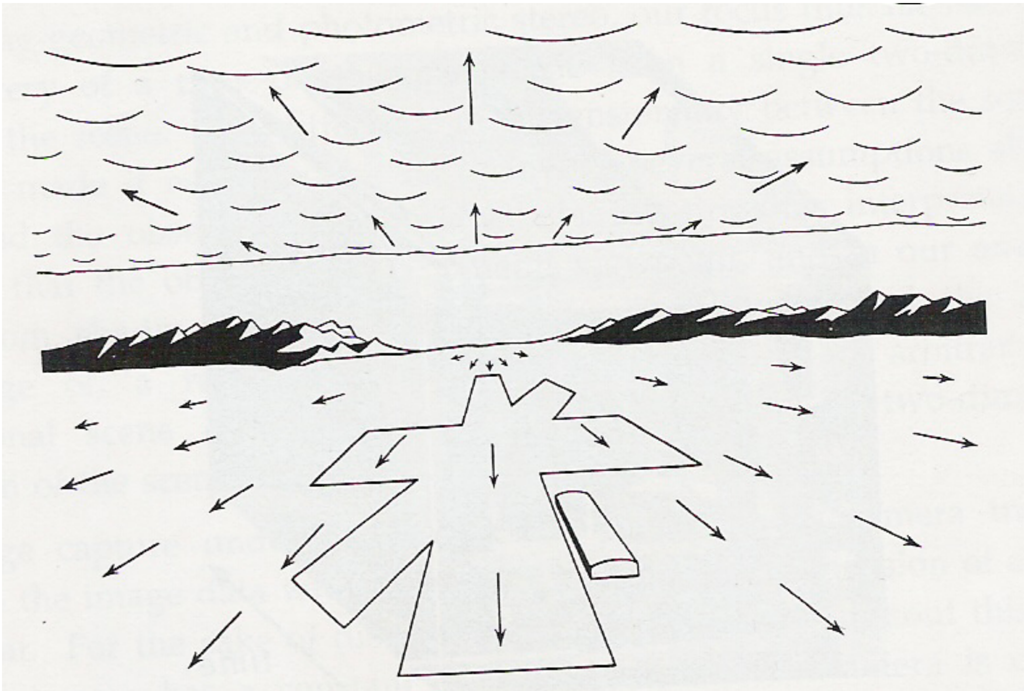
SHAPE FROM X

One image:
- Texture
- Shading

Two images or more:
- Stereo
- Contours
- **Motion**

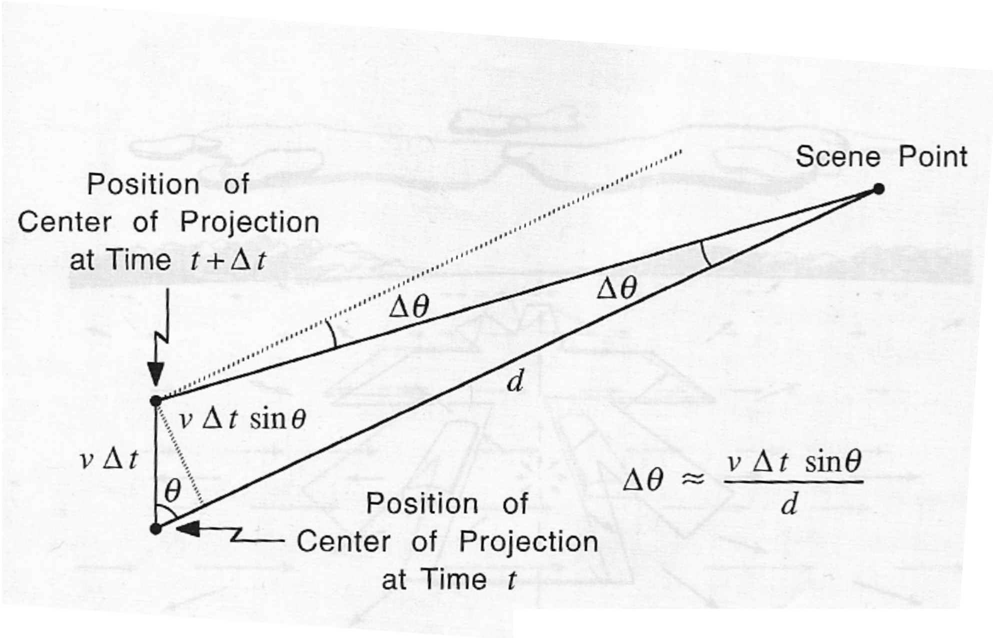
MOTION



When objects move at equal speed, those more remote seem to move more slowly.

Euclid, 300 BC

VELOCITY vs DISTANCE



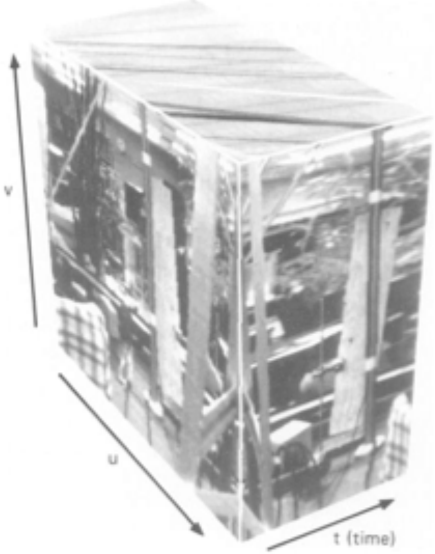
Velocity is:

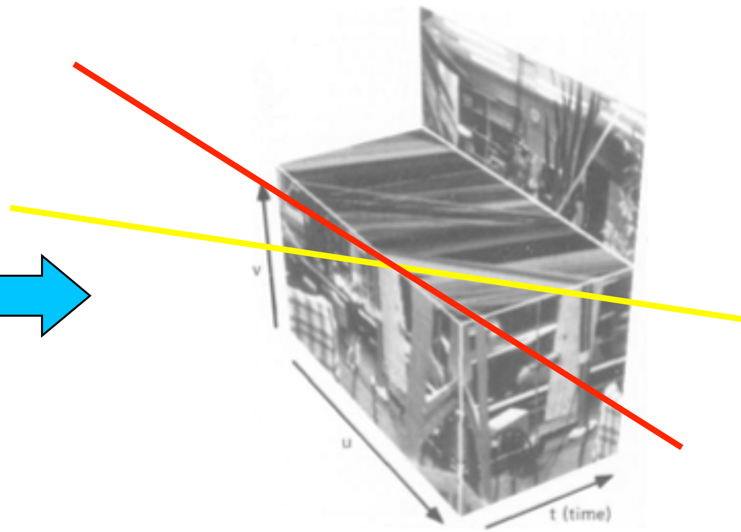
- Inversely proportional to the distance of the point to the observer.
- Proportional to the sine of the angle between the line of sight and the direction of translation.

