EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mathematisch-
Naturwissenschaftliche
Fakultät

Kognitionswissenschaften

Bachelorarbeit

Depth Based Homing to Remembered Place in Kite-Shaped Room

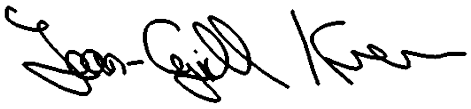
Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Kognitionswissenschaften
Jean-Cyrill Kreuer, jean-cyrill.kreuer@student.uni-tuebingen.de, 2020

Bearbeitungszeitraum: von 01.09.2020 bis 31.12.2020

Betreuer/Gutachter: Prof. Dr. H. Mallot, Universität Tübingen

Selbstständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.



Jean-Cyrill Kreuer (Matrikelnummer 4096845), December 21, 2020

Abstract

The faculty of place recognition is a central condition for successful navigation. It involves the comparison of two elements: A set of remembered characteristics of a place (place code) and a representation of the current surroundings cached in spatial working memory. Different models of the internal representation of places give rise to different predicted error patterns of homing, which can be compared to behavioral data. Previous work on visual homing using this approach had some success in determining properties of the internal representation used by humans. While in certain situations models based on features extracted directly from views perform best, in other situations models based on reconstruction type representations showed the best performance. Here, we address the question of how depth information is used in humans for representing places in closed room environments. We discriminate between a relatively unprocessed depth picture and a more abstract egocentric reconstruction as possible content of spatial working memory. For this purpose, we extend two simple maximum likelihood models of place recognition by Mallot and Lancier (2017) to closed room environments and compare the predicted errors to experimental data by Halfmann (2015). In the Halfmann experiment, subjects performed a simple return to cued condition task in a kite-shaped room in immersive virtual reality. The walls of the room had special properties reducing visual cues to depth. The experimental error patterns significantly deviate from error patterns expected for the model with place codes based on visual angles. The present analysis excludes this model and favors the reconstruction based model. The model is based on the shortest distance to walls as place code and an egocentric spatial working memory featuring the current shortest distance to walls.

Acknowledgments

I especially thank my supervisor Prof. Dr. Hanspeter A. Mallot for introducing me to this subject and for his support during preparing this thesis. I would also like to thank my Grandmother for the occasional supply of delicacies that really helped to keep me going.

Contents

1 Introduction

Animals including us humans have no sensing organ for their absolute position in space. Yet, efficient navigation is an important faculty that a lot of animals have. To estimate their position in space, humans use visual cues and path integration among several other sources of information. A self experiment in a completely dark room easily shows that visual cues are crucial for human navigation.
Due to limited resources, space cannot be processed as the continuous coordinate frame that it is. If a location is processed in animals, this location probably has a meaning and it is characterized by this meaning. The context in which the place was important and properties of the environment probably also characterize places. In this work we will treat places as previously learned, and therefore meaningful, locations in space that an agent can navigate to.
The ability to navigate to a previously learned location implies the existence of spatial memory. Research on simpler organisms like bees was quite successful in determining the information content of this spatial memory. In experiments with bees, visible landmarks were moved away from a previously learned location. When the bees tried to return to that location, their searching behavior indicated that they find places by minimizing the difference of the sensed image and the image that would be sensed at the goal location. The distance of the goal to the landmarks did not seem to play a role [Cartwright and Collett, 1983]. Therefore it is assumed that bees memorize places in form of panoramic images that have undergone little processing. In this work the information used to memorize a place will be called a place code.

The strategy to approach a goal by minimizing the difference of the perceived surroundings and the memory of those surroundings is also conceivable for more complex animals like humans [Kuipers, 2000]. Of course, human place codes are probably very different from the raw panoramic view that bees use. Humans are different from bees in many ways. They have their eyes on the front, hence they don't have access to panoramic views. Place codes and the content of spatial working

memory must therefore be composed of information from multiple views. Humans obtain depth information through stereo vision in addition to motion parallax. They are also known to identify prominent points such as church towers and use them for navigation and are capable of navigating precisely in small scale environments like closed rooms and large scale environments like cities. All characteristics of a place that humans perceive, and cache in spatial working memory could be part of place codes, including simple unprocessed views, the distance to landmarks, inter landmark angles and all kinds of other descriptions of the surroundings.

There even could be differences in the representation of places in long term memory and the representation of places in working memory that is being used while navigating to a place. This work deals with the properties of spatial working memory because this can be examined in behavioral experiments. The recognition of a place will be treated in the sense that it is successful if the agent was able to navigate to that place. Apart from behavioral approaches also physiological research has contributed to the understanding of the representations used in spatial working memory. In the following, I will therefore briefly summarize findings from both types of approaches with special emphasize on those publications which have some relevance for the present study.

1.1 Physiological Research

Physiological research gives several clues on what is encoded in spatial memory. Visual information has a retinal reference frame, i.e. a retinotoptic depth map exists in the parietal cortex [Gardner et al., 2008]. This means that in early stages of processing, depth is represented, like an additional dimension to the three color intensities. A more recent study has found evidence for a spatiotopic map sensitive to movement, in the parietal cortex, specifically in MT [Gardner et al., 2008]. MT is therefore suspected to be involved in building a representation of space with world centered coordinates. Unfortunately, it is not yet known how this representation is built, and how it is used to form spatial memory.
The most direct neural correlate of place memory are cells that encode specific places on a map. Those cells have been found through extracellular recordings in the hippocampus of rats [O'Keefe and Dostrovsky, 1971]. Cells that encode head direction at a certain view point [Taube, 1998] and the proximity to boundaries

have also been found in the hippocampus and neighboring areas of the rat brain [Solstad et al., 2008].
This clearly indicates that rats have access to bearing information, and that boundaries and their proximity to the rat play a role, but it does not tell us which information is stimulating the firing of those neurons. The firing of neurons when the rat has a certain location and head direction could simply be representing a constellation of visual features in the field of view, or the rat's position and orientation relative to a constellation of recognized objects obtained through all senses available to the rat, just to name two very different possibilities.

