SCIENTIFIC DATA

OPEN

Data Descriptor: A dataset of stereoscopic images and ground-truth disparity mimicking human fixations in peripersonal space

Received: 05 September 2016
Accepted: 13 January 2017
Published: 28 March 2017

Andrea Canessa[1,*], Agostino Gibaldi[1,*], Manuela Chessa[1], Marco Fato[1], Fabio Solari[1] & Silvio P. Sabatini[1]

Binocular stereopsis is the ability of a visual system, belonging to a live being or a machine, to interpret the different visual information deriving from two eyes/cameras for depth perception. From this perspective, the ground-truth information about three-dimensional visual space, which is hardly available, is an ideal tool both for evaluating human performance and for benchmarking machine vision algorithms. In the present work, we implemented a rendering methodology in which the camera pose mimics realistic eye pose for a fixating observer, thus including convergent eye geometry and cyclotorsion. The virtual environment we developed relies on highly accurate 3D virtual models, and its full controllability allows us to obtain the stereoscopic pairs together with the ground-truth depth and camera pose information. We thus created a stereoscopic dataset: *GENUA PESTO—GENoa hUman Active fixation database: PEripersonal space STereoscopic images and grOund truth disparity*. The dataset aims to provide a unified framework useful for a number of problems relevant to human and computer vision, from scene exploration and eye movement studies to 3D scene reconstruction.

Design Type(s)	benchmarking objective
Measurement Type(s)	depth perception
Technology Type(s)	stereoscopic imaging
Factor Type(s)	observer position • visual stimulus
Sample Characteristic(s)	

[1]DIBRIS—University of Genoa, Genoa, GE 16145, Italy. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to A.G. (email: agostino.gibaldi@unige.it).

Background & Summary

Stereopsis is commonly dealt with as a static problem, because the disparity map obtained by a fixed-geometry stereo camera pair with parallel axes, is sufficient to reconstruct the 3D spatial layout of the observed scene. However, the capability of scanning the visual scene with the eyes, as for active vision systems, provides a number of contingent cues about the 3D layout of objects in a scene that could be used for planning and controlling goal-directed behaviors. Specifically, a purposeful approach can take into account attention mechanisms and a gaze-centric analysis of the scene. This is particularly true while interacting with peripersonal workspace, i.e., the space at short (reaching) distances, representing the next challenge of advanced cognitive robotic systems. A systematic collection of stereoscopic image pairs under vergent geometry, with ground-truth depth/disparity information, would thus be an ideal tool to characterize the problem of purposeful 3D vision. Nevertheless, these kinds of datasets are, unfortunately, rare or nearly absent.

Existing public datasets mainly provide scene images and range data: e.g., see 2.5D/3D Datase[1], CIN 2D+3D[2], Cornell RGB-D[3,4], LIVE Color+3D[5,6], B3DO[7], Sun3D[8]. Among them, few provide stereoscopic images with disparity data: e.g., see Middelbury[9], IMPART[10], KITTI[11,12], and SYNS[13] datasets. Yet, they mainly follow a standard machine vision approach, i.e., with parallel optical axes for the two cameras (*off-axis* technique). In such a situation, vertical disparity is identically equal to zero over the whole field of view, thus reducing the search zone for stereo correspondence to horizontal lines, i.e., to a 1D problem. Moreover, horizontal disparity is proportional to the depth of the points in space.

A binocular visual system, belonging it to a human or a robotic system (e.g., see iCub[14]), is required to correctly fixate with the two eyes, i.e., to have both optical axes converging on the same point in space (*toe-in* technique). The resulting disparity patterns are considerably different from those derived by the *off-axis* technique[15], and the search for stereo correspondence turns into a 2D problem. The horizontal disparity is close to zero at fixation, and assumes positive values for points closer than the fixation point and negative values for points farther away. Besides, vertical disparity takes on a small but significant value, providing a characteristic 'cross' shape.

The dataset presented in this paper is meant to mimic the actual 6 degree-of-freedom eye pose of an active binocular visual system in peripersonal space, i.e., with camera pose exhibiting vergence and cyclotorsion as a function of the gaze direction[16], (i.e., direction from camera to fixation point). Our approach relies on two complementary parts: 3D virtual models of natural scenes composed of in peripersonal space, which consist of accurate depth information and natural textures for the scenes (see Figs 1 and 2), and a graphic stereo vision simulator, to mimic the natural eye position of the human visual system (see Fig. 3).

The resulting dataset has a number of outstanding qualities, required for stereo benchmark[17]: (1) high spatial accuracy (≈0.1 mm), (2) realistic color texture with high resolution, (3) ground-truth disparity maps, (4) ground-truth occlusions and depth edges, (5) ground-truth position of stereo cameras, and (6) large number of stereo pairs.

Accordingly, this dataset has a large potential for use in different fields of research. In *Human Vision*, ground-truth data of natural scenes allow for a quantitative characterization of human behaviour, providing a useful means to create fully controlled and accurate visual stimulation for psychophysics and neuroscience experiments[18]. The stereo pairs can be used to investigate the influence of depth cue on: eye movement[19–24], visual saliency and attention[19,25–32], surface orientation perception[33–40] and object recognition[41–45]. Considering recent widespread of 3D visualization methodologies, from low-cost TV monitors to head-mounted displays, the vergent geometry can be also useful to improve eye-hand coordination[46–48], or for stereoscopic perception[49–51] and image quality assessment[52–54]. The large number of stereo pairs can be used to collect retinal disparity statistics, for a direct comparison with the known binocular visual functionalities[55–62]. Specifically, the *ground-truth* quality of the data allows for an in-depth comparison that would not be possible otherwise[63]. Furthermore, the dataset can be used to learn monocular and binocular receptive fields[56,64–66].

In *Machine Vision*, the ground-truth information included in the dataset, seldom provided in 3D databases[9], is an optimal instrument to perform unconventional analyses of the problem, in order to develop and benchmark the algorithms themselves. The high accuracy and resolution of the ground-truth disparity data, makes the dataset optimal for an extensive evaluation of disparity estimation (for a recent review, see[67]). Specifically, since no databases of images with ground-truth *vector* disparity are available, the proposed database is unique in its kind. Exemplifying, it allows deriving quantitative performance indexes for horizontal and vertical disparity estimation algorithms, both on a pixel and a local basis[9,68–70]. Moreover, it can be used for image segmentation[71,72] and object recognition[73–75], surface orientation estimation[76,77] and *Occlusion* and depth discontinuities estimation[17,71,77–84], and visual saliency models[85–87]. The ground-truth position of the cameras can be used to benchmark stereo calibration[88], epipolar geometry and camera pose estimation[80,89–95].

The dataset can be also useful for algorithms not directly related to stereovision, like: structure from parallax[73,96], three views geometry and trifocal tensor[97,98], and multiple view geometry[97,99], as well as feature matching[100] and affine transformation[101].

Summarizing, the present dataset of stereoscopic images provides a unified framework useful for many problems relevant to human and computer vision, from visual exploration and attention to eye movement studies, from perspective geometry to depth reconstruction.

Figure 1. Examples of 3D model acquisition and registration. For each presented object, the insets on the left show the different 3D raw scans used to build the complete object model.

Figure 2. Renderings of the 3D models used: an office desk (left) and a kitchen table (right).

Methods

In order to render naturalistic stereoscopic images together with ground-truth disparity information, our methodology consists of two complementary parts. From the one side, it requires 3D virtual models of natural scenes, which provide both the natural texture of the objects composing the scene (see Fig. 2), and the accurate geometric information about the objects' shape. Those models are required for the stereo vision simulator, in order to render naturalistic stereoscopic images accompanied by ground truth disparity information. Specifically, the simulator has been implemented to mimic the viewing posture of the human visual system in terms of vergent geometry and cyclotorsion. In this section, we describe the technical details about the 3D models acquisition and registration, and about the implementation of the stereo vision simulator.

3D model acquisition and composition

Aiming to study the peripersonal space, the scenes that we considered were bounded inside a workspace 1 m × 1 m. The scenes were composed of real-world objects arranged in a cluttered way in order to have a high complexity structure. The scenes tried to replicate every-day life situations.

