Computer stereo vision

Computer stereo vision is the extraction of 3D information from digital images, such as those obtained by a CCD camera. By comparing information about a scene from two vantage points, 3D information can be extracted by examining the relative positions of objects in the two panels. This is similar to the biological process of stereopsis.

Outline

In traditional stereo vision, two cameras, displaced horizontally from one another, are used to obtain two differing views on a scene, in a manner similar to human binocular vision. By comparing these two images, the relative depth information can be obtained in the form of a disparity map, which encodes the difference in horizontal coordinates of corresponding image points. The values in this disparity map are inversely proportional to the scene depth at the corresponding pixel location.

For a human to compare the two images, they must be superimposed in a stereoscopic device, with the image from the right camera being shown to the observer's right eye and from the left one to the left eye.

In a computer vision system, several pre-processing steps are required.[1]

1. The image must first be undistorted, such that barrel distortion and tangential distortion are removed. This ensures that the observed image matches the projection of an ideal pinhole camera.
2. The image must be projected back to a common plane to allow comparison of the image pairs, known as image rectification.
3. An information measure which compares the two images is minimized. This gives the best estimate of the position of features in the two images, and creates a disparity map.
4. Optionally, the received disparity map is projected into a 3d point cloud. By utilising the cameras' projective parameters, the point cloud can be computed such that it provides measurements at a known scale.

Active stereo vision

The active stereo vision is a form of stereo vision which actively employs a light such as a laser or a structured light to simplify the stereo matching problem. The opposed term is passive stereo vision.

- Conventional structured-light vision (SLV) employs a structured light or laser, and finds projector-camera correspondences.[2][3]

- Conventional active stereo vision (ASV) employs a structured light or laser, however, the stereo matching is performed only for camera-camera correspondences, in the same way as the passive stereo vision.
- Structured-light stereo (SLS) is a hybrid technique, which utilizes both camera-camera and projector-camera correspondences.[4]

Applications

3D stereo displays find many applications in entertainment, information transfer and automated systems. Stereo vision is highly important in fields such as robotics to extract information about the relative position of 3D objects in the vicinity of autonomous systems. Other applications for robotics include object recognition,[5] where depth information allows for the system to separate occluding image components, such as one chair in front of another, which the robot may otherwise not be able to distinguish as a separate object by any other criteria.

Scientific applications for digital stereo vision include the extraction of information from aerial surveys, for calculation of contour maps or even geometry extraction for 3D building mapping, photogrammetric satellite mapping, or calculation of 3D heliographical information such as obtained by the NASA STEREO project.

Detailed definition

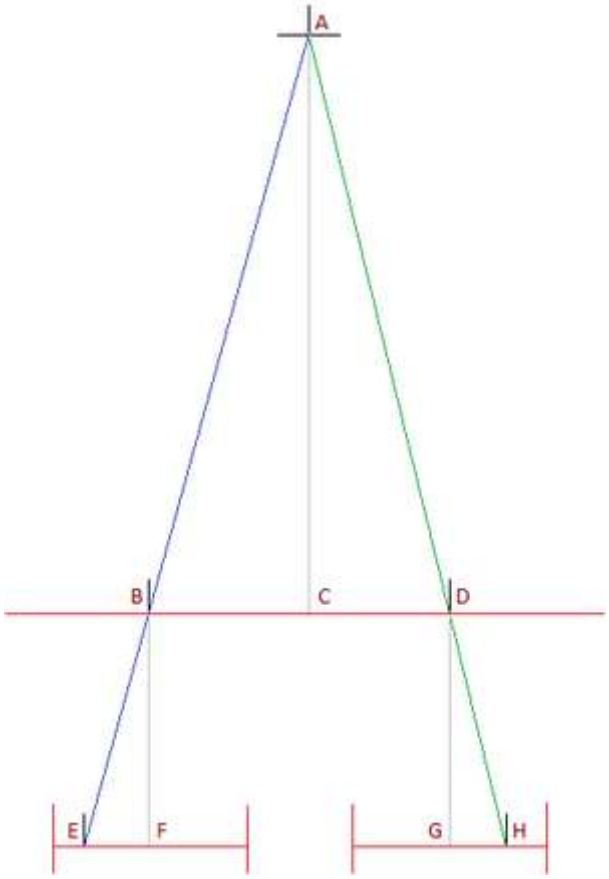

Diagram describing relationship of image displacement to depth with stereoscopic images, assuming flat co-planar images

A pixel records color at a position. The position is identified by position in the grid of pixels (x, y) and depth to the pixel *z*.

Stereoscopic vision gives two images of the same scene, from different positions. In the adjacent diagram light from the point *A* is transmitted through the entry points of pinhole cameras at *B* and *D*, onto image screens at *E* and *H*.

