Depth Perception Through the Use of Image Sensor Arrays

Victor TOPORAN

1 STATE OF THE ART

1.1 DEPTH PERCEPTION

Obtaining depth perception data from an image sensor can be a daunting task, due to the nature of the media used in the process. In order to be able to assess 2-dimensional image from a 3-dimensional perspective, the best example that can be used is human vision. While cameras operate in a similar fashion to our eyes, with light information being converted into electrical impulses, the difference in understanding 3-dimensional space comes from our ability to interpret several visual cues that we associate with a difference in depth.

Those cues can relate to the relative positioning of the objects within a given scene (occlusion and size difference), their appearance (atmospheric perspective and texture gradient), the dynamic of the scene (motion parallax), or the perspective of the viewer (convergence and stereopsis) [1].

For the case of occlusion and size difference, although the importance of the cues are relevant for humans due to their experience, programming them into an autonomous system could lead to potential false positive results, such as the following case:

Figure 1.1: Occlusion mislead. [1]

In order to make use of atmospheric perspective, the scene depth needs to be in the order of kilometers, such as the following:

Figure 1.2: Atmospheric perspective. [1]

and for the texture gradient to be relevant it needs to be readily noticeable:

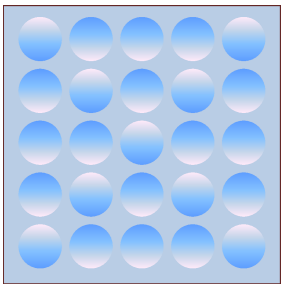

Figure 1.3: Texture gradient. [1]

Thus, in order to make use of those two cues, the scene composition needs to be curated accordingly, ruling out broader applications.

Motion parallax refers to the difference in perceived motion, based on the proximity to the viewer. Closer entities appear to be moving faster, in relation to background objects that seem slower or even still, depending on the distance to the camera. In the case of still images, this phenomenon manifests as different degrees of blur, depending on the velocity of each object [1]:

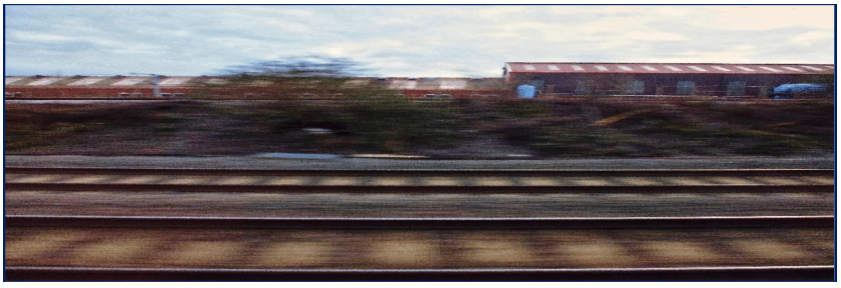

Figure 1.4: Parallax effect. [1]

While the blur could be quantized in order to obtain image depth, the use cases are again fringe, as the scene needs to be dynamic.

Used by architects and artists alike, perspective convergence refers to the use of converging lines in order to denote the depth of a scene. This leads to convergence being a relevant clue in most cases, as perspective line can be derived from the scene composition:

Figure 1.5: Convergence. [1]

However, convergence can interfere with the interpretation of other cues, such as size difference, leading to optical illusions [1]:

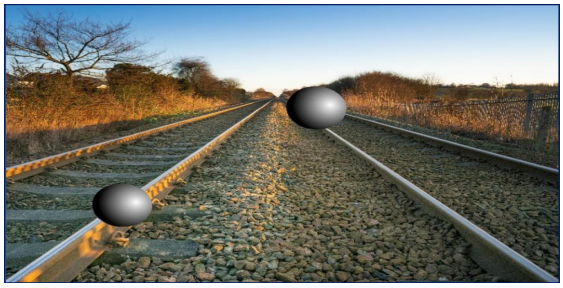

Figure 1.6: Convergence and size illusion. [1]

This leads to stereopsis as the sole viable candidate for depth perception, as it relies on the comparison of two slightly different images, with the disparity between them being interpreted as 3-dimensional cues [1]:

Figure 1.7: Stereopsis. [1]

While it may not be reliable for cases where image depth is so great that the disparity becomes

irrelevant, stereopsis can provide valid data for a myriad of applications, and at the same time be easily implementable as an embedded system, through the use of separate image sensors with a common control module, similar in may ways to human vision.

