2020-10-11

# Development of a bionic interactive interface for Owl robot using stereo vision algorithms

Rogers, J



10.1002/adc2.54
Advanced Control for Applications: Engineering and Industrial Systems
Wiley

# Development of a Bionic Interactive Interface for Owl Robot using Stereo Vision Algorithms

James Rogers[1] | Philip Culverhouse[1] | Benjamin Wickenden[1] | Chunxu Li*[1,2]

[1]Centre for Robotics and Neural Systems, University of Plymouth, Devon, UK
[2]School of Automation and Electrical Engineering, Qingdao University of Science and Technology, Shandong, China

**Correspondence**
*Chunxu Li is the corresponding author.
Email: chunxu.li@plymouth.ac.uk

**Present Address**
University of Plymouth, Drake Circus, Plymouth, PL4 8AA.

**Summary**

With the requirements for improving life quality, companion robots have gradually become a hotspot of application for healthy home living. In this paper, a novel bionic human-robot interaction (HRI) strategy using stereo vision algorithms has been developed to imitate the animal vision system on the Owl robot. Depth information of a target is found via two methods, vergence and disparity. Vergence requires physical tracking of the target, moving each camera to align with a chosen object, and through successive camera movements (saccades) a sparse depth map of the scene can be built up. Disparity however requires the cameras to be fixed and parallel, using the position of the target within the field of view, of a stereo pair of cameras, to calculate distance. As disparity does not require the cameras to move, multiple targets can be chosen to build up a disparity map, providing depth information for the whole scene. In addition, a salience model is implemented imitating how people explore a scene. This is achieved with feature maps, which apply filtering to the scene to highlight areas of interest, for example colour and edges, which is purely a bottom-up approach based on Itti and Koch's saliency model. A series of experiments have been conducted on Plymouth Owl robot to validate the proposed interface.

**KEYWORDS:**
Human-robot interaction; bionic; stereo vision; disparity map; Itti and Koch's saliency model.



## 1 | INTRODUCTION

In 2017, around 1.4 million people above the age of 50 reported to feel lonely in England and this number is predicted to increase in the next couple of years. Loneliness can manifest in multiple forms[1]. Social loneliness is described as a feeling where an individual is lacking connections and the feeling of belonging to a group or community. Emotional loneliness is where an individual is lacking deeper connection to specific figures like a partner or a close friend. No matter the form of loneliness, all have long term effects both mentally and physically. While it is not a direct cause, loneliness has been shown to have a correlation with malnutrition, sleep problems and depression[2]. There have been many kinds of interventions developed to try and tackle both of these kinds of loneliness. In the UK, The National Health Service provides an online support page with basic information and the organisation age UK has created a telephone befriending service where users are matched up with another person or volunteer and have regular weekly phone calls. Both of these interventions are considered low cost but the effectiveness of these methods are limited. Statistics shows that around 61% of the people aged 75+ never used the internet and the success of the telephone befriending system depends on the number of volunteers available[3].

[0]**Abbreviations:** HRI, human-robot interaction; HSV, Hue, Saturation and Value; DoG, Difference of Gaussians; FOV, Field of View; IPD, Pupillary Distance

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/adc2.54

Another, highly successful intervention method is Animal Assisted Therapy (AAT) and Pet Ownership (PO). Studies showed that petting animals can reduce anxiety and spending regular time with pets have long term mental health benefits[4]. Besides these, animal sessions can initiate interaction between patients, hence giving an indirect solution to social loneliness. One thing that is however overlooked is that these solutions can bring a wide variety of negative effects: patients who are allergic cannot participate in the activities; the animals can spread infections; the animals can give more attention to specific people which can leave the other patients feel left out; some people might not have the physical capacity to walk their pets every day in PO. Using social intervention robots, it is possible to combat these effects, still provide similar qualities-of-life a traditional pet would offer and in addition add extra features that could help the users and the carer's everyday life[5].

The rapid development of automation technology and artificial intelligence control algorithms provides sufficient development momentum for robots, enabling robot systems to solve more and more engineering and production applications[6]. Human-robot interaction (HRI) technology is an important part of robot research, not only as a key technology that is urgently needed for the development of high-tech fields such as aerospace (maintenance of space stations, planetary exploration such as the moon), marine exploration, atomic energy applications, and military warfare[7]. And it plays an important role in leisure, entertainment and medical treatment. HRI technology can replace workers in a complex, dangerous and unknown working environment, and can work in an environment that humans cannot reach, ensuring the safety and efficiency of operations.

Many advances in engineering today, particularly robotics, have been inspired by animals. By observing their behaviour, their physical structure and how they process information, we are able to develop systems that can take full advantage of the world we live in. Whether it be the kingfisher inspired bullet train[8] or Boston Dynamics's big dog[9], we are always looking to the animal kingdom for inspiration in our designs so that we may ultimately overcome real world problems. Animals have an amazing ability to quickly and efficiently extract useful information from their environment. Humans for example, can react to visual stimuli with a mean reaction time of 180 to 200 milliseconds[10]. To explore how this could be done, a robotic analogue for the human visual system has been developed to assess algorithms, and begin to mimic observed human eye behaviours. This paper focuses on animal perception, and the method of generating a sense of depth to the environment, to develop a bionic HRI interface for our own made social robot - Owl. Multiple experiments have been conducted, which allowed to evaluate the effectiveness of vengeance and disparity methods used to obtain depth from a target. Finally, an application was developed that combines the previous vergence techniques with Itti & Koch's saliency model, to demonstrate how human eyes can freely examine an unfamiliar environment.