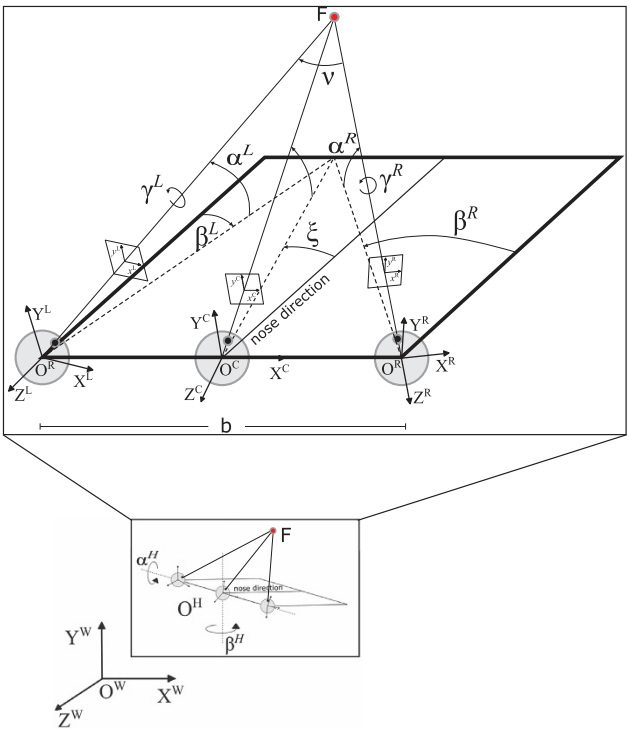

Figure 3. Schematic representation of the geometry of the binocular active vision system. *F* is the fixation point, *C* is the cyclopic position (halfway between the eyes), *L* and *R* are the left and right camera positions, separated by a baseline $b = 60$ mm. The α, β and γ stand for the elevation (pitch), azimuth (yaw) and torsion (roll) angles of the left *L* and right *R* eye. The nose direction is the line orthogonal to the baseline and lying in a transverse plane passing through the eyes. The angles ε and ν stands for the binocular azimuth and vergence (see text for detailed explanation).

The 3D laser scanner. For the simulations shown in the following, we first captured 3D data from a real-world scene of the peripersonal space, by using a 3D laser scanner. To this purpose we selected a Konica Minolta Vivid 910, a fully portable sensor specifically designed for the acquisition of 3D images at close distance (in the range 0.6 ~ 2.5 m). The scanner combines a laser range-finder with a camera, providing digitized images with accurate distance as well as luminance information for every pixel in the scene.

The device emits a laser beam that scans the physical scene in equi-angular steps of elevation and azimuth relative to the center of the scanner. The distance of each point in the scene from the scanner is calculated by measuring the return time of the laser signal. The resulting range measurements generate a 3D representation of the scene, from the vantage point of the laser range finder. The device identifies the visible object surface at distances from 0.6 to 1.2 m, with an accuracy of 0.1 mm. The software provided by the vendor allows to acquire the same environment from different viewpoints, and to align and merge the different scans, in order to construct a full 3D representation of the scene. Each scan contained up to 307200 points acquired in 2.5 s. The device, providing three interchangeable lenses (focal distance: TELE 25 mm, MIDDLE 14 mm and WIDE 8 mm), allows for a variable angular field of view from $\approx$10 cm^2 to $\approx$1 m^2 (computed as the angular image area at near depth). Moreover, the device provides not just the point cloud, but also a polygonal mesh created with all connectivity information retained, thereby eliminating geometric ambiguities and improving detail capture. Furthermore, the scene camera also provides the color texture registered to the polygonal mesh, at a resolution of 640×480 pixels.

The acquisition of the single objects. Each object used for the scenes was first acquired separately, in order to obtain a full 360° model with no occlusions. The the laser scanner was used in TELE modality to obtain a high spatial resolution (approximately 30.000 3D points per cm^2). Each object was scanned from different positions and then the scans were aligned and merged to obtain the full model. The number of scans for each object ($\approx$20) varied according to the complexity and size of the object, and the position and orientation of the laser scanner is decided to reduce holes and occluded points in the scan data. The unconnected scans are aligned, first with a point-to-point manual procedure and then with an automated global registration procedure that minimizes the distance among the different scans. Finally, the merge procedure allows us to obtain the full connected model of the whole object. In general, a post-processing

phase is required to obtain the final model. We perform hole filling or smoothing where required, using a custom region selection tool. The â€œreduce noise' function smooths the data to reduce the acquisition noise, while preserving the shape and the original texture of the object. Then, the polygons that are not connected to the object are manually trimmed, and the 'fill holes' command is performed to ensure that no remaining holes are present in the model. The final model has a number of points ranging between 500.000 and 2.000.000, where each 3D point has its own color associated. Figure 1 shows some examples of the final object models.

The acquisition of the complete scenes. In order to recreate everyday living environments, we considered two cluttered scenes, an office table and a kitchen table (see Fig. 2). The real objects were mounted on a rotating table to facilitate the acquisition. The scenes were illuminated using a set of lamps at 6,500° K, to create diffuse illumination with white light, so to obtain texture objects as much similar as possible to the real one.

We proceeded following a simple protocol to generate the final scene models. First of all, we take 8 scans of the entire scenes with the scanner in WIDE modality, rotating the table by step of 45°, thus obtaining a full 360° model. The scans were then aligned (manual and automated registration procedure) and then merged, as for the single objects. The obtained raw model is characterized by a lower spatial and texture resolution with respect to the single object models, and by a large number of holes, due to the high complexity of the scene, where objects occlude each other.

After these steps, each single object model was aligned within the raw scene, with a procedure similar to the one used to align the different scans of a single object. Each object was first aligned within the full scene with a point-to-point manual procedure. The alignment was then refined using an automated global registration procedure that minimizes the distance between the high resolution object and its low resolution version in the full scene. Finally, the points belonging to the low resolution were manually selected and removed from the scene, recreating it as a composition of high resolution and high precision object models. The scene was then aligned with a right hand world coordinate system with metric unit, where the x axis is directed from left to right, the y axis is directed from down to up, perpendicular to the table surface, and the z axis points toward the viewer. The origin of the three axes of the coordinate system was located in the center of the table surface.

The scene, was exported as a clean VRML file. It is worth noting that, in this file format, the single objects are separated, allowing us to set specific material properties for each object.

The presented methodology was used to compose the two virtual environments, as in Fig. 2.

Stereo vision simulator of images and ground-truth data

The simulator generates the sequence of stereo image pairs, the depth maps with respect to both the cyclopic position and the left camera, and the exact pose of both the left and the right camera, by considering a binocular head moving in the 3D space and fixating a 3D point (X^F, Y^F, Z^F). The geometry of the system is shown in Fig. 3.

Once the initial position of the head is fixed, then different behaviours are possible:

- to move the eyes by keeping the head (position and orientation) fixed;
- to change the head orientation, thus mimicking neck movements;
- to change both the head orientation and position, thus generating more complex motion patterns.

Head movement. The head is characterized by the following parameters (each expressed with respect to the world reference frame (X^W, Y^W, Z^W)):

- head position $\mathbf{O}^{\mathbf{H}}$;
- nose direction $\mathbf{n} = \mathbf{n}(\alpha^H, \beta^H)$, function of the elevation α^H and azimuth β^H angles of the neck;
- fixation point $\mathbf{F} = (X^F, Y^F, Z^F)$.

Specifically, the nose direction is the line orthogonal to the baseline and lying in a transverse plane passing through the eyes. Note that the rotation of a rigid body, generally expressed by yaw, pitch and roll for rotations about the X, Y and Z axis, respectively, is here expressed as azimuth (yaw), elevation (pitch) and torsion (roll), to maintain a notation more familiar to eye movements studies. Note that the elevation (pitch), azimuth (yaw) and torsion (roll) define a rotation about the X, Y and Z axis, respectively. For the head, the elevational and azimuthal orientation are described by the following rotation matrices:

$$\mathbf{R}_{\alpha}^{H} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha^{L/R} & -\sin\alpha^{L/R} \\ 0 & \sin\alpha^{L/R} & \cos\alpha^{L/R} \end{bmatrix}$$
$$\mathbf{R}_{\beta}^{H} = \begin{bmatrix} \cos\beta^{L/R} & 0 & \sin\beta^{L/R} \\ 0 & 1 & 0 \\ -\sin\beta^{L/R} & 0 & \cos\beta^{L/R} \end{bmatrix} \quad (1)$$

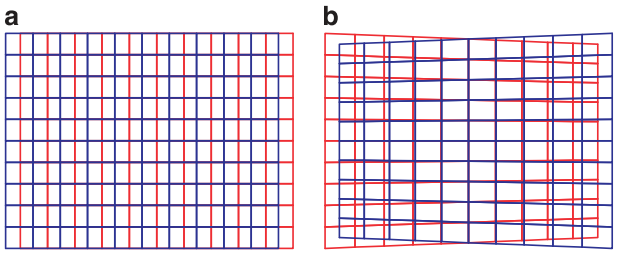

Figure 4. The projections of a fronto-parallel rectangle onto the image planes, drawn in red for the left image and blue for the right. The texture applied to the rectangle is a regular grid. (**a**) The projection obtained with the off-axis technique: only horizontal disparity is introduced. (**b**) The projection obtained with the toe-in technique: both vertical and horizontal disparities are introduced.

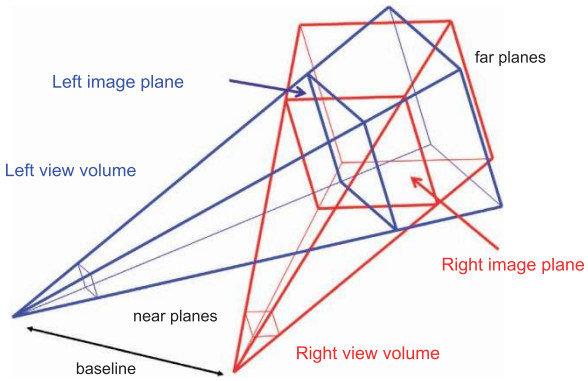

Figure 5. The two view volumes of the stereo cameras for the toe-in technique.