1.2 STEREO VISION

As it is cumbersome to extract 3-dimensional information from a monocular perspective without specialized artificial intelligence tools [2], as well as the same perspective leading to different results on different cameras, stereopsis became one of the most popular methods for depth perception. [3]

The baseline principle for stereo vision is "epipolar geometry", relying on the input of two cameras with different positions observing the same scene, resulting in two separate perspectives of the same subject. Under those conditions, "epipolar lines" can be derived from the analysis of a certain point in space and its projections onto the two images citewithApplications:

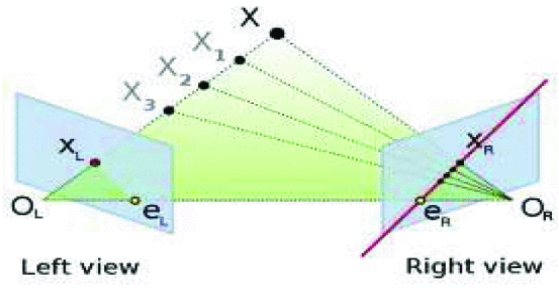

Figure 1.8: Epipolar geometry. [3]

In the case above, the "epipolar lines are a function of the 3D point X, and for every 3D point, there is a set of epipolar lines in both images". O_L and O_R represent the center of projection for each of the cameras, with X_L and X_R respectively being the image points corresponding to X. The projection of each center onto the other camera's image represent the "epipolar points", eL and e_R, and they are aligned with the centers of projection on a 3-dimensional line, outside of the images. Due to the alignment of points O_L, X_L and X, the representation of the line $O_L - X_L$ onto the left image is a single point, while in the right images plane it is observed as the epipolar line $e_R - X_R$. The opposite also holds true, with the line $O_R - X_R$ corresponding to the epipolar line $e_L - X_L$ [3].

Due to those properties, the following observations can be made, as long as "the relative translation and rotation of the two cameras is known":

- If the projection point X_L is known, then the epipolar line $e_R - X_R$ is also known.
- If both projection points X_L and X_R are known, then their projection lines are also known.

Those observations provide "epipolar constraints", allowing for the check of corresponding points in order to determine wether or not they represent the same spatial point. They also allow for the triangulation of the 3-dimensional point, as long as both its projections are known. Using

those principles, image depth can be calculated using the disparity between scenes captured by different cameras [3].

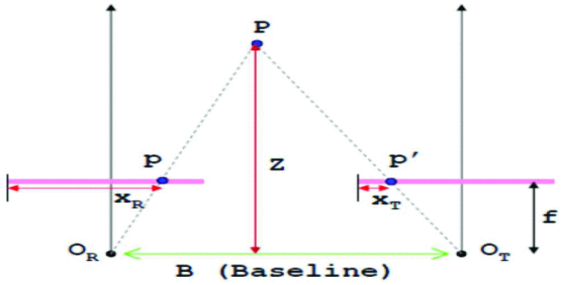

Figure 1.9: Depth calculation. [3]

Disparity refers to "the relative shift between two matching pixels", and scales inversely proportional to the depth, meaning that points closer to the camera will be more shifted than those further away. In the case of (1.9) the depth can be calculated using epipolar geometry, with the left image being used as a reference, and the right one as the target image. Points O_R and O_T the optical centers of each camera, P the target point in the scene, and X_R and X_T representing the relative coordinates of the projection points p and p', using the following relations [3]:

$$\frac{Z}{Z-f} = \frac{B}{(B + X_T - X_R)} \tag{1.1}$$

$$Z = \frac{B \times f}{X_R - XT} \tag{1.2}$$

$$Z = \frac{B \times f}{d} \tag{1.3}$$

where d denotes the disparity $X_R - X_T$, z represents the depth, with coefficients B and F denoting the distance between the cameras, and the focal lengths, respectively.

As an additional step that reduces scene clutter, an edge detection filter, such as Sobel or Canny is applied. This step leaves in the scene only those points that are relevant for the disparity calculation.