## 2 | PRELIMINARIES

The Owl robot is a stereo camera host designed for the exploration of verging camera stereopsis. It has five degrees of freedom offering Neck rotate with Stereo eyes with pan and tilt. Local processing is by the Raspberry Pi dual camera compute board located at the base of the robot. An additional PCB provides an interface to the Pi for Servo control and an audio codec. The cameras are Pi HD cameras set to deliver stereo pairs at VGA resolution and streamed using RTP protocol web streaming at 30 fps. Fig.1 shows a photo of the robots, camera separation is 65mm. A pair of MKS DS65k high-speed digital servos moves the eyes with a 333Hz period. Normal drive Pulse Width Modulation (PWM) is 3 ms period, with a pulse width between $850\mu s$ - $2150\mu s$ (ie. 1300 PWM value range). The centre position is set at approximately $1500\mu s$ and the servos have a dead band of one microsecond. The PWM range allows for 160° rotation, with one PWM step being 0.113°.

The Pi compute board was chosen as it offers dual CSI format (DMA) camera inputs, that facilitate the high-speed streaming of camera images to the internet. The processor is BCM2835 (the same installed in Raspberry Pi B+). The eyes of the robot are OV5647 camera modules, each capable of 2592×1944 pixels in a 4:3 aspect ratio. The cameras have been set to display a lower 640×480 pixels, as higher resolutions are not required and this reduces image processing time. The two video streams are combined into a single 1280×480 stream, which is then communicated to a host computer via an RTP interface over USB in the MJPEG format. They are not synchronised at present. The cameras can pan and tilt independently via the four high-speed MKS DS65K servos, each capable of a no-load angular velocity of 0.203 sec/60° at 4.8V. This equals to approximately $300°s^{-1}$. Observational data has shown that human eye saccades average $160°s^{-1}$ [11], but can reach a peak angular velocity of $900°/s$ [12], thus the robotic analogue cannot capture the full speed of the human visual system. However, $300°s^{-1}$ is satisfactory for the experiments planned.

The neck of the robot can rotate about one axis with a Corona DS558HV servo, offering a range of 160 degrees and a peak no-load angular velocity of $300°s^{-1}$. The velocity will be lower due to the mass of the Owl head, but still satisfactory for target


**FIGURE 1** A parliament of Plymouth OWL robots.

tracking experiments. Servos are controlled by the Pi compute module using PiGPIO module, the PWM position values are generated by the host computer. The Pi compute module runs an IP server program which creates an IP socket over the host USB connection using the TCP protocol. The script waits for a 24-byte packet, which contains five 4-digit decimal integer numbers in an ASCII string separated by spaces, these are the new servo positions instructed by the host computer.

Software limits are applied to these new positions so that the servos do not over actuate to a state where the cameras or cables are damaged. The host computer supports the main software, programmed in C++ with the OpenCV 3.2 library. This arrangement allows for faster computations, compared to just using the on-board Pi compute module, and offers the potential to use GPU arrays for vision and deep learning in the future. Cameras were calibrated using OpenCV functions to correct for intrinsic distortions.

## 3 | METHODOLOGIES

### 3.1 | Vision Depth Approximation of a Single Target

When focusing on a target, both eyes move to centre the target about each fovea, which is a 64x64 patch at the centre of the field of view of each camera. This behaviour is known as Vergence, and provides depth information via the trigonometric relationship between the angles of the eyes and their distance from each other.

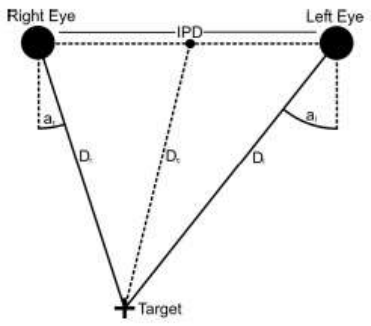



**FIGURE 2** Diagram of the depth information to a target using vergence.

Fig.2 better describes the geometric problem. The known variables are $\alpha_r$, $\alpha_l$ and Inter Pupillary Distance (IPD). From this, Dc is to be calculated, which is the distance to the target from the centre position of both eyes. To find this distance, all internal angles of the triangle drawn from both eyes to the target have to be calculated. Assuming angles are measured clockwise, the internal angles of the eyes are expressed in (1) and (2):

$$\begin{aligned} \theta_{Ir} &= 90 + \alpha_r \\ \theta_{Il} &= 90 - \alpha_l \end{aligned} \tag{1}$$

where, $\theta_{Ir}$ and $\theta_{Il}$ are the right and left internal angles of the triangle shown in Fig.2. As all angles of a triangle are equal to 180 degrees, the angle at the target, created by the alignment of the eyes, is shown as follows:

$$\theta_T = 180 - (90 - \alpha_l) - (90 + \alpha_r) = \alpha_l - \alpha_r \tag{2}$$

where, $\theta_T$ is the angle at the target, made by the two vectors from each eye and the target and by using the sine rule, the distance from the eyes to the target can be found, for example the left eye shown as follows:

$$\begin{aligned} \frac{sin(a_l - a_r)}{IPD} &= \frac{sin(90 + a_r)}{D_l} \\ \frac{IPDcos(a_r)}{sin(a_l - a_r)} &= D_l \end{aligned} \tag{3}$$

Distance to the target is defined as the length between the vergence point of the eyes, and the centre point of the eyes. If a line is drawn between these points, the triangle would be cut in two. If the right triangle is inspected, the one drawn between the target, the centre position, and the left eye, the length of one of its sides is the distance to the target. As the other two side lengths are known ($IPD/2$ and $D_l$), and the angle between them is known (90 - $D_l$), the cosine rule can be used to find the final triangle length:

$$\begin{aligned} D_c^2 &= D_l^2 + (\frac{IPD}{2})^2 - 2 \cdot D_l \cdot \frac{IPD}{2} \cdot cos(90 - a_l) \\ D_c &= \sqrt{D_l^2 + \frac{IPD^2}{4} - D_l \cdot IPD \cdot sin(a_l)} \end{aligned} \tag{4}$$