When human subjects are presented with visual stimuli relevant for navigation, such as roads, rooms and landscapes, a stronger fMRI Repetition Suppression is measured in the Hippocampal Place Area (PPA) and the Retrosplenial Cortex (RSP) compared to when visual stimuli unsuitable for navigation, such as objects and people, are presented [Epstein, 2008]. The fMRI Repetition Suppression (fMRI-RS) is a phenomenon of reduced fMRI response when a Stimulus is presented more than once. It is believed that when a region of the brain responds to a stimulus with a stronger fMRI-RS than the rest of the brain, it is involved in processing that stimulus. Consequently, two different stimuli are considered to be represented in similar ways, if their subsequent presentation give rise to an fMRI-RS [Epstein, 2008]. For example, two different photos of the same face will show an fMRI-RS [Larsson and Smith, 2011]. With this type of experiment, it was shown that the processing of scenes in the PPA is mostly dependent on the spatial layout of the scene. Individual objects in the scene and the position of the scene in the field of view seemed to be irrelevant [Epstein, 2008]. No indication of a scene representation independent of the point of view could be found, which indicates that information encoded in the PPA has an egocentric reference frame. Similar results were found in the RSP, albeit with stronger responses to known scenes. Therefore, the RSP is suspected of being involved in the retrieval of spatial memories [Epstein, 2008].

From physiological research it is not known which aspects of a view are used to form this "spatial layout" that the PPA reacts to. These could be visual features like color, texture and depth, but also descriptions like "a barrier to the right and a passage to the left" and spatial relations between landmarks and between landmarks and the observer could be used for forming a "spatial layout".

1.2 Behavioral Research

Behavioral research on navigation in space tries to infer information about the internally used representations from behavioral data. A common limitation of such type of aproaches is the fact that behavior can often be explained by fundamentally different models. To give a typical example: In rooms, rats confuse the goal location with geometrically equivalent points. This behavior was first interpreted as giving evidence for the use of geometric information like distances and angles [Stürzl et al., 2008]. However, Later work showed that this behavior is in line with the predictions of a model based on simple panoramic views with edge detection [Cheng et al., 2007, Cheng, 2008]. In this particular case, more sophisticated experiments are needed for discriminating between the two explanations or even coming up with another one. Generally speaking it is hard to show that a certain representation of the surroundings for one situation is also used in any other case, or can be used by an agent in any given situation. Nevertheless, behavioral experiments help to narrow down the properties of the internal representation of the environment.

Evidence for a representation of indoor spaces with an intrinsic frame of reference has been found. In an imagined pointing experiment, subjects had to learn the layout of objects in a room, and then they were asked to point at objects without them being visible. Subjects where most accurate when the imagined viewing direction was parallel to the intrinsic axis of the room [Mou and McNamara, 2002]. This could mean that the spatial layout of landmarks is perceived and memorized in relation to an intrinsic axis of the room: A form of representation with a world centered reference frame. Humans are able to recognize a place in the absence of landmarks, with the help of simple unprocessed views [Gillner et al., 2008]. This was shown in virtual reality in a round room with a color gradient on the wall. Even though the distance to the wall could be obtained through motion parallax and the elevation of the wall in the field of view, the radial component of the goal position seemed to be obtained from the color gradient on the wall, as radial accuracy decreased when the contrast in the color gradient was reduced. The authors concluded that the subjects used a simple view to remember the goal location. Depth information alone enables humans to recognize a place [Halfmann, 2015]. This was shown in virtual reality environment with a kite-shaped room presented through stereoscopic random dots with limited life time. Instead of colored polygons the walls were composed of

dynamically generated random dots with an uniform distribution. The number of visible dots was always the same and their lifetime was randomly chosen between 100ms and 200ms. The dots all had the same size in the field of view. Since the pattern of dots always changed, it could not be used for navigation. Depth information from stereo disparity and motion parallax was the only visual cues available. Nevertheless the participants were able to navigate precisely, and their performance was not significantly increased when additional texture cues were added to the environment [Halfmann, 2015]. The authors concluded that humans can build a representation of a local environment from depth information alone[Halfmann, 2015]. Reconstructions with world centered reference frame as well as texture and depth information are probably present in spatial working memory.

All following experimental work deals with situations where subjects have to find back to a previously learned place. Such tasks are called homing tasks. The systematic errors in the endpoint locations are informative about the internal representations that were used to navigate. A common assumption is that subjects recognize a place by matching the memorized place code of that place with the currently perceived surroundings, the errors in the endpoint positions depend on the errors that lie in the place code and the errors that lie in the perception of the surroundings. This can easily be illustrated with an example in a one dimensional world where a place was memorized in form of a place code consisting of the distance to one landmark. It is assumed that the memorized distance and the distance perceived by the one dimensional agent have the same underlying error distribution because both measurements where made in the same way. If this error distribution is Gaussian with standard deviation σ, the predicted error distribution of the endpoint locations will be Gaussian as well with a standard deviation of 2σ and mean at the true goal location. Evidence for the use of a particular cue can be obtained by modeling the errors that occur during perception and how they propagate to the endpoint locations, and then compare the resulting error distribution to experimental data. This method can be used to study the representation used to remember locations when landmarks are available.

In a series of experiments with similar conditions, Pickup et al. tested the predictions of three different models of visual navigation. In an immersive virtual reality setup, participants viewed a room with three thin long vertical poles with different colors

positioned close to each other, so they could all be seen at once. Participants could only walk on a straight line of defined length, allowing them to have access to depth information through motion parallax. They had to remember a certain position on that line, before being teleported to a new location. The participants had to walk back to the previously learned place and indicate when they thought that they had reached the goal position. Different pole configurations were tested. The pole configurations were chosen in a way that the predicted probability distributions of the different models differ as much as possible. In an additional experimental condition, several cues like furniture and objects were placed in the room. Indeed different error distributions were observed for different pole configurations [Pickup et al., 2013, Gootjes-Dreesbach et al., 2017]: In the cue rich condition, the errors were very small, and therefore uninformative with respect to the strategy that the participants used. The first model had place codes consisting of position estimates of the three poles in an egocentric coordinate system. The modeled errors were physically plausible for position estimates through motion parallax [Pickup et al., 2013]. The second model was an extension of the first. Place codes consisted of the three distances between the poles. The fact that errors propagate from the egocentric position estimates to the relative position estimates makes this model a model with a world centered reference frame [Pickup et al., 2013]. Both the first and the second model can be described as reconstruction type models, because position estimates are reconstructed from views. The third model was a view based model. Many different monocular and binocular features can be extracted from a view. In previous work the authors had identified the two most predictive ones. The place code consisted of the largest of the three visual angles spanned between the poles and the disparity gradient between the two poles standing closest to each other. The disparity gradient describes the inclination of a plane relative to the viewing direction. This measure of relative depth is obtained from the change in vergence angle between two poles divided by the angle between them. The errors of all angular measurements where assumed to be Gaussian [Gootjes-Dreesbach et al., 2017].