The complete rotation of the head is defined by composing in cascade the above rotations following a Fick gimbal system[102,103]:

$$\mathcal{R}^H = \mathbf{R}^H_\beta \mathbf{R}^H_\alpha \tag{2}$$

Since $\mathcal{R}^H$ is composed of rotations about the horizontal and vertical axes, only, the resulting head pose is characterized by no torsional component, according to the kinematics of the Fick gimbal system. From this, it is possible to write the pose matrix for the head as:

$$\mathcal{P}^H_W = \left[\begin{array}{c|c} \mathcal{R}^H & \mathbf{O^H} \\ \hline 0 & 1 \end{array}\right] \tag{3}$$

Toe-in technique to simulate eye movements. As anticipated, to render stereoscopic stimuli on a screen, two techniques exist: the *off-axis* and the *toe-in* techniques. The first one is used to obtain stereo 3D for a human observer, as for the cinema and television stereo 3D[104]. The disparity patterns produced by the off-axis and toe-in techniques, when observing a frontoparallel plane, are shown in Fig. 4a,b, respectively. In this context, the correct way to create the stereo pairs is the toe-in method: the projection direction of each camera is set to the target point (the fixation point **F**) through a proper rotation. Then the left and the right views project onto two different planes, (see Fig. 5). The cameras are characterized by the following parameters (each expressed with respect to the head reference frame):

- camera position $\mathbf{O}^{L/R/C}$;
- camera orientation $\mathcal{R}^{L/R/C} = \mathcal{R}^{L/R/C}(\alpha^{L/R/C}, \beta^{L/R/C})$, function of the elevation α and azimuth β angles;

Moreover, the cameras have pinhole optics with unitary focal length. The origin of the left and the right view volume is fixed at

$$\mathbf{T}^{L/R} = \left[\pm\frac{b}{2}\ 0\ 0\right]^T \tag{4}$$

while the cyclopic view volume is located at the origin of the head reference frame. The elevational and azimuthal orientation of the cameras are described by the rotation matrices in equation (2). To emulate the behavior of a couple of verging pan-tilt cameras the complete rotation of each camera is defined composing in cascade the above rotations following an Helmholtz gimbal system[103]:

$$\mathcal{R}^{L/R/C} = \mathbf{R}_{\alpha}^{L/R/C}\mathbf{R}_{\beta}^{L/R/C} \tag{5}$$

In this way, it is possible to insert a camera in the scene (e.g., a perspective camera), to obtain a stereoscopic representation with convergent axes and to decide the location of the fixation point.

Binocular coordination of eye/camera movements. A single eye/camera, like any rigid body, has three rotational degrees of freedom. Though, in a head centered reference frame, only two angles are sufficient to determine the gaze direction: namely the azimuth (yaw) and the elevation (pitch) of the target, as by equation (5). This implies that the eye/camera could, in principle, assume an infinite number of torsional poses for any gaze direction, while correctly fixating a given target.

Considering a human observer, the complete 3D pose of a single eye is defined (limited) by the Listing's law (LL), which specifies the amount of the torsional angle with respect to the gaze direction[105]. The ecological reason subtending LL is to provide a *motor advantage* to the visual system, through an optimization of the oculomotor control. In fact, according to LL, each eye would move always rotating along the shortest geodetic path from the primary position (i.e., straight ahead gaze direction).

The situation differs if we consider a binocular visual system, because a thoughtful motor coordination of the two eyes/cameras has a meaningful perceptual significance. The relative torsion of the two eyes, defined as the cyclotorsion, has in fact the goal of reducing the search zone for retinal correspondence[106–108], thus facilitating the problem of stereopsis[109]. Empirical evidences showed how the kinematics of the eyes slightly deviated from LL. To obtain perceptual advantages in binocular coordination, the torsional posture of each eye changes with eye convergence, particularly in close viewing[16,110–115]. This difference is defined as the binocular extension of Listing's Law (L2)[16,111,115,116]. The resulting posture provides an alignment of the horizontal meridians of the left and right eyes. This alignment reduces the search zone for retinal correspondence, thus providing a *perceptual optimization* of eye pose for stereopsis[106–108].

During convergence movements, once defined the starting vergence angle, the eyes' rotation axes remain confined to planes that rotate temporally and roughly symmetrically by ϕ_l and ϕ_r angle, for the left and the right eye, respectively. These convergence dependent changes of torsional positions (i.e., orientation of Listing's plane) have been referred to as the binocular extension of LL or, in brief, L2 (ref. 117). The grater is the convergence, the more the planes rotate temporally, implying that during convergence, there is a relative excyclotorsion on upgaze, and a relative incyclotorsion on downgaze. To mimic the natural eye pose, taking into account the tilting of the Listing's plane, we had to consider a torsion of the camera $\gamma^{L/R}$ given by[118]:

$$\frac{\tan\gamma^{L/R}}{2} = -\frac{\tan\alpha^{L/R}}{2}\left[\frac{\tan\phi^{L/R} + \tan\beta^{L/R}}{1 + \tan\phi^{L/R}\tan\beta^{L/R}}\right] \tag{6}$$

with

$$\phi^{L} = -\phi^{R} = \frac{\delta}{2}\arcsin\frac{\sin\frac{\nu}{2}}{\cos\frac{\xi}{2}} \tag{7}$$

where $\nu = \beta^R - \beta^L$ and $\xi = \frac{\beta^L + \beta^R}{2}$ are the vergence and azimuth angles, respectively. The parameter δ controls the balance between the motor advantage provided by LL, and the perceptual optimization for stereopsis, provided by L2. In all the following simulations we decided to follow previous literature and to adopt $\delta = 0.8$ (refs 62,63,108,114). Thus, the angle γ provide the amount of torsion (roll) to be applied to the left and the right cameras for a given azimuth and vergence distance, in order to obtain a camera pose compliant with the L2 law.

For taking into account the torsion γ of left and right cameras we had to pre-multiply the rotation matrices $\mathcal{R}^{L/R}$ by a torsional rotation matrix:

$$\mathbf{R}_{\gamma}^{L/R} = \begin{bmatrix} \cos\gamma^{L/R} & -\sin\gamma^{L/R} & 0 \\ \sin\gamma^{L/R} & \cos\gamma^{L/R} & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{8}$$

The complete rotation of the view-volumes is described by the following relation:

$$\mathcal{R}_{L2}^{L/R} = \mathcal{R}^{L/R}\mathbf{R}_{\gamma}^{L/R} \tag{9}$$

where the subscript $L2$ is used to indicate that it is Listing's law compliant matrix.

It is now possible to obtain the pose matrix for the cameras in head reference frame:

$$\mathcal{P}_H^{L/R/C} = \left[\begin{array}{c|c} \mathcal{R}_{L2}^{L/R/C} & \mathbf{O^{L/R/C}} \\ \hline 0 & 1 \end{array}\right] \tag{10}$$

and to compose the matrices to obtain the pose matrix in world reference frame:

$$\mathcal{P}_W^{L/R/C} = \mathcal{P}_W^H \mathcal{P}_H^{L/R/C} \tag{11}$$

Rendering of stereo pairs and ground-truth data. The scene is rendered in an on-screen OpenGL context. The available SoOffScreenRenderer class is used for rendering scenes in off-screen buffers and to save to disk the sequence of stereo pairs.

The renderer engine allows us to produce stereo images of different resolution and acquired by cameras with different field of views. In particular, the following 'optical' parameters can be set:

- resolution of the cameras (the maximum possible resolution depends on texture resolution and on the 3D point cloud density);
- horizontal and vertical field of view (HFOV and VFOV, respectively);
- distance from camera position to the near clipping plane in the camera's view volume, also referred to as a viewing frustum, (`nearDistance`);
- distance from camera position to the far clipping plane in the camera's view volume (`farDistance`);
- distance from camera position to the point of focus (`focalDistance`).

To compute the ground-truth data it is necessary to exploit the resources available from the graphics engine by combining them through the computer vision relationships that describe the geometry of two views, typically used to obtain a 3D reconstruction.

Formally, by considering two static views, the two camera reference frames are related by a rigid body transformation described by the rotation matrix $\mathcal{R}_{L2}^{L/R}$ and the translation $\mathcal{T} = \mathbf{T}^R - \mathbf{T}^L$. The x^L homogeneous coordinates on the left image plane are back-projected on the 3D scene, and the obtained 3D points are then re-projected on the right image plane, obtaining the associated x^R homogeneous coordinates by perspective division, in the following way[119]:

$$\lambda^R \mathrm{x}^R = \mathcal{R}_{L2}^R \left[\left(\mathcal{R}_{L2}^L\right)^{-1} \lambda^L \mathrm{x}^L - \mathcal{T} \right] \tag{12}$$

where λ^L and λ^R are the depth values.