To implement this equation, both eyes must be programmed to track a target in order to measure the angle of each eye. The right eye will be following a colour target, whereas the left eye will be using normalised cross correlation (5) to find what the right eye is pointed at, by matching a small window of pixels in the centre of the right eye across the entirety of the left FOV. Programming the eyes in this fashion makes for a robust tracking system, as the left eye is always tracking what the right eye is looking at, regardless of the target. The equation of template matching is shown as follows[13]:

$$R(x, y) = \frac{\sum_{x',y'}(T'(x', y') \cdot I'(x + x', y + y'))}{\sqrt{\sum_{x',y'} T'(x', y')^2 \cdot \sum_{x',y'} I'(x + x', y + y')^2}} \tag{5}$$

where $T'$ and $I'$ are template and image respectively, both normalised by size and by average intensity, x and y are the positions of the template pixel, $x'$ and $y'$ are the corresponding displacement increment of the pixel templates.

## 3.2 | Stereo Calibration and Disparity Calculation

Although each camera has been calibrated for lens distortions and other intrinsic errors, they need to be calibrated together as a stereo pair, correcting for extrinsic distortions and establishing a common three-dimensional baseline, which is done by an example program provided by OpenCV. The software similarly uses an XML file to input parameters, this time requiring the file locations of any number of stereo pair images. A simple program was written to take images when a key was pressed, and save them as a pair of JPG files. Nineteen images pairs were taken, each containing the checkerboard calibration target at different positions and orientations. After processing the program stated that the RMS error was -0.606984, and the average epipolar error was -0.553882, which was acceptable. The distortion correction was applied to the calibration images and displayed (shown in Fig.3), note that the green guides on the images align with the exact same areas on the checkerboard in both camera views, this indicates that the calibration worked as intended.

The compute function of StereoSGBM in OpenCV was used to create the disparity map, which was then normalised so it could be viewed in a visualizer. Fig.4 shows the generated disparity map. For the relatively low resolution, the map looked

**FIGURE 3** Two stereo pairs corrected for extrinsic distortions, showing epipolar lines.

surprisingly good, given no additional noise reduction. As expected, this is very error prone, but for a rough measure of depth, a disparity map may be satisfactory. The left side of the image has been seemingly removed. However, this is just a result of the reduced FOV when overlapping two cameras.

**FIGURE 4** Stereo images and resulting disparity map.

### 3.3 | Saccadic Eye Motion and Saliency

The human eye has a central region of high resolution, called the fovea. This is so that the body's limited resources are focused where they are needed, providing a high-fidelity view of what is being attended to, with a more general, lower resolution, view in the periphery. Saccadic eye motions make use of this small area, by rapidly moving the fovea over a set of salient targets, allowing the brain to build up a more detailed view of an object or scene much larger than this small area of high resolution. The motions are ballistic, and cannot be corrected for the duration of the movement. If the desired target moves as the eye is saccading, the eye will miss the target and a second saccade will be required to correct, stated by [14].

Saccades can occur both voluntarily and reflexively, demonstrating a combination of bottom-up and top-down processing. As reflexive eye saccades are based solely on the information contained within the scene, computer models can attempt to imitate how this is done in humans. Itti & Koch proposed a model based on salience (Fig.5), where feature maps are created from the environment, and combined into what can be described as an "interest map" with which a vision system can sequentially direct its attention, according to [15]. A feature map will highlight a single property about the scene, such as colour or orientation. Multiple feature maps means that the visual model will take into account more information about the visual stimulus. A saliency map is the linear combination of these feature maps, each weighted on their importance. Peaks in the salience map are targets that a human's visual system may find interesting and saccade to (Fig.13). One issue with this model is that it's stable, and with a static image as stimuli, the model would saccade to the most salient target and not anywhere else. The model proposed by Itti solves this problem by deleting the region in the stimuli, where the model found the last salient target. This would make the second most salient target the next location to saccade, and so on. A limitation with this solution is that it assumes that salient targets are only looked at once.

