Common Classification Tasks

- Recognition of individual objects/faces
  - Analyze object-specific features (e.g., key points)
  - Train with images from different viewing angles
- Recognition of object classes
  - Analyze features that are consistent within class and differ between them as much as possible.
  - Train with many exemplars from each class.
- Recognition of scene types
  - Find and analyze common features, objects, or layouts within scene classes.
  - Use large variety of scene photos.

Example: AlexNet

AlexNet (Krizhevsky et al. 2012)

The class with the highest likelihood is the one the DNN selects.

When AlexNet is processing an image, this is what is happening at each layer.

Another Example: Inception

9 Inception modules

Network in a network in a network...

Another Example: Inception

Depth

Estimating the distance of a point from the observer is crucial for scene and object recognition.
There are many monocular cues such as shading, texture gradient, and perspective.
However, just like biological systems, computer vision systems can benefit a lot from stereo vision, i.e.
using two visual inputs from slightly different angles.
Computer vision systems also have the option to use completely different approaches such as radar,
laser range finding, and structured lighting.
Let us first talk about stereo vision.
Stereo Vision

In the simplest case, the two cameras used for stereo vision are identical and separated only along the x-axis by a **baseline distance** $b$.

Then the image planes are **coplanar**. Depth information is obtained through the fact that the same feature point in the scene appears at **slightly different positions** in the two image planes.

This displacement between the two images is called the **disparity**.

The plane spanned by the two camera centers and the feature point is called the **epipolar plane**.

Its intersection with the image plane is the **epipolar line**.

---

Stereo Vision

Ideally, every feature in one image will lie in the **same vertical position** in the other image.

However, due to distortion in the camera images and imperfect geometry, there usually is also some **vertical disparity**.

While many algorithms for binocular stereo do not account for such vertical disparity, it **increases the robustness** of our system if we allow at least a few pixels of deviation from the epipolar line.

---

Stereo Vision

Comparing the similar triangles $PMC$ and $pL_{C_p}$, we get:

$$\frac{x}{z} = \frac{x'}{f}$$

Similarly, for $PNC$, and $pR_{C_p}$, we get:

$$\frac{x - b}{z} = \frac{x'}{f}$$

Combining gives us:

$$z = \frac{bf}{x' - x'}$$

---

Stereo Vision

Due to the limited resolution of images, increasing the baseline distance $b$ gives us a **more precise** estimate of depth $z$.

However, the greater $b$, the more different are the two viewing angles, and the **more difficult** it can become to determine the correspondence between the two images.

This brings us to the main problem in stereo vision:

How can we find the conjugate pairs in our stereo images?

This problem is called **stereo matching**.
Stereo Matching

In stereo matching, we have to solve a problem that is still under investigation by many researchers, called the correspondence problem. It can be phrased like this: For each point in the left image, find the corresponding point in the right image. The idea underlying all stereo matching algorithms is that these two points should be similar to each other. So we need a measure for similarity. Moreover, we need to find matchable features.
Stereo Matching

[Diagram showing left and right camera images with occlusions marked]

Fig. 10.8: Occlusions along an epipolar line in the example stereo-image pair caused by parallax. The image content enclosed by each pair of vertical, dashed lines is only visible in one of the images, making naïve stereo matching impossible.

Stereo Matching

[Diagram showing depth map and disparity map]

Fig. 10.9: Depth map for one row of a labeled stereo image. The image shows a desk with a clock, with a textured background. Each cell contains an integer distance in meters, with black lines indicating the associated cost per pixel for each row.

Generating Interesting Points

Using interpolation to determine the depth of points within large homogeneous areas may cause large errors.

To overcome this problem, we can generate additional interesting points that can be matched between the two images.

The idea is to use structured light, i.e., project a pattern of light onto the visual scene.

This creates additional variance in the brightness of pixels and increases the number of interesting points.
Generating Interesting Points

Shape from Shading
Besides binocular disparity, there are many different ways of depth estimation based on monocular information. For example, if we know the reflective properties of the surfaces in our scene and the position of the light source, we can use shape from shading techniques: Basically, since the amount of light reflected by a surface depends on its angle towards the light source, we can estimate the orientation of surfaces based on their intensity. More sophisticated methods also use the contours of shadows cast by objects to estimate the shape and orientation of those objects.

Photometric Stereo
To improve the coarse estimates of orientation derived from shape from shading methods, we can use photometric stereo. This technique uses three light sources that are located at different, known positions. Three images are taken, one for each light source, with the other two light sources being turned off. This way we determine three different intensities for each surface in the scene. These three values put narrow constraints on the possible orientation of a surface and allow a rather precise estimation.

Shape from Texture
As we discussed before, the texture gradient gives us valuable monocular depth information. At any point in the image showing texture, the texture gradient is a two-dimensional vector pointing towards the steepest increase in the size of texture elements. The texture gradient across a surface allows a good estimate of the spatial orientation of that surface. Of course, it is important for this technique that the image has high resolution and a precise method of texture size (granularity) measurement is used.

Shape from Motion
The shape from motion technique is similar to binocular stereo, but it uses only one camera. This camera is moved while it takes images from the visual scene. This way two images with a specific baseline distance can be obtained, and depth can be computed just like for binocular stereo. We can even use more than two images in this computation to get more robust measurements of depth. The disadvantages of shape from motion techniques are the technical overhead for moving the camera and the reduced temporal resolution.

Range Imaging
If we can determine the depth of every pixel in an image, we can make this information available in the form of a range image. A range image has exactly the same size and number of pixels as the original image. However, each pixel does not specify color or intensity, but the depth of that pixel in the original image, encoded as grayscale. Usually, the brighter a pixel in a range image is, the closer the corresponding pixel is to the observer. By providing both the original and the range image, we basically define a 3D image.
Range Imaging Through Triangulation

How can we obtain precise depth information for every pixel in the image of a scene?

One precise but slow method uses a laser that can rotate around its vertical axis and can also assume different vertical positions. This laser systematically and sequentially illuminates points in the image. A scene camera determines the position of every single point in its picture. The trick is that this camera looks at the scene from a different direction than does the laser pointer. Therefore, the depth of every point can be easily and precisely determined through triangulation: