equations without some constraints on the cameras. The key question is what constraints can be used. A couple of constraints are investigated below, followed with a general result.

Identical Intrinsic Parameters (K)

$$\mathbf{K} = \begin{bmatrix} \alpha_u & s & u_o \\ 0 & \alpha_v & v_o \\ 0 & 0 & 1 \end{bmatrix}$$

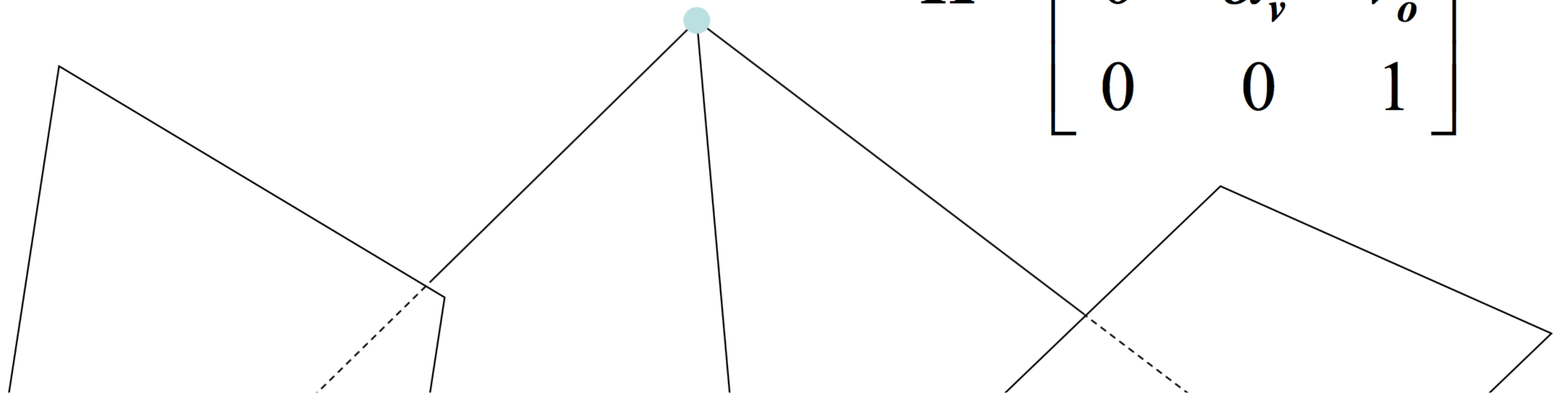


$\mathbf{M}_m=\mathbf{K}[\mathbf{R}_m\ \mathbf{t}_m]$

$\mathbf{M}_i=\mathbf{K}[\mathbf{R}_i\ \mathbf{t}_i]$

Identical Intrinsic Parameters (K)

$$\mathbf{K} = \begin{bmatrix} \alpha_u & s & u_o \\ 0 & \alpha_v & v_o \\ 0 & 0 & 1 \end{bmatrix}$$



Case 1: Identical Intrinsic Parameters:

Let us suppose now that we do not know the intrinsic parameters of the cameras, but that we do know that they are all identical, that is, $\mathbf{K}_i = \mathbf{K}$ for all cameras i. For all the cameras, we have:

$$\mathbf{M}_i\mathbf{L}\mathbf{M}_i^T \equiv \boldsymbol{\omega} \quad i=1,..,m \quad \text{(with the equality up to scale)}$$

Where $\boldsymbol{\omega}$ is computed from the first image:

$$\mathbf{M}_i\mathbf{L}\mathbf{M}_i^T \equiv \mathbf{M}_1\mathbf{L}\mathbf{M}_1^T \quad i=2,..,m$$

This gives us $5(m\text{-}1)$ independent equations for 8 unknowns (in $\mathbf{Q}_3$). Therefore, we can solve the reconstruction problem in this case if:

$$5(m\text{-}1) >= 8 \rightarrow m >= 3$$

$$\mathbf{K} = \begin{bmatrix} \alpha_u & s & u_o \\ 0 & \alpha_v & v_o \\ 0 & 0 & 1 \end{bmatrix}$$

$$\downarrow$$

$$\omega = \mathbf{K}\mathbf{K}^T = \begin{bmatrix} \alpha_u^2 + s^2 + u_o^2 & s\alpha_v + u_o v_o & u_o \\ & \alpha_v^2 + v_o^2 & v_o \\ & & 1 \end{bmatrix}$$

$$\omega = \mathbf{K}\mathbf{K}^{T} = \begin{bmatrix} \alpha_u^2 + s^2 + u_o^2 & s\alpha_v + u_o v_o & u_o \\ & \alpha_v^2 + v_o^2 & v_o \\ & & 1 \end{bmatrix}$$

Case 2: Principal point at origin

In that case, $u_o = v_o = 0$, which implies that $\omega_{13} = \omega_{23} = 0$. Therefore, going back to the original equation, we can write that:

$(\mathbf{M}_i\mathbf{L}\mathbf{M}_i^T)_{13} = 0$ and $(\mathbf{M}_i\mathbf{L}\mathbf{M}_i^T)_{23} = 0$

Those are two equations in $\mathbf{L}$ that are independent of $\mathbf{K}_i$. We have $2m$ such equations for m views for 8 unknowns in $\mathbf{L}$ (meaning, 8 unknowns in $\mathbf{Q}_3$). These equations are independent of $\mathbf{K}$. Therefore:

If the principal point is at the origin, a metric reconstruction can be obtained from a minimum of 4 views. $m >= 4$

$$\omega = \mathbf{K}\mathbf{K}^{T} = \begin{bmatrix} \alpha_u^2 + s^2 + u_o^2 & s\alpha_v + u_o v_o & u_o \\ & \alpha_v^2 + v_o^2 & v_o \\ & & 1 \end{bmatrix}$$

Case 3: Zero-Skew

If the skew is zero but all the other parameters are allowed to vary, then we have the constraint:

$\omega_{12}\omega_{33} = \omega_{13}\ \omega_{23}$

This provides *m* constraints. Therefore, we must have *m*>=8, thus 8 images are necessary.