EPIPOLAR PLANE ANALYSIS

Sequence:

Image cube:

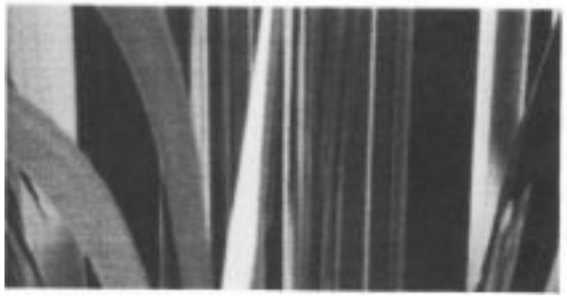




Bolles et al. , IJCV'87

GENERALIZED MOTION

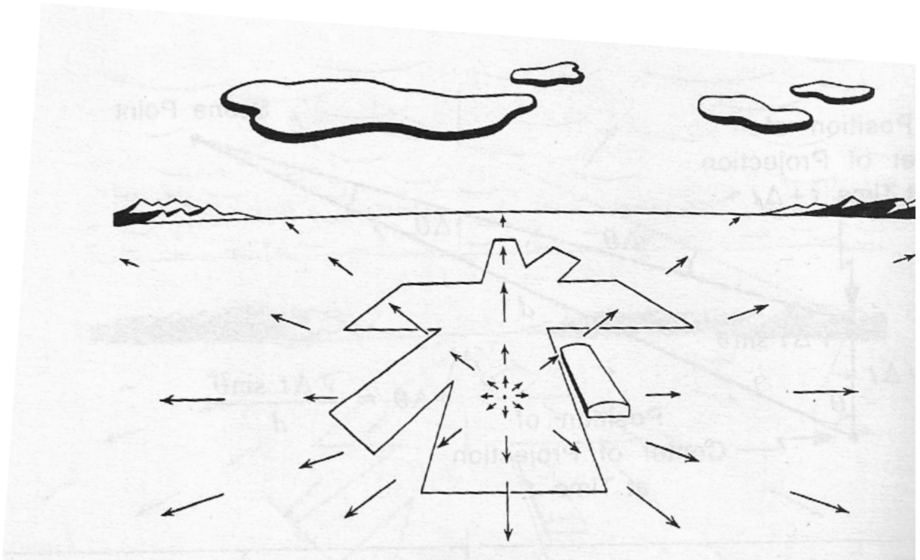
Orthogonal viewing



Non-orthogonal viewing

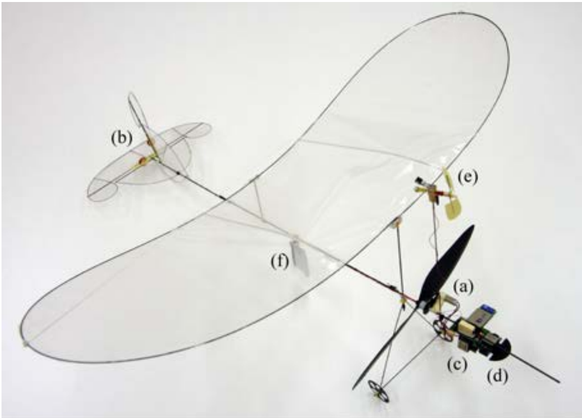
View direction varying

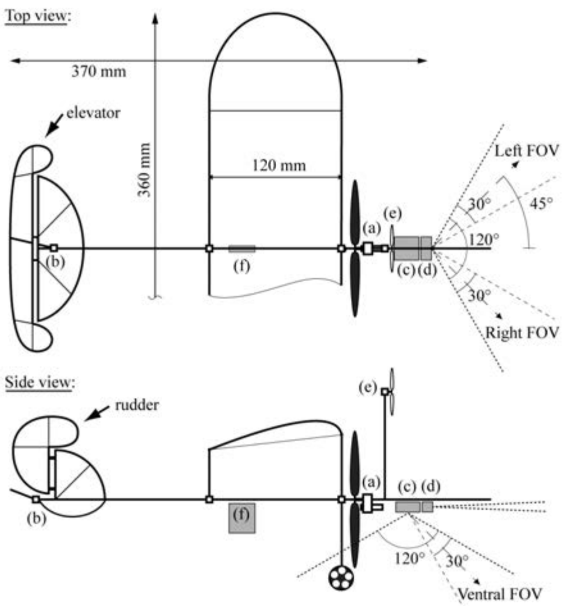
FOCUS OF EXPANSION

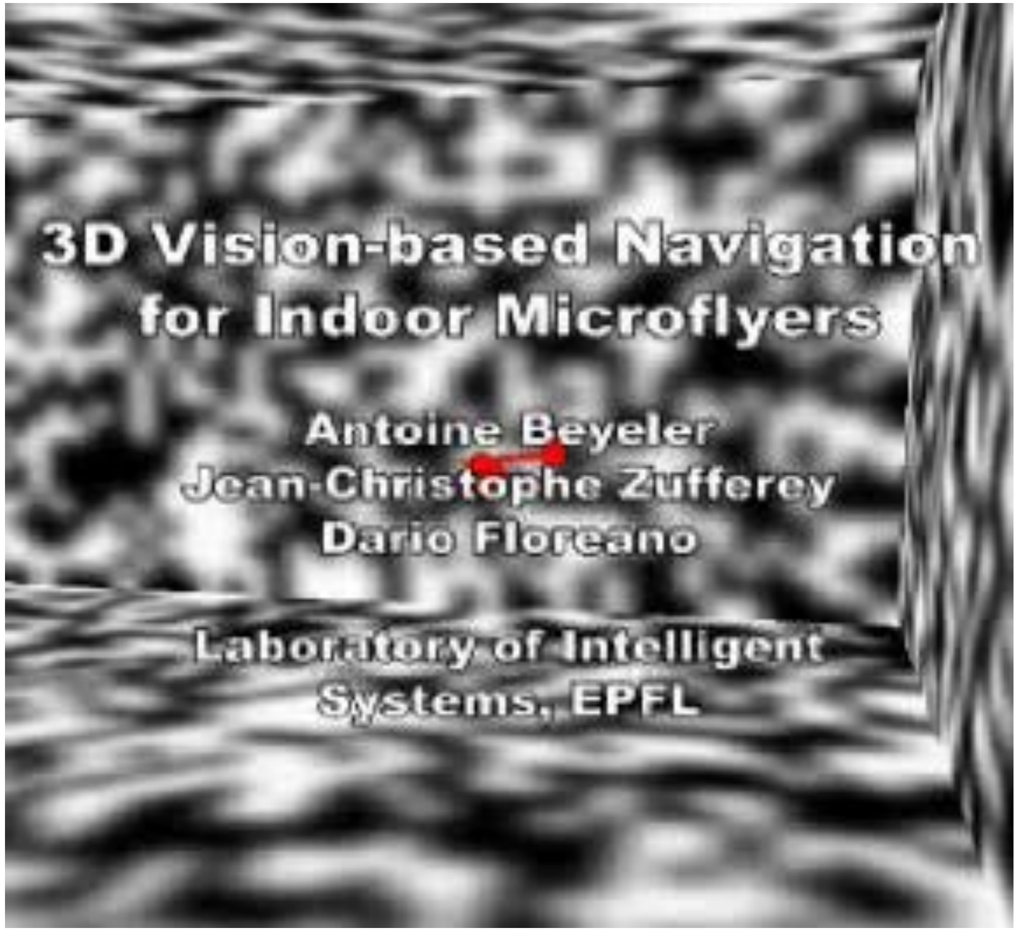


For a translational motion of the camera, all the **motion-field** vectors converge or diverge from a single point: the focus of expansion (FOE) or contraction (FOC).

MICROFLYER







Zufferey et al. , IJMAR 2010.

MOTION FIELD ESTIMATION

Approaches classified with respect to the assumptions they make about the scene:

- Images properties are preserved under relative motion between the camera and the scene.
- Feature points can be tracked across frames.