In the attached diagram the distance between the centers of the two camera lens is $BD = BC + CD$. The triangles are similar,

- *ACB* and *BFE*
- *ACD* and *DGH*

$$\begin{aligned}\text{Therefore displacement } d &= EF + GH \\ &= BF(\frac{EF}{BF} + \frac{GH}{BF}) \\ &= BF(\frac{EF}{BF} + \frac{GH}{DG}) \\ &= BF(\frac{BC + CD}{AC}) \\ &= BF\frac{BD}{AC} \\ &= \frac{k}{z}, \text{ where}\end{aligned}$$

- *k* = *BD BF*
- *z* = *AC* is the distance from the camera plane to the object.

So assuming the cameras are level, and image planes are flat on the same plane, the displacement in the y axis between the same pixel in the two images is,

$$d = \frac{k}{z}$$

Where *k* is the distance between the two cameras times the distance from the lens to the image.

The depth component in the two images are z_1 and z_2, given by,

$$z_2(x,y) = \min\left\{v : v = z_1(x, y - \frac{k}{z_1(x,y)})\right\}$$
$$z_1(x,y) = \min\left\{v : v = z_2(x, y + \frac{k}{z_2(x,y)})\right\}$$

These formulas allow for the occlusion of voxels, seen in one image on the surface of the object, by closer voxels seen in the other image, on the surface of the object.

Image rectification

Where the image planes are not co-planar, image rectification is required to adjust the images as if they were co-planar. This may be achieved by a linear transformation.

The images may also need rectification to make each image equivalent to the image taken from a pinhole camera projecting to a flat plane.

Smoothness

Smoothness is a measure of the similarity of colors. Given the assumption that a distinct object has a small number of colors, similarly-colored pixels are more likely to belong to a single object

than to multiple objects.

The method described above for evaluating smoothness is based on information theory, and an assumption that the influence of the color of a voxel influences the color of nearby voxels according to the normal distribution on the distance between points. The model is based on approximate assumptions about the world.

Another method based on prior assumptions of smoothness is auto-correlation.

Smoothness is a property of the world rather than an intrinsic property of an image. An image comprising random dots would have no smoothness, and inferences about neighboring points would be useless.

In principle, smoothness, as with other properties of the world, should be learned. This appears to be what the human vision system does.

Information measure

Least squares information measure

The normal distribution is

$$P(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Probability is related to information content described by message length *L*,

$$P(x) = 2^{-L(x)}$$
$$L(x) = -\log_2 P(x)$$

so,

$$L(x, \mu, \sigma) = \log_2(\sigma\sqrt{2\pi}) + \frac{(x-\mu)^2}{2\sigma^2}\log_2 e$$

For the purposes of comparing stereoscopic images, only the relative message length matters. Based on this, the information measure *I*, called the Sum of Squares of Differences (SSD) is,

$$I(x, \mu, \sigma) = \frac{(x-\mu)^2}{\sigma^2}$$

where,

$$L(x, \mu, \sigma) = \log_2(\sigma\sqrt{2\pi}) + I(x, \mu, \sigma)\frac{\log_2 e}{2}$$

Because of the cost in processing time of squaring numbers in SSD, many implementations use Sum of Absolute Difference (SAD) as the basis for computing the information measure. Other

methods use normalized cross correlation (NCC).

Information measure for stereoscopic images

The least squares measure may be used to measure the information content of the stereoscopic images,[6] given depths at each point $z(x,y)$. Firstly the information needed to express one image in terms of the other is derived. This is called I_m.

A color difference function should be used to fairly measure the difference between colors. The color difference function is written *cd* in the following. The measure of the information needed to record the color matching between the two images is,

$$I_m(z_1,z_2)=\frac{1}{\sigma_m^2}\sum_{x,y}\mathrm{cd}(\mathrm{color}_1(x,y+\frac{k}{z_1(x,y)}),\mathrm{color}_2(x,y))^2$$

An assumption is made about the smoothness of the image. Assume that two pixels are more likely to be the same color, the closer the voxels they represent are. This measure is intended to favor colors that are similar being grouped at the same depth. For example, if an object in front occludes an area of sky behind, the measure of smoothness favors the blue pixels all being grouped together at the same depth.