Assumption	Fixed f=	Known k=	Constraints	Images m=
Constant **K**	5	0	$\omega_{ij}/\omega_{33} = \omega^1_{ij}/\omega^1_{33}$	3
Principal point known	0	2	$\omega_{13}=\omega_{23}=0$	4
Aspect ratio and skew constant	2	0		5
Zero Skew	0	1	$\omega_{12}\omega_{33}=\omega_{13}\omega_{23}$	8
P.P. known + Zero skew	0	3	$\omega_{12}=0$ $\omega_{13}=\omega_{23}=0$	3

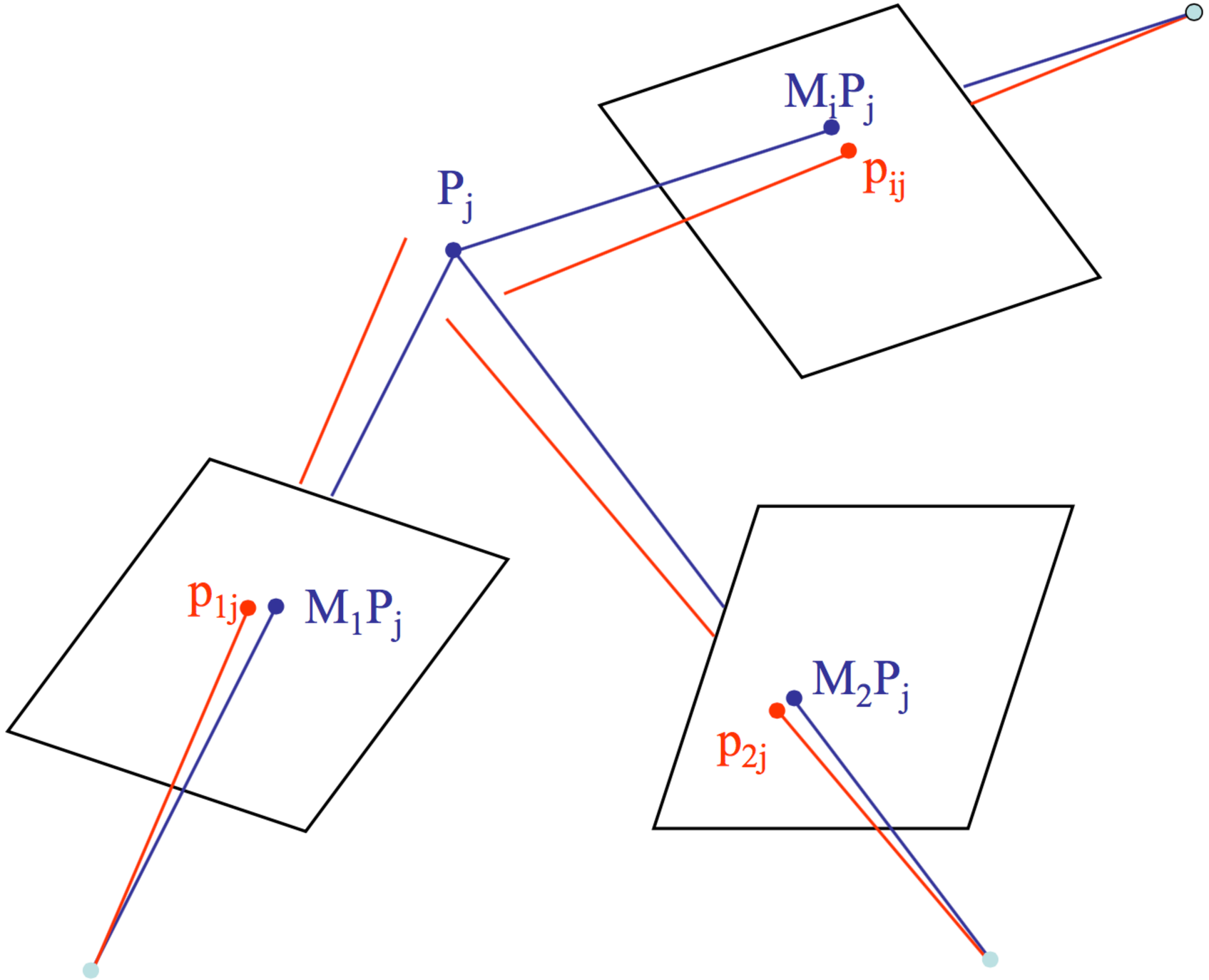
$$mk + (m\text{-}1)f >= 8$$

Issues