ASSUMPTION 1: BRIGHTNESS CONSTANCY

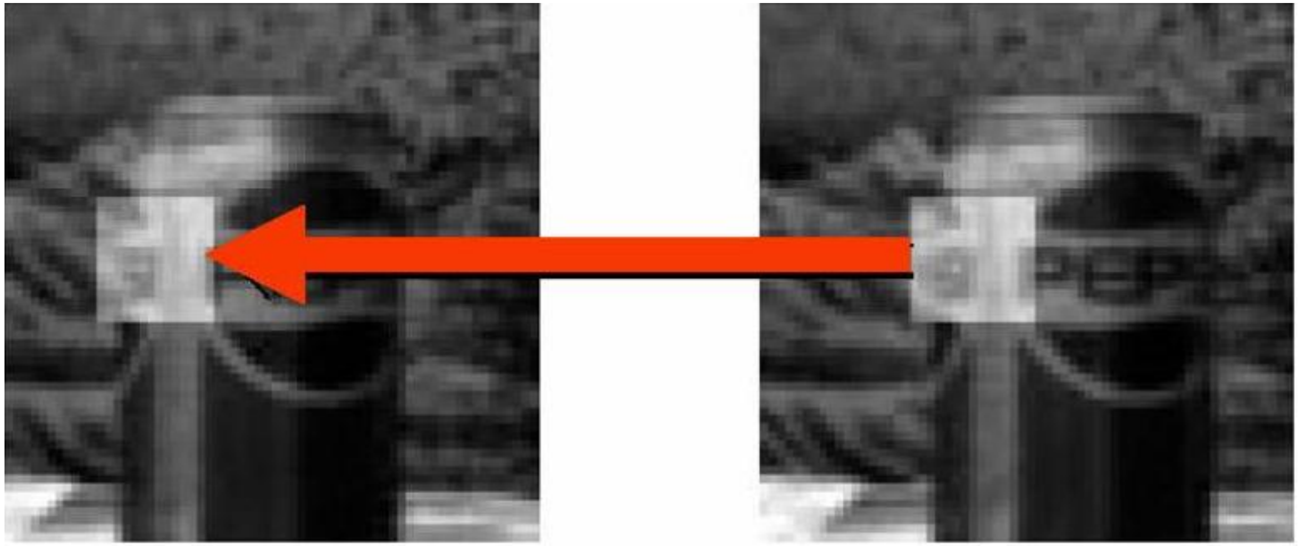
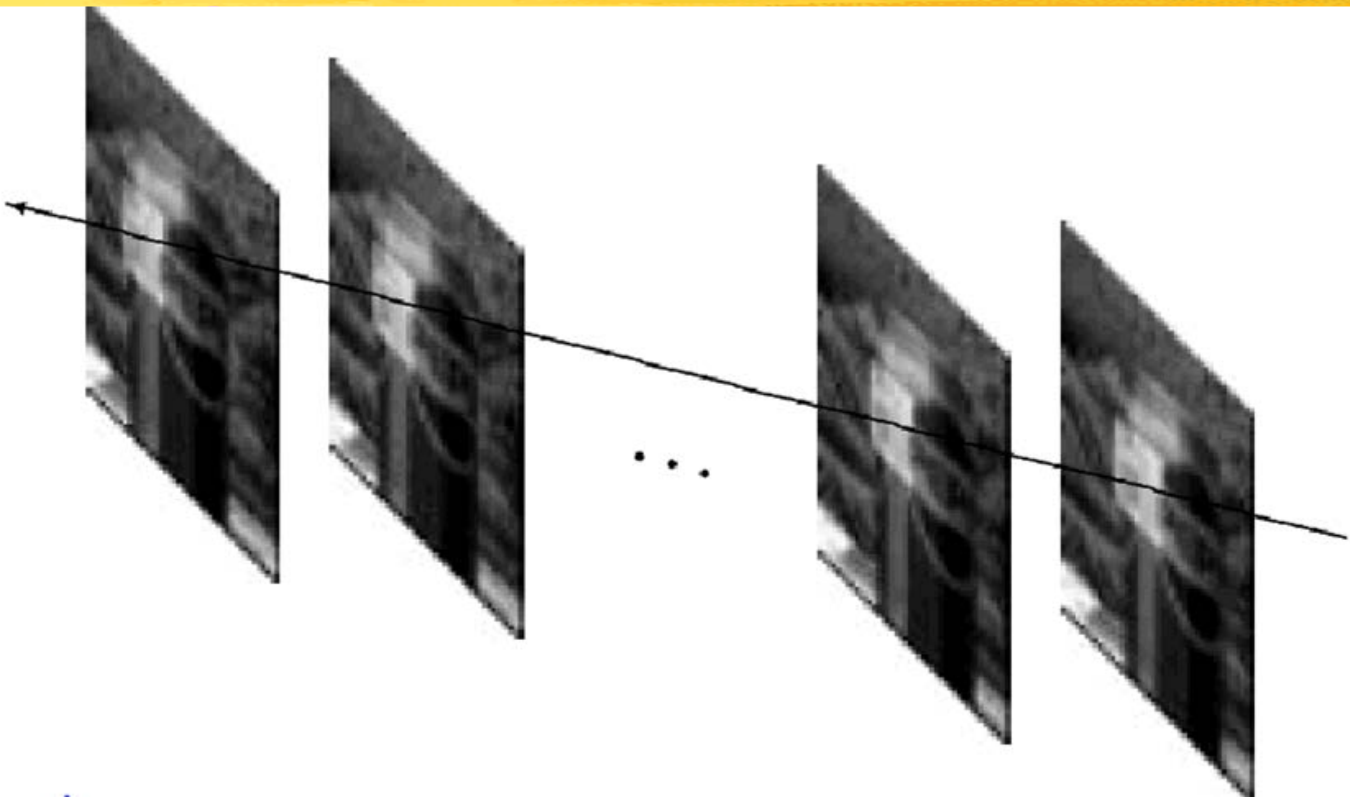


Image measurements (e.g. brightness) in a small region remain the same although its location may change.

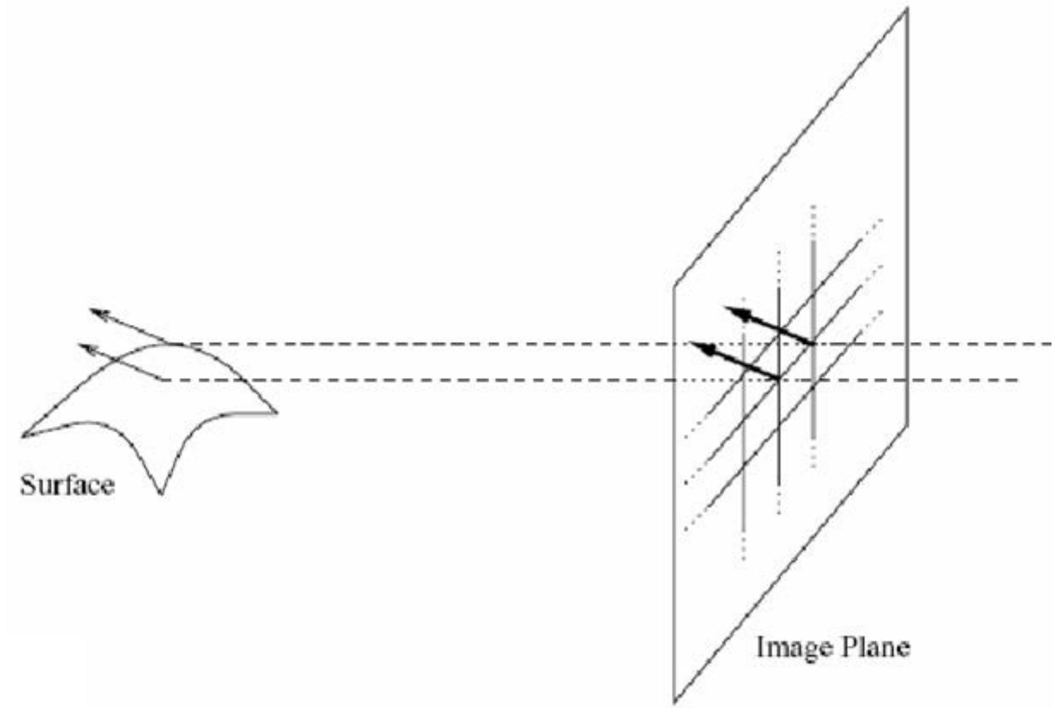
$$I(x+u, y+v, t) = I(x, y, t)$$

ASSUMPTION 2: TEMPORAL CONSISTENCY



The image motion of a surface patch changes gradually over time.

