Visual Cues for Depth Perception in Natural Scenes

Aude Oliva

9.912
Depth Perception

**Stimulus ambiguity**: the three cubes produce the same retinal image. Monocular information cannot give absolute depth measurements. Only relative depth information such as shape from shading and junctions (occlusions) can be obtained.

**Mean depth** refers to a global measurement of the mean distance between the observer and the main objects and structures that compose the scene.
Depth Perception

• Basic fact: images are produced by perspective projection. Reconstruction process seeks to recover the 3D information lost during perspective projection.

• Early stage of visual perception can be viewed as trying to solve the inverse problem: the projection from environment to image goes from 3D to 2D: each point in the scene maps into a unique point in the image.

• The inverse mapping goes from 2D to 3D: each point in the image could map into an infinite number of points in the environment.

Illustration of inverse projection: a single line segment on the retina can be the projection of an infinite variety of lines in the environment.
Depth Perception

- To resolve the ambiguity, the visual system makes **assumptions** about the **nature of the environments** and the conditions under which it is viewed.

- The assumptions **constrain** the inverse problem enough to make it solvable most of the time.

- Vision is thus a heuristic process. Under most everyday circumstances, the assumptions are true.

Illustration of inverse projection: a single line segment on the retina can be the projection of an infinite variety of lines in the environment.
Depth Perception

- Sources of information about depth. There are 5 important characteristics about depth information.

<table>
<thead>
<tr>
<th>INFORMATION SOURCE</th>
<th>Ocular/Optical</th>
<th>Binocular/Monocular</th>
<th>Static/Dynamic</th>
<th>Relative/Absolute</th>
<th>Qualitative/Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accommodation</td>
<td>ocular</td>
<td>monocular</td>
<td>static</td>
<td>absolute</td>
<td>quantitative</td>
</tr>
<tr>
<td>Convergence</td>
<td>ocular</td>
<td>binocular</td>
<td>static</td>
<td>absolute</td>
<td>quantitative</td>
</tr>
<tr>
<td>Binocular Disparity</td>
<td>optical</td>
<td>binocular</td>
<td>static</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Motion Parallax</td>
<td>optical</td>
<td>monocular</td>
<td>dynamic</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Texture Accretion/Deletion</td>
<td>optical</td>
<td>monocular</td>
<td>dynamic</td>
<td>relative</td>
<td>qualitative</td>
</tr>
<tr>
<td>Convergence of Parallels</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Position relative to Horizon</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Relative Size</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Familiar Size</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>absolute</td>
<td>quantitative</td>
</tr>
<tr>
<td>Texture Gradients</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Edge Interpretation</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>qualitative</td>
</tr>
<tr>
<td>Shading and Shadows</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>qualitative</td>
</tr>
<tr>
<td>Aerial Perspective</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>qualitative</td>
</tr>
</tbody>
</table>
David Marr: 2.5 D sketch

How might visual information about the layout of surfaces in depth be represented?

A 2.5 D surface-based representation is thought to be computed from a 2D image-based representation by a set of parallel and quasi independent processes that extract information about surface orientation and depth from a variety of sources. In computer vision community, these processes are referred to as “shape_from_X” where X is the source of depth information.

Marr’s 2.5 D sketch: it summarizes the many converging outputs of different processes that extract information about surface orientation and depth from a variety of sources (see table).
Gibson View: Textural Gradient

• Gibson argued that perception is based primarily on the structure of the environment (*ecological* perception).
• The sensory data of the world already possessed sufficient order.
• What is this “*order*”?
• Order is in the *visual texture* of the material world. We perceive “surface” through the visual textures of which materials are made.
Variety in Texture Gradient

Texture gradient describes the correspondence between the pattern of a surface and the structure of the 3D world.

There are several signature textural gradient: e.g. frontal surface project uniform gradients. Longitudinal surfaces such as floors and streets project gradient that diminish with greater distance from the observer.
Texture Gradient in natural scenes
Categories of Texture Gradient
Frontal or Ground Surface?
Depth Cues

• Depth cues are conditions of the visual field (e.g. one object overlapping another) than commonly yield perception of depth.

• Depth of surfaces can be estimated using pictorial information
• Pictorial information are potentially available in static, monocular viewed pictures (so they are interesting for machine vision).

• **Gibson** breaks the list of depth cues into two groups:

  1) Primary Cues : (needs two eyes). These cues are only effective within short distance from the observer. Accommodation, Disparity vision, convergence.

  2) Secondary cues (7 cues). The secondary cues are *pictorial* cues.
1- Distance from the horizon line

- Based on the tendency of objects to appear nearer the horizon line with greater distance to the horizon.
- Objects approach the horizon line with greater distance from the viewer.
- The base of a nearer column will appear lower against its background floor and further from the horizon line.
- Conversely, the base of a more distant column will appear higher against the same floor, and thus nearer to the horizon line.
2 & 3- Overlap and Motion parallax

- **Overlap**: Based on the tendency of near objects to overlap far objects (more striking effect if the nearest object is bigger).
- According to Gibson, overlapping edges arise in abrupt shifts of textural gradients.
- Overlap is related to **motion parallax**: when we move, nearer things tend to move more rapidly across the visual field (Godzilla will look scarier to you), and distance things move more slowly.
4- Shade and Shadows

- Based on 3 dimensional modeling of objects in light, shade and shadows.