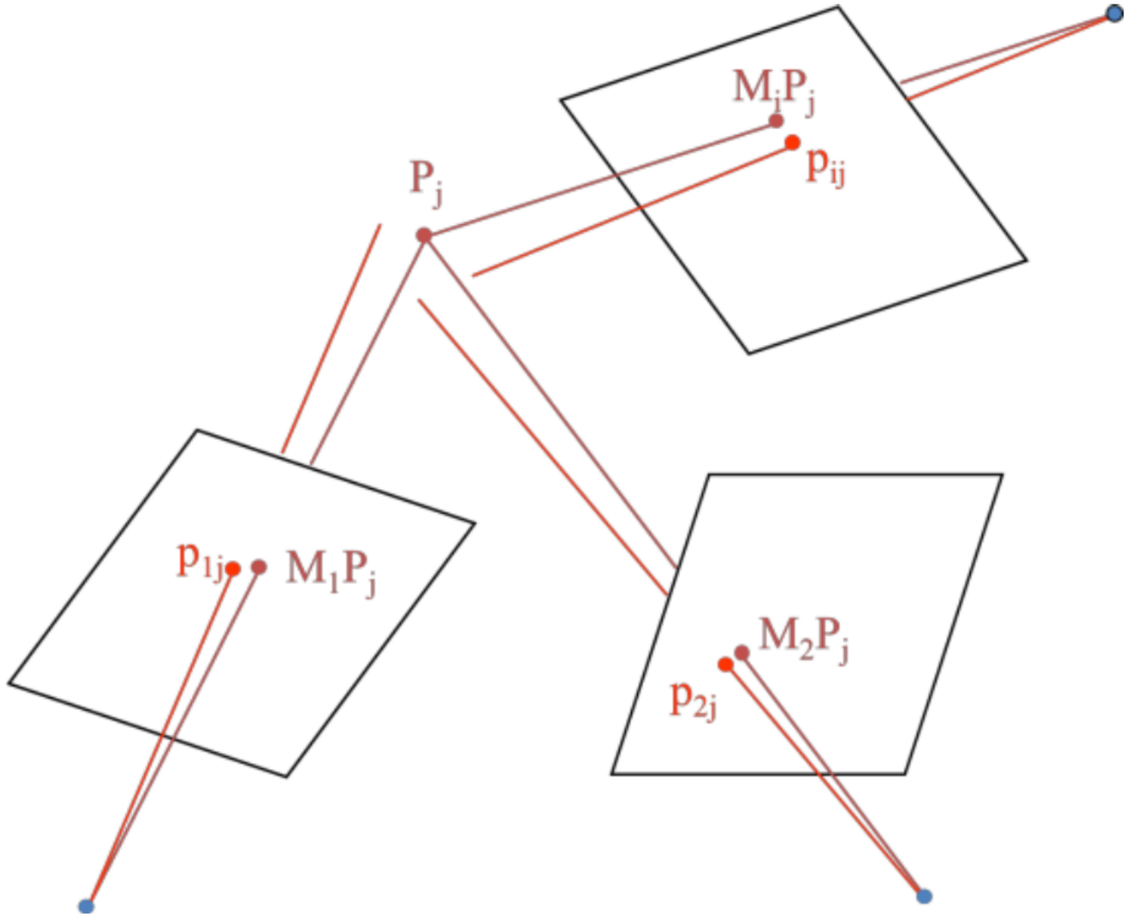
- We know how to compute the camera geometry from 2 images*
- Remaining issues:
  1. What about m > 2 images?
  2. Projective reconstruction is not unique: How can we find the metrically correct reconstruction?
  3. How can we find the “optimal” reconstruction?

*: (or 3, not shown in class)

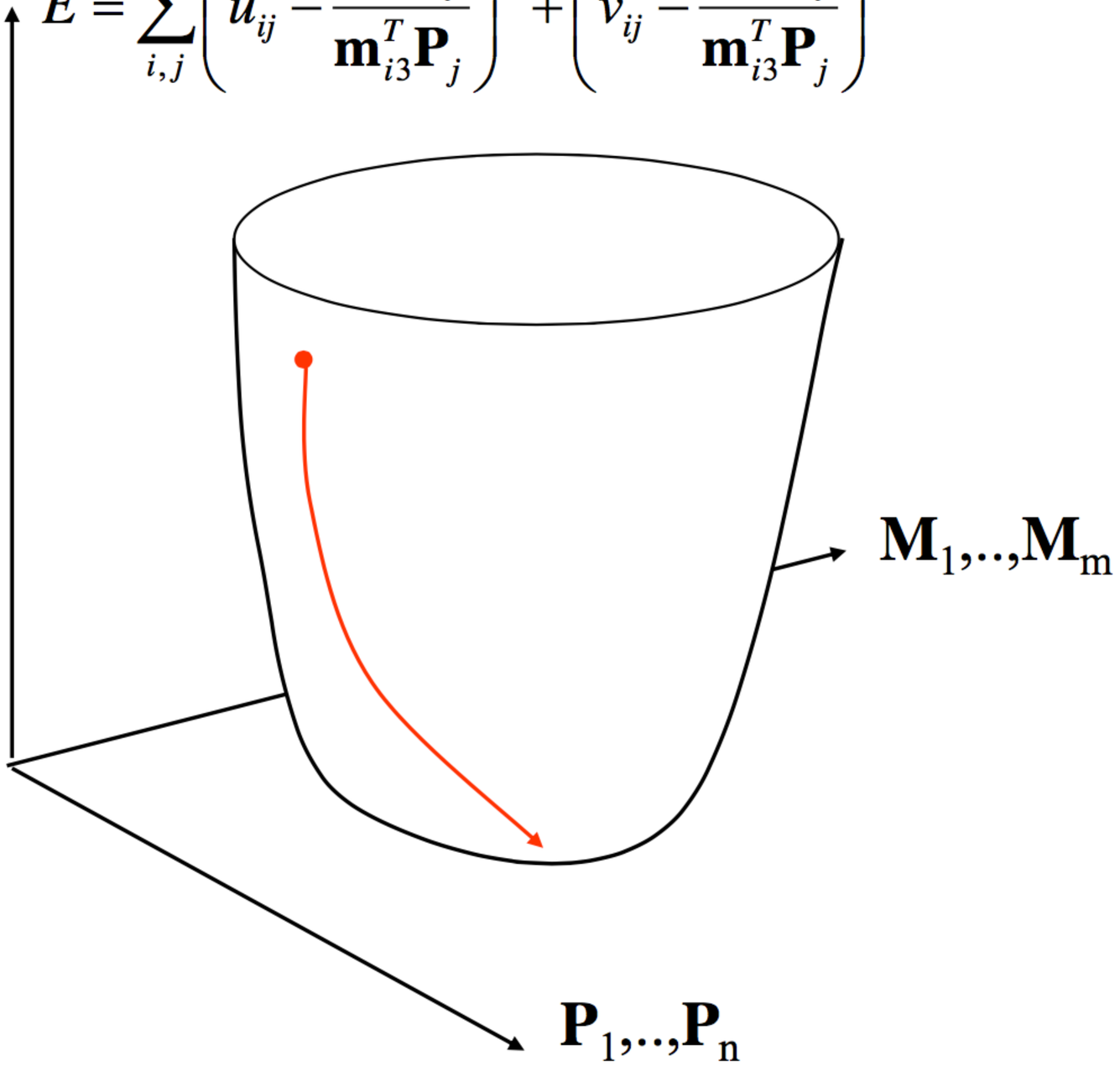
Non-Linear Refinement: Bundle Adjustment
to be continued…