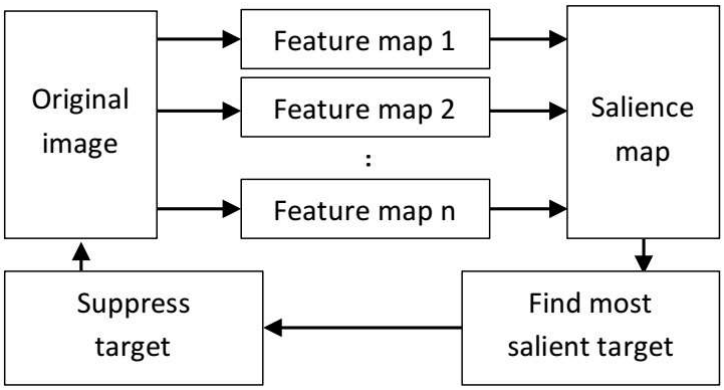
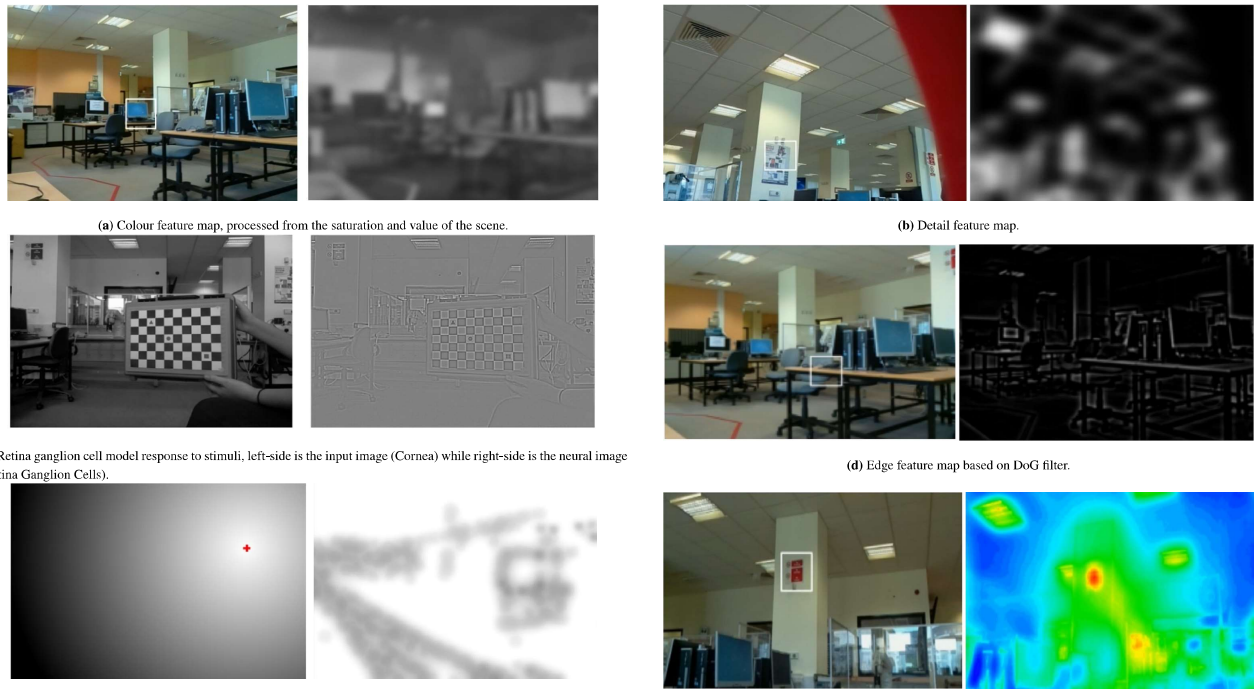



**FIGURE 5** Salience based visual attention model of the proposed human-robot interface.



**(a)** Colour feature map, processed from the saturation and value of the scene.

**(b)** Detail feature map.

**(c)** Retina ganglion cell model response to stimuli, left-side is the input image (Cornea) while right-side is the neural image (Retina Ganglion Cells).

**(d)** Edge feature map based on DoG filter.

**(e)** Local map for a target located at the red mark (left); familiarity map of a target after 150 seconds of saccading. Note darker regions are not saccaded to, but gradually return to white and the Owl can become interested in them again (right).

**(f)** Salience map.

**FIGURE 6** Different type of feature maps outputted by the streaming of owl robot: (a) colour map, (b) detail map, (c) computer model of ganglion response, (d) edge map, (e) local map and familiarity map, (f) combined salience map.

### 3.3.1 | Feature Maps

**Colour Map** To evaluate the colour within the room, it would be a lot easier if the colour format of the video stream was encoded as Hue, Saturation and Value (HSV). This format arranges colour information in a convenient way, as the "strength" of the colour is contained in the saturation and value variables. For a colour to stand out in a scene, it has to be bright compared to the background. However a bright shade of white is not that striking compared to a bright green or other colour. Thus saturation is important. To create a map that rewards a combination of value and saturation, these variables for each pixel are multiplied together. As hue is not important in this map, the spare channel in the matrix was used to store the product of saturation and value. Fig.6a shows the result of this filtering. The orange wall in the background and the computer monitors are quite colourful, which correspond to high values on the colour map.

**Detail Map** The detail map is to reward high concentrations of edges in small areas, such as a keyboard or text. To do this a high-frequency Canny edge detector created a binary view of the scene, populated with edges. This was morphologically filtered using the dilate and erode commands, this has the effect of filling in the gaps when edges are close to one another. With the view filled with blotches of white wherever there are high concentrations of edges, the image is blurred. Small groups of edges will blend into the background black, however larger groups will persist, and gain a radial gradient with a peak at the centre.

The resulting feature map is shown in Fig.6b. Note that the air vent, poster and signage appear as bright spots in the processed map. With this feature map incorporated into the attention system, highly detailed objects like these will have greater chance of being salient targets.

**Edge Map** The salience map so far is a combination of colour and detail, and it can highlight low-resolution areas of interest. However, specifically where to look within these salient patches is not evaluated. Adding an edge map will give the attention system a preference for observing contrasting areas. If a colourful and detailed sign was in frame for example, instead of simply looking at its centre, the edge map will reward looking at the text or symbol on the sign.

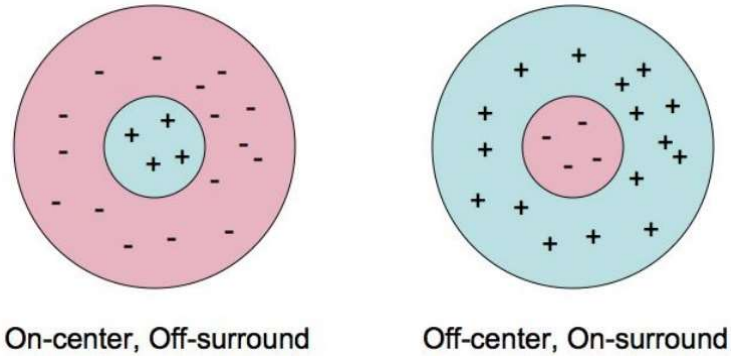



**FIGURE 7** Centre surround structure of a ganglion cells receptive field, both on and off centre types (adapted from [16].