To apply the relationship described in equation (12), we first read the depth map (w) of the camera through a specific method added in the `SoOffScreenRenderer` class, then we obtain the depth values with respect to the reference frame of the camera in the following way:

$$\lambda = \frac{f\ n}{w(f-n)-f} \tag{13}$$

where f and n represent the values of the far and the near planes of the virtual camera, respectively.

The simulator provides:

- the images saved from the off-screen buffers, for the left, right and cyclopic postions.
- the depth values, computed from the depth map by following equation (13), for the cyclopic position and for the left camera.

Though in general the disparity map of a stereo pair would be calculated with respect to one of the two image planes (left or right image)[9], to avoid asymmetry problems we decided to refer also to the image plane of a cyclopic camera[120], located in the mid point between the left and right cameras, pointing along the gaze line at the selected fixation point. Given the projection of a 3D virtual point on the cyclopic image plane, the disparity maps were computed by the correspondent projections in the left and right image planes. Starting from the image coordinate x^C of the cyclopic image and the depth values λ^C obtained by the depth map, the image coordinate x^L and x^R of the left and right view are obtained in the following way:

$$\begin{aligned} \lambda^L \mathrm{x}^L &= \mathcal{R}_{L2}^L \left[\left(\mathcal{R}^C\right)^{-1} \lambda^C \mathrm{x}^C - \mathbf{T}^L \right] \\ \lambda^R \mathrm{x}^R &= \mathcal{R}_{L2}^R \left[\left(\mathcal{R}^C\right)^{-1} \lambda^C \mathrm{x}^C - \mathbf{T}^R \right]. \end{aligned} \tag{14}$$

By exploiting the relationship of equation (12) and equation (14), given the knowledge of the depth values for a camera, and of the relative pose of the two cameras, we compute the image coordinates for both the views, and then we define the horizontal $d_x = x^R - x^L$ and the vertical $d_y = y^R - y^L$ disparities.

Due to the different position of the left and right cameras, some points in one image may happen not to be visible on the other image, depending on the 3D structure of the scene. Those points are defined as *occlusions*, and can be computed by the ground-truth disparity map, since the forward-mapped disparity would land at a location with a larger (nearer) disparity. Similarly, the ground-truth disparity can be used

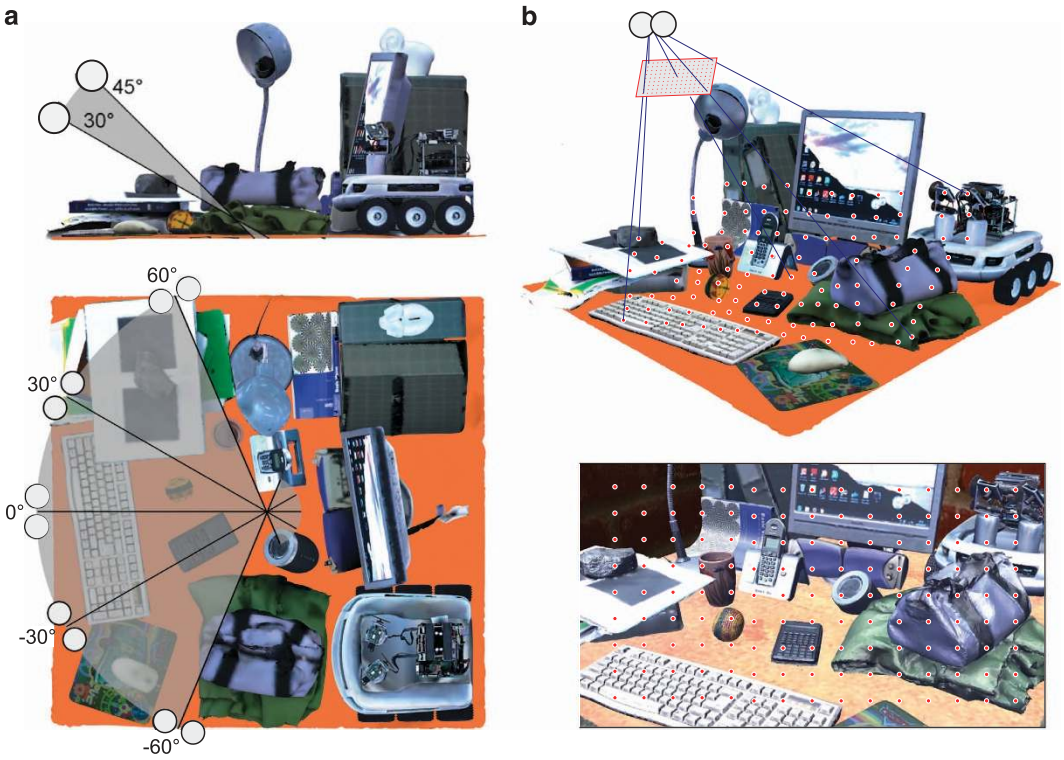

Figure 6. Representation of head position and 3D fixation point in the virtual scene. (**a**) Side and top view of the position of the 10 vantage points used to place the head within the 3D scene. The solid thick lines represent the nose direction for each head position. (**b**) Image acquired by the cyclopean camera from one of the ten vantage points (top), and geometrical configuration of the camera system with respect to the 3D scene (bottom). The red dots represent the 9×15 grid of points equally spaced on the image plane of the cyclopean camera (top), that are used to compute the actual 3D fixation points in the scene (bottom). The black solid line represents the nose direction, while the green lines represent the four most lateral fixations within the grid of fixation points.

to compute the depth discontinuity regions, i.e., those regions whose neighboring disparities differ by more than a defined threshold. On this basis, we computed two binary maps, one for the occluded points and one for the depth discontinuities.

3D fixation database

Once we obtained the VRML models of our 3D environment, we need to model a visual system able to explore this virtual environment, making fixation movements on the objects' surfaces.

For the two virtual worlds we considered 10 different vantage points for the subject head, corresponding to different positions and orientations of the nose direction of the head (see Fig. 6a). Facing the scene, the head was put with the nose direction pointing to its center. The head position orbited uniformly around center of the scene with an azimuth angle in the range $[-60°\ 60°]$ by steps of 30° and 2 elevation angles of 30° and 45°. In each of the vantage points, the head was then put at a fixed vergence distance of $\approx 3°$ ($= 155$ cm) from the closest object along the nose direction. The distance between the nodal points of the left and right, i.e., the baseline, is 60 mm. While stereoscopic algorithms are generally insensitive to the baseline, which works as a *scaling* factor, this baseline has been selected to be close to the interpupillary distance of a human being[121]. The fixation points were obtained by means of the cyclopic virtual camera centered in between the left and right cameras, and oriented along the same gaze line direction. A grid of 9×15 equally spaced image points was projected on the 3D scene (see Fig. 6b). The 3D coordinates of the fixation point in the scene were thus computed as the intersection between the binocular line of sight and the closest visible surface of the scene, i.e., the closest triangle of the model's mesh. The procedure was repeated for the two virtual scenes considered. The proposed approach is intended to provide an even and complete sampling of the visual space, thus we considered uniform spacing for both head position and gaze direction. Nevertheless, it is worth considering that human fixations are neither evenly nor randomly distributed within the environment. Generally, human fixation strategy is preferentially directed towards behaviorally significant points[122]. Specifically in a 3D scene, the distribution of vergence angle is biased towards closer points with respect to the

possible fixations within the visual scene[62], because close targets more likely and more immediately attract our gaze[20,22,23].

Active fixations on the scanned 3D scenes were simulated by accessing to the image textures and to the depth map. The depth ranged from ≈500 mm to ≈2,200 mm. For each of the $2 \times 10 \times 9 \times 15$ given camera poses we obtained the left and right retinal images and the horizontal and vertical cyclopic disparity maps. To eliminate bias due to the disposition of the objects in the scenes, we decided also to calculate, for each fixation, the mirrored cyclopic disparity maps. Mirrored disparity maps were always obtained from equation (14) mirroring the depth λ^C along the middle vertical line. Accordingly, we obtained a dataset of 5,400 binocular fixations, constituted by the left and right views and the associated disparity patterns.

Code availability

Together with the present dataset, the necessary code for its proper use will be made available at `http://www.pspc.unige.it/Code/index.html`, but can be also downloaded at `https://sourceforge.net/projects/genua-pesto-usage-code/`. The Matlab code has been developed on R2011a version and is compatible with all the subsequent versions. The C/C++ code has been developed in Unix environment and has been tested with Windows OS (Microsoft Visual Studio 2010). This code requires the *libpng* library (`http://www.libpng.org`) in order to load the *png* images of the database.