M_iP_j
p_{ij}
P_j
p_{1j}
M_1P_j
M_2P_j
p_{2j}

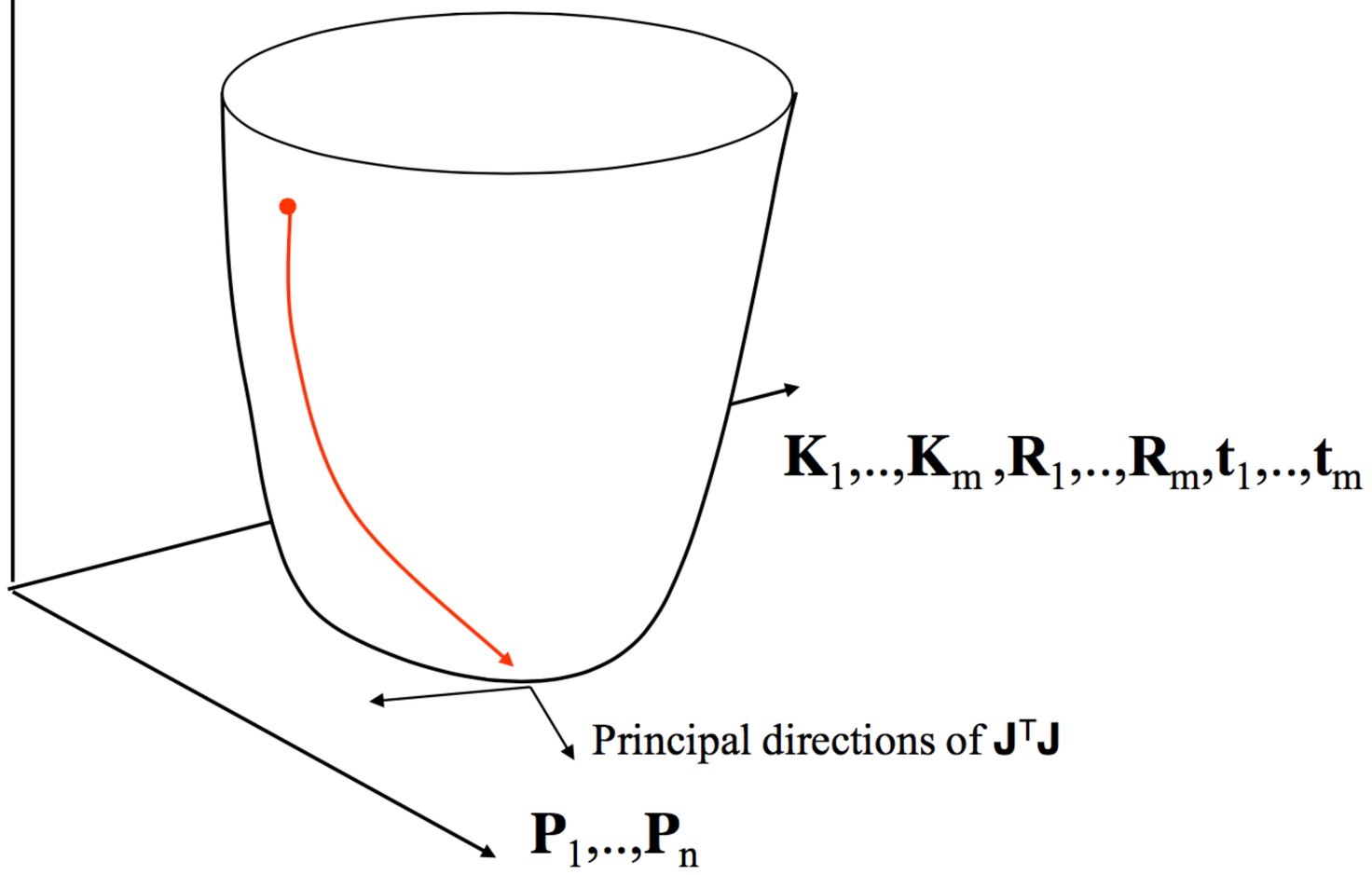
$$E = \sum_{i,j} \left(u_{ij} - \frac{\mathbf{m}_{i1}^T \mathbf{P}_j}{\mathbf{m}_{i3}^T \mathbf{P}_j} \right)^2 + \left(v_{ij} - \frac{\mathbf{m}_{i2}^T \mathbf{P}_j}{\mathbf{m}_{i3}^T \mathbf{P}_j} \right)^2$$



$$E = \sum_{i,j} \left(u_{ij} - \frac{\mathbf{m}_{i1}^T \mathbf{P}_j}{\mathbf{m}_{i3}^T \mathbf{P}_j} \right)^2 + \left(v_{ij} - \frac{\mathbf{m}_{i2}^T \mathbf{P}_j}{\mathbf{m}_{i3}^T \mathbf{P}_j} \right)^2$$