ASSUMPTION 3: SPATIAL COHERENCE



- Neighboring points in the scene typically belong to the same surface and hence have similar motions.
- Since they also project to nearby image locations, we expect spatial coherence of the flow.

SPATIO TEMPORAL DERIVATIVES

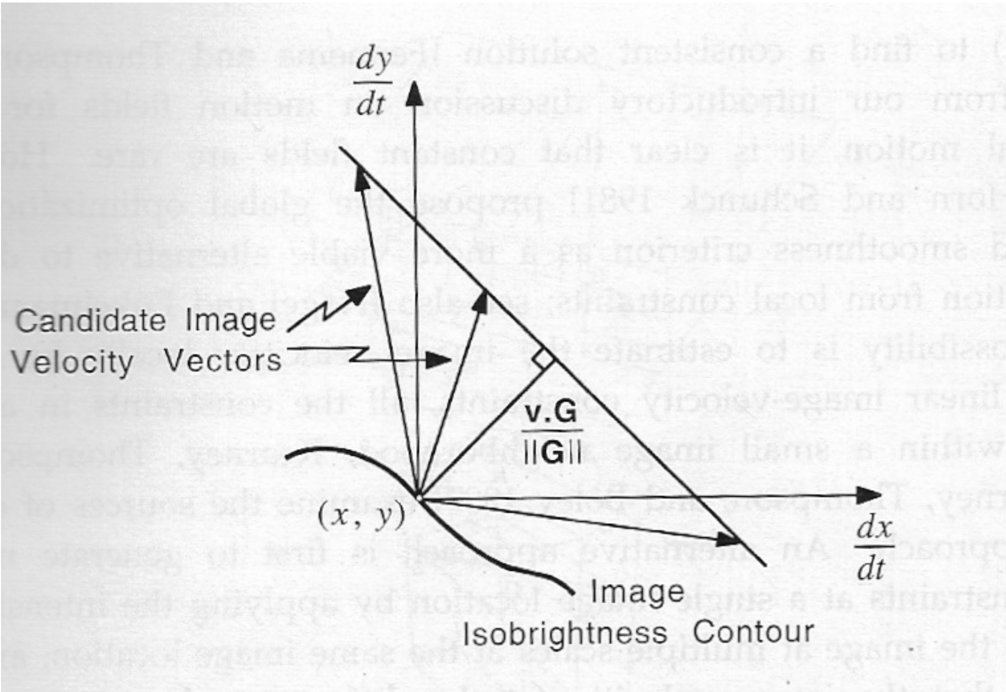
Under the assumptions of

- Brightness constancy,
- Temporal consistency,

we write: $cte = I(x(t), y(t), t)$

$$0 = \frac{dI}{dt} = \frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t}$$

NORMAL FLOW EQUATION



$$v\frac{G}{\|G\|} = -\frac{\dfrac{\partial I}{\partial t}}{\sqrt{\dfrac{\partial I}{\partial x}^2 + \dfrac{\partial I}{\partial y}^2}}$$

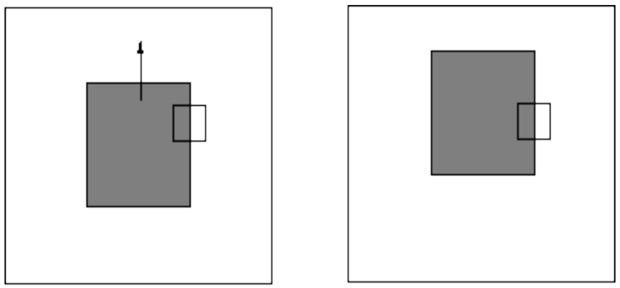
$$G = \left[\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}\right]$$

$$v = \left[\frac{dx}{dt}, \frac{dy}{dt}\right]$$

AMBIGUITIES

At each pixel, we have one equation and two unknowns.

--> Only the flow component in the gradient direction can be determined locally.



The motion is parallel to the edge, and it cannot be determined.

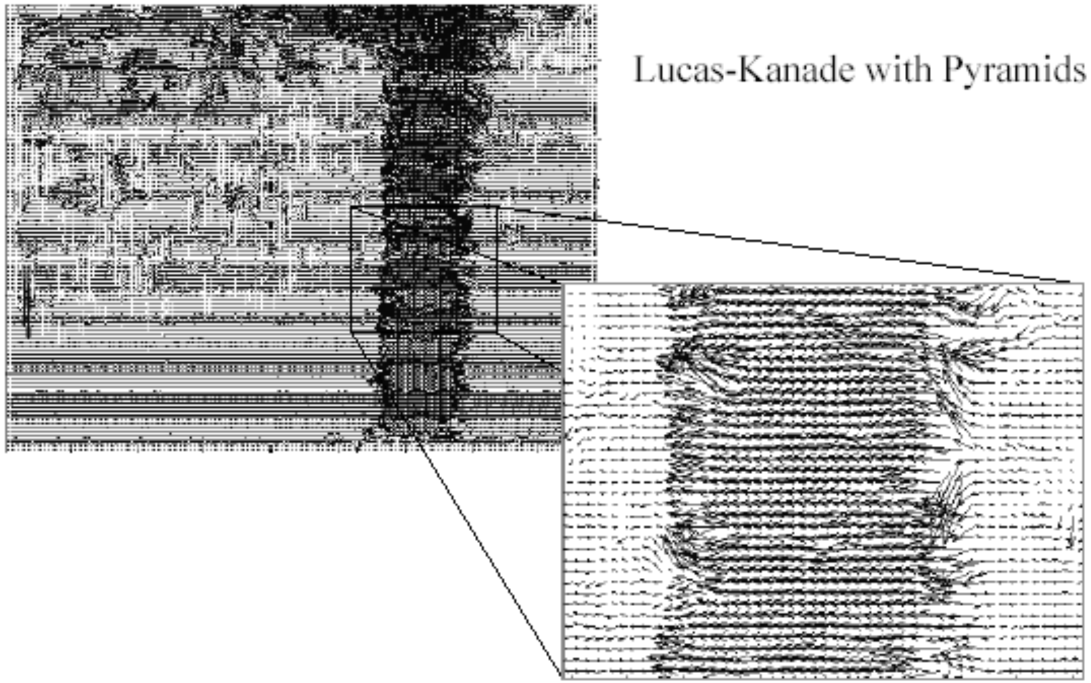
LOCAL CONSTANCY

Assume the flow to constant is a 5x5 window:

$$\begin{bmatrix} I_x(\mathbf{p_1}) & I_y(\mathbf{p_1}) \\ I_x(\mathbf{p_2}) & I_y(\mathbf{p_2}) \\ \vdots & \vdots \\ I_x(\mathbf{p_{25}}) & I_y(\mathbf{p_{25}}) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} I_t(\mathbf{p_1}) \\ I_t(\mathbf{p_2}) \\ \vdots \\ I_t(\mathbf{p_{25}}) \end{bmatrix}$$

--> 25 equations for 2 unknown, which can be solved in the least squares sense.