The final step is the "depth map generation", in which the disparity between all corresponding pixels is computed and stored into a matrix, holding the relative depth of each pixel as a grey scale value. [3]

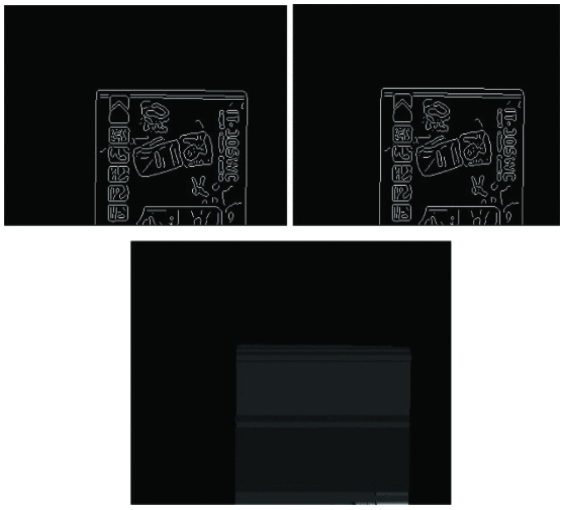

Figure 1.10: Canny filter and depth map. [3]

1.2.1 Applications

In practice, this process is used for tasks that require precise data about the distance from the viewer to the points in the scene. The most prominent case of such a use case is autonomous driving, due to the myriad of potential hazards surrounding an unmanned vehicle.

In the particular case of an "Autonomous Surface Vehicle (ASV)", the proposed solution relies on the coordination between stereoscopic inputs, as well as a GPS signal. The system has the following hardware architecture [4]:

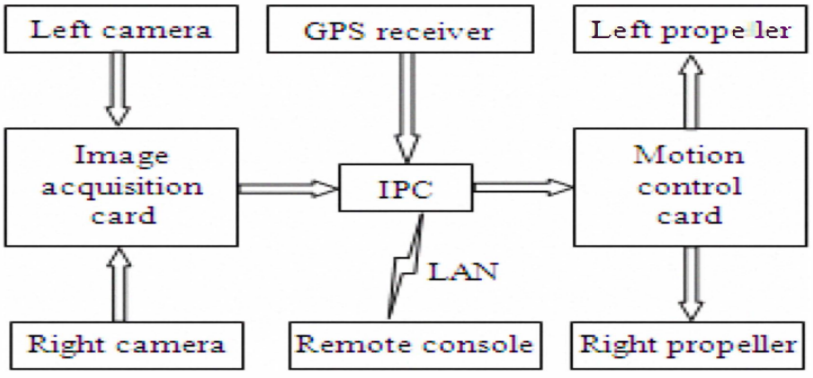

Figure 1.11: ASV hardware architecture. [4]

The images provided by the cameras are fed into an image processing module, where several filtering steps are carried out, before the relevant features are extracted and used to create the disparity map. From there, the camera parameters are taken into account, and the position and distance of potential obstacles are computed. This information is passed to a path finding

algorithm that determines the adequate control signal for each propeller.

Another implementation improves upon the initial algorithm, increasing its reliability such that it could be used for fast obstacle detection on the road. The most relevant addition is an "adaptive thresholding scheme", used to reduce the amount of noise that can affect data from the furthest points in the scene. Additionally, two enhancement steps are taken, both for the horizontal as well as the vertical disparity maps, in order to correctly assess wether a group of pixels is an obstacle. [5]

In the case of vertical enhancement, the disparity map is scanned both top-down and bottom-up, with pixels that are determined to be part of an obstacle generating a scan of their neighborhood, and any other pixels of similar disparity being considered as part of the same obstacle. For horizontal enhancement, the disparity map is scanned row by row, with the maximum being recorded and used to analyze the following row. Those maximums are then used to generate the "road profile", based on the assumption that obstacle pixels have a higher disparity then road pixels. [5]

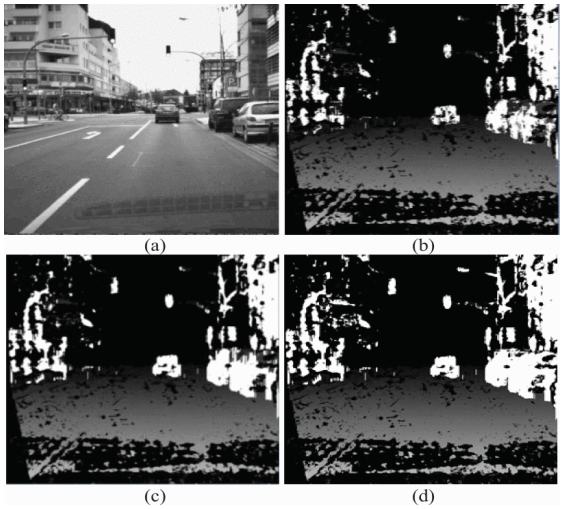

Figure 1.12: Processing steps. [5]