The two reconstruction based models were each better than the other model under conditions with different pole configurations [Pickup et al., 2013]. The view based model outperformed the egocentric reconstruction model under all conditions. The world centered reconstruction based model and the view based model are not directly compared in this study. The authors concluded that a view based representation of visual and depth information is more likely to be present in spatial working memory,

than an egocentric one, and that it is possible to construct experiments that can differentiate between view based and egocentric reconstruction based models of visual homing [Gootjes-Dreesbach et al., 2017].

Mallot and Lancier found evidence for the presence of egocentric distance estimates in spatial working memory [Mallot et al., 2017]. In an immersive virtual reality setup, participants had to navigate with the help of four distant sphere shaped landmarks. The most important difference of the situation in this experiment compared to the situation in the three pole experiments by Pickup et al. was that the place the participants had to navigate to, was located within the square formed by the four landmarks. The participants therefore could only see two out of the four landmarks at a time. The navigation errors where recorded for different landmark configurations, and compared to the predictions of a model with place codes based on egocentric distance estimates and the predictions of a model with place codes based on visual angles. The recorded error distributions had two characteristic properties: An elongation toward the most distant landmarks and a systematic bias away from the most distant landmarks. The view based model failed to predict both properties of the recorded error distributions. The distance based model predicted the elongation towards the most distant landmark [Mallot et al., 2017]. A systematic bias suggests that either the place code or the perception of the surroundings is biased. If both the place code and the perception of the surroundings were biased, this bias would cancel out and no systematic bias would be observed in the navigation errors [Mallot et al., 2017]. As the place code is constructed from information gathered over a longer period of time than the momentary perception of the surroundings, one may assume that place codes are unbiased but the content of spatial working memory is [Mallot et al., 2017]. By adding a parabolic compression to the perceived distance estimates, the distance based model was able to predict both the shape of the error distributions and the systematic bias [Mallot et al., 2017]. A parabolic compression of perceived depth is plausible as it was already observed by Gilinski [Gilinsky, 1951]. A common explanation why egocentric reconstruction based strategies (as opposed to view based strategies) seem to be used in situations where landmarks can't be seen at once, is that visual angels between landmarks have to be estimated from multiple views cashed in spatial working memory. Therefore, they are less precise and less useful for navigation than distance estimates [Gootjes-Dreesbach et al., 2017]. An additional reason why depth through parallax is more informative than visual angles is sensitivity. The sensitivity of visual angles and parallax to distance decreases

with the inverse square of the distance. While the sensitivity of visual angles stays the same, the sensitivity of parallax can be increased by increasing the distance between the two observation points. This advantage especially kicks in for situations with small and distant landmarks, like in this experiment by Mallot and Lancier.

While in some situations world centered representations may be important, in other situations there is more evidence for the use of views or egocentric reconstructions of the surroundings [Mou and McNamara, 2002, Gillner et al., 2008, Gootjes-Dreesbach et al., 2017, Mallot et al., 2017]. Pickup et al. showed that a view based model including depth information outperforms an egocentric reconstruction based model in situations where all landmarks can be seen at once. As described above Mallot and Lancier showed that a egocentric reconstruction based model outperforms a view based model not including depth information in situations with distant landmarks that can't be seen at once.

The results of Mallot and Lancier suggest that representations in spatial working memory have an egocentric reference frame and the results of Pickup et al. suggest that image and depth information in spatial working memory has a view centered reference frame. In both experiments image and depth information was available to the subjects. In order to investigate if depth information is used as a feature of a view or if it is used to build a reconstruction of the environment, a situation is needed where only depth information is available. This is the case for the experimental environment presented by Halfmann [Halfmann, 2015]. In the present thesis, we have used the Halfmann data for comparing the performance of the two models presented by Mallot and Lancier.

For two reasons, we hypothesize the egocentric reconstruction based model based on distances performing better than the view-based model based on visual angles:

i) In the present experiment, visual angles can only be obtained from multiple views which is a complication leading subjects preferring the use of distances for navigation in similar experiments [Waller et al., 2000, Mallot et al., 2017].

ii) In addition, there is indication that corners of a room (the only cues defining visual angles) are poorly useful for navigation in this particular experimental setup [Halfmann, 2015].

Hence, if the view-based model still performed better under the given conditions,

this would be a strong argument for processing of depth in a view centered reference frame.
Unfortunately, distances in the present experimental setup are too small and goal positions too close to walls for any systematic bias to occur because of compression of long distances. Therefore, there is virtually no information from the bias of endpoint distributions. Apart from the shape of endpoint distributions, bias information actually provided strong evidence in favor of the reconstruction-based model in the work of Mallot and Lancier. In the present work, however, the only source of information used for discriminating between the two models is the shape of endpoint distributions.

2 Methods

2.1 Subjects and Procedure

Before starting with the experiment, the 40 subjects had to do a preliminary test, in order to verify that they were able to perceive stereoscopic depth. In the main experiment, the subjects were asked to perform a "return to cued condition task" [Gillner et al., 2008] in a kite-shaped room. During the first phase of the experiment, the participants were placed at one of the three goal positions where they could look around and perform small translational movements. The second phase commenced by setting the subjects back to a start position from where they returned back to the goal position by using a joystick. They indicated recognition of the goal position by pressing a button and were then teleported to the true goal position. The next trial started from there with the goal being one out of the two remaining goals. Each subject completed a total of twelve trials i.e. two trials for each possible transition between the three goal locations.