- *Data loading* (Matlab and C/C++): correct loading of images, depth maps and head/eye position;
- *Disparity computation* (Matlab and C/C++): computation of binocular disparity form the depth map;
- *Occlusion computation* (Matlab): computation of the ground-truth occlusion map from the disparity map;
- *Depth edges computation* (Matlab): computation of depth edges from disparity map;
- *Validation indexes computation* (Matlab): compute the indexes for validation and testing of disparity algorithm. The horizontal and vertical disparity estimation indexes are computed as the mean absolute error and standard deviation, with respect to the ground-truth disparity[9]. The image-based indexes are the MAE, NCC and SSI, described in the Technical Validation section.

Data Records

The database is available at `Dryad Digital Repository` (see Data Citation 1). We divided it in two sub-datasets, one for each of the two virtual worlds (see Fig. 2). Each sub-dataset is firstly divided according to the 10 different vantage points for the head. These are organized and stored in separate zip files using two numeric indexes, according to the following file name format: #scn HP #a #e, where:

- #scn defines the virtual world, K for the kitchen and O for the office,
- HP stands for *head position* (head azimuth and elevation),
- #a the head azimuth index, which is an integer from −2 to 2 that corresponds to angles in the range [−60°, 60°], with step of 30°,
- #e the head elevation index, which can assume value 1 and 2 corresponding to angles of 30° and 45°, respectively.

Within the above folders, for each of the 9×15 cyclopean image points the following data are stored:

- Stereo-pair images (left and right camera images) as PNG files ($1{,}921 \times 1{,}081$ pixels).
- Cyclopic images as PNG files ($1{,}921 \times 1{,}081$ pixels).
- Cyclopic camera depth map as PNG files ($1{,}921 \times 1{,}081$ pixels).
- Left camera depth map as PNG files ($1{,}921 \times 1{,}081$ pixels).
- Info file is provided in TXT format with all the geometrical information regarding the virtual stereo head for the actual fixation, i.e., the head position, head target and head orientation (world reference frame); the camera position and camera orientation (both world and head reference frame) for the left cyclopean and right cameras; the binocular gaze direction, expressed as version elevation and vergence (head reference frame), or as quaternion (both world and head reference frame). The file also contains the normalization values for the conversion of the depth map from PNG format to real depth value in mm.

The file names are defined according to the following format:

XX_HP_#*a*_#*e*_*H*_#*h*_*V*_#*v*;

where XX denotes the data name (e.g., *cycdepth* for the cyclopic depth map), HP_#a_#e recalls the information about the head position, as for the folder name. The indexes #h and #v describe the azimuth (from −7 to 7) and elevation (from −4 to 4) of the binocular fixation point within the 9×15 grid of gaze directions. The database is accompanied by two functions. The first, *Disparity_computation*, available both in Matlab and C++, takes as arguments a.txt info file and the associated cyclopic and left depth map PNG images and returns the following data (see Fig. 7):

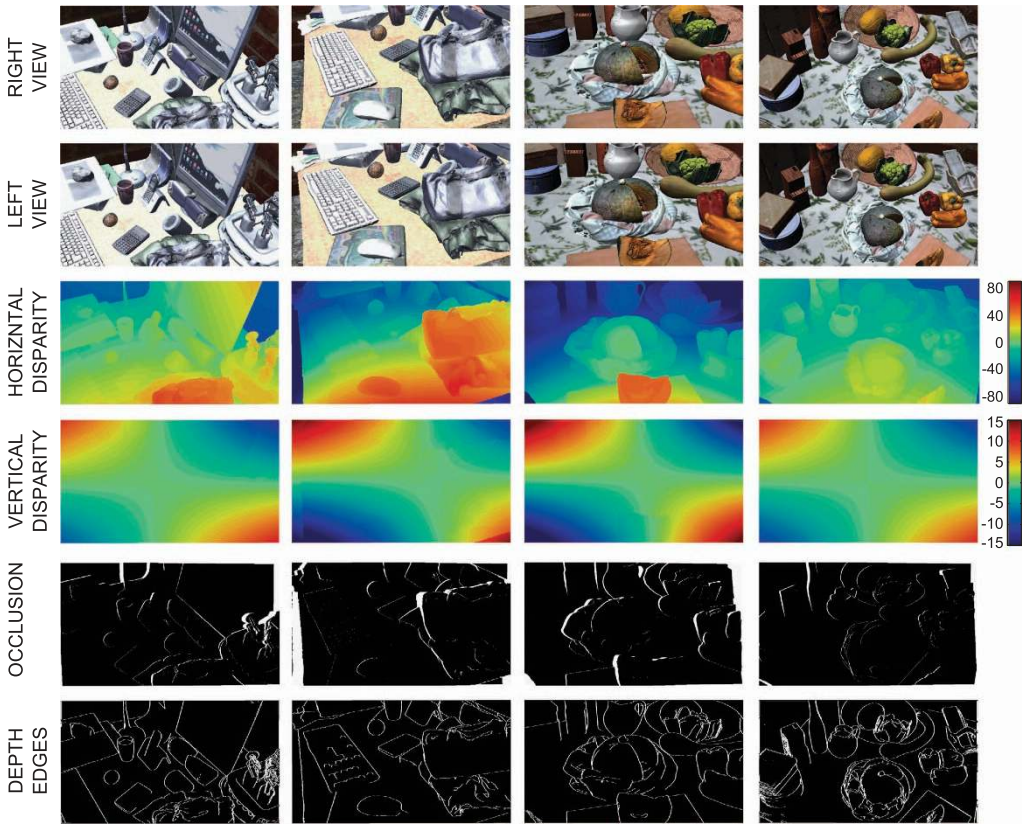

Figure 7. Example of stereoscopic pairs from the dataset, including, from top to bottom, the left and right views, the horizontal and vertical ground-truth disparity maps, and the occlusion and edge maps. In the disparity maps, reported in pixel, hot colors represent crossed horizontal disparity and right-hyper vertical disparity, whereas blue colors represent uncrossed horizontal disparities and left-hyper vertical disparities, according to the colorbars on the right.

- horizontal and vertical cyclopic disparity, stored in two separated binary files of $1{,}921 \times 1{,}081$ floating point values.
- horizontal and vertical left disparity, stored in two separated binary files of $1{,}921 \times 1{,}081$ floating point values.
- Cyclopic camera depth map in mm, stored in a binary file of $1{,}921 \times 1{,}081$ floating point values.
- Left camera depth map in mm, stored in a binary files of $1{,}921 \times 1{,}081$ floating point values.

The function, *compute_occlusion*, available in Matlab, takes as arguments the stereo pair and the associated horizontal and vertical ground-truth disparity, and returns the ground-truth occlusion mask for the actual stereo-pair, stored as a binary image, and the right image of the stereo pair, warped by the ground-truth disparity and removed the occluded pixels (see Fig. 7) The function, *compute_depth_edges* takes as arguments the horizontal and vertical ground-truth disparity maps and the depth maps, and computed the and depth edge map, stored as a binary image.

Technical Validation

In order to provide a technical validation of the released dataset, the two complementary parts that constitute the approach have to be considered separately: the 3D virtual models and the VR simulator. The former has to be tested with respect to the accuracy of the 3D models obtained, whereas for the latter we will provide a validation of the correctness of the geometrical projection and rendering of the stereoscopic pair.

3D virtual models: acquisition and registration error

On each single scan, the manufacturer of the used laser scanner guarantees a root mean square error of 1 mm. The post-processing procedure allows us to resolve every possible problem in the different scans, due to unwanted surface reflection and to acquisition noise. The manual alignment and automated registration of the different scans might introduce a larger error, if not performed accurately.

	Moneybox	Lamp	Pentray	Jug	Melon	Bottle	Mean
Maximum	2.971	6.599	1.260	2.757	2.654	4.030	3.578
Average	0.066	0.100	0.051	0.085	0.084	0.110	0.112
Std. Dev.	0.103	0.176	0.054	0.128	0.098	0.128	0.124

Table 1. Table summarizing registration errors, expressed in mm, on single 3D object models. The table reports the maximum, average and standard deviation of the error after the global registration procedure, for six 3D object models (see Fig. 1), randomly selected among those used to construct the virtual worlds. The last column reports the mean values over all object models.

Table 1 reports the maximum, average and standard deviation of the registration error in mm, measured on six objects randomly selected among all the scanned objects (see Fig. 1), and the mean values computed over all the object models used to assemble the scenes. While few 3D points with notable error might remain in the final model, the low average error and standard deviation provide a measurement of the quality of the obtained object models. In fact, the use of the TELE modality for the acquisition of the single scan, combined with a proper post-processing procedure allows us to obtain a spatial accuracy that is comparable to the accuracy of the device.

VR simulator: disparity reconstruction error

The obtained virtual worlds are used within the virtual simulator to generate the stereoscopic pairs as well as the depth and ground-truth disparity maps. The disparity map can be interpreted as the transformation to obtain, from a pixel on the left image, the corresponding pixel on the right image. From this perspective, in order to assess the correctness of our approach, it is possible to use the image quality indexes that are commonly used to evaluate the errors between a distorted image and a reference image (see[69], as review).