There is evidence that the human eye processes edges before the signal reaches the brain. Such processing is done by ganglion cells within the retina, and are the first cells to signal an action potential from the sensor cells to the brain [17]. Each cell responds to stimuli with their receptive fields, which is a small region of cone (colour) and rod (monochromatic) sensor cells. The receptive field is split into two sub-regions, centre and surround (Fig.7).

These cells do not respond to featureless stimuli, as subtracting one sub region from the other results in zero, because the average activation of each will be the same. The result is approximated as a Difference of Gaussians (DoG), [16]. However, something differing over receptive field, such as an edge, will elicit a response (Fig.6c shows an image and a DoG filtered result). Such edge detection can be derived from the application of a Difference of Gaussians (DoG) filter over an image, according to the discussion in [16]. This map (Fig.6d) was deemed to highlight edges sufficiently, and be satisfactory for use in the salience map.

**Local Map** The local map needs to appear like a spotlight, with a gradient of 255 at the current salient target, trailing off to 0 at the extremes. To achieve this, a large white filled circle is drawn where the previous salient target was. A strong Gaussian blur smooths this out into a spotlight with a single peak (Fig.6e). This map will boost salience around the current target, making more frequent smaller saccades more likely, as is observed in humans.

**Familiarity Map** This map (Fig.6e) keeps track of the most visited areas of the attention system, darkening areas of the map that are observed. Over time, frequently observed regions will get darker, making the subsequent observation of these areas more unlikely. This has the effect of suppressing salient targets, without preventing multiple observations of the same point.

Dimming a region will require opacity, which can be somewhat achieved with the “addWeighted” command. A blank white map is initialized. When a salient target is saccaded to, this map is duplicated. One version gets a solid black mark, and blurred, whereas the other version is unchanged. The final updated map is a weighted sum of the two versions, and the strength of the memory effect can be adjusted by changing these weights. This should appear as semi-transparent blurred circles plotting onto a white background, to create an accumulative history of what has been observed. The memory can have a variable persistence over time, forgetting the oldest saccades plotted on the global map.

**Salience Map** With all the feature maps processed, they are linearly combined into a salience map using a weight, which is shown in Fig.6f. The familiarity map is different as it suppresses salient targets, and a zero in this map must be a zero in the salience map. So in this case, this feature map is multiplied to the weighted sum of the other maps. This map is scaled and normalised to convert the greyscale salience map into a hue-scale visualisation, blue being the lowest salience, and red the highest. Interestingly the visual attention system finds a hazard sign to be the most salient target. By design these signs attempt to catch the attention of humans, so the fact the computer model also responded in the same way is a promising start.

### 3.4 | Owl Robot Deployment

Minimal modification is required to make the software ready for the owl, since the image processed is taken from the left camera rather than from a pre-loaded static image, and the most salient target coordinates are used to move the cameras and neck. To find the right corrective servo movement needed to centre the next salient target, the difference between the centre of the image and the new target is to be calculated. This is the trajectory of the next saccade. The differences are converted from pixels to degrees, and sent to a function that moves the servos. The right camera is moved so that it is always parallel to the left camera, using the lefts absolute position to generate a corrective trajectory. The neck loosely tries to keep the eyes at the centre of their horizontal movement range. The global position of the left eye is subtracted from the calibration point, which is the motor position for looking straight ahead relative to the Owl head. A number is generated that indicates how off centre the eyes are.

The feature maps will have to change as salient targets in view of the camera will not stay in the same position on the screen as the cameras move. A solution to this is to create a map that takes into account the motor positions. Assuming that the Owl body is not moved, and assuming the environment is mostly static, salient targets will always be in the same physical location. To create this new global map, not much is changed apart from the larger size. New updates to the map are offset by the current servo positions to put them into a global or panoramic frame. The only change that was made was to create a “GlobalPos” variable that stored the offset to the salient target position. The weighted-add function was given a minor edit to make the opacity variable. This global position is based on the servo position, converted to pixel units. When this larger familiarity map is used for the salience map, a smaller window is cut from it, representing the current view of the cameras. The local map was changed so that the spotlight gradient was permanently centred on the screen, as the fovea of the camera doesn’t move relative to the camera. The robot could now freely view the environment, saccading once every 400-500 ms. Again, the magnitude of each saccade was recorded, so that it the frequency of chosen eye movements could be measured against a human’s.

## 4 | EXPERIMENTAL STUDIES

In the experiment section of this paper, three tests are conducted on the Owl robot to validate the proposed human-robot interface, which are vergencing and tracking an moving item, stereo calibration for the dual eyes of the Owl robot and the saccadic eye motions, respectively. The indoor experiment environment is of sufficient illumination.

### 4.1 | Experimental Setup and Discussion

#### 4.1.1 | Vergence Control Experiment

In order to test the quality of the vergence control, an experiment was setup that aimed to measure a target over a range of distances. Fig.8 shows the target placed inside a vice and positioned in front of the owl. A metre rule was setup parallel to the owls FOV so that an accurate measurement of the targets distance was obtainable. In our experiment, the target began 150mm directly in front of the Owl and left a few seconds for the eyes to verge onto it. Once verged, the average distance estimate would be displayed to the screen for us to record. The target was then shifted 50mm away from the Owl and the distance was noted once again. The test was repeated every 50mm until 1000mm was reached. At that point, the test was reset and repeated two more times so that an average of the three could be calculated. This allows us to produce a more accurate estimate and account for any unreliable results should we encounter any.