2.2 Virtual Environment

In the results reported here, the virtual environment was presented to the subjects via an Oculus-Rift stereoscopic head-mounted display. The environment was specially designed to allow for depth perception but exclude all other visual cues [Sperling et al., 1989]. The walls of the kite-shaped room were defined by dynamic random dots distributed uniformly in the field of view. The dots had a lifespan that was randomly chosen between 100 and 200ms. After this time, the dots disappeared, and they were instantaneously replaced by other dots keeping the total number of dots constant. Because the position of the dots always changed, they could not be used for pattern recognition. The only cue available to the subjects was depth through stereo disparity and motion parallax [Sperling et al., 1989]. As a consequence, the

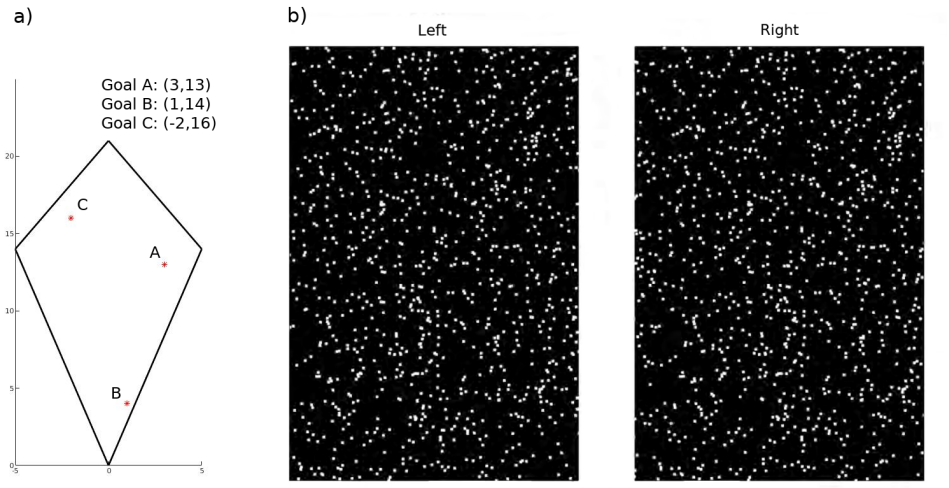


Figure 2.1: **a)** Layout of the virtual environment with the three goal points. The goal points are also used as starting points. **b)** Exaple view on the virtual environment [Halfmann, 2015].

structure of the room was barely visible without moving or turning the head. When moving/turning however, the structure of the room suddenly became visible with depth information stemming from parallax only.

2.3 Models

The models used in this work are kept as close as possible to the models presented by Mallot and Lancier (2017) and were only modified for closed rooms. The distance-based model assumes that depth information is used to build an egocentric reconstruction of the surroundings. One of the simplest variants of egocentric representation was chosen: Places are represented by the shortest distance to each of the four walls. This model is related to the reconstruction based model proposed by Pickup et al. (2013). The view-based model assumes that places are represented by views, and that depth is processed in a view-centered reference frame. Views are described by four visual angles, one for each of the four walls. This representation is similar to the representation used in models based on panoramic views with edge detection [Cheung et al., 2008] and to the view-based model proposed by Pickup et al. (2013).

2.3.1 Distance Based Model

In the original reconstruction based model by Mallot and Lancier (2017), the place code was composed of an egocentric distance measurement and an allocentric bearing measurement for each landmark. It was assumed that the agent had access to the identity of the landmarks and its own allocentric bearing. Because there is not one distance and bearing of a wall in a closed room, we have chosen to use just the shortest distance to each of the four walls. This set of four egocentric distance measurements formed the referential place code $\boldsymbol{c}_i$. At position $\boldsymbol{x} = (x, y)$ the agent will take four measurements $\boldsymbol{m}_i$ on the representation of their surroundings in spatial working memory. Those measurements are subjected to errors and therefore considered random variables drawn from a Gaussian distribution with expected value $\boldsymbol{m}_i$ and variance σ_i^2. The likelihood that the agent recognizes the goal at a given location is the likelihood that it takes a measurement equal to the remembered place code. This probability is described by the following equation:

$$L(x) = \prod_{i=1}^{4} \mathcal{N}(m_i | c_i, \sigma_i^2)$$

The perceived distance $\boldsymbol{m}_i$ was compressed as proposed by Gilinsky (1951):

$$\boldsymbol{m}_i = \frac{||l_i - x_i|| * 60}{||l_i - x_i|| + 60}$$

$\boldsymbol{l}_i$ is the closest point on the wall. The distance of "infinity" at $60m$ was taken from Mallot and Lancier (2017). Under the conditions of small distances investigated here, compression has only a minor effect on the results. But for the purpose of comparison, we still have adopted compression. The error distribution of a distance measurement through parallax is not Gaussian, but has a tail on the side away from the observer [Bailer-Jones, 2015]. The error modeled here is the resulting error from depth perception and the processing of egocentric reconstruction of the surroundings. Here the resulting error is approximated by a Gaussian distribution with variance dependent on the distance that is measured. The dependence of variance and distance is described by this quadratic equation:

$$\sigma_i^2 = s * ||l_i - x_i||^2 + a$$

As the sensitivity of parallax decreases with the inverse square of the distance, the variance should increase with the square of the distance. The original model by Mallot and Lancier (2017) had no y-intercept $\boldsymbol{a}$. This additional free parameter was included because errors are not expected to approach zero for small distances. When the measurement error is small, other sources of error become important. Errors could for example arise from the retrieval from memory, the process of comparison with the remembered place code or even from the limited precision of movement by the subject.
The free parameters $\boldsymbol{s}$ and $\boldsymbol{a}$ were fitted using the standard simplex algorithm, in order to maximize the likelihood to observe the experimental data under the predicted distribution. If an endpoint was better explained by the distance-based model with rotated wall identities, it was considered a rotation error and excluded from the optimization process (see *figure 2.4*). The optimal parameters were:

$$s = 0.1092$$

$$a = 0.3483$$

2.3.2 View Based Model

The view-based model proposed by Mallot and Lancier has place codes composed of eight visual angles, two for each side of the four landmarks. The agent was assumed to have access to the identity of the landmarks and thus access to its own allocentric bearing. In the model presented here, the place code $\boldsymbol{c}_i$ is composed of the four visual angles that the four walls take up. We decided against using the bearings of the four corners, because results by Halfmann (2015) suggest that corners are not used for place recognition in this experimental setup. At position $\boldsymbol{x} = (x, y)$ the agent will take four angle measurements $\boldsymbol{m}_i$ on the panoramic view-based representation of the surroundings in spatial working memory. The likelihood of place recognition L(x) is described by the following formula:

$$L(x) = \prod_{i=1}^{4} \mathcal{N}(m_i | c_i, \sigma^2)$$

The angle measurements were modeled with a Gaussian error and a fixed variance σ^2. Visual angle estimates are independent of the distance to objects. The main