ENFORCING CONSISTENCY

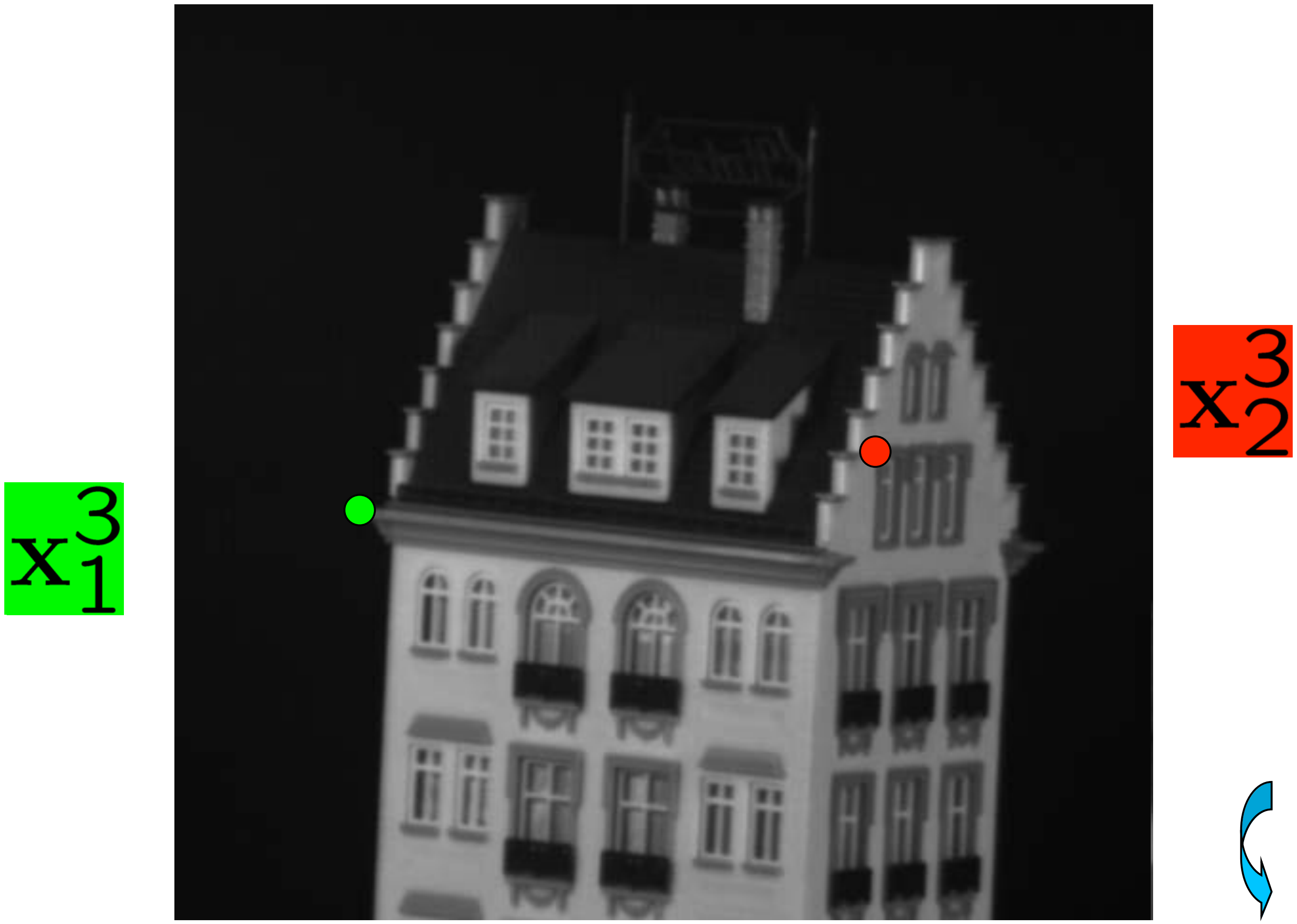


Under the assumption of spatial coherence:

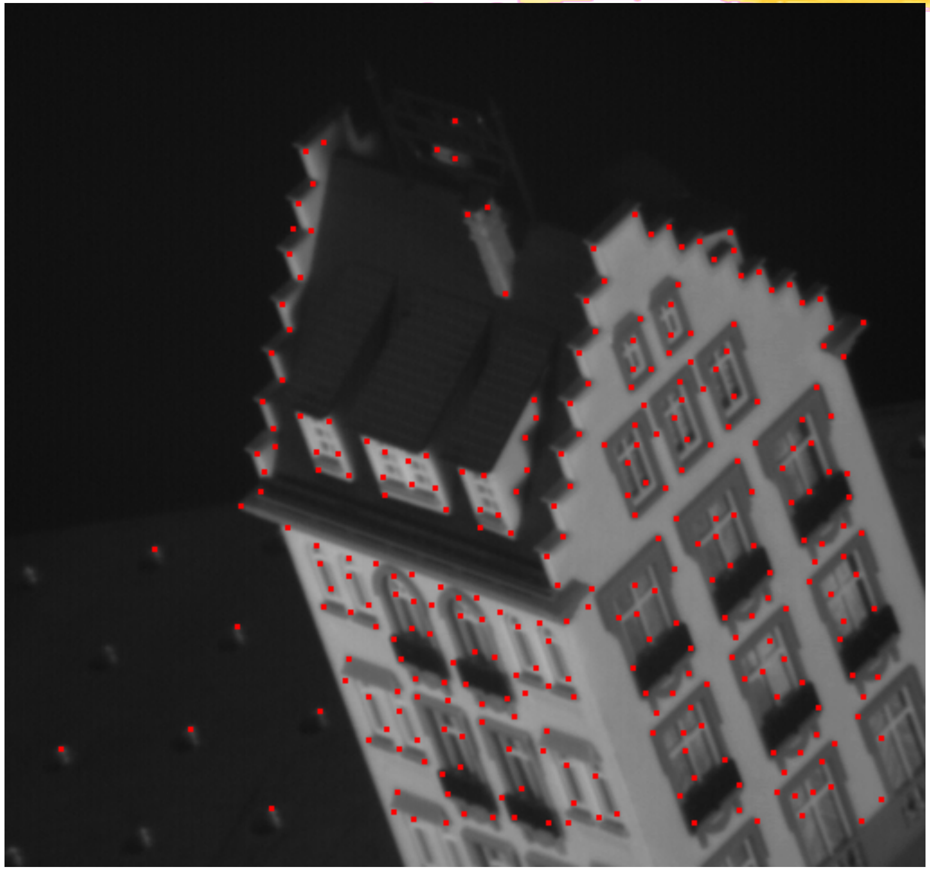
- Hough Transform on the motion vectors.
- Regularization of the motion field.
- Multi scale approach.

