Binocular Stereo

Left Image

Right Image
There are various different methods of extracting relative depth from images, some of the “passive ones” are based on
(i) relative size of known objects,
(ii) texture variations,
(iii) occlusion cues, such as presence of T-Junctions,
(iv) motion information,
(v) focusing and defocusing,
(vi) relative brightness

Moreover, there are active methods such as
(i) Radar, which requires beams of sound waves or
(ii) Laser, uses beam of light

Stereo vision is unique because it is both passive and accurate.
Julesz’s Random Dot Stereogram. The left image, a black and white image, is generated by a program that assigns black or white values at each pixel according to a random number generator.

The right image is constructed from by copying the left image, but an imaginary square inside the left image is displaced a few pixels to the left and the empty space filled with black and white values chosen at random. When the stereo pair is shown, the observers can identify/match the imaginary square on both images and consequently “see” a square in front of the background. It shows that stereo matching can occur without recognition.
Human Stereo: Illusory Contours

Stereo matching occurs in the presence of illusory.

Here not only illusory figures on left and right images don’t match, but also stereo matching yields illusory figures not seen on either left or right images alone.

Not even the identification/matching of illusory contour is known a priori of the stereo process. These pairs gives evidence that the human visual system does not process illusory contours/surfaces before processing binocular vision. Accordingly, binocular vision will be thereafter described as a process that does not require any recognition or contour detection a priori.
Human Stereo: Half Occlusions

An important aspect of the stereo geometry are half-occlusions. There are regions of a left image that will have no match in the right image, and vice-versa. Unmatched regions, or half-occlusion, contain important information about the reconstruction of the scene. Even though these regions can be small they affect the overall matching scheme, because the rest of the matching must reconstruct a scene that accounts for the half-occlusion.

Leonardo DaVinci had noted that the larger is the discontinuity between two surfaces the larger is the half-occlusion. Nakayama and Shimojo in 1991 have first shown stereo pair images where by adding one dot to one image, like above, therefore inducing occlusions, affected the overall matching of the stereo pair.
Projective Camera

Let \( P = (X, Y, Z) \) be a point in the 3D world represented by a “world” coordinate system. Let \( O \) be the center of projection of a camera where a camera reference frame is placed. The camera coordinate system has the \( z \) component perpendicular to the camera frame (where the image is produced) and the distance between the center \( O \) and the camera frame is the focal length, \( f \). In this coordinate system the point \( P = (X, Y, Z) \) is described by the vector \( \vec{P}_o = (X_o, Y_o, Z_o) \) and the projection of this point to the image (the intersection of the line \( PO \) with the camera frame) is given by the point \( \vec{P}_o = (x_o, y_o, f) \)\(^\top\), where

\[
\vec{P}_o = \frac{f}{Z} \vec{P}_0
\]
Projective Camera and Image Coordinate System

We have neglected to account for the radial distortion of the lenses, which would give additional intrinsic parameters. Equation above can be described by the linear transformation

\[
\begin{align*}
\vec{p}_o &= Q^{-1} \vec{q}_o \\
Q^{-1} &= \begin{pmatrix}
s_x & 0 & -s_x o_x \\
0 & -s_y & s_y o_y \\
0 & 0 & f
\end{pmatrix}
\end{align*}
\]

where the intrinsic parameters of the camera, \((s_x, s_y); (o_x, o_y); f\), represent the size of the pixels (say in millimeters) along x and y directions, the coordinate \(q_x, q_y\) in pixels of the image (also called the principal point) and the focal length of the camera.

\[
x_0 = (q_{x0} - o_x) s_x \quad ; \quad y_0 = -(q_{y0} - o_y) s_y
\]
Two Projective Cameras

A 3D point $P$, view in the cyclopean coordinate system, projected on both cameras. The same point $P$ described by a coordinate system in the left eye is $P_l$ and described by a coordinate system in the right eye is $P_r$. The translation vector $T$ brings the origin of one the left coordinate system to the origin of the right coordinate system.
The transformation of coordinate system, from left to right is described by a rotation matrix $R$ and a translation vector $T$. More precisely, a point $P$ described as $P_l$ in the left frame will be described in the right frame as

$$R \vec{P}_r = \vec{P}_l - \vec{T}$$
Each 3D point $P$ defines a plane $PO_lO_r$. This plane intersects the two camera frames creating two corresponding epipolar lines. The line $O_lO_r$ will intersect the camera planes at $\bar{e}_l$ and $\bar{e}_r$, known as the epipoles. The line $O_lO_r$ is common to every plane $PO_lO_l$ and thus the two epipoles belong to all pairs of epipolar lines (the epipoles are the “center/intersection” of all epipolar lines.)
Estimating Epipolar Lines and Epipoles