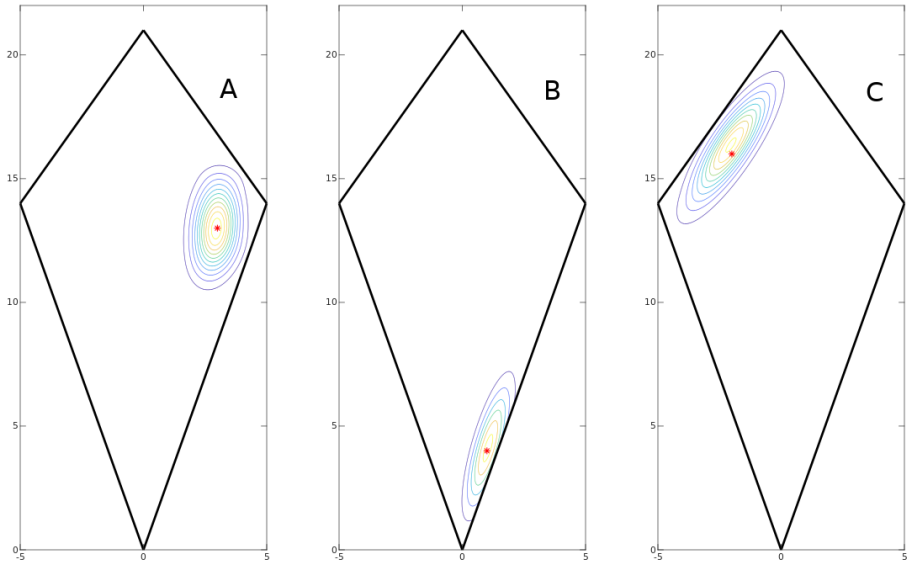


Figure 2.2: Contour plot of the likelihood $L(x)$ of place recognition when at position $\boldsymbol{x} = (x, y)$ as predicted by the **distance-based model**. The maximum point of $L(x)$ does not coincide with the true goal position marked by the star (**red**). This systematic bias arises from the parabolic compression of distance estimates.

component of the errors modeled here is the error that arises from building a panoramic view from multiple views. The retrieval of the place code from memory, and the comparison process also contribute to the error. In this experiment visual angles are retrieved from depth information which is subject to errors and bias from depth perception. It can be shown that the compression of depth estimates and the errors that arise from depth perception do not affect the angles at which the corners of the room appear in the visual field (for explanation see supplementary material). The free parameter σ^2 was fitted using a simplex algorithm, in order to maximize the likelihood to observe the experimental data under the predicted distribution. If an endpoint was better explained by the view-based model with rotated wall identities, it was considered a rotation error and excluded from the optimization process (see *figure 2.4*). The optimal parameter was:

$$\sigma^2 = 20.48$$

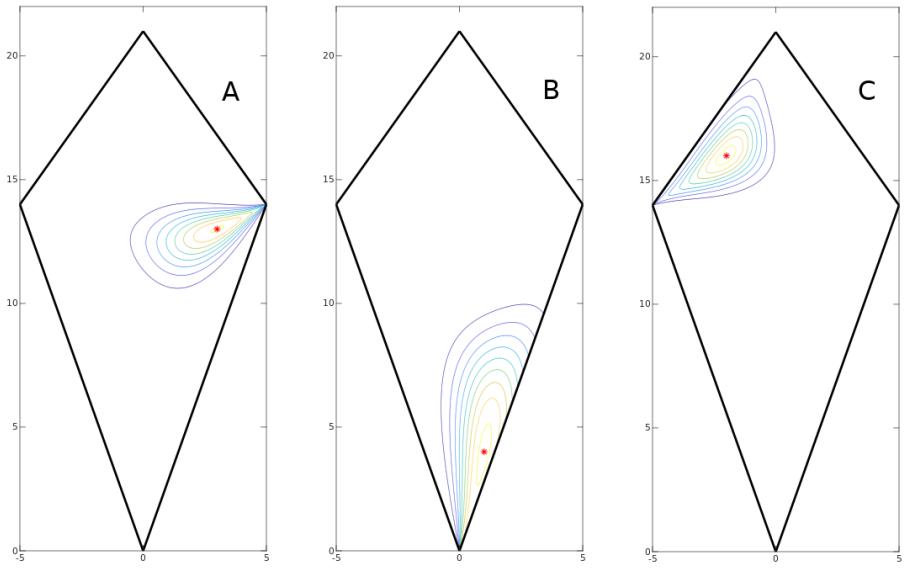


Figure 2.3: Contour plot of the likelihood $L(x)$ of place recognition when at position $x = (x, y)$ as predicted by the **view-based model**. The goal position is marked by the star (**red**).

2.4 Description of Experimental Data

The subjects showed a performance way above chance level [Halfmann, 2015] i.e. the endpoint locations only varied moderately around the true goal location. The distribution of the endpoint locations has a distinct shape for each of the three goal locations. In some trials the subjects seem to have confused the goal with completely different positions. Those qualitative errors occurred especially in trials with goal position A and C. Halfmann (2015) suggested that those errors are "rotational errors" meaning that the identity of the walls were confused while their order remained the same. Unfortunately, the low number of this type of error occurring in this experiment does not allow for deciding whether or not this claim corresponds to the predictions made by the models presented here. We considered the maximum point

of the likelihood functions with permutated place code as the points of confusion.

$$x^* = argmax_x(L(x))$$

We included the points equivalent to the goal location as predicted by the respective model with rotated wall identities to *figure 2.4*.

2.5 Statistical Testing

A bivariate goodness-of-fit test is needed to quantitatively determine which of the two models better predicts the data. Due to boundary effects caused by the walls, the use of statistical procedures with normality assumption are excluded from the outset. We therefore decided for the multivariate Kolmogorov-Smirnov test. This test essentially applies the Kolmogorov-Smirnov statistic to every conditional distribution of the empirical distribution function and corrects the global significance level through Bonferrony-Correction [Justel et al., 1997]. The Kolmogorov-Smirnov test can reject the null-hypothesis that two samples are derived from the same distribution at a certain significance level.