In stereo vision, two common approaches are used to evaluate the performance of disparity estimation algorithms (see[9]), a disparity-based and an image-based approach. The former relies on a ground-truth knowledge of the disparity, and computes indexes like the mean absolute error or the standard deviation with respect to the estimated map. If the ground truth disparity is not available, it is possible to warp the right image by the binocular disparity, in order to 'reconstruct' the left image, and compute indexes of similarity between the left (original) and right (warped) images.

Considering that our methodology provides the ground-truth disparity, if the geometrical projection and the rendering engine are correct, it should be possible to effectively reconstruct the left image by the right one, with negligible error. We thus used a bilinear interpolation algorithm, and we evaluated the obtained results over the whole dataset, using three error parameters sensitive to different image features:

- the mean absolute error (MAE)[9,70] is computed at pixel level, by averaging the absolute intensity differences of original and reconstructed images,
- the normalized cross correlation (NCC)[9], being the 2D version of the Pearson product-moment correlation coefficient, it provides a similarity metric that is invariant to changes in location and scale in the two images,
- the structure similarity index (SSIM) has been conceived to mimic the visual quality assessment methods of the human visual system, and is insensitive also to local luminance and contrast variations between the two images[68,69].

Figure 8 shows the median (solid thick line) of the three indexes computed over the whole stereo pair dataset, together with the first and third quartile (box), and the range (whiskers). Although the images have been rendered in color (see Fig. 7), the indexes have been computed on the gray level images (from 0 to 255), for the sake of simplicity. The indexes were first computed on the original stereo pair (ORIG), in order to obtain a reference value, and between the left original image and the warped right image (WARP). It is clearly evident how the warping reduces the absolute error and increases the correlation and the structure similarity between the pairs. In order further to validate the approach, it is worth considering that in stereoscopic vision some region on one image are occluded in the other one, and should be removed from the computation[9]. From this perspective, also the depth edges should not be considered since the image might suffer of rendering problem alongside the edges[9]. Hence, the computation was also performed on three different image regions over the warped pair: not considering the occluded areas (NO OCC), not considering both the occluded areas and the depth edges (NO DE), and finally considering only the pixels corresponding to the occluded areas and the depth edges (OCC). Since occluded regions and depth edges reasonably suffer of large errors, removing them provides a marked positive effect on the reconstructed stereo pair. In fact, the MAE is considerably reduced, and both the NCC and SSIM tend to their maximum. The removal of depth edges results in a further improvement, providing reconstructed images with a negligible error (< 0.7 gray levels), high correlation (>0.997) and large structure similarity (>0.95).

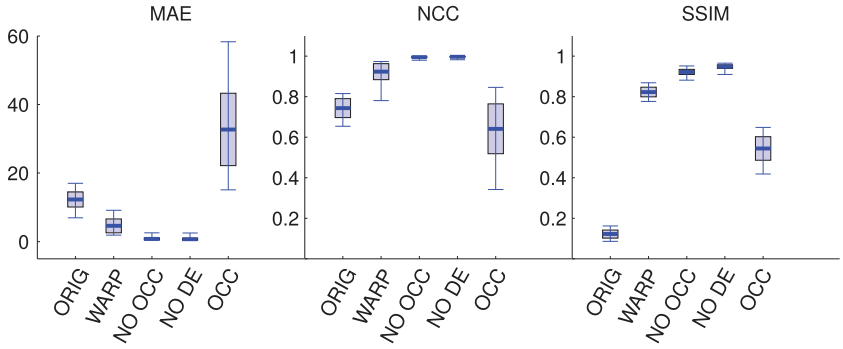

Figure 8. Disparity reconstruction error, computed on three different stereo quality indexes over the whole dataset, and represented as median value (horizontal thick line), first and third quartile (rectangle) and range (whiskers). Three different indexes are represented, from left to right: the Mean Absolute Error (MAE[9,70]), the Normalized Cross Correlation (NCC[9]) and the Structure Similarity Index (SSIM[68,69]). Each index has been computed for the original stereo pair (ORIG), not considering the occluded areas (NO OCC), not considering both the occluded areas and the depth edges (NO DE), and finally considering only the pixels corresponding to the occluded areas and the depth edges (OCC).

Summarizing, the proposed methodology allows us to: (1) obtain 3D geometric models of real scenes with high spatial fidelity, and (2) render realistic stereoscopic pairs with ground-truth disparity characterized by accurate perspective geometry, thus assessing the correctness of the released dataset and validating the approach.

References

1. Mian, A. S., Bennamoun, M. & Owens, R. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28,** 1584–1601 (2006).
2. Browatzki, B., Fischer, J., Graf, B., Bülthoff, H. H. & Wallraven, C. Going into depth: Evaluating 2d and 3d cues for object classification on a new, large-scale object dataset. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1189–1195 (IEEE, 2011).
3. Anand, A., Koppula, H. S., Joachims, T. & Saxena, A. Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research*, **32,** 19–34 (2012).
4. Koppula, H. S., Anand, A., Joachims, T. & Saxena, A. Semantic labeling of 3d point clouds for indoor scenes. In *Advances in Neural Information Processing Systems*, pages 244–252 (2011).
5. Su, C., Bovik, A. C. & Cormack, L. K. Natural scene statistics of color and rangeIn *2011 18th IEEE International Conference on Image Processing*, pages 257–260 (IEEE, 2011).
6. Su, C., Cormack, L. K. & Bovik, A. C. Color and depth priors in natural images. *IEEE Transactions on Image Processing* **22,** 2259–2274 (2013).
7. Janoch, A. *et al.* A category-level 3d object dataset: Putting the kinect to workIn *Consumer Depth Cameras for Computer Vision*, 141–165 (Springer, 2013).
8. Xiao, J., Owens, A. & Torralba, A. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1625–1632 (2013).
9. Scharstein, D. & Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision* **47,** 7–42 (2002).
10. Kim, H. & Hilton, A. Influence of colour and feature geometry on multi-modal 3d point clouds data registrationIn *3D Vision (3DV), 2014 2nd International Conference on* volume 1, pages 202–209 (IEEE, 2014).
11. Geiger, A., Lenz, P., Stiller, C. & Urtasun, R. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* **32,** 1231–1237 (2013).
12. Geiger, A. N., Lenz, P. & Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361 (IEEE, 2012).
13. Adams, W. J. *et al.* The southampton-york natural scenes (syns) dataset: Statistics of surface attitude. *Scientific Reports* **6,** 35805 (2016).
14. Beira, R. *et al.* Design of the robot-cub (icub) head. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 94–100 (IEEE, 2006).
15. Hansard, M. & Horaud, R. Patterns of binocular disparity for a fixating observerIn *International Symposium on, Vision, and Artificial IntelligenceBrain*, pages 308–317 (Springer, 2007).
16. Mok, D., Ro, A., Cadera, W., Crawford, J. D. & Vilis, T. Rotation of Listing's plane during vergence. *Vision Research* **32,** 2055–2064 (1992).
17. Xu, J., Yang, Q. & Feng, Z. Occlusion-aware stereo matching. *International Journal of Computer Vision*, pages 1–16 (2016).
18. Bohil, C. J., Alicea, B. & Biocca, F. A. Virtual reality in neuroscience research and therapy. *Nature Reviews Neuroscience* **12,** 752–762 (2011).
19. Gautier, J. & Le Meur, O. A time-dependent saliency model combining center and depth biases for 2d and 3d viewing conditions. *Cognitive Computation* **4,** 141–156 (2012).
20. Jansen, L., Onat, S. & König, P. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision* **9,** 29 (2009).
21. Liu, Y., Bovik, A. C. & Cormack, L. K. Disparity statistics in natural scenes. *Journal of Vision* **8,** 19 (2008).
22. Wexler, M. & Ouarti, N. Depth affects where we look. *Current Biology* **18,** 1872–1876 (2008).
23. Wismeijer, D. A., Erkelens, C. J., van Ee, R. & Wexler, M. Depth cue combination in spontaneous eye movements. *Journal of Vision* **10,** 25 (2010).