1.3 MULTIPLE VIEW STEREOPSIS

Also taking inspiration from nature, this time however from the way insects perceive their environment, micro-lens arrays focus light in several spots of the image sensor. Through this, several similar images are created, and those disparities can be analyzed in order to assess the depth of the scene, similar to stereo vision systems.

The design is based on the study of "Xenos pekii", insects that are capable of "multi-view stereopsis with high visual acuity through chunk sampled images, which is created by multiple photoreceptors on a single facet lens". Similarly, "multi-aperture systems" are capable of capturing a series of perspective scenes around the same target object, all in a singe shot. [6]

Although there are limitations in recreating the visual patterns of a Xenos pekii with traditionally

flat image sensors, the "ultrathin microlens array camera (MAC)" aims to emulate the insect view through the use of "constant FOV microlens arrays". The lens array allows for the capture of "partial images with all-in-focus", allowing for more accurate disparity calculations. [6]

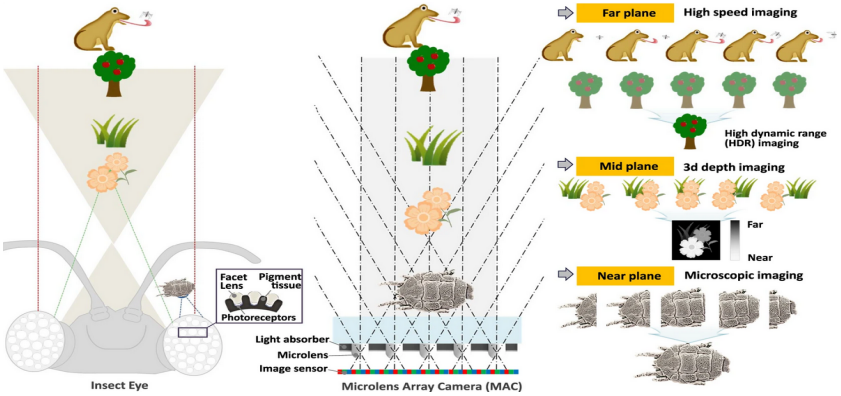

Figure 1.13: Parallel between insect stereopsis and MAC. [6]

1.4 CURVED IMAGE SENSOR

The idea originates from the need for "soft biometric devices", meant to act as implantable devices meant to assist with diminishing organ functions, in this case those of the retina. Several such image sensor arrays were proposed, among which is the "hemispherically curved image sensor (CurvIS)", that can achieve "aberration-free imaging and a wide field-of-view". Its design is based on "ultrathin MoS_2", a novel 2-dimensional nanomaterial, whose advantages are its "superb photo-absorption coefficient, photoresponsivity, and high fracture strain". [7]

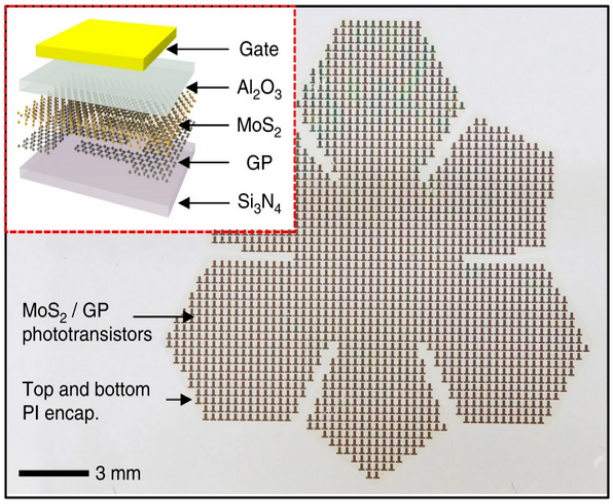

Figure 1.14: Proposed truncated icosahedron design for phototransistor array. [7]

Due to the requirement for the hemispherical construction of the image sensor, the use of a classical "film-type image sensor array" is inadequate, due to the mechanical strain that could induce failures in the device. Thus, the sensor is composed of a "MoS_2-graphene heterostructure" as well as other sublayers, leading to a final device thickness of "$51nm$", thicker than the average silicon-based sensor, but with significantly greater fracture strain and higher photo-absorption coefficient.