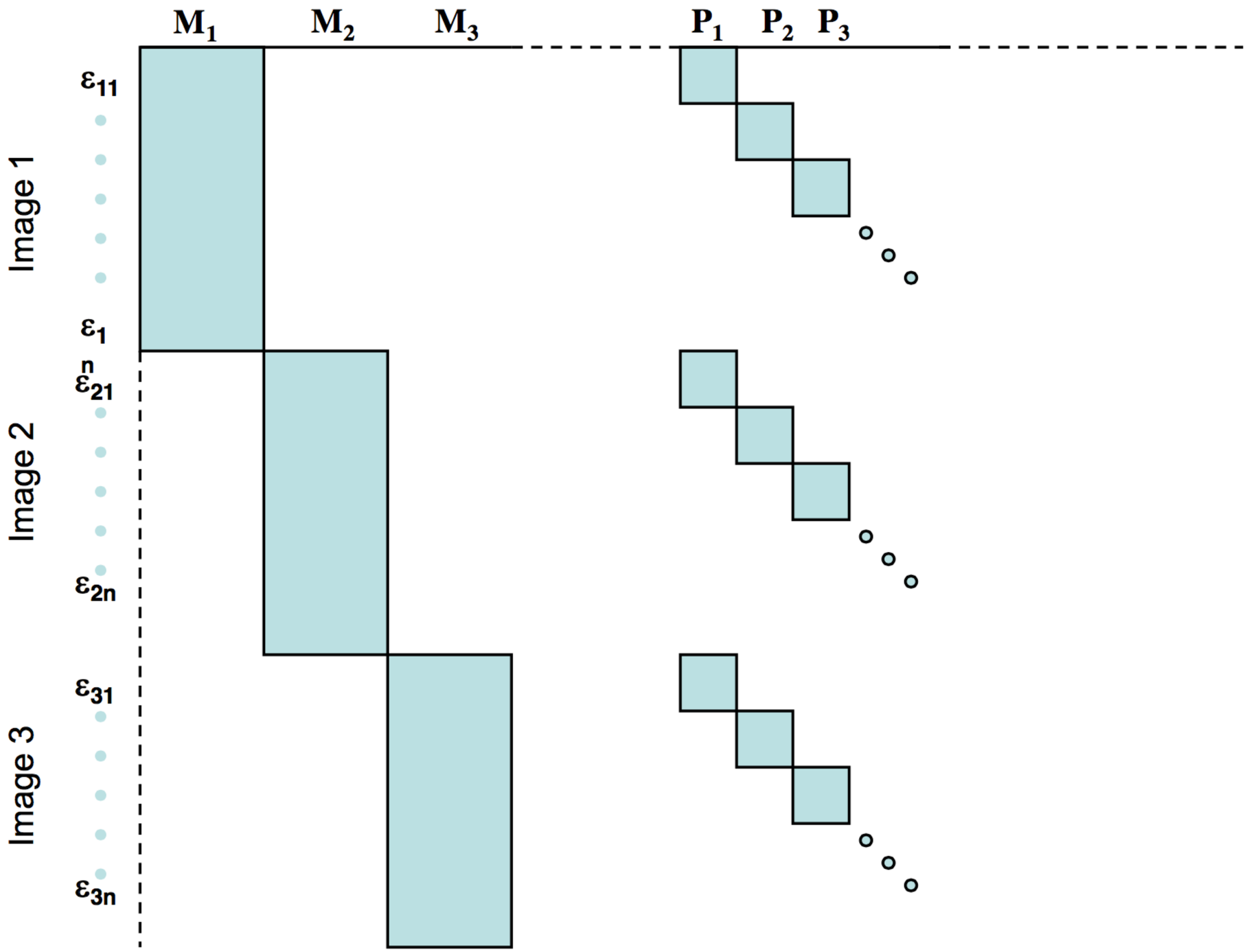
$$E = \sum_{i,j} \left(u_{ij} - \frac{\mathbf{m}_{1i}^T \mathbf{P}_j}{\mathbf{m}_{3i}^T \mathbf{P}_j} \right)^2 + \left(v_{ij} - \frac{\mathbf{m}_{2i}^T \mathbf{P}_j}{\mathbf{m}_{3i}^T \mathbf{P}_j} \right)^2$$