24. Yang, Z. & Purves, D. A statistical explanation of visual space. *Nature neuroscience* **6,** 632–640 (2003).
25. Fang, Y., Wang, J., Narwaria, M., Le Callet, P. & Lin, W. Saliency detection for stereoscopic images. *Image Processing, IEEE Transactions on* **23,** 2625–2636 (2014).
26. Huynh-Thu, Q. & Schiatti, L. Examination of 3d visual attention in stereoscopic video contentIn *IS&T/SPIE Electronic Imaging*, pages 78650J–78650J (International Society for Optics and Photonics, 2011).
27. Khaustova, D., Fournier, J., Wyckens, E. & Le Meur, O. How visual attention is modified by disparities and textures changes?In *IS&T/SPIE Electronic Imaging*, pages 865115–865115 (International Society for Optics and Photonics, 2013).
28. Kollmorgen, S., Nortmann, N., Schröder, S. & König, P. Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Computational Biology* **6,** e1000791 (2010).
29. Lang, C. *et al.* Depth matters: Influence of depth cues on visual saliency. In *Computer Vision-ECCV 2012*, pages 101–115 (Springer, 2012).
30. Onat, S., Açk, A., Schumann, F. & König, P. The contributions of image content and behavioral relevancy to overt attention. *PloS One* **9,** e93254 (2014).
31. Wang, J., Le Callet, P., Tourancheau, S., Ricordel, V. & Da Silva, M. P. Study of depth bias of observers in free viewing of still stereoscopic synthetic stimuli. *Journal of Eye Movement Research* **5,** pp-1 (2012).
32. Wolfe, J. M. & Horowitz, T. S. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience* **5,** 495–501 (2004).
33. Ban, H. & Welchman, A. E. fmri analysis-by-synthesis reveals a dorsal hierarchy that extracts surface slant. *The Journal of Neuroscience* **35,** 9823–9835 (2015).
34. Girshick, A. R. & Banks, M. S. Probabilistic combination of slant information: weighted averaging and robustness as optimal percepts. *Journal of Vision* **9,** 8–8 (2009).
35. Gumming, B. G., Johnston, E. B. & Parker, A. J. Effects of different texture cues on curved surfaces viewed stereoscopically. *Vision Research* **33,** 827–838 (1993).
36. Knill, D. C. Robust cue integration: A bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *Journal of Vision* **7,** 5–5 (2007).
37. Murphy, A. P., Ban, H. & Welchman, A. E. Integration of texture and disparity cues to surface slant in dorsal visual cortex. *Journal of Neurophysiology* **110,** 190–203 (2013).
38. Rogers, B. & Cagenello, R. Disparity curvature and the perception of three-dimensional surfaces. *Nature* **339,** 135–137 (1989).
39. Rosenberg, A., Cowan, N. J. & Angelaki, D. E. The visual representation of 3d object orientation in parietal cortex. *The Journal of Neuroscience* **33,** 19352–19361 (2013).
40. van Ee, R. & Erkelens, C. J. Temporal aspects of stereoscopic slant estimation: An evaluation and extension of howard and kaneko's theory. *Vision Research* **38,** 3871–3882 (1998).
41. Durand, J. *et al.* Anterior regions of monkey parietal cortex process visual 3d shape. *Neuron* **55,** 493–505 (2007).
42. Murata, A., Gallese, V., Luppino, G., Kaseda, M. & Sakata, H. Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area aip. *Journal of Neurophysiology* **83,** 2580–2601 (2000).
43. Orban, G. A., Janssen, P. & Vogels, R. Extracting 3d structure from disparity. *Trends in Neurosciences* **29,** 466–473 (2006).
44. Van Dromme, I. C., Premereur, E., Verhoef, B., Vanduffel, W. & Janssen, P. Posterior parietal cortex drives inferotemporal activations during three-dimensional object vision. *PLoS Biol* **14,** e1002445 (2016).
45. Verhoef, B., Bohon, K. S. & Conway, B. R. Functional architecture for disparity in macaque inferior temporal cortex and its relationship to the architecture for faces, color, scenes, and visual field. *The Journal of Neuroscience* **35,** 6952–6968 (2015).
46. Sherstyuk, A., Dey, A., Sandor, C. & State, A. Dynamic eye convergence for head-mounted displays improves user performance in virtual environments. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 23–30 (ACM, 2012).
47. Sherstyuk, A. & State, A. Dynamic eye convergence for head-mounted displaysIn *Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology*, pages 43–46 (ACM, 2010).
48. State, A., Ackerman, J., Hirota, G., Lee, J. & Fuchs, H. Dynamic virtual convergence for video see-through head-mounted displays: maintaining maximum stereo overlap throughout a close-range work space. In *Augmented Reality, 2001. Proceedings. IEEE and ACM International Symposium on*, pages 137–146 (IEEE, 2001).
49. Canessa, A., Chessa, M., Gibaldi, A., Sabatini, S. P. & Solari, F. Calibrated depth and color cameras for accurate 3d interaction in a stereoscopic augmented reality environment. *Journal of Visual Communication and Image Representation* **25,** 227–237 (2014).
50. Chessa, M., Maiello, G., Borsari, A. & Bex, P. J. The perceptual quality of the oculus rift for immersive virtual reality. *Human-Computer Interaction*, 1–32 (2016).
51. Chessa, M. *et al.* Veridical perception of 3d objects in a dynamic stereoscopic augmented reality system. In *Computer Vision, Imaging and Computer Graphics. Theory and Application*, pages 274–285 (Springer, 2013).
52. Hanhart, P. & Ebrahimi, T. Subjective evaluation of two stereoscopic imaging systems exploiting visual attention to improve 3d quality of experience. In *IS&T/SPIE Electronic Imaging*, 90110D–90110D (International Society for Optics and Photonics, 2014).
53. Moorthy, A. K., Su, C., Mittal, A. & Bovik, A. Subjective evaluation of stereoscopic image quality. *Signal Processing: Image Communication* **28,** 870–883 (2013).
54. Shao, F. *et al.* Binocular energy response based quality assessment of stereoscopic images. *Digital Signal Processing* **29,** 45–53 (2014).
55. Hibbard, P. B. A statistical model of binocular disparity. *Visual Cognition* **15,** 149–165 (2007).
56. Hunter, D. W. & Hibbard, P. B. Distribution of independent components of binocular natural images. *Journal of Vision* **15,** 6–6 (2015).
57. Liu, Y., Cormack, L. K. & Bovik, A. C. Dichotomy between luminance and disparity features at binocular fixations. *Journal of Vision* **10,** 23 (2010).
58. Prince, S. J. D. & Eagle, R. A. Weighted directional energy model of human stereo correspondence. *Vision Research* **40,** 1143–1155 (2000).
59. Read, J. Early computational processing in binocular vision and depth perception. *Progress in Biophysics and Molecular Biology* **87,** 77–108 (2005).
60. Read, J. C. A bayesian approach to the stereo correspondence problem. *Neural Computation* **14,** 1371–1392 (2002).
61. Read, J. C. A. & Cumming, B. G. Understanding the cortical specialization for horizontal disparity. *Neural Computation* **16,** 1983–2020 (2004).
62. Sprague, W. W., Cooper, E. A., Tosić, I. & Banks, M. S. Stereopsis is adaptive for the natural environment. *Science Advances* **1,** e1400254 (2015).
63. Gibaldi, A., Canessa, A. & Sabatini, S. P. The Active Side of Stereopsis: Fixation Strategy and Adaptation to Natural Environments. *Scientific Reports*, doi:10.1038/srep44800 (2017).
64. Hoyer, P. O. & Hyvärinen, A. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation In Neural Systems* **11,** 191–210 (2000).