2 PAPER DESCRIPTION

For this project, "Fast Obstacle Detection Using U-Disparity Maps with Stereo Vision" was chosen as it represents a relevant application of the stereo vision principle, as well as adding improvements to the iterpretation of disparity maps in order to determine scene depth.

The paper present the base techniques used in determining depth from as stereoptic camera system, and provides a variable thresholding step for determining disparity, as well as two distinct depth map enhancement for greater accuracy in obstacle classification.

2.1 MOTIVATION

As obstacle detection is a key step towards fully autonomous driving vehicles, the accuracy of determining the 3-dimensional positions of surrounding objects becomes a necessary hurdle to overcome. For this purpose, stereovision systems provide "dense 3D data", with good accuracy for driving assistance applications. At the same time, further exploitation of the resulting depth information can lead to more intricate applications, such as obstacle classification. [5]

The paper presents an algorithm based on the "widely used u- and v-disparity concepts", with the caveat that the threshold for longer distances does not need to be calibrated a priori, but is instead adapted dynamically to scene conditions. It also implements a vertical enhancement step, meant to improve obstacle classification, and a horizontal enhancement step, used to separate the road surface from the rest of the environment. [5]

2.2 IMPLEMENTATION

The detection algorithm begins with the formation of the U-Disparity map, denoted U, recording the amount of "appearances of disparity values in a column wise manner". This step involves "iterating through the disparity image and incrementing the u-disparity map location", with the row position in U representing the disparity value of the given pixel, and the column position of U corresponding to the column location of the pixel. Likewise, the V-Disparity map, V, is formed in a row wise manner, with the disparity value being encoded inside each column and the pixel rows being translated. [5]

Figure 2.1: Recorded image, disparity image and UV-disparity maps. [5]

As shown in Figure, due to the constrains of the disparity maps, the horizontal one will have the same amount of rows as the original image, and the number of columns will represent the total amount of possible disparity values. The vertical map on the other hand is flipped, with the disparity values being encoded as each row of the matrix.

2.2.1 Adaptive thresholding

This technique makes use of the stereo vision system characteristics, in order to classify obstacle pixels from the disparity map not through a fixed threshold, but through one tailored specifically for the detection capabilities of the cameras. For this, the following characteristics are taken into account:

- Y_{3dmin} - minimum height of the physical obstacle
- y_{min} - minimum projection height
- F - focal length of the cameras
- Z - obstacle depth
- B - baseline distance between the cameras
- d - disparity value of the pixel

In order to determine the threshold, the following equalities are taken into account [5]:

$$\frac{y_{min}}{F} = \frac{Y_{3dmin}}{Z} \tag{2.1}$$

$$y_{min} = \frac{F * Y_{3dmin}}{Z} \tag{2.2}$$

$$Z = \frac{F * B}{d} \tag{2.3}$$

Resulting in the formula for determining y_{min} as:

$$y_{min} = \frac{Y_{3dmin} * d}{B} \tag{2.4}$$

From this, the threshold value for distant pixels can be computed based on the values of y_{min} and Y_{3dmin}, such that if the disparity value is small, it cannot be influenced by perceived noise.

2.2.2 Vertical enhancement

This step refers to the scan of each disparity column, both top-down and bottom-up, and identifying obstacle pixels. If such a pixel is found, its pixel neighborhood is scanned, and similar disparities are considered as part of the same obstacle.