For each condition (goal location A, B, C) a random sample of $N = 2000$ was drawn from the two distributions predicted by the two models [Ursel, 2020]. A multivariate Kolmogorov-Smirnov test was then performed to test whether the sample and the experimental data were derived from the same distribution. Data from trials that were considered rotation errors by the respective model were not included in the test. After removal of those trials the cardinality of the set was $N_A = 85$, $N_B = 90$ and $N_C = 62$ for the distance-based model and $N_A = 79$, $N_B = 90$ and $N_C = 55$ for the view-based model.

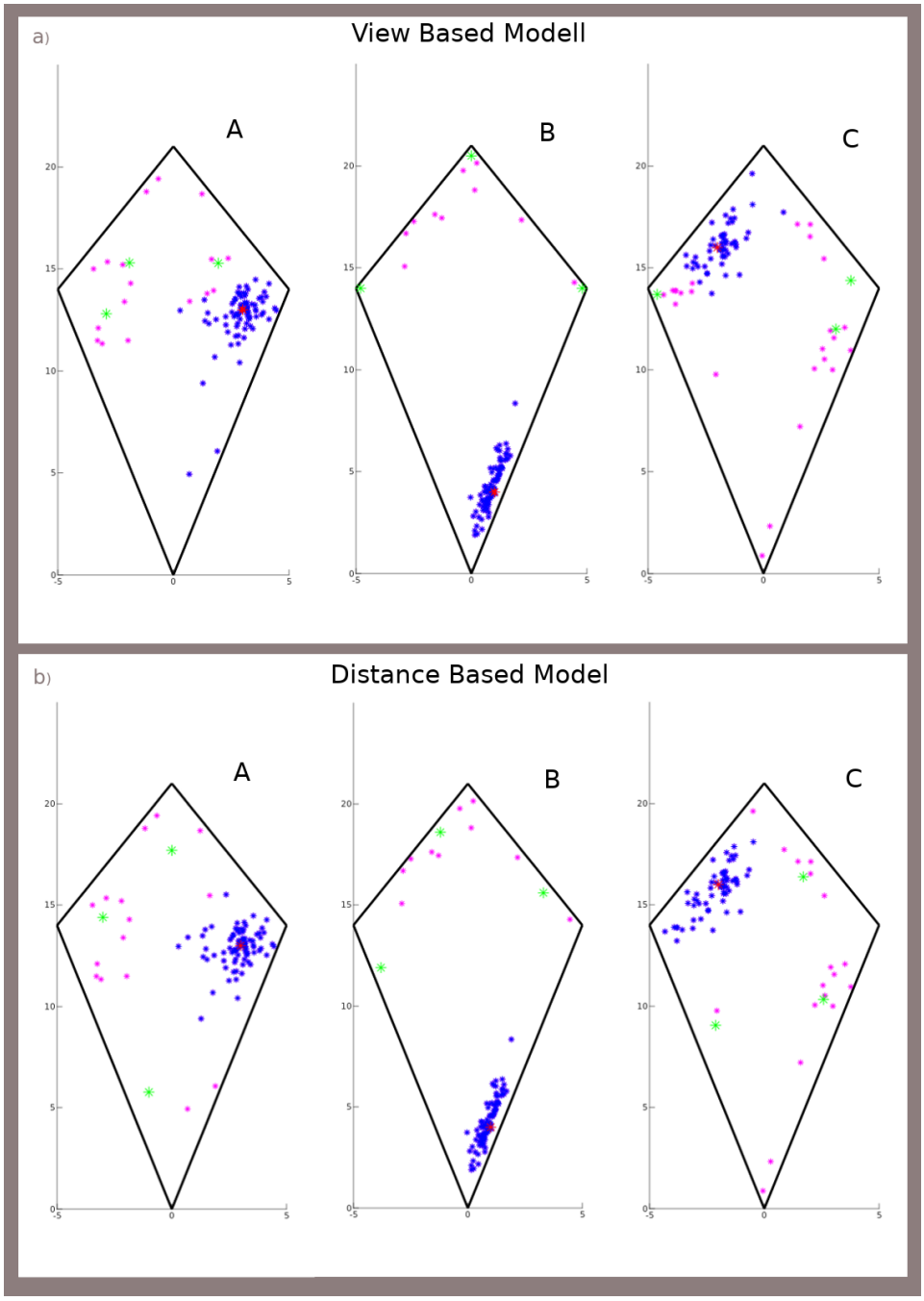


Figure 2.4: Decision points (**blue**) for the three goal positions (**red**). The decision points marked in **magenta** were better explained by the respective model with rotated place code, and were not used during the optimization of the model and statistical testing. The locations marked in **green** are the points of confusion predicted by the respective model.

3 Results and Discussion

The results of the bivariate Kolmogorov-Smirnov tests for both models and all three goal positions (A, B, C), as summarized in table 3.1 and figure 3.1, provide some support for our hypothesis that the distance-based model performs better than the view-based model. Especially for goal position B (close to one of the short walls), the p-value for the view-based model falls short of the significance level of $\alpha = .05$, and it is only slightly above this level for position A (close to one corner). On the other hand, for the distance-based model, the p-value for all three goal positions remain well above this level. As the bivariate Kolmogorov-Smirnov test is a conservative test, this in itself is a strong case against the view-based model presented here.

To be more specific: the view-based model predicts by far too much variance for goal position B (figure 3.1.). For goal position A and C one can clearly see that the classification of endpoints as rotation errors has an influence on the shape of their distribution. This weakness of the experiment presented here can be resolved by designing the environment in such a way that the points of confusion as predicted by both models do not lie too close to the true goal location.