Fig.9 shows how the distance estimate compares to the true distance. The blue line is the ideal characteristic that the algorithm would achieve, and the green line is the average response of the distance estimate. Error between the two increases over distance, however expressed as a percentage the error is more consistent. This is evidence for the IPD measurement being incorrect, as

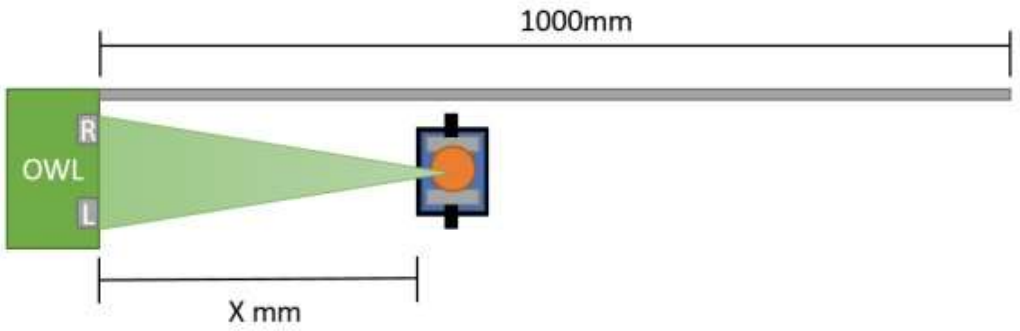



**FIGURE 8** Vergence experiment setup.

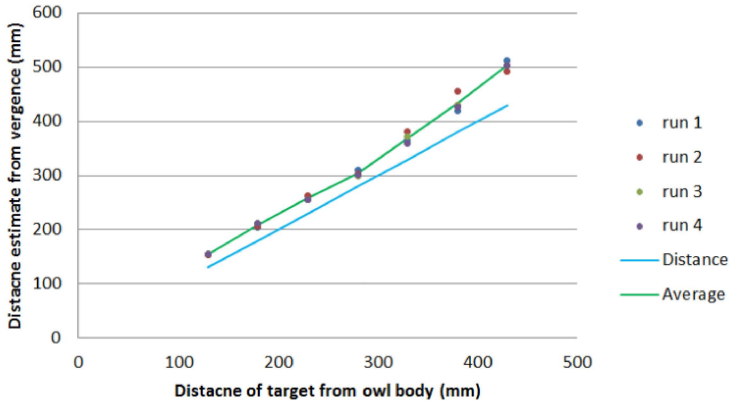



**FIGURE 9** Distance estimation from initial vergence experiment.

this is the scale that the algorithm uses to define a millimetre. The average percentage error was calculated to be 14%, thus the IPD constant within the program was modified from 66.5 mm to 58.3 mm. With this correction, the experiment was repeated.

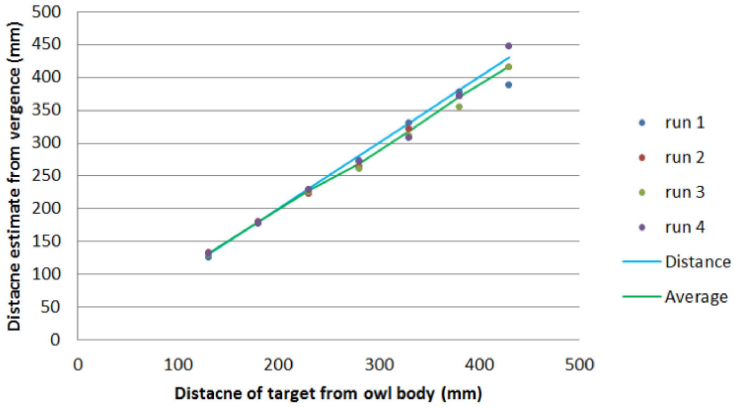



**FIGURE 10** Distance estimation via vergence experimental results 2.

Compared to the previous experiment, Fig.10 shows improvement, with an average percentage error of 1.96% below the true distance. This error was due to the cameras being mounted with the lens being at the centre of eye rotational axis, and not the silicon sensor. And an error of smaller than 2% is within the bounds of tracking/image resolution, which is reasonable.

### 4.1.2 | Experiment of Stereo Calibration and Disparity Generation

The brightness of each pixel represents the disparity of that physical object in both camera views, very similar to how a single target was tracked in the previous experiments. To see how this disparity relates to depth, an experiment was set up. A tape measure was laid out perpendicularly to the robot, and a target placed at set distances. The data in Fig.11 shows that the rela-

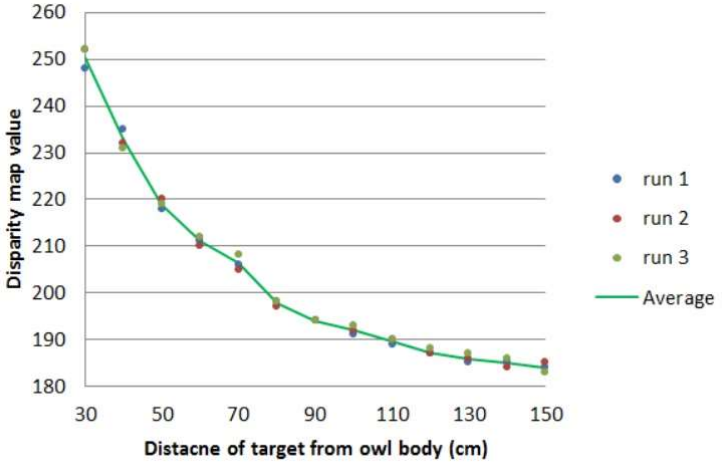



**FIGURE 11** Disparity of a target vs its distance from the owl.

tionship between distance and disparity is non-linear. Disparity and distance are inversely related. As the units are ambiguous a model will attempt to estimate the disparity for a given distance, as there exists experimental data to test against. Once a model has been calibrated, it can be reversed to calculate distance for a given disparity. Disparity values were inverted to start from 0 instead of 255, and an arctangent (Distance) function was plotted, see Fig.12. From the graph it is clear that the model fits closely with the experimental data. With this model calibrated, the final conversion between disparity and distance is as so:

$$Distance = 26.4842 \cdot (tan(\frac{Distance + 54.6682}{91.4178}) + 0.3815). \quad (6)$$