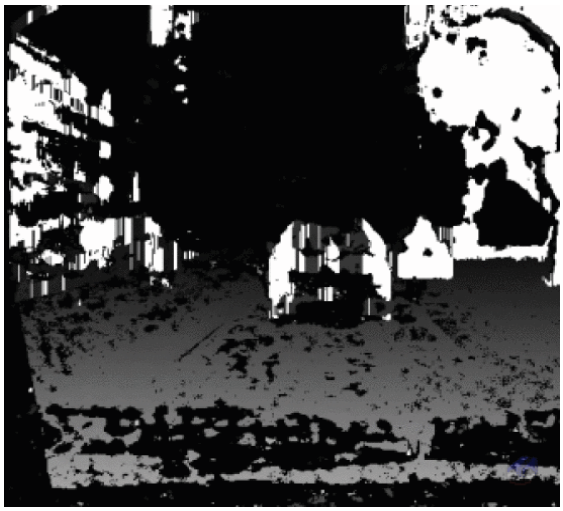

Figure 2.2: Vertical enhancement result. [5]

2.2.3 Horizontal enhancement

As the final step of the algorithm, it scans the v-disparity map in order to separate the road surface from any potential obstacles. As seen in Figure uv disparity, the road surface is most visible in the v-disparity map, with the maximum value in each row being most likely a road pixel.

In order to extract the road profile from the v-disparity map, the following steps are taken:

- Find the bottom-most row with a nonzero maximum.
- On the next row, add the row maximum to the road surface if its position is adjacent to the previous maximum.

- Stop the process if the profile becomes vertical, as it might overlap a vertical obstacle surface.
- If the current profile is beneath a threshold value, it is discarded and the process starts with the next disparity map row that was not visited

Figure 2.3: Horizontal enhancement result. [5]

2.3 RESULTS

Implemented in C++, the algorithm was tested on an Intel I5 processor and resulted in a time of detection of $4ms$, including the determination of the U and V disparity maps. As seen in the bellow images, each step of the algorithm improves upon the result of its antecedent, resulting in an efficient and reliable detection of roadside obstacles for autonomous vehicles.

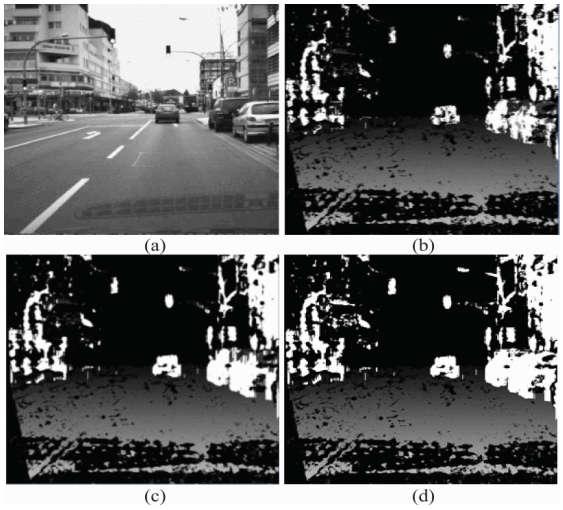

Figure 2.4: Final result after all processing steps. [5]

REFERENCES

[1] M. Hickton, "Exploring depth perception," vol. 35, pp. 16–21, Jul. 2020.

[2] J. Liu, S. Tsujinaga, S. Chai, *et al.*, "Single image depth map estimation for improving posture recognition," *IEEE Sensors Journal*, vol. 21, no. 23, pp. 26 997–27 004, 2021. DOI: 10.1109/JSEN.2021.3122128.

[3] D. N. Rajurkar, "Canonical applications of stereo vision and depth sensing," in *2023 16th International Conference on Sensing Technology (ICST)*, 2023, pp. 1–6. DOI: 10.1109/ICST59744.2023.10460833.

[4] J. Wang, P. Huang, C. Chen, W. Gu, and J. Chu, "Stereovision aided navigation of an autonomous surface vehicle," in *2011 3rd International Conference on Advanced Computer Control*, 2011, pp. 130–133. DOI: 10.1109/ICACC.2011.6016382.

[5] F. Oniga, E. Sarkozi, and S. Nedevschi, "Fast obstacle detection using u-disparity maps with stereo vision," in *2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2015, pp. 203–207. DOI: 10.1109/ICCP.2015.7312630.

[6] Kim, K., Jang, KW., Bae, SI. et al., "Multi-functional imaging inspired by insect stereopsis.," *Commun Eng*, vol. 1, 2022. DOI: https://doi.org/10.1038/s44172-022-00039-y.

[7] Choi, C., Choi, M.K., Liu, S. et al., "Human eye-inspired soft optoelectronic device using high-density mos2-graphene curved image sensor array.," vol. Nat Commun 8, 2017. DOI: https://doi.org/10.1038/s41467-017-01824-6.