Generally speaking, goodness of fit tests like the bivariate Kolmogorov-Smirnov test cannot prove that experimental data are the result of an underlying theoretical distribution. This type of tests can only exclude (disprove) a hypothetical model at a certain significance level. In order to get more certainty about whether or not a model of behavior is accurate, it is necessary to study whether it can predict multiple properties of behavior. In the open field, e.g., bias away from the landmarks that are further away was used as additional property [Mallot et al., 2017] (see also end of Introduction). This bias could be predicted by the distance-based model with compressed distance estimates in spatial working memory and a place code composed of accurate distance estimates. The view-based model was not able to predict this bias. Like on the open field, in closed rooms and with perception limited

Goal position	view-based model	distance-based model
A	D = .5220, p = .0799	D = .4150, p = .2606
B	D = .5783, p = .0492	D = .3961, p = .3661
C	D = .4566, p = .2039	D = .9466, p = .1396

Table 3.1: The bivariate Kolmogorov-Smirnov statistic D and the corresponding p-value for the thee experimental conditions. With a significance level of $\alpha = .05$ the null hypothesis was rejected for the view-based model and goal position B e.g. it is unlikely that the endpoints presented here are the result of the distribution predicted by the view based model. It can be stated that the distance-based model outperforms the view-based model for goal position A and B.

to depth, a view-based model does not predict a systematic bias in the endpoint locations (see supplementary material). For making use of systematic bias as an argument for the distance-based model in a closed room environment, the room needs to be scaled up as to increase the effect of parabolic compression and to reduce boundary effects caused by walls.

The broader context of the present work actually is to learn more about the reference frame in which depth information is processed in spatial working memory when performing a homing task. In this study, we considered an egocentric and a view centered reference frame. In particular, we compared the ability of two simple models based on the comparison of a remembered place code and the content of spatial working memory for predicting the endpoint distribution in a visual homing task with visual perception reduced to depth. We showed that, under the given conditions, the model with place codes based on egocentric estimates of the distance to landmarks performs better than a model with place codes based on visual angles.

Since only those trials were considered in which the subjects successfully performed the homing task and only data close to the true goal location were used for the statistical tests, above conclusion only hold for the final approach of the goal.

Successful homing requires the correct identification of landmarks (here walls) which, in a closed room, is equivalent to knowing the bearing of the view that is currently perceived. In the models presented here the wall identities were as-

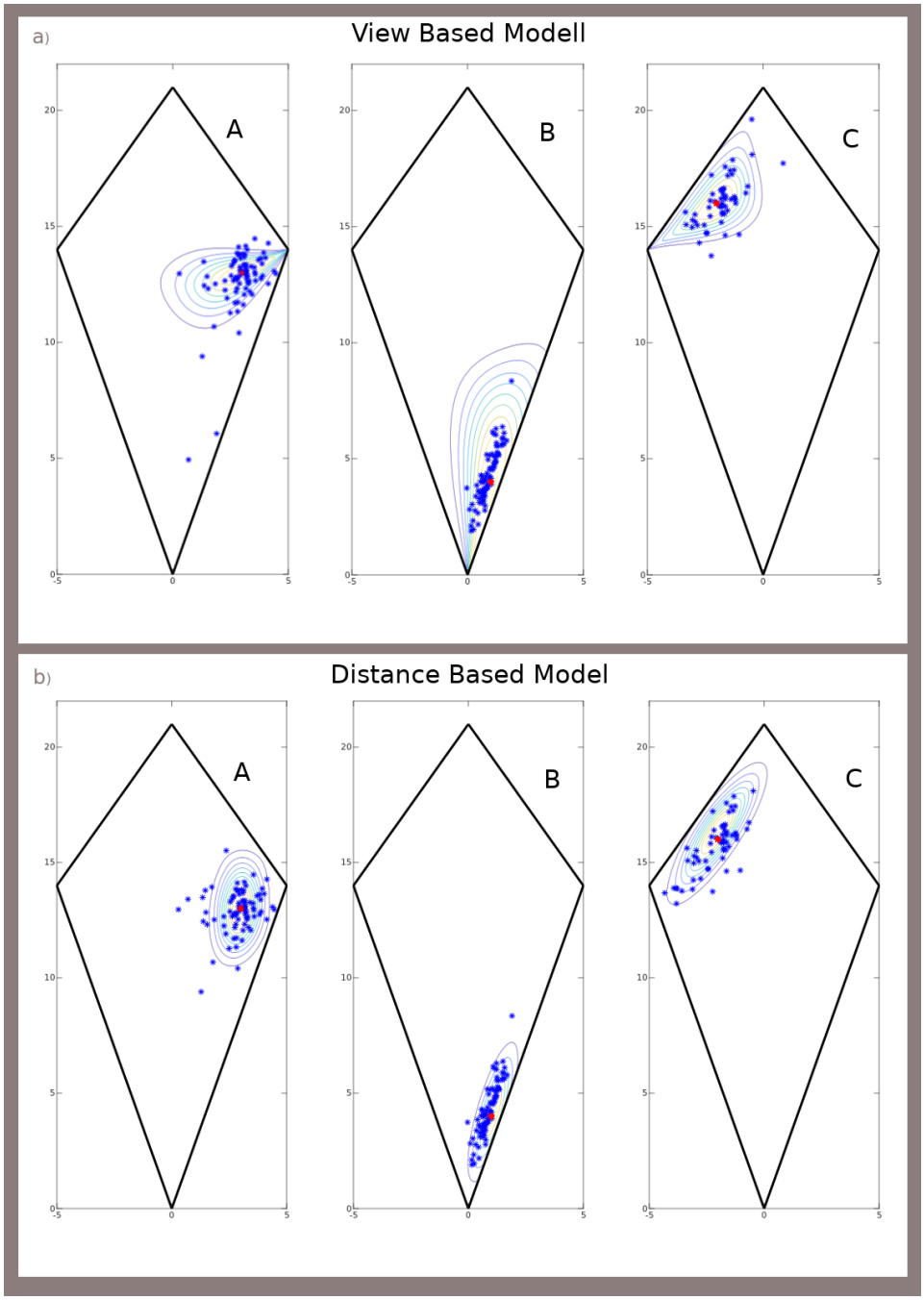


Figure 3.1: Endpoint locations that were not considered a rotation error by the respective model (**blue**). Contour plot of the likelihood $L(x)$ of place recognition at position $x = (x, y)$ as predicted by the respective model.

sumed to be known to the agent. Homing involves identifying the own position an bearing in some kind of reference frame, and a process of relating that position to the goal position [Pickup et al., 2013] e.g. the approach of the goal is preceded by the recognition of the own position. Confusion of landmarks (walls) during comparing the remembered place code to the content of spatial working memory lead to rotational errors. The identification of the own relation to the remembered representation of the surroundings at the goal (place code) may happen several times during the navigation as subjects sometimes changed their heading completely [Halfmann, 2015]. The work presented here has provided evidence that place codes contain egocentric distance estimates to identified landmarks, however our results do not exclude that the process of identifying the own orientation relative to the remembered place code happens with the help of depth information processed in a view centered reference frame. It still has to be clarified how the orientation of the world is deduced from depth pictures and egomotion during the first moments of a homing task. A more comprehensive model of place recognition in closed rooms should also be able to predict the errors that occur during the entire process of homing e.g. the directions in which subjects erroneously walk, the rotational errors they make as well as the precision errors around the true goal location.