### 4.1.3 | Saccadic Eye Motion Experiment

The third experiment is hard to assess. Here, an example is taken from a research work [18]: a participant was given an image of a building in a free viewing condition, and their eye movements were recorded for 5 seconds, the eye motion path is shown in Fig.13. Highly salient points such as a window reflection are visited more than once, as regions of interest may be too detailed for a single glance to capture. Whatever the reason, simply preventing a visual attention system from visiting the same spot twice is not true to what is observed in humans. A solution is to implement a familiarity map, where the locations of salient targets are added as faint blobs with a low opacity. Areas commonly looked at will get darker as many of these blobs overlap. This map could be seen as memory, where light regions indicate little knowledge of an area. If this is included as a feature map, highly salient targets will be supressed as they are observed. However, if a target is sufficiently interesting, this model allows for the visual attention system to re-visit a target. The visual attention system was developed on static images, allowing saccades over static scenes, so that the stimuli are more consistent and changes to the program can more easily be monitored. The maps that were decided upon were edges, colour, and detail. Colour is based on how saturated and bright the colours are in the scene, detail activates when many edges are detected in a small space, and edges highlight areas of contrast as its unlikely that smooth featureless areas will become saccade targets.

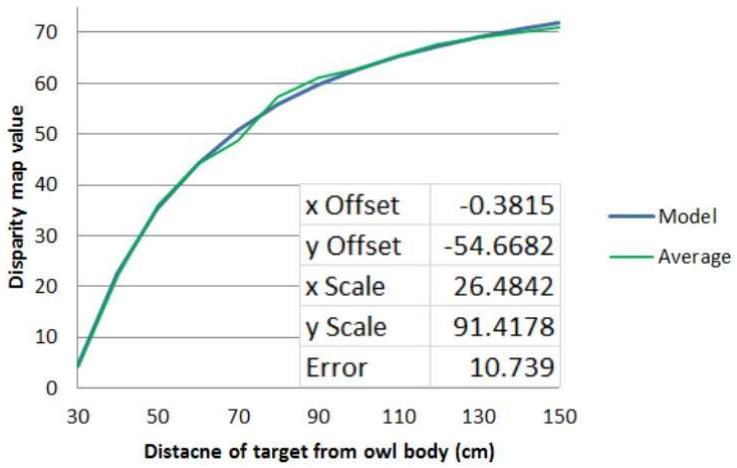



**FIGURE 12** Model for estimating disparity for a given target.

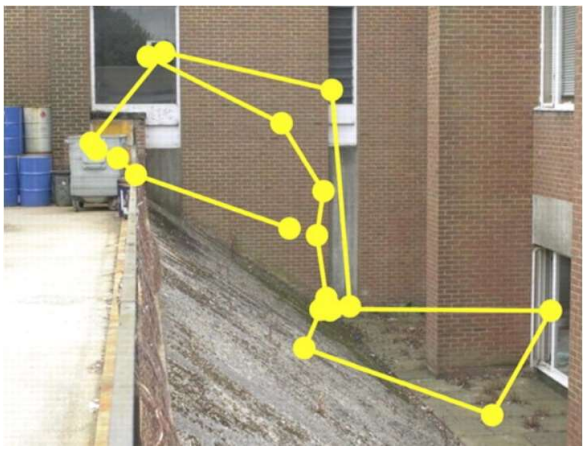


**FIGURE 13** Saccadic movement from a single participant, observing the stimuli for 5 seconds, retrieved from [18].

An important characteristic about human saccades is a magnitude versus frequency plot, and is an easy test to apply to the synthetic attention system. This characteristic made adjusting the local map easy, as changing this weight affected the size of the saccades. After calibrating, this is how the human (Fig.14a) and the current visual model compare (Fig.14b). The graph has a clear peak of low magnitude saccades and a linear drop off in amplitude. This linear region in steeper for the Owl compared a human, though this could be due to the restrictive FOV of the cameras compared to a human eye. There is a strange peak of high amplitude saccades at around 30 degrees of visual angle. Observing the video footage, the Owl seems to catch its own eye-socket every now and then. Since the Owl frame was bright red, this can trigger a salient target at the corner of the FOV, accumulating in a higher than average peak at this amplitude. Over a large enough sample set, one would assume that there would be little relation between the size of a saccade and its frequency, when using the current proposed salience model for visual attention. This is due to the current point of attention having no effect on the next (excluding the familiarity map which is a very local effect). Fig.14 shows that with human saccades, there is a high bias for looking at salient targets near the fovea. To implement this into the model, a spotlight-like feature map will be created to track the previous area of attention. This will have the effect of amplifying the salience of features near to the fovea of the attention model. The general shape of the curve is very similar to that of a human. Small motions are more likely than large ones, tiny motions however are less likely, and the frequency drop-off has a linear region. To test the similarity of the models saccade targets to a human's, experimental data of a human's response