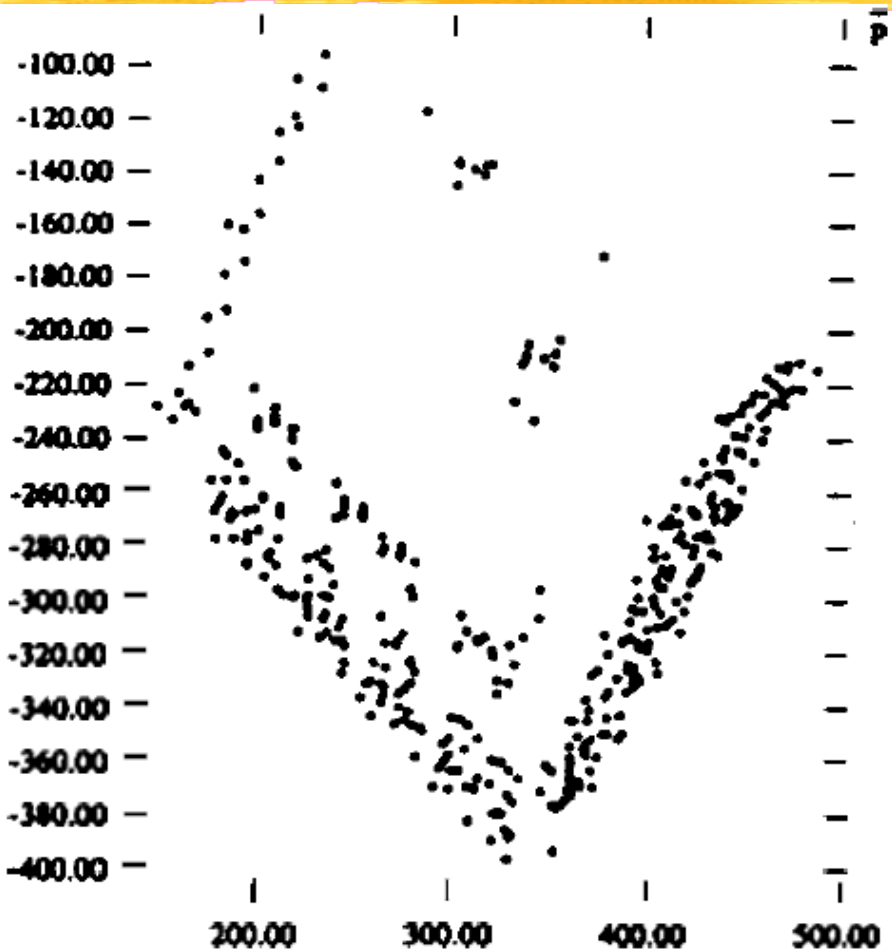
But, the world is neither Lambertian nor smooth → Assumptions rarely valid.

TRACKING POINTS ACROSS IMAGES



SHAPE RECONSTRUCTION





Factoring Image Sequences into Shape and Motion, C. Tomasi and T. Kanade, Proc. IEEE Workshop on Visual Motion (1991).

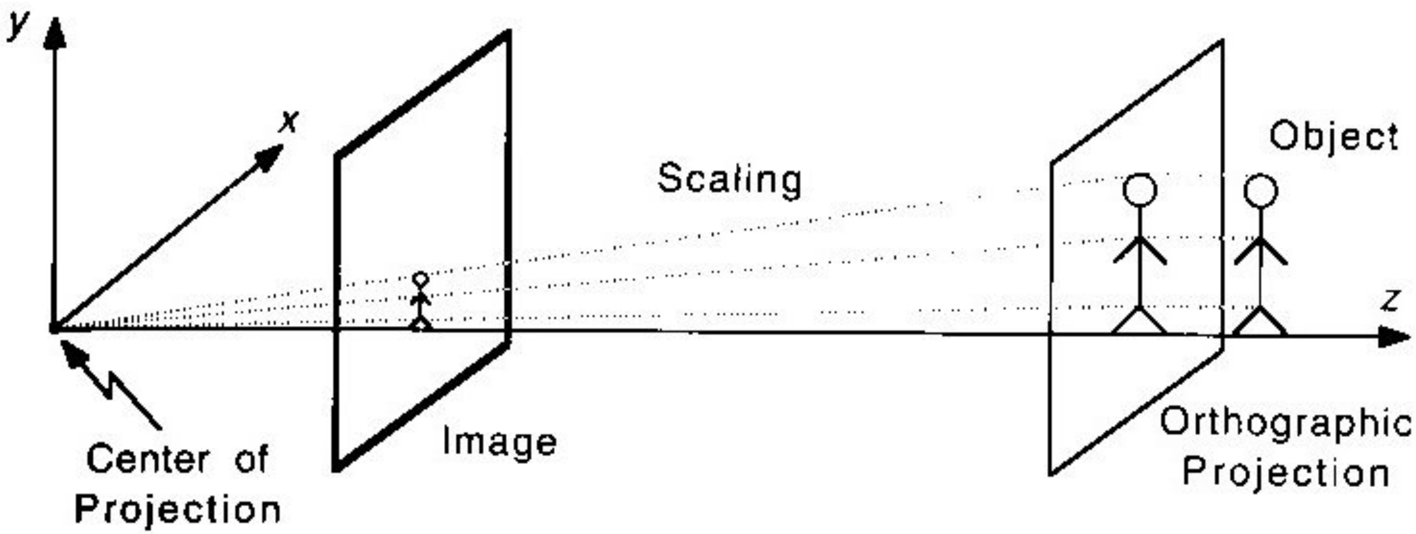
MULTI-VIEW PROJECTION

n image points are projected from 3-D scene points over m views via

$$\mathbf{x}_j^i = \mathbf{P}^i \mathbf{X}_j$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Here each $\mathbf{P}^i$ is a 3 x 4 matrix and each $\mathbf{X}_j$ is a homogeneous 4-vector

ORTHOGRAPHIC PROJECTION



$$u = sx$$

$$v = sy$$

Special case of perspective projection:

- Large f
- Objects close to the optical axis

→Parallel lines mapped into parallel lines.

MULTI-VIEW ORTHOGRAPHIC PROJECTION

The last row of each $\mathbf{P}^i$ is (0, 0, 0, 1) for affine cameras, so we can "ignore" it and write the orthographic projection as:

$$\mathbf{x}_j^i = \mathbf{M}^i \mathbf{X}_j + \mathbf{t}^i$$

where each $\mathbf{X}_j$ is now an inhomogeneous 3-vector, each $\mathbf{M}^i$ a 2 x 3 matrix, and each $\mathbf{t}^i$ a 2-vector.

RECONSTRUCTION PROBLEM

Estimate affine cameras $\mathbf{M}^i$, translations $\mathbf{t}^i$, and 3-D points $\mathbf{X}_j$ that minimize the geometric error in image coordinates:

$$\min_{\mathrm{M}^i,\mathbf{t}^i,\mathbf{X}_j} \sum_{i,j} \left(\mathbf{x}_j^i - (\mathrm{M}^i \mathbf{X}_j + \mathbf{t}^i) \right)^2$$

SIMPLIFYING THE PROBLEM

Normalization: We can eliminate the translation vectors $\mathbf{t}^i$ by choosing the centroid of the image points in each image as the coordinate system origin

$$\mathbf{x}_j^i \leftarrow \mathbf{x}_j^i - \frac{1}{n}\sum_j \mathbf{x}_j^i$$

Working in "centered coordinates", the minimization problem becomes:

$$\min_{\mathbf{M}^i, \mathbf{X}_j} \sum_{i,j} \left(\mathbf{x}_j^i - \mathbf{M}^i \mathbf{X}_j\right)^2$$

This works because the centroid of the 3-D points is preserved under affine transformations

MATRIX FORMULATION

Let the *measurement* matrix be:

$$\mathbf{W} = \begin{pmatrix} \mathbf{x}_1^1 & \mathbf{x}_2^1 & \cdots & \mathbf{x}_n^1 \\ \mathbf{x}_1^2 & \mathbf{x}_2^2 & \cdots & \mathbf{x}_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1^m & \mathbf{x}_2^m & \cdots & \mathbf{x}_n^m \end{pmatrix}$$