The two vectors, $\vec{T}, \vec{P}_l$, span a 2 dimensional space. Their cross product $(\vec{T} \times \vec{P}_l)$, is perpendicular to this 2 dimensional space. Therefore

\[ \gamma = \frac{\alpha}{\beta} \]

\[ \vec{P}_l = \gamma \vec{P}_l \]

\[ \vec{P}_r = R^{-1} (\vec{P}_l - \vec{T}) \]

\[ (\alpha \vec{P}_l - \beta \vec{T})^T \cdot (\vec{T} \times \vec{P}_l) = 0 \Rightarrow \beta (\gamma \vec{P}_l - \vec{T})^T \cdot (\vec{T} \times \vec{P}_l) = 0 \Rightarrow (\vec{P}_l - \vec{T})^T \cdot (\vec{T} \times \vec{P}_l) = 0 \Rightarrow \]

\[ (R \vec{P}_r)^T \cdot (\vec{T} \times \vec{P}_l) = 0 \Rightarrow \vec{P}_r^T R^T S(T) \vec{P}_l \Rightarrow \vec{P}_r^T E(R,T) \vec{P}_l = 0 \Rightarrow \vec{p}_r^T E(R,T) \vec{p}_l = \frac{f^2}{Z_l Z_r} \]

where

\[
S(T) = \begin{bmatrix}
0 & -T_z & T_y \\
T_z & 0 & -T_x \\
-T_y & T_x & 0
\end{bmatrix}
\]

and $E(R,T) = R^T S(T)$ is the essential matrix

\[ \vec{p}_r^T E(R,T) \vec{p}_l = 0 \Rightarrow \vec{q}_r^T Q_r^{-T} E(R,T) Q_l^{-1} \vec{q}_l = 0 \Rightarrow \vec{q}_r^T F(R,T, i_l, i_r) \vec{q}_l = 0 \]

$F$ is known as the fundamental matrix and needs to be estimated.
“Eight point algorithm”:

(i) Given two images, we need to identify eight points or more on both images, i.e., we provide \( n \geq 8 \) points with their correspondence. The points have to be non-degenerate.

(ii) Then we have \( n \) linear and homogeneous equations \( \tilde{q}_r^T F(R,T,i_l,i_r) \tilde{q}_l = 0 \) with 9 unknowns, the components of \( F \). We need to estimate \( F \) only up to some scale factors, so there are only 8 unknowns to be computed from the \( n \geq 8 \) linear and homogeneous equations.

(i) If \( n=8 \) there is a unique solution (with non-degenerate points), and if \( n > 8 \) the solution is overdetermined and we can use the SVD decomposition to find the best fit solution.
Each potential match is represented by a square. The black ones represent the most likely scene to “explain” the images, but other combinations could have given rise to the same images (e.g., red).

What makes the set of black squares preferred/unique is that they have similar disparity values, the ordering constraint is satisfied and there is a unique match for each point. Any other set that could have given rise to the two images would have disparity values varying more, and either the ordering constraint violated or the uniqueness violated. The disparity values are inversely proportional to the depth values.
Stereo Correspondence: Matching Space

In the matching space, a point (or a node) represents a match of a pixel in the left image with a pixel in the right image.

Note 1: Depth discontinuities and very tilted surfaces can/will yield the same images (with half occluded pixels).

Note 2: Due to pixel discretization, points A and C in the right frame are neighbors.
The Uniqueness-Opaque Constraint

There should be only one disparity value, one depth value, associated to each cyclopean coordinate $x$ (see figure). The assumption is that objects are opaque and so a 3D point $P$, seen by the cyclopean coordinate $x$ and disparity value $w$ will cause all other disparity values not to be allowed. Closer points than $P$, along the same $x$ coordinate, must be transparent air and further away points will not be seen since $P$ is already been seen (opaqueness). However, multiple matches for left eye points or right eye points are allowed. This is indeed required to model tilt surfaces and occlusion surfaces as we will later discuss. This constraint is a physical motivated one and is easily understood in the cyclopean coordinate system.

Given that the $l=3$ and $r=5$ are matched (blue square), then the red squares represent violations of the uniqueness-opaqueness constraint while the yellow squares represent unique matches, in the cyclopean coordinate system but multiple matches in the left or right eyes coordinate system.
**Surface Constraints I**

**Smoothness**: In nature most surfaces are smooth in depth compared to their distance to the observer, but depth discontinuities also occur.

Given that the $l=3$ and $r=5$ are matched (blue square), then the red squares represent violations of the ordering constraint while the yellow squares represent smooth matches.
Discontinuities: Note that in these cases, some pixels will not be matched to any pixel, e.g., “l+1”, and other pixels will have multiple matches, e.g., “r-1”. In fact, the number of pixels unmatched in the left image is the same as the number of multiple matches in the right image.
Neighborhood structure for a node \((e, xw)\) consisting of flat, tilt, or occluded surfaces. Note that when an occlusion/discontinuity occurs, the contrast matches on the front surface. Jumps “at the right eye” are from back to front, while jumps “at the left eye” are from front to the back.
Discontinuities: Note that in these cases, some pixels will not be matched to any pixel, e.g., “\(l+1\)”, and other pixels will have multiple matches, e.g., “\(r-1\)”. In fact, the number of pixels unmatched in the left image is the same as the number of multiple matches in the right image.
Limit Disparity

The search is within a range of disparity: $2D+1$, i.e.,

The rational is:

(i) Less computations

(ii) Larger disparity matches imply larger errors in 3D estimation.

(iii) Humans only fuse stereo images within a limit. It is called the Panum’s limit.

We may start the computations at $x=D$ to avoid limiting the range of $w$ values. In this case, we also limit the computations to up to $x=2N-2-D$. 

\[ |w| = |r-l| \leq D \]
Image Matching (occlusions)

*Left* and *Right* windows centered at a pixel can accommodate for occlusions.

\[ WeL[I^L](x^L) \sim WeL[I^R](x^R) \quad ? \]

\[ WeR[I^L](x^L) \sim WeR[I^R](x^R) \quad ? \]