Shape from X

- One image:
  - Texture
  - Shading
- Two images or more:
  - Stereo
  - Contours
  - **Motion**

[Image of a forest scene]
Motion

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

Euclid, 300 BC
Velocity vs Distance

Apparent 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.

\[ \Delta \theta = \frac{v \Delta t \sin \theta}{d} \]
Epipolar Plane Analysis

Image sequence

Image cube

Bolles et al., IJCV’87
Generalized Motion

Orthogonal viewing

Non-orthogonal viewing

View direction varying
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).
The plane detects POEs and uses them to avoid collisions.
Approaches can be classified with respect to the assumptions they make about the scene:

- Images properties remain invariant 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.

\[ I(x + dx, y + dy, t + dt) = I(x, y, t) \]
Assumption 2: Temporal Consistency

The image speed of a surface patch only changes gradually over time.
• 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:

\[
\text{cst } = I(x(t), y(t), t)
\]

\[
\Rightarrow 0 = \frac{\delta I}{\delta x} \frac{dx}{dt} + \frac{\delta I}{\delta y} \frac{dy}{dt} + \frac{\delta I}{\delta t}
\]
Normal Flow Equation

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

\[ G = \begin{bmatrix} \frac{\partial I}{\partial x} & \frac{\partial I}{\partial y} \end{bmatrix} \]

\[ v = \begin{bmatrix} \frac{dx}{dt} & \frac{dy}{dt} \end{bmatrix} \]
Ambiguities

- At each pixel, we have 1 equation and 2 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 be constant is a 5x5 window:

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

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

Under the assumption of spatial consistency:
• Hough Transform on the motion vectors.
• Regularization of the motion field.
• Multi scale approach.

But, the world is neither Lambertian nor smooth.

→ These assumptions are rarely valid.
Tracking Points across Images
3D Shape Reconstruction

Multi-View Projection

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

$$x_j^i = P^i X_j$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, n$.

- Here each $P^i$ is a $3 \times 4$ matrix and each $X_j$ is a homogeneous 4-vector.
Orthographic Projection

Special case of perspective projection:

- Large $f$
- Objects close to the optical axis

- Parallel lines mapped into parallel lines.

$u = sx$

$v = sy$
• The last row of each $P^i$ is $(0, 0, 0, 1)$ for affine cameras, so we can “ignore” it and write the orthographic projection as:

$$x^i_j = M^i X_j + t^i$$

where each $X_j$ is now an inhomogeneous 3-vector.

• Here, each $M^i$ a $2 \times 3$ matrix, and each $t^i$ a 2-vector.
Reconstruction Problem

• Estimate affine cameras $M^i$, translations $t^i$, and 3-D points $X_j$ that minimize the geometric error in image coordinates:

$$\min_{M^i, t^i, X_j} \sum_{i,j} \left( x^i_j - (M^i X_j + t^i) \right)^2$$
Simplifying the Problem

• Normalization: We can eliminate the translation vectors \( t_i \) by choosing the centroid of the image points in each image as the coordinate system origin

\[
    x_{j}^{i} \leftarrow x_{j}^{i} - \frac{1}{n} \sum_{j} x_{j}^{i}
\]

• Working in “centered coordinates”, the minimization problem becomes:

\[
    \min_{M^i, X_j} \sum_{i,j} \left( x_{j}^{i} - M^i X_j \right)^2
\]

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

- Let the measurement matrix be:

\[ W = \begin{pmatrix}
  x_1^1 & x_2^1 & \ldots & x_n^1 \\
  x_1^2 & x_2^2 & \ldots & x_n^2 \\
  \vdots & \vdots & \ddots & \vdots \\
  x_1^m & x_2^m & \ldots & x_n^m
\end{pmatrix} \]

- Since \( x_j^i = M^i X_j \), this means solving

\[ W = \begin{bmatrix}
  M^1 \\
  \vdots \\
  M^m
\end{bmatrix} \begin{bmatrix}
  X_1, \ldots, X_n
\end{bmatrix} \]

in the least squares sense.
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 $2m \times 3$ motion matrix $M'$ and a $3 \times n$ structure matrix $X'$

• Use singular value decomposition to get rank 3 matrix $W'$ closest to $W$

  Let SVD of $W = UDV^T$

  Then $W' = U_{2mx3} D_{3x3} V_{nx3}^T$, where

  $U_{2mx3}$ is the first 3 columns of $U$, $D_{3x3}$ is an upper-left 3 x 3 submatrix of $D$,

  $V_{nx3}^T$ is first three columns of $V$. 

• Set stacked camera matrix as

\[ M' = U_{2mx3} \sqrt{D_{3x3}} \]

• Set stacked 3-D structure matrix as

\[ X' = \sqrt{D_{3x3}} V_{nx3}^T \]

so that \[ W' = M'X' \]
Metric Upgrade

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

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

• 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: $m_i . m_i = 1$.
  – Rows of $M=M'A$ are orthogonal: $m_i . m_j = 0$.

• Everything relies on linear algebra but is limited to orthographic cameras.
Simultaneous Localization And Mapping

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

Steedly et al., ICCV’03
Sequential Structure from Motion

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

argmin_{R_i, t_i, M_j} \sum_i \sum_j \|\text{proj}(R_i, t_i, M_j) - m_j^i\|^2
Global Non-Linear Optimization

\[ \arg\min_{R_i, t_i, M_j} \sum_i \sum_j \| \text{proj}(R_i, t_i, M_j) - m^i_j \|^2 \]

- 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, \( \Delta R \) and \( \Delta T \) 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
Simultaneous Localization And Mapping

A robot can reconstruct its environment and position itself at the same time.

Engel et al., ECCV’14
Fusing Depth Maps

- Both the depth camera and the person are moving.
- Use a deformable model to combine the data over time.
- Real-time implementation.

Newcombe et al., CVPR’15
Facebook buys British virtual reality start-up Surreal Vision
Surreal Vision aims to make a computerised version of the world so real that users are unable to distinguish between the two.
Into the Commercial World

Microsoft Hololens

... and they are both being worked on in Zurich!
Strengths And Limitations

Strengths:
• Combine information from many images.

Limitations:
• Requires multiple views.
• Requires either texture or a depth camera.