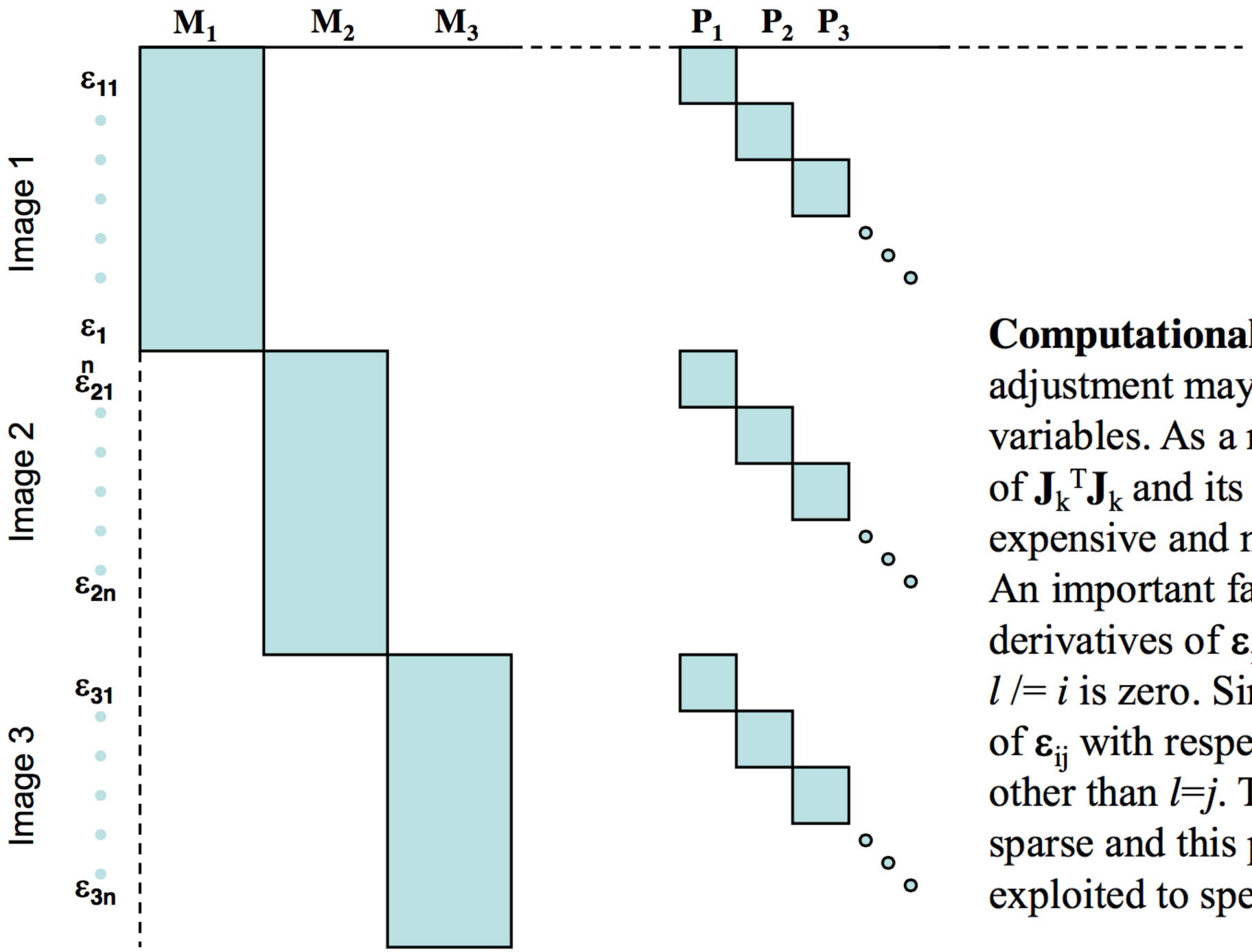


The directions of uncertainty are the eigenvectors of $\mathbf{J}^T\mathbf{J}$

The magnitudes of the uncertainty are the inverse eigenvalues of $\mathbf{J}^T\mathbf{J}$

Formally: The covariance matrix of the error on all the variables is $\mathbf{\Sigma} = (\mathbf{J}^T\mathbf{J})^{-1}$

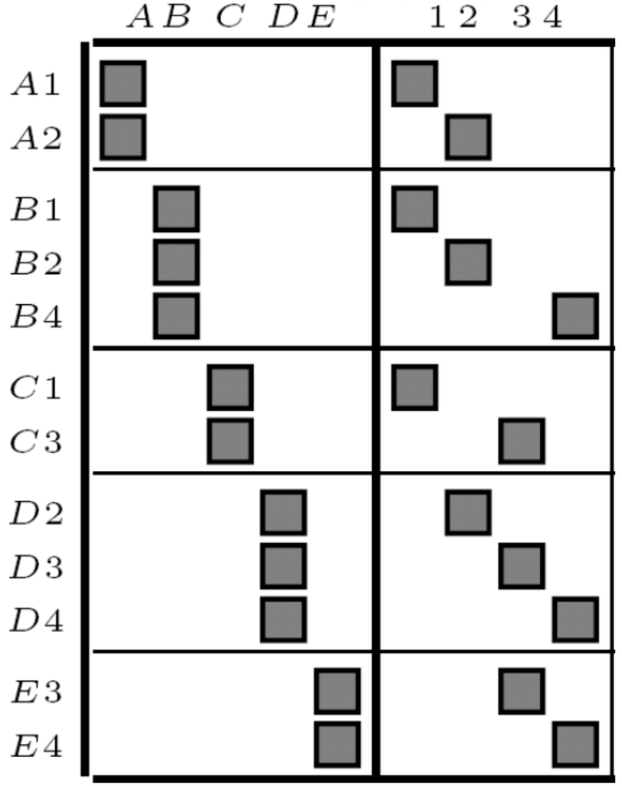

M1
M2
M3
P1
P2
P3
Image 1
Image 2
Image 3
ε11
ε1
n
ε21
ε2n
ε31
ε3n



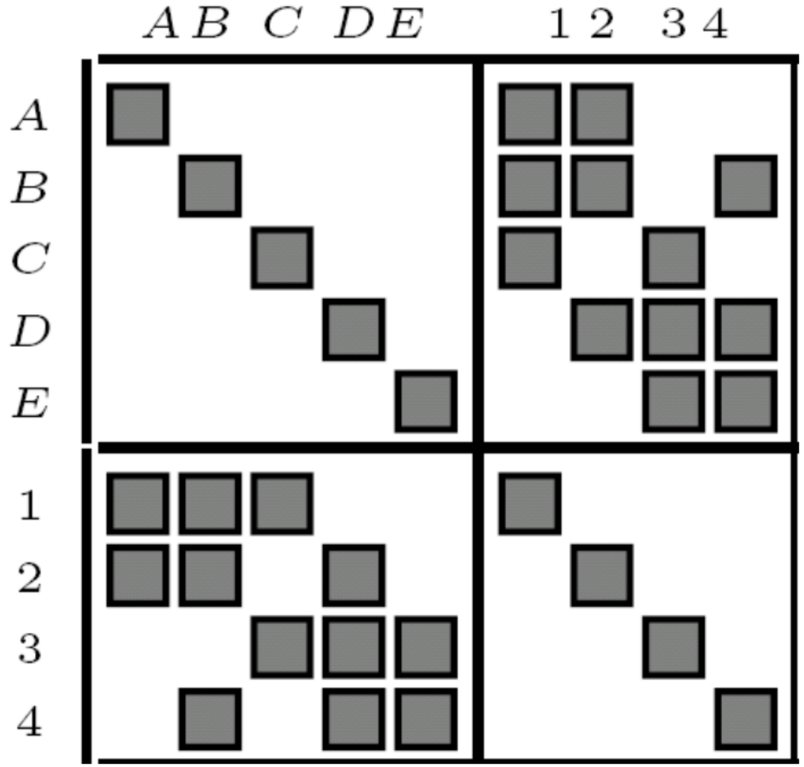
Computational issues: Bundle adjustment may involve hundreds of variables. As a result, the computation of $\mathbf{J}_k^T\mathbf{J}_k$ and its inversion may be expensive and numerically unstable. An important fact to note is that the derivatives of $\boldsymbol{\varepsilon}_{ij}$ with respect to $\mathbf{M}_l$ for $l \neq i$ is zero. Similarly, the derivative of $\boldsymbol{\varepsilon}_{ij}$ with respect to $\mathbf{P}_l$ is zero for any l other than $l=j$. Therefore, $\mathbf{J}_k^T\mathbf{J}_k$ is very sparse and this property can be exploited to speed up the iterations.

