3D Image Analysis

Extraction of 3D information from an image (sequence) is important for
- vision in general (= scene reconstruction)
- many tasks (e.g. robot grasping and navigation, traffic analysis)
- not all tasks (e.g. image retrieval, quality control, monitoring)

Basic problem:
Projection from 3D to 2D causes loss of depth information!

Recovery of 3D information is possible
• by multiple cameras (e.g. binocular stereo)
• by a monocular image sequence with motion + weak assumptions
• by a single image + strong assumptions or prior knowledge about
  the scene

Single image 3D analysis

Humans exploit various cues for a tentative (heuristic) depth analysis:

- size of known objects
- texture gradient
- occlusion
- colour intensities
- angle of observation
- continuity assumption
- generality assumption
Texture gradient

Assume that texture does not mimic projective effects

Interpret texture gradient as a 3D projection effect
(Witkin 81)

Shape from texture

Assume
- homogeneous texture on 3D surface and
- 3D surface continuity

Reconstruct 3D shape from perspective texture variations
(Barrow and Tenenbaum 81)
Surface shape from contour

Assume "non-special" illumination and surface properties

3D surface shape maximizes probability of observed contours and minimizes probability of additional contours

2D image contour

possible 3D reconstructions

a b c

3D line shape from 2D projections

Assume that lines connected in 2D are also connected in 3D

Reconstruct 3D line shape by minimizing spatial curvature and torsion

2D collinear lines are also 3D collinear
3D shape from multiple lines

Assume that similar line shapes result from similar surface shapes.

Parallel lines lie locally on a cylinder

(Stevens 81)

3D junction interpretation

rules for junctions of curved lines

(Binford 81)

a not behind b

rules for blocks-world junctions

(Waltz 86)

a, b and c meet

"general" ensemble

"special" ensemble

Kantenmarkierung
From the laws of perspective projection:
The projections of 3D parallel straight lines intersect in a single point, the vanishing point.

Assume that more than 2 straight lines do not intersect in a single point by coincidence.

If more than 2 straight lines intersect, assume that they are parallel in 3D.

Generality principle

Assume that
- observer position
- illumination
- surface properties
are general, i.e. without special coincidences.

Choose interpretation based on generality assumption.
Principle of 3D motion analysis

2D displacements of points observed on an unknown moving rigid body may provide information about
- the 3D structure of the points
- the 3D motion parameters

Cases of interest:
- stationary camera, moving object(s)
- moving camera, stationary object(s)
- moving camera, moving object(s)

3D analysis of point displacements

- relative motion of one rigid object and one camera
- observation of P points in M views

For each point \( v_p \) in 2 consecutive images we have:
\[
\begin{align*}
v_{p,m+1} &= R_m v_{pm} + t_m & \text{motion equation} \\
v_{pm} &= \lambda_{pm} v_{pm} & \text{projection equation}
\end{align*}
\]

For P points in M images we have
- 3MP unknown 3D point coordinates \( v_{pm} \)
- 6(M-1) unknown motion parameters \( R_m \) and \( t_m \)
- MP unknown projection parameters \( \lambda_{pm} \)
- 3(M-1)P motion equations
- 3MP projection equations
- 1 arbitrary scaling parameter

\[\# \text{equations} \geq \# \text{unknowns} \Rightarrow P \geq 3 + \frac{2}{2M - 3} \Rightarrow\]
Essential matrix

Geometrical constraints derived from 2 views of a point in motion

- motion between image \( m \) and \( m+1 \) may be decomposed into
  1) rotation \( R_m \) about origin of coordinate system (= optical center)
  2) translation \( t_m \)
- observations are given by direction vectors \( n_m \) and \( n_{m+1} \) along projection rays

\[ R_m n_m, t_m \text{ and } n_m \text{ are coplanar: } [ n_m \times R_m n_m]^T n_{m+1} = 0 \]

After some manipulation:

\[ n_m^T E_m n_{m+1} = 0 \]

with \( E_m = \begin{bmatrix} l_m x_1 & l_m x_2 & l_m x_3 \\ n_p x_1 & n_p x_2 & n_p x_3 \end{bmatrix} \) and \( R_m = \begin{bmatrix} f_1 & f_2 & f_3 \end{bmatrix} \)

Nagel-Neumann constraint

Consider 2 views of 3 points \( v_{pm}, \) \( p = 1..3, m = 1,2 \)

The planes through \( R_m n_{pm} \) and \( v_{p,m+1} \) all intersect in \( t_m \)

\[ \text{=> the normals of the planes are coplanar} \]

Coplanarity condition for 3 vectors \( a, b, c \): \( (a \times b)^T c = 0 \)

\[ ([R_m n_{1m} \times v_{1,m+1}] \times [R_m n_{2m} \times v_{2,m+1}])^T [R_m n_{3m} \times v_{3,m+1}] = 0 \]

Nonlinear equation with 3 unknown rotation parameters.

\[ \text{=> Observation of at least 5 points required to solve for the unknowns.} \]
Homogeneous coordinates

- (N+1)-dimensional notation for points in N-dimensional Euclidean space
- allows to express projection and translation as linear operations

Normal coordinates: \( y^T = [x \ y \ z] \)
Homogeneous coordinates: \( v^T = [wx \ wy \ wz \ w] \)
\( w \neq 0 \) is arbitrary constant

Rotation and translation in homogeneous coordinates:
\[ v' = Av \] with 
\[ A = \begin{bmatrix} R & I \\ 0 & 1 \end{bmatrix} \]

Projection in homogeneous coordinates:
\[ v' = Bv \] with 
\[ B = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \]

Divide the first N components by the (N+1)rst component to recover normal coordinates