Since $\mathbf{x}_j^i = \mathrm{M}^i \mathbf{X}_j$, minimizing means solving:

$$\mathbf{W} = \underbrace{\begin{bmatrix} \mathbf{M}^1 \\ \vdots \\ \mathbf{M}^m \end{bmatrix}}_{2m \times 3} \underbrace{[\mathbf{X}_1, \ldots, \mathbf{X}_n]}_{3 \times n}$$

SOLVING WITH SVD

- There will be no exact solution with noisy points, so we want the nearest **W'** to **W** that is an exact solution
  - **W'** is rank 3 since it's the product of a 2m x 3 motion matrix **M'** and a 3 x n structure matrix **X'**
- Use singular value decomposition to get rank 3 matrix **W'** closest to **W**
  - Let SVD of $\mathbf{W} = \mathbf{UDV}^T$
  - Then $\mathbf{W}' = \mathbf{U}_{2m\times3}\mathbf{D}_{3\times3}\mathbf{V}_{n\times3}^T$, where $\mathbf{U}_{2m\times3}$ is the first 3 columns of **U**, $\mathbf{D}_{3\times3}$ is an upper-left 3 x 3 submatrix of **D**, and $\mathbf{V}_{n\times3}^T$ is first three columns of **V**.

STRUCTURE AND MOTION

Set stacked camera matrix as

$$\mathbf{M}' = \mathbf{U}_{2m\text{x}3}\,\text{sqrt}(\mathbf{D}_{3\text{x}3})$$

and stacked 3-D structure matrix as

$$\mathbf{X}' = \text{sqrt}(\mathbf{D}_{3\text{x}3})\mathbf{V}_{n\text{x}3}{}^{T}$$

so that $\mathbf{W}' = \mathbf{M}'\mathbf{X}'$

METRIC UPGRADE

There is an affine ambiguity since an arbitrary 3 x 3 rank 3 matrix **A** can be inserted as:

$$\mathbf{W'} = (\mathbf{M'A})(\mathbf{A}^{-1}\mathbf{X'})$$

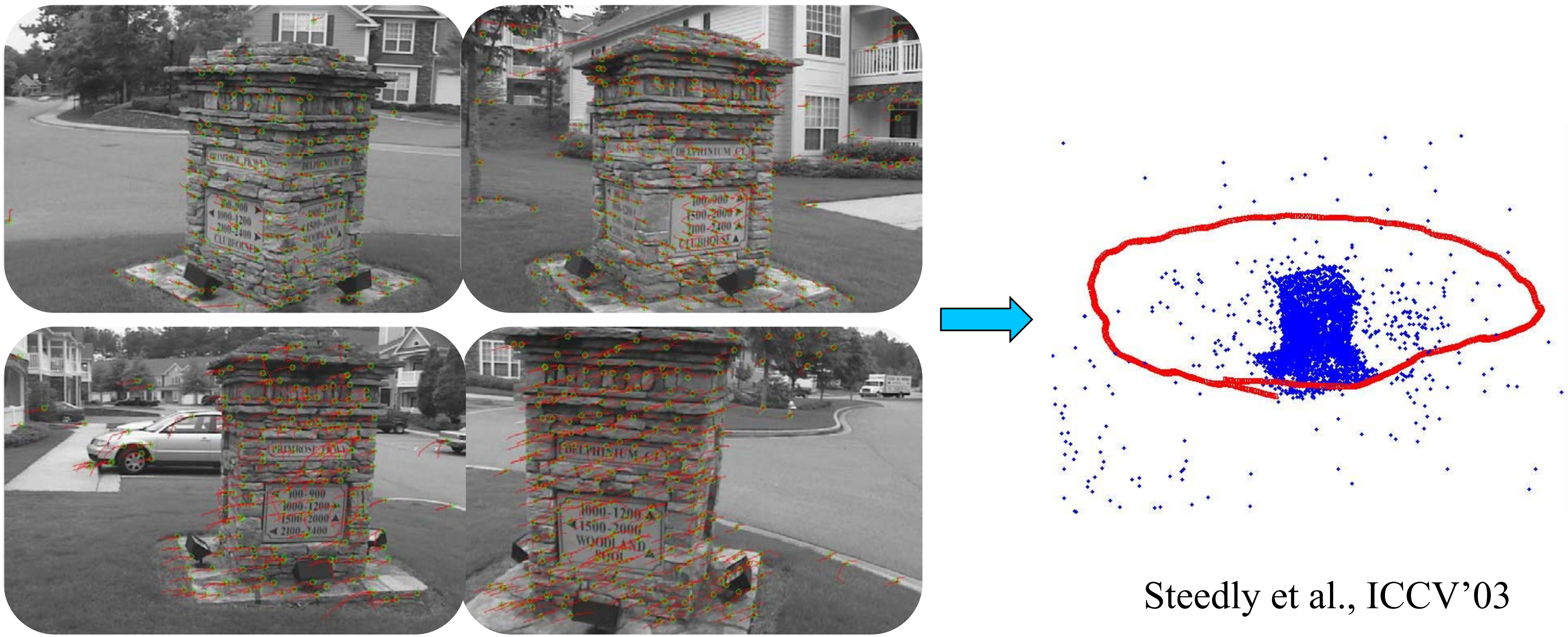
Get rid of ambiguity by finding **A** that performs "metric rectification"

Affine camera provides orthonormality constraints on **A**:

Rows of **M=M'A** are unit vectors: $\mathbf{m}_i \cdot \mathbf{m}_i = 1$.

Rows of **M=M'A** are orthogonal: $\mathbf{m}_i \cdot \mathbf{m}_j = 0$.