The total measure of smoothness uses the distance between voxels as an estimate of the expected standard deviation of the color difference,

$$I_s(z_1,z_2)=\frac{1}{2\sigma_h^2}\sum_{i:\{1,2\}}\sum_{x_1,y_1}\sum_{x_2,y_2}\frac{\mathrm{cd}(\mathrm{color}_i(x_1,y_1),\mathrm{color}_i(x_2,y_2))^2}{(x_1-x_2)^2+(y_1-y_2)^2+(z_i(x_1,y_1)-z_i(x_2,y_2))^2}$$

The total information content is then the sum,

$$I_t(z_1,z_2)=I_m(z_1,z_2)+I_s(z_1,z_2)$$

The z component of each pixel must be chosen to give the minimum value for the information content. This will give the most likely depths at each pixel. The minimum total information measure is,

$$I_{\min}=\min\{i:i=I_t(z_1,z_2)\}$$

The depth functions for the left and right images are the pair,

$$(z_1,z_2)\in\{(z_1,z_2):I_t(z_1,z_2)=I_{\min}\}$$

Methods of implementation

The minimization problem is NP-complete. This means a general solution to this problem will take a long time to reach. However methods exist for computers based on heuristics that

approximate the result in a reasonable amount of time. Also methods exist based on neural networks.[7] Efficient implementation of stereoscopic vision is an area of active research.

See also

- 3D reconstruction from multiple images
- 3D scanner
- Autostereoscopy
- Computer vision
- Epipolar geometry
- Semi-global matching
- Structure from motion
- Stereo camera
- Stereophotogrammetry
- Stereopsis
- Stereoscopic depth rendition
- Stixel
- Trifocal tensor - for trifocal stereoscopy (using three images instead of two)

References

1. Bradski, Gary; Kaehler, Adrian. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly.
2. Je, Changsoo; Lee, Sang Wook; Park, Rae-Hong (2004). "High-Contrast Color-Stripe Pattern for Rapid Structured-Light Range Imaging". *Computer Vision - ECCV 2004*. Lecture Notes in Computer Science. Vol. 3021. pp. 95–107. arXiv:1508.04981 (https://arxiv.org/abs/1508.04981) . doi:10.1007/978-3-540-24670-1_8 (https://doi.org/10.1007%2F978-3-540-24670-1_8) . ISBN 978-3-540-21984-2. S2CID 13277591 (https://api.semanticscholar.org/CorpusID:13277591) .
3. Je, Changsoo; Lee, Sang Wook; Park, Rae-Hong (2012). "Colour-stripe permutation pattern for rapid structured-light range imaging" (https://dx.doi.org/10.1016/j.optcom.2012.01.025) . *Optics Communications*. **285** (9): 2320–2331. Bibcode:2012OptCo.285.2320J (https://ui.adsabs.harvard.edu/abs/2012OptCo.285.2320J) . doi:10.1016/j.optcom.2012.01.025 (https://doi.org/10.1016%2Fj.optcom.2012.01.025) .

4. Jang, Wonkwi; Je, Changsoo; Seo, Yongduek; Lee, Sang Wook (2013). "Structured-light stereo: Comparative analysis and integration of structured-light and active stereo for measuring dynamic shape" (https://dx.doi.org/10.1016/j.optlaseng.2013.05.001) . *Optics and Lasers in Engineering*. **51** (11): 1255–1264. Bibcode:2013OptLE..51.1255J (https://ui.adsabs.harvard.edu/abs/2013OptLE..51.1255J) . doi:10.1016/j.optlaseng.2013.05.001 (https://doi.org/10.1016%2Fj.optlaseng.2013.05.001) .
5. Sumi, Yasushi; Kawai, Yoshihiro; Yoshimi, Takashi; Tomita, Fumiaki (2002). "3D Object Recognition in Cluttered Environments by Segment-Based Stereo Vision" (https://link.springer.com/article/10.1023/A:1013240031067) . *International Journal of Computer Vision*. **46** (1): 5–23. doi:10.1023/A:1013240031067 (https://doi.org/10.1023%2FA%3A1013240031067) . S2CID 22926546 (https://api.semanticscholar.org/CorpusID:22926546) .
6. Lazaros, Nalpantidis; Sirakoulis, Georgios Christou; Gasteratos1, Antonios (2008). "Review of Stereo Vision Algorithms: From Software to Hardware" (https://doi.org/10.1080%2F15599610802438680) . *International Journal of Optomechatronics*. **2** (4): 435–462. doi:10.1080/15599610802438680 (https://doi.org/10.1080%2F15599610802438680) . S2CID 18115413 (https://api.semanticscholar.org/CorpusID:18115413) .
7. WANG, JUNG-HUA; HSIAO, CHIH-PING (1999). "On disparity matching in stereo vision via a neural network framework". *Proc. Natl. Sci. Counc. ROC(A)*. **23** (5): 665–678. CiteSeerX 10.1.1.105.9067 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.9067) .

External links

- Tutorial on uncalibrated stereo vision (http://pages.cs.wisc.edu/~chaol/cs766/)
- Learn about stereo vision with MATLAB (http://www.mathworks.com/discovery/stereo-vision.html)
- Stereo Vision and Rover Navigation Software for Planetary Exploration (http://www2.ece.ohio-state.edu/ion/documents/IEEE_aero.pdf)