From homogeneous world coordinates to homogeneous image coordinates

\( x, y, z = \) scene coordinates
\( x_p, y_p = \) image coordinates
\[
\begin{bmatrix} wx_p \\ wy_p \\ w \end{bmatrix} = \begin{bmatrix} KR & K1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = M \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}
\]

\( K = \begin{bmatrix} fa & fb & x_{p0} \\ 0 & fc & y_{p0} \\ 0 & 0 & 1 \end{bmatrix} \) intrinsic camera parameters

\( R, t \) extrinsic camera parameters

\( M = 3 \times 4 \) projective matrix
Camera calibration

Determine intrinsic and/or extrinsic camera parameters for a specific camera-scene configuration. Prior calibration may be needed:
- to measure unknown objects
- to navigate as a moving observer
- to perform stereo analysis
- to compensate for camera distortions

Important cases:
1. Known scene
   Each image point corresponding with a known scene point provides an equation $v_p = Mv$
2. Unknown scene
   Several views are needed, differing by rotation and/or translation
   a. Known camera motion
   b. Unknown camera motion ("camera self-calibration")

Calibration of one camera from a known scene

- "known scene" = scene with prominent points, whose scene coordinates are known
- prominent points must be non-coplanar to avoid degeneracy
Projection equation $v_p = Mv$ provides 2 linear equations for unknown coefficients of $M$:

\[
\begin{align*}
x_p (m_{31}x + m_{32}y + m_{33}z + m_{34}) &= m_{11}x + m_{12}y + m_{13}z + m_{14} \\
y_p (m_{31}x + m_{32}y + m_{33}z + m_{34}) &= m_{21}x + m_{22}y + m_{23}z + m_{24}
\end{align*}
\]

Taking $N$ points, $N>6$, $M$ can be estimated with a least-square method from an overdetermined system of $2N$ linear equations.

From $M = [\begin{bmatrix} K & R \end{bmatrix}] = [\begin{bmatrix} A & b \end{bmatrix}]$ one gets $K$ and $R$ by QR decomposition of $A$ and $\hat{t}$ from $\hat{t} = K^{-1}b$. 
Fundamental matrix

The fundamental matrix $F$ generalizes the essential matrix $E$ by incorporating the intrinsic camera parameters of two (possibly different) cameras.

Essential matrix constraint for 2 views of a point:

\[ \mathbf{n}^T E \mathbf{n'} = 0 \]

From $\mathbf{v}_p = K\alpha \mathbf{n}$ and $\mathbf{v}_p' = K'\beta \mathbf{n'}$ we get:

\[ \mathbf{v}_p (K')^T E (K')^{-1} \mathbf{v}_p' = \mathbf{v}_p F \mathbf{v}_p' = 0 \]

Note that $E$ and hence $F$ have rank 2.

For each epipole of a 2-camera configuration we have $\mathbf{e}^T F = 0$ and $F \mathbf{e'} = 0$.

Epipolar plane

The epipolar plane is spanned by the projection rays of a point $\mathbf{v}$ and the baseline $CC'$ of a stereo camera configuration.

The epipoles $\mathbf{e}$ and $\mathbf{e'}$ are the intersection points of the baseline with the image planes. The epipolar line $l$ and $l'$ mark the intersections of the epipolar plane in the left and right image, respectively.

Search for corresponding points in stereo images may be restricted to the epipolar lines.

In a canonical stereo configuration (optical axes parallel and perpendicular to baseline) all epipolar lines are parallel:
Correspondence problem

For multiple-view 3D analysis it is essential to find corresponding images of a scene point - the correspondence problem.

Difficulties:
• scene may not offer enough structure to uniquely locate points
• scene may offer too much structure to uniquely locate points
• geometric features may differ strongly between views
• there may be no corresponding point because of occlusion
• photometric features differ strongly between views

Note that difficulties apply to multiple-camera 3D analysis (e.g. binocular stereo) as well as single-camera motion analysis.

Constraining search for correspondence

The ambiguity of correspondence search may be reduced by several (partly heuristic) constraints.
• Epipolar constraint reduces search space from 2D to 1D
• Uniqueness constraint a pixel in one image can correspond to only one pixel in another image
• Photometric similarity constraint intensities of a point in different images may differ only a little
• Geometric similarity constraint geometric features of a point in different images may differ only a little
• Disparity smoothness constraint disparity varies only slowly almost everywhere in the image
• Physical origin constraint points may correspond only if they mark the same physical location
• Disparity limit constraint in humans disparity must be smaller than a limit to fuse images
• Ordering constraint corresponding points lie in the same order on the epipolar line
• Mutual correspondence constraint correspondence search must succeed irrespective of order of images
Neural stereo computation

Neural-network inspired approach to stereo computation devised by Marr and Poggio (1981)

Exploitation of 2 constraints:
• each point in the left image corresponds only to one point in the right image
• depth varies smoothly

Relaxation procedure:
Modify correspondence values $c(x, y, d)$ iteratively until values converge.

$$c_{n+1}(x,y,d) = w_1 \sum_{S_1} c_n(x',y',d') - w_2 \sum_{S_2} c_n(x',y',d'') + w_0 c_0(x,y,d)$$

$S_1 = \{ \text{neighbours of } (x, y) \text{ with } d' = d \}$
$S_2 = \{ \text{neighbours of } (x, y) \text{ with } |d' - d| = 1 \text{ and } (x, y) = (x', y') \}$

Model-based 3D analysis

Principle:
• generate hypotheses from incomplete evidence
• predict additional evidence
• verify or falsify hypotheses
3D Models vs. 2D Models

1. Requirement:
object models must represent invariant class properties
=> 3D models, properties independent of views

   e.g.

2. Requirement:
object models must support recognition
=> 2D models, view-dependent properties