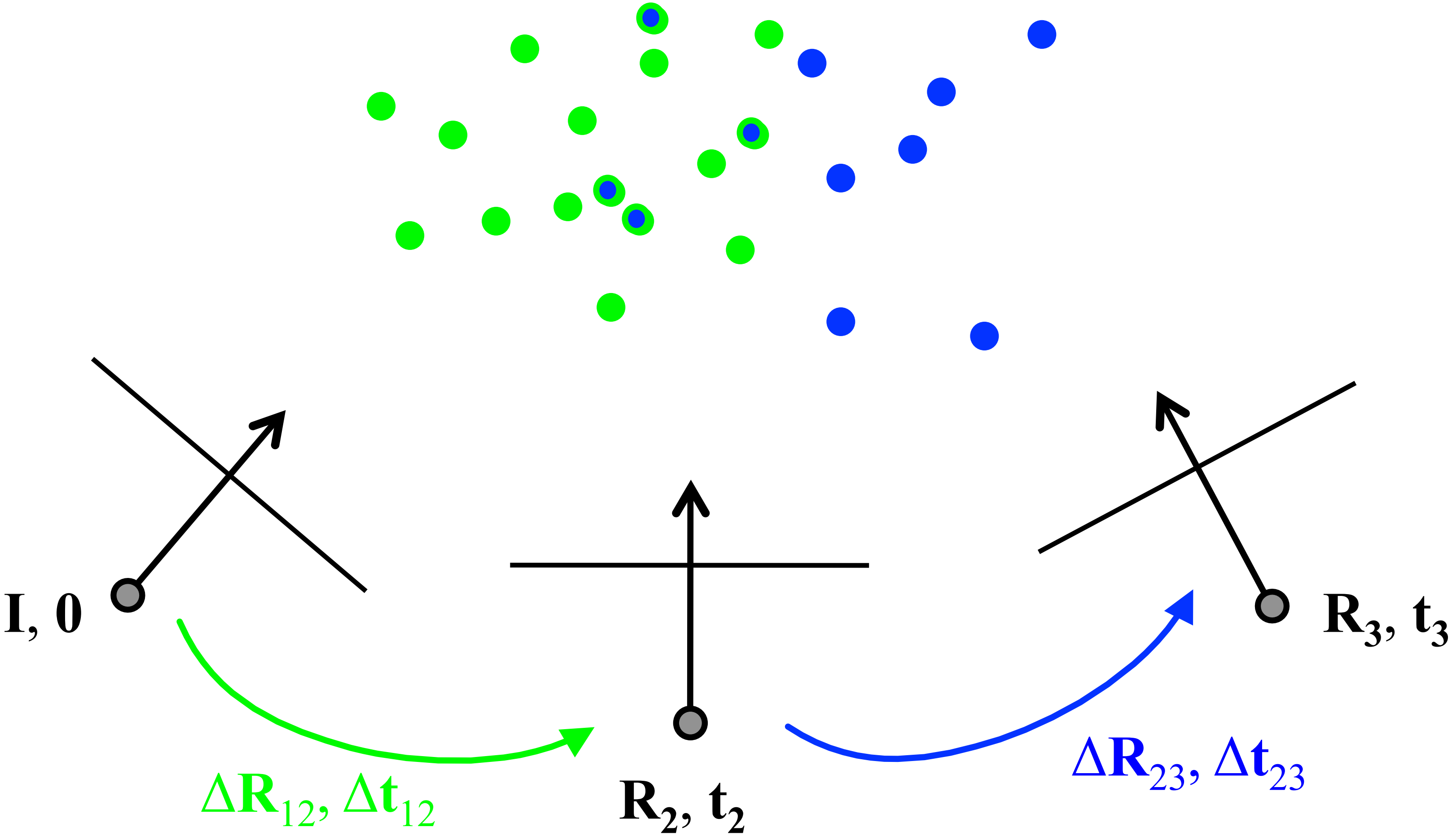
SIMULTANEOUS LOCALIZATION AND MAPPING



Steedly et al., ICCV'03

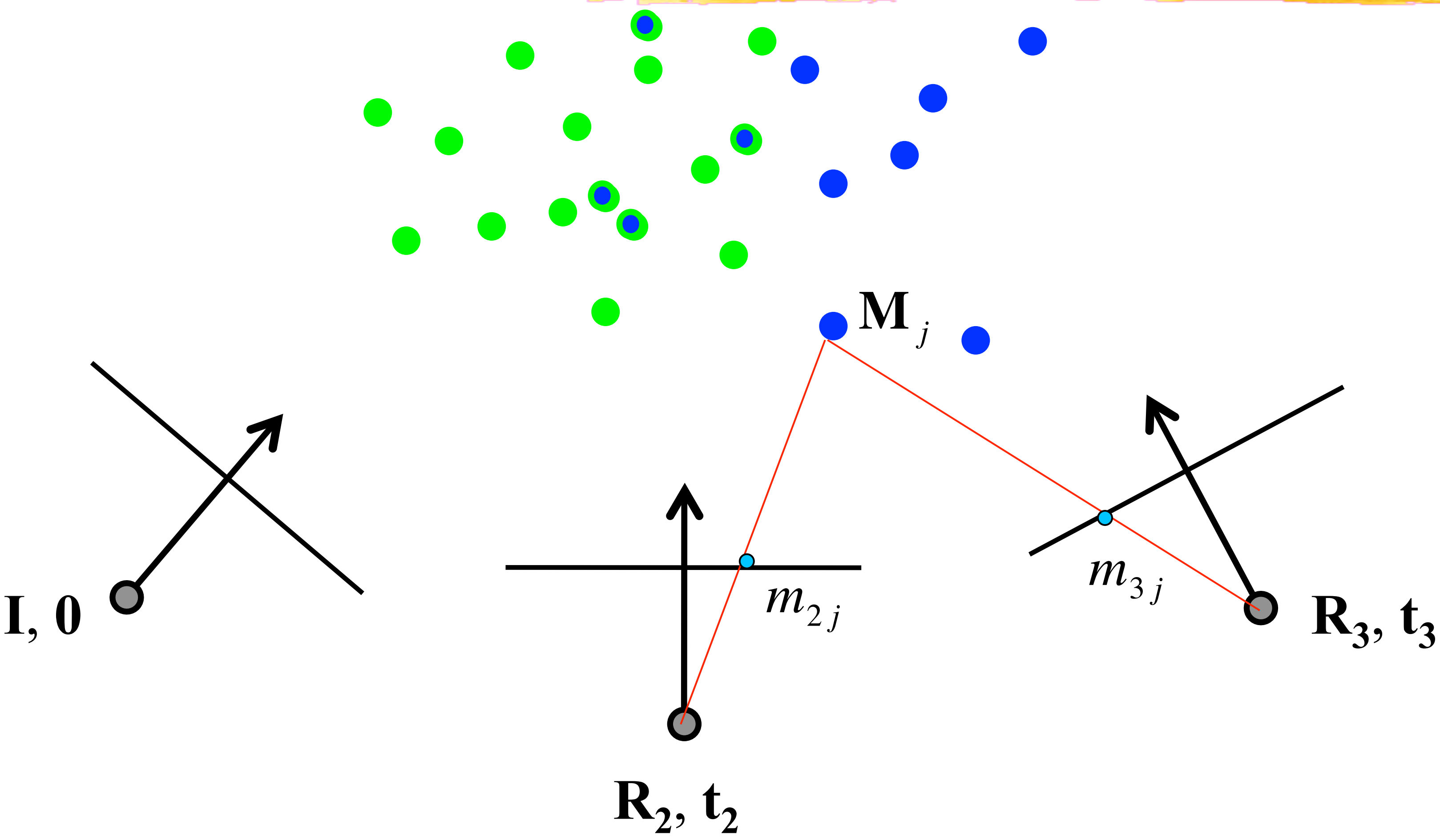
- Compute point tracks.
- Infer both camera motion and 3D structure.

SEQUENTIAL STRUCTURE FROM MOTION



-> Trajectory and 3D points defined up to a Euclidean motion and scale

BUNDLE ADJUSTMENT



$$\underset{\{\mathbf{R}_i, \mathbf{t}_i, \mathbf{M}_j\}}{\arg\min} \sum_i \sum_j \left\| \operatorname{proj}\left(\mathbf{R}_i, \mathbf{t}_i, \mathbf{M}_j\right) - \mathbf{m}_{ij} \right\|^2$$

GLOBAL OPTIMIZATION

$$\underset{\{\mathbf{R}_i,\mathbf{t}_i,\mathbf{M}_j\}}{\arg\min} \sum_i \sum_j \left\| \operatorname{proj}\left(\mathbf{R}_i,\mathbf{t}_i,\mathbf{M}_j\right) - \mathbf{m}_{ij} \right\|^2$$

- Often performed using the Levenberg-Marquardt algorithm.
- Many parameters to estimate, but sparse Jacobian matrix.
- Initial estimates computed using the eight point algorithm:

Given 8 point correspondences between a pair of images, ΔR and Δt can be estimated in closed form by solving an SVD.

AUGMENTED REALITY

Parallel Tracking and Mapping
for Small AR Workspaces

Extra video results made for
ISMAR 2007 conference

Georg Klein and David Murray
Active Vision Laboratory
University of Oxford

STRENGTHS AND LIMITATIONS

Strengths:

- Combine information from many images.

Limitations:

- Requires multiple views.
- Involves strong assumptions.