General case: Example

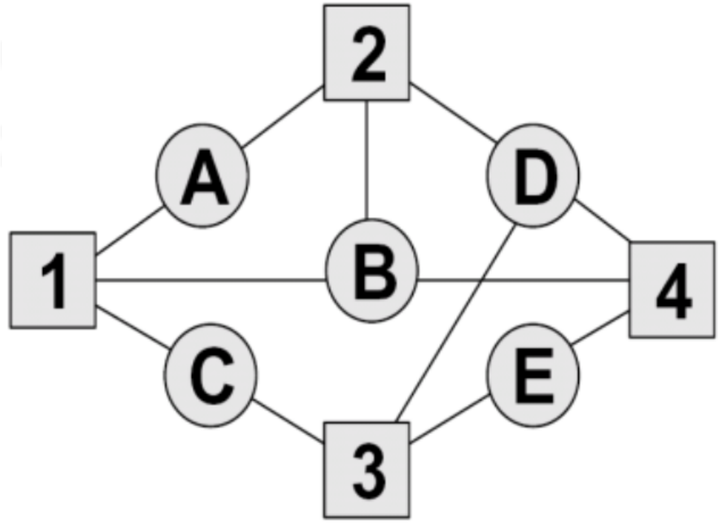
$J =$



$J^{\mathrm{T}}J =$



- Features: A,B,C,D,E
- Images: 1,2,3



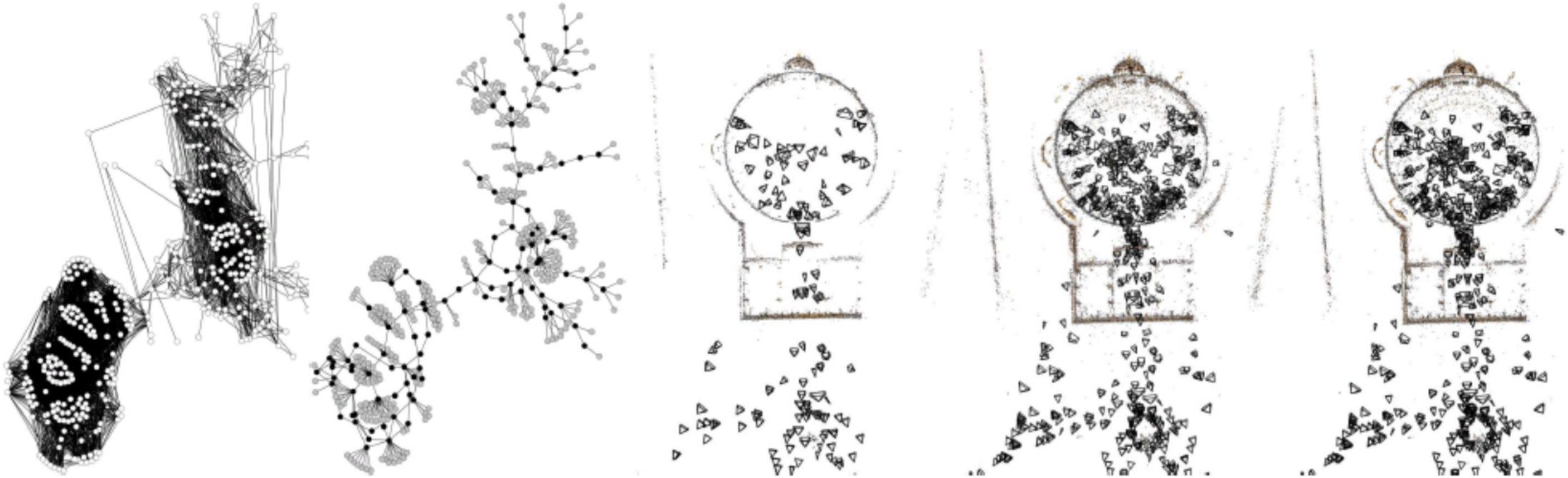
Open Source Bundle Adjustment

- http://phototour.cs.washington.edu/bundler/
- factor graph optimization - GTSAM, G2O, Others

Examples

- http://grail.cs.washington.edu/projects/rome/
- Rome: 150,000 images
  – Trevi Fountain, 1,936 images, 656,699 points
  – St. Peter's Basilica, 1,294 images, 530,076 points
- Venice 250,000 images
  – The Grand Canal, 3,272 images, 561,389 points
  – San Marco Square, 14,079 images, 4,515,157 points
- Dubrovnik 58,000images
- Building Rome in a Day, International Conference on Computer Vision, 2009
- Building Rome on a Cloudless Day,European Conference on Computer Vision, 2010

Dealing with Huge Numbers of Images

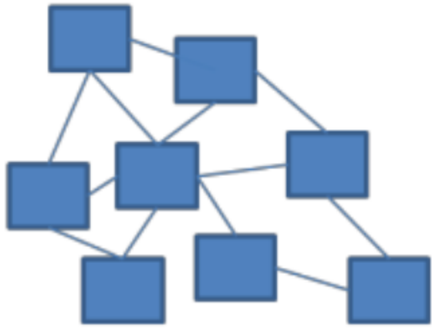


Full image graph Skeletal graph Reconstruction from skeletal Reconstruction from full

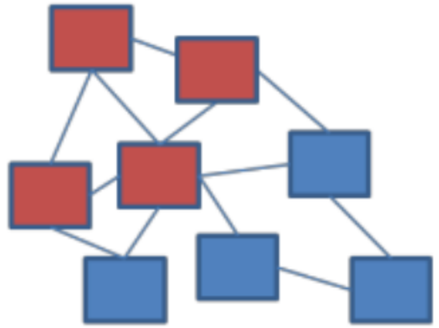
- Create graph of images with weights on edges = uncertainty of 2-image reconstruction
- Construct skeletal graph = lots of leaf + interior nodes
- More leaves = faster bundle adjustment but higher uncertainty
- Trade-off between speed and accuracy
- Noah Snavely, Steven M. Seitz, Richard Szeliski. Skeletal Sets for Efficient Structure from Motion. Proc. Computer Vision and Pattern Recognition (CVPR), 2008.