65. Hunter, D. W. & Hibbard, P. B. Ideal binocular disparity detectors learned using independent subspace analysis on binocular natural image pairs. *PLoS ONE* **11,** e0150117 (2016).
66. Okajima, K. Binocular disparity encoding cells generated through an infomax based learning algorithm. *Neural Networks* **17,** 953–962 (2004).
67. Tippetts, B., Lee, D. J., Lillywhite, K. & Archibald, J. Review of stereo vision algorithms and their suitability for resource-limited systems. *Journal of Real-Time Image Processing* **11,** 5–25 (2016).
68. Sheikh, H. R. & Bovik, A. C. Image information and visual quality. *IEEE Transactions on Image Processing* **15,** 430–444 (2006).
69. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13,** 600–612 (2004).
70. Xia, Y., Zhi, J., Huang, M. & Ma, R. Reconstruction error images in stereo matching. In *Automation and Logistics, 2008. ICAL 2008. IEEE International Conference on*, pages 460–463 (IEEE, 2008).
71. Bleyer, M. & Gelautz, M. Graph-based surface reconstruction from stereo pairs using image segmentation. In *Electronic Imaging 2005*, pages 288–299 (International Society for Optics and Photonics, 2005).
72. Zitnick, C. L. & Kang, S. B. Stereo for image-based rendering using image over-segmentation. *International Journal of Computer Vision* **75,** 49–65 (2007).
73. Antonelli, M., Del Pobil, A. P. & Rucci, M. Depth estimation during fixational head movements in a humanoid robot. In *International Conference on Computer Vision Systems*, pages 264–273 (Springer, 2013).
74. Beuth, F., Wiltschut, J. & Hamker, F. Attentive stereoscopic object recognition. In *Workshop New Challenges in Neural Computation 2010*, page 41 (Citeseer, 2010).
75. Rasolzadeh, B., Björkman, M., Huebner, K. & Kragic, D. An active vision system for detecting, fixating and manipulating objects in the real world. *The International Journal of Robotics Research* **29,** 133–154 (2010).
76. Devernay, F. & Faugeras, O. D. Computing differential properties of 3-d shapes from stereoscopic images without 3-d models. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 208–213 (IEEE, 1994).
77. Hoff, W. & Ahuja, N. Surfaces from stereo: Integrating feature matching, disparity estimation, and contour detection. *IEEE Transactions On Pattern Analysis And Machine Intelligence* **11,** 121–136 (1989).
78. Baek, E. & Ho, Y. Occlusion and error detection for stereo matching and hole-filling using dynamic programming. *Electronic Imaging* (2016); 1–6 (2016).
79. Huq, S., Koschan, A. & Abidi, M. Occlusion filling in stereo: Theory and experiments. *Computer Vision and Image Understanding* **117,** 688–704 (2013).
80. Ishikawa, H. & Geiger, D. Occlusions, discontinuities, and epipolar lines in stereo. In *European Conference on Computer Vision*, pages 232–248 (Springer, 1998).
81. Min, D. & Sohn, K. Cost aggregation and occlusion handling with wls in stereo matching. *IEEE Transactions on Image Processing* **17,** 1431–1442 (2008).
82. Nakamura, Y., Matsuura, T., Satoh, K. & Ohta, Y. Occlusion detectable stereo-occlusion patterns in camera matrix. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on*, pages 371–378 (IEEE, 1996).
83. Sun, J., Li, Y., Kang, S. B. & Shum, H. Symmetric stereo matching for occlusion handling. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* volume 2, pages 399–406 (IEEE, 2005).
84. Zhu, Z., Stamatopoulos, C. & Fraser, C. S. Accurate and occlusion-robust multi-view stereo. *ISPRS Journal of Photogrammetry and Remote Sensing* **109,** 47–61 (2015).
85. Lei, J., Zhang, H., You, L., Hou, C. & Wang, L. Evaluation and modeling of depth feature incorporated visual attention for salient object segmentation. *Neurocomputing* **120,** 24–33 (2013).
86. Wang, J., Fang, Y., Narwaria, M., Lin, W. & Le Callet, P. Stereoscopic image retargeting based on 3d saliency detection. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 669–673 (IEEE, 2014).
87. Wang, J., DaSilva, M. P., LeCallet, P. & Ricordel, V. Computational model of stereoscopic 3d visual saliency. *Image Processing, IEEE Transactions on* **22,** 2151–2165 (2013).
88. Dang, T., Hoffmann, C. & Stiller, C. Continuous stereo self-calibration by camera parameter tracking. *IEEE Transactions on Image Processing* **18,** 1536–1550 (2009).
89. Bjorkman, M. & Eklundh, J. Real-time epipolar geometry estimation of binocular stereo heads. *IEEE Transactions on pattern analysis and machine intelligence* **24,** 425–432 (2002).
90. Chai, J. & De Ma, S. Robust epipolar geometry estimation using genetic algorithm. *Pattern Recognition Letters* **19,** 829–838 (1998).
91. Lu, J., Cai, H., Lou, J. & Li, J. An epipolar geometry-based fast disparity estimation algorithm for multiview image and video coding. *IEEE Transactions on Circuits and Systems for Video Technology* **17,** 737–750 (2007).
92. Papadimitriou, D. V. & Dennis, T. J. Epipolar line estimation and rectification for stereo image pairs. *IEEE Transactions On Image Processing* **5,** 672–676 (1996).
93. Zhang, Z. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision* **27,** 161–195 (1998).
94. Torr, P. & Zisserman., A. Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding* **78,** 138–156 (2000).
95. Zhang, Z., Deriche, R., Faugeras, O. R. & Luong, Q. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence* **78,** 87–119 (1995).
96. Santini, F. & Rucci, M. Active estimation of distance in a robotic system that replicates human eye movement. *Robotics and Autonomous Systems* **55,** 107–121 (2007).
97. Hartley, R. & Zisserman., A. *Multiple View Geometry in Computer Vision* (Cambridge university press, 2003).
98. Tang, C., Medioni, G. & Lee, M. N-dimensional tensor voting and application to epipolar geometry estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23,** 829–844 (2001).
99. Faugeras, O., Luong, Q. & Papadopoulo, T. *The geometry of multiple images: the laws that govern the formation of multiple images of a scene and some of their applications* (MIT press, 2004).
100. Han, J. & Park, J. Contour matching using epipolar geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22,** 358–370 (2000).
101. Mikolajczyk, K. *et al.* A comparison of affine region detectors. *International Journal of Computer Vision* **65,** 43–72 (2005).
102. Glenn, B. & Vilis, T. Violations of Listing's law after large eye and head gaze shifts. *Journal of Neurophysiology* **68,** 309–318 (1992).
103. Haslwanter, T. Mathematics of three-dimensional eye rotations. *Vision Research* **35,** 1727–1739 (1995).
104. Held, R. T. & Banks, M. S. Misperceptions in stereoscopic displays: a vision science perspective. In *Proceedings of the 5th symposium on Applied Perception in Graphics and Visualization*, pages 23–32 (ACM, 2008).
105. Tweed, D. & Vilis, T. Geometric relations of eye position and velocity vectors during saccades. *Vision Research* **30,** 111–127 (1990).
106. Schreiber, K., Crawford, J. D., Fetter, M. & Tweed, D. The motor side of depth vision. *Nature* **410,** 819–822 (2001).

107. Schreiber, K. M., Hillis, J. M., Filippini, H. R., Schor, C. M. & Banks., M. S. The surface of the empirical horopter. *Journal of Vision* **8,** 7 (2008).
108. Schreiber, K. M., Tweed, D. B. & Schor., C. M. The extended horopter: Quantifying retinal correspondence across changes of 3d eye position. *Journal of Vision* **6,** 6 (2006).
109. Gibaldi, A., Vanegas, M., Canessa, A. & Sabatini, S. P. A portable bio-inspired architecture for efficient robotic vergence control. *International Journal of Computer Vision*, pages 1–22 (2016).
110. Bruno, P. & Van den Berg, A. V. Relative orientation of primary position of the two eyes. *Vision Research* **37,** 935–947 (1997).
111. Minken, A. W. H. & Van Gisbergen, J. A. M. A three dimensional analysis of vergence movements at various level of elevation. *Exp. Brain Res.* **101,** 331–345 (1994).
112. Porrill, J., Ivins, J. P. & Frisby, J. P. The variation of torsion with vergence and elevation. *Vision Research* **39,** 3934–3950 (1999).
113. Somani, R. A. B., Desouza, J. F. X., Tweed, D. & Vilis, T. Visual test of Listing's Law during vergence. *Vision Research* **38,** 911–923 (1998).
114. Tweed, D. Visual-motor optimization in binocular control. *Vision Research* **37,** 1939–1951 (1997).
115. Van Rijn, L. J. & Van den Berg, A. V. Binocular eye orientation during fixations: Listing's law extended to include eye vergence. *Vision Research* **33,** 691–708 (1993).
116. Maxwell, J. S. & Schor, C. M. The coordination of binocular eye movements: Vertical and torsional alignment. *Vision Research* **46,** 3537–3548 (2006).
117. Wong, A. Listing's law: clinical significance and implications for neural control. *Surv. Ophthalmol.* **49,** 563–575 (2004).
118. Gibaldi, A., Canessa, A., Chessa, M., Solari, F. & Sabatini, S. P. A neural model for coordinated control of horizontal and vertical alignment of the eyes in three-dimensional space. In *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 955–960 (IEEE, 2012).
119. Ma, Y., Soatto, S., Kosecka, J. & Sastry, S. *An Invitation to 3D Vision. From Images to Geometric Models* (Springer-Verlag, 2004).
120. Erkelens, C. J. & Van Ee, R. The role of the cyclopean eye in vision: sometimes inappropriate, always irrelevant. *Vision Research* **42,** 1157–1163 (2002).
121. Dodgson, N. A. Variation and extrema of human interpupillary distance. In *Electronic Imaging 2004*, pages 36–46 (International Society for Optics and Photonics, 2004).
122. Land, M. F. & Hayhoe, M. In what ways do eye movements contribute to everyday activities? *Vision Research* **41,** 3559–3565 (2001).

Data Citation

1. Canessa, A. *et al. Dryad Digital Repository*. http://dx.doi.org/10.5061/dryad.6t8vq (2016).

Acknowledgements

The authors would like to gratefully thank Prof. Marty Banks for the technical and expository comments. This work has been partially supported by the EC Project FP7-ICT-217077 'EYESHOTS—Heterogeneous 3D perception across visual fragments'.

Author Contributions

A.C.: study design, 3D data acquisition and processing, extended software framework to vergent geometry, writing of manuscript. A.G.: study design, 3D data acquisition and processing, technical validation, writing and revision of manuscript. A.C. and A.G. contributed equally to this work. M.C.: developed software framework, writing of manuscript. M.F.: provided conceptual discussion, revision of manuscript, obtained funding. F.S.: developed software framework, writing of manuscript. S.P. S.: conceived research, study design, writing and revision of manuscript, provided conceptual discussion, obtained funding.

Additional Information

Competing interests: The authors declare no competing financial interests.

How to cite this article: Canessa, A. *et al.* A dataset of stereoscopic images and ground-truth disparity mimicking human fixations in peripersonal space. *Sci. Data* 4:170034 doi: 10.1038/sdata.2017.34 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.