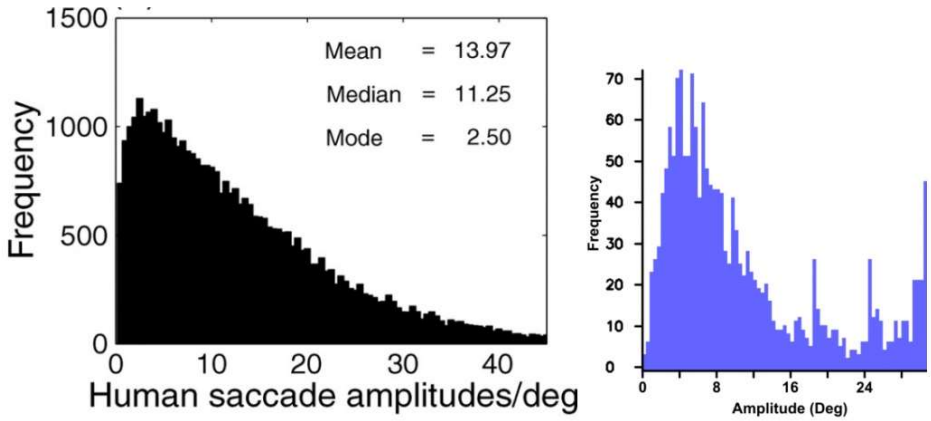



**FIGURE 14** Comparison of saccading visual systems: (a) human (left), from [18], and (b) the Owl model (right).

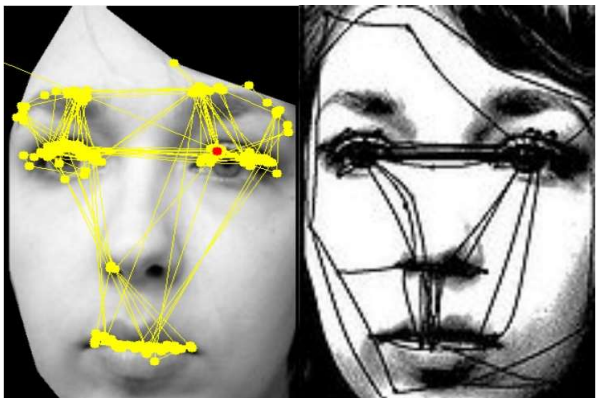


**FIGURE 15** Comparison of saccade paths and targets between (a) the OWL visual attention system (left) and (b) a pre-recorded path from the participant in [18] (right).

to stimuli is required, along with the original stimuli. Online research returned few examples, but the best that could be found were tests done on faces (Fig.15).

## 5 | CONCLUSIONS

In this paper, a computer vision algorithms based interface has been developed for the Owl robot to interact with human operators, which is inspired from bionics. The OWL robot offers a simple TCP/IP interface for the exploration of stereo verged and static eye motions for distance estimation from a host computer. Software has been developed that demonstrates distance measurements and target tracking using the OpenCV computer vision library. A simple model of visual attention has been developed that applies saccadic eye control which is driven by bottom-up analysis of scene features including colour saturation, edge density and orientation. Motion and other feature maps can be added to enrich the model. The Plymouth OWL robot offers an open programming interface to explore advanced topics in cognitive vision. Three experiments have been conducted to validate our proposed method. In addition, it has been used at Plymouth University to explore the computer vision and behavioural computing as a source of inspiration for robot vision systems regarding to the teaching purposes.

## 6 | ACKNOWLEDGEMENTS

We thank Bill Stephenson, Martin R Simpson and Clare Simpson from the School of Computing, Electronics & Mathematics, for their design skills in creating the OWL robot to be such an engaging robot for teaching robotics visual perception engineering at Plymouth University.

## References

1. Age U. All the lonely people: loneliness in later life. *London, England: Author* 2018.
2. Singh A, Misra N. Loneliness, depression and sociability in old age. *Industrial psychiatry journal* 2009; 18(1): 51.
3. Day P, Gould J, Hazelby G. A public health approach to social isolation in the elderly.. *Journal of Community Nursing* 2020; 34(3).
4. Age U. The Internet and older people in the UK–Key Statistics. 2016.
5. Chen J, Glover M, Yang C, Li C, Li Z, Cangelosi A. Development of an immersive interface for robot teleoperation. In: Springer. ; 2017: 1–15.
6. Li C, Yang C, Wan J, Annamalai A, Cangelosi A. Neural learning and kalman filtering enhanced teaching by demonstration for a baxter robot. In: IEEE. ; 2017: 1–6.
7. Li C, Yang C, Ju Z, Annamalai AS. An enhanced teaching interface for a robot using DMP and GMR. *International journal of intelligent robotics and applications* 2018; 2(1): 110–121.
8. Wiltgen B, Goel A. Functional model simulation for evaluating design concepts. *Advances in Cognitive Systems* 2016; 4: 151–168.
9. Raibert M, Blankespoor K, Nelson G, Playter R. Bigdog, the rough-terrain quadruped robot. *IFAC Proceedings Volumes* 2008; 41(2): 10822–10825.
10. Shelton J, Kumar G. Comparison between Auditory and Visual Simple Reaction Times. Neuroscience & Medicine, 01 (01), 30–32. 2010.
11. Abrams RA, Meyer DE, Kornblum S. Speed and accuracy of saccadic eye movements: characteristics of impulse variability in the oculomotor system.. *Journal of Experimental Psychology: Human Perception and Performance* 1989; 15(3): 529.
12. Wilson SJ, Glue P, Ball D, Nutt DJ. Saccadic eye movement parameters in normal subjects. *Electroencephalography and clinical neurophysiology* 1993; 86(1): 69–74.
13. Marengoni M, Stringhini D. High level computer vision using opencv. In: IEEE. ; 2011: 11–24.
14. Purves D, Augustine GJ, Fitzpatrick D, et al. Circuits within the basal ganglia system. In: Sinauer Associates. 2001.
15. Itti L, Koch C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research* 2000; 40(10-12): 1489–1506.
16. Marr D. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY* 1982; 2(4.2).
17. Markaryan A, Nelson EG, Hinojosa R. Major arc mitochondrial DNA deletions in cytochrome c oxidase-deficient human cochlear spiral ganglion cells. *Acta oto-laryngologica* 2010; 130(7): 780–787.
18. Tatler BW, Hayhoe MM, Land MF, Ballard DH. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision* 2011; 11(5): 5–5.