Dealing with Huge Numbers of Images

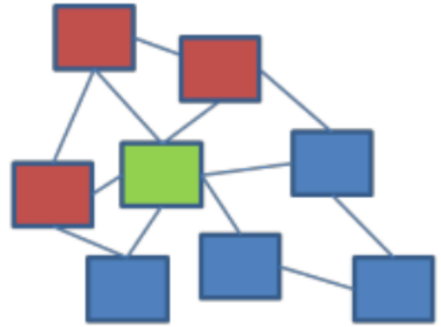
Cluster based on appearance using gist features

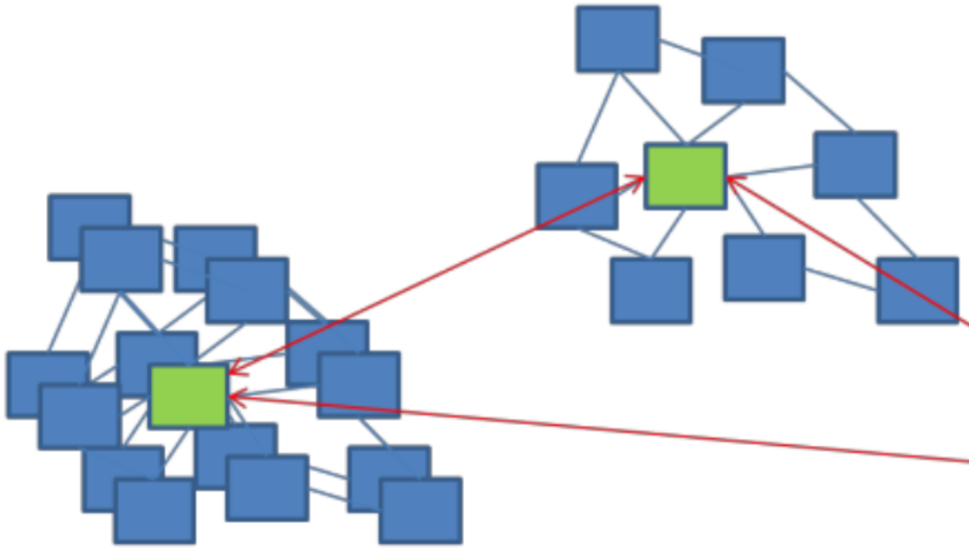


Retain geometrically valid clusters = at least n geometrically valid images

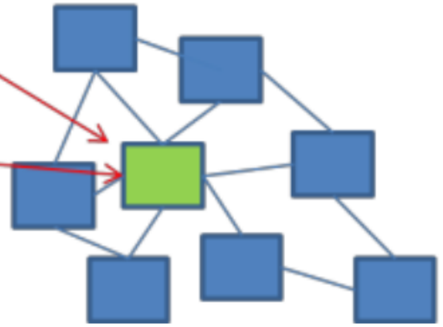


Select iconic image = image with the largest number of feature matches (SIFT)





Group iconics into 3D models



Dealing with Huge Numbers of Images

Seed cluster

Geometrically consistent cluster

COMPLETE SYSTEM

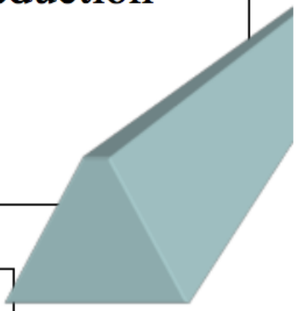
images + features

PROJECTIVE

Epipolar geometry: Fundamental matrix estimation (min. 2 images + 7 correspondences)

Linear eight-point + RANSAC $Min \sum_i (\mathbf{p}_i^T \mathbf{F} \mathbf{p}'_i)^2$ + rank-2 SVD reduction

Non-linear refinement $Min \sum_i d^2(\mathbf{p}_i, \mathbf{F}\mathbf{p}'_i) + d^2(\mathbf{p}'_i, \mathbf{F}^T\mathbf{p}_i)$



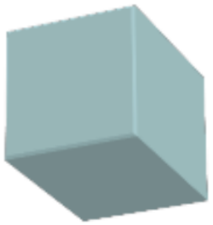
Projective reconstruction:

$$\mathbf{F} \rightarrow \mathbf{A} = [\mathbf{b}]_\times \mathbf{F} \quad \mathbf{F}^T\mathbf{b} = \mathbf{0} \rightarrow \quad \mathbf{M} = [\mathbf{A} \mid \mathbf{b}]$$

METRIC

Self-calibration (intrinsic parameter matrix **K**):

$$\mathbf{M}_i \mathbf{Q}_3 \mathbf{Q}_3^T \mathbf{M}_i^T \equiv \mathbf{K}_i \mathbf{K}_i^T$$



Metric reconstruction:

$$\mathbf{M}_i \mathbf{Q} = \mathbf{K}_i [\mathbf{R}_i \; \mathbf{t}_i] \quad \mathbf{P}_j \leftarrow \mathbf{Q}^{-1}\mathbf{P}_j$$

REFINEMENT

Bundle adjustment:

$$\underset{\mathbf{P}_j, \mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i}{Min} \sum_{i,j} \left(u_{ij} - \frac{\mathbf{m}_i^{1T}\mathbf{P}_j}{\mathbf{m}_i^{3T}\mathbf{P}_j} \right)^2 + \left(v_{ij} - \frac{\mathbf{m}_i^{2T}\mathbf{P}_j}{\mathbf{m}_i^{3T}\mathbf{P}_j} \right)^2$$

Building Rome on a Cloudless Day