- Perception of depth through shading alone is always subject to the concave/convex inversion. The pattern shown can be perceived as stairsteps receding towards the top and lighted from above, or as an overhanging structure lighted from below.
5- Size Perspective

- Apparent reduction in size of objects at a greater distance from the observer

Size perspective is thought to be conditional, requiring knowledge of the objects.
6- Atmospheric perspective

• Based on the effect of air on the color and visual acuity of objects at various distances from the observer.
• Consequences:
  • Distant objects appear bluer
  • Distant objects appear less distinct.
7- Perspective

• Based on the apparent convergence of parallel lines to common vanishing points with increasing distance from the observer.

In Gibson’s term, perspective is a characteristic of the visual field rather than the visual world. It approximates how we see (the retinal image) rather than what we see, the objects in the world.

Perspective: a representation that is specific to one individual, in one position in space and one moment in time.

Is perspective a universal fact of the visual retinal image? Or is perspective something that is learned?
# Depth Perception

<table>
<thead>
<tr>
<th>INFORMATION SOURCE</th>
<th>Ocular/Optical</th>
<th>Binocular/Monocular</th>
<th>Static/Dynamic</th>
<th>Relative/Absolute</th>
<th>Qualitative/Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accommodation</td>
<td>ocular</td>
<td>monocular</td>
<td>static</td>
<td>absolute</td>
<td>quantitative</td>
</tr>
<tr>
<td>Convergence</td>
<td>ocular</td>
<td>binocular</td>
<td>static</td>
<td>absolute</td>
<td>quantitative</td>
</tr>
<tr>
<td>Binocular Disparity</td>
<td>optical</td>
<td>binocular</td>
<td>static</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Motion Parallax</td>
<td>optical</td>
<td>monocular</td>
<td>dynamic</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Texture Accretion/Deletion</td>
<td>optical</td>
<td>monocular</td>
<td>dynamic</td>
<td>relative</td>
<td>qualitative</td>
</tr>
<tr>
<td>Convergence of Parallels</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Position relative to Horizon</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Relative Size</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>quantitative</td>
</tr>
<tr>
<td>Familiar Size</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>absolute</td>
<td>quantitative</td>
</tr>
<tr>
<td>Texture Gradients</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>qualitative</td>
</tr>
<tr>
<td>Edge Interpretation</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>qualitative</td>
</tr>
<tr>
<td>Shading and Shadows</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>qualitative</td>
</tr>
<tr>
<td>Aerial Perspective</td>
<td>optical</td>
<td>monocular</td>
<td>static</td>
<td>relative</td>
<td>qualitative</td>
</tr>
</tbody>
</table>

**Statistical structure**

optical  monocular  static  absolute  quantitative
However, nature (and man) do not build in the same way at different scales.

If $d_1 >> d_2 >> d_3$ the structures of each view strongly differ. **Structure** provides monocular information about the scale (mean depth) of the space in front of the observer.
Mean Depth in Natural Scenes
Spectral signature of man-made environments

Evolution of the slope of the global magnitude spectrum with respect to the mean depth of the scene

We observe a decreasing slope indicating an increasing of the global clutter of the scene with increasing distance: close-up views on objects have large and homogeneous regions. When increasing the size of the space, the scene “surface” breaks down in smaller pieces (objects, walls, windows, etc).

Torralba & Oliva (2002). Depth estimation from Image Structure. PAMI, 24, 9, 1225.
Image Statistics and Scene Scale
Mean Depth Statistics

Spectral signature of natural environments at Various scales

Evolution of the slope of the global magnitude spectrum with respect to the mean depth of the scene

When increasing the size of the space, structures in natural images become larger and smoother.

Torralba & Oliva (2002). Depth estimation from Image Structure. PAMI, 24, 9, 1225.
Image Statistics and Scene Scale

1-5 m  |  5-50 m  |  50-500 m  |  > 500 m

Images of different scales showing variations in texture and detail.
Image Statistics and Scene Scale
Image Statistics and Scene Scale

Close-up views

On average, low clutter
Point view is unconstrained

Large scenes

On average, highly cluttered
Point view is strongly constrained
Figure 5. Polar plots of responses of multiscale oriented Gabor filters. The magnitude of each orientation corresponds to the total output energy averaged across the entire image. The energies are normalized across image scale by multiplying by a constant so that noise with $1/f$ amplitude spectrum has the same polar plots at all image scales.
Basic level categories from scene attributes…
Depth Perception

Same Surface structure

“A close-up view on a beach”

“A large scene scale, a coastline”

Oliva et al (in preparation)
Depth Perception

We got wrong:
• 3D shape (mainly due to assumption of light from above)
• The absolute scale (due to the wrong recognition).

The inversion affected the perception of concavities and convexities due to the assumption of light from above. But also, the wrong recognition has affected the size of the perceived space of the scene (Torralba & Oliva, 2003).
Depth *Misperception*
Mean Depth from Image Structure

We learn the relationship between image structure and the mean depth of the scene

\[ v = \text{structure} \]

\[ P(\text{scale} \mid x, v) \]

\[ \prod_x P(\text{scale} \mid x, v) \]

We can learn what is the relationship between the global depth of the scene (e.g. 10 meters, 50 meters, etc) and the local structures in the scene. So, here, for each local patch, we provide what is the probability of each global depth.

Then, you can put all the local curves together by doing the product and it will give you one unique curve that tells you what is the most likely depth for this scene given all the local evidences.

Torralba & Oliva, 2002, PAMI
Mean Depth from Image Structure

76% images with correct mean depth estimation. (Torralba & Oliva, 2002)
Performance in depth estimation

Torralba & Oliva, 2002, PAMI