Broader models of spatial working memory that can explain the whole process of homing have been developed already. They range from graphs putting views into their allocentric relation [Röhrich et al., 2014] to egocentric reconstructions of the world with identified landmarks [Loomis et al., 2014]. But those models are not suitable for the type of quantitative predictions like the models we presented in this work.
Both an allocentric graph of views and an egocentric reconstruction of the surroundings could be present in spatial working memory (simultaneously). The results presented here only provide evidence for the availability and use of the latter during the final approach of the goal in homing tasks in closed rooms where depth is the only visual cue available.

Bibliography

[Bailer-Jones, 2015] Bailer-Jones, C. A. L. (2015). Estimating Distances from Parallaxes. *Publications of the Astronomical Society of the Pacific*, 127(956):994–1009.

[Cartwright and Collett, 1983] Cartwright, B. and Collett, T. S. (1983). Landmark learning in bees. *Journal of comparative physiology*, 151(4):521–543.

[Cheng, 2008] Cheng, K. (2008). Whither geometry? Troubles of the geometric module. *Trends in Cognitive Sciences*, 12(9):355–361.

[Cheng et al., 2007] Cheng, K., Shettleworth, S. J., Huttenlocher, J., and Rieser, J. J. (2007). Bayesian Integration of Spatial Information. *Psychological Bulletin*, 133(4):625–637.

[Cheung et al., 2008] Cheung, A., Stürzl, W., Zeil, J., and Cheng, K. (2008). The Information Content of Panoramic Images II: View-Based Navigation in Nonrectangular Experimental Arenas. *Journal of Experimental Psychology: Animal Behavior Processes*, 34(1):15–30.

[Epstein, 2008] Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in Cognitive Sciences*, 12(10):388–396.

[Gardner et al., 2008] Gardner, J. L., Merriam, E. P., Movshon, J. A., and Heeger, D. J. (2008). Maps of visual space in human occipital cortex are retinotopic, not spatiotopic. *Journal of Neuroscience*, 28(15):3988–3999.

[Gilinsky, 1951] Gilinsky, A. S. (1951). Perceived size and distance in visual space. *Psychological review*, 58(6):460.

[Gillner et al., 2008] Gillner, S., Weiß, A. M., and Mallot, H. A. (2008). Visual homing in the absence of feature-based landmark information. *Cognition*, 109(1):105–122.

[Gootjes-Dreesbach et al., 2017] Gootjes-Dreesbach, L., Lyndsey, C., Fitzgibbon, P.

A. W., and Glennerster, A. (2017). Comparison of view-based and reconstruction-based models of human navigational strategy. *Journal of Vision*, 17(9):1–19.

[Halfmann, 2015] Halfmann, M. (2015). Place Recognition and Navigation in Virtual Environments.

[Justel et al., 1997] Justel, A., Pefia, D., and Zamar, R. (1997). STATISTICS and PROBABILITY LETTERS A multivariate Kolmogorov-Smimov test of goodness. *Statistics and Probability Letters*, 35:251–259.

[Kuipers, 2000] Kuipers, B. (2000). Spatial semantic hierarchy. *Artificial Intelligence*, 119(1):191–233.

[Larsson and Smith, 2011] Larsson, J. and Smith, A. T. (2011). fMRI Repetition Suppression: Neuronal Adaptation or Stimulus Expectation? *Cerebral Cortex*, 22(3):567–576.

[Loomis et al., 2014] Loomis, J. M., Lacey, S., and Lawson, R. (2014). Multisensory imagery. *Multisensory Imagery*, pages 1–435.

[Mallot et al., 2017] Mallot, H. A., Lancier, S., and Halfmann, M. (2017). Place recognition from distant landmarks.

[Mou and McNamara, 2002] Mou, W. and McNamara, T. P. (2002). Intrinsic Frames of Reference in Spatial Memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, 28(1):162–170.

[O'Keefe and Dostrovsky, 1971] O'Keefe, J. and Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*.

[Pickup et al., 2013] Pickup, L. C., Fitzgibbon, A. W., and Glennerster, A. (2013). Modelling human visual navigation using multi-view scene reconstruction. *Biological Cybernetics*, 107(4):449–464.

[Röhrich et al., 2014] Röhrich, W. G., Hardiess, G., and Mallot, H. A. (2014). View-based organization and interplay of spatial working and long-term memories. *PLoS ONE*, 9(11).

[Solstad et al., 2008] Solstad, T., Boccara, C. N., Kropff, E., Moser, M. B., and Moser,

E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science*, 322(5909):1865–1868.

[Sperling et al., 1989] Sperling, G., Landy, M. S., Dosher, B. A., and Perkins, M. E. (1989). Kinetic depth effect and identification of shape. *Journal of Experimental Psychology: Human Perception and Performance*, 15(4):826.

[Stürzl et al., 2008] Stürzl, W., Cheung, A., Cheng, K., and Zeil, J. (2008). The Information Content of Panoramic Images I: The Rotational Errors and the Similarity of Views in Rectangular Experimental Arenas. *Journal of Experimental Psychology: Animal Behavior Processes*, 34(1):1–14.

[Taube, 1998] Taube, J. S. (1998). Head direction cells and the neurophysiological basis for a sense of direction. *Progress in neurobiology*, 55(3):225–256.

[Ursel, 2020] Ursel, T. (2020). Generate random numbers from a 2d discrete distribution. *MATLAB Central File Exchange*.

4 Supplementary Material

In the following, we will explain why errors of depth perception do not affect the angles at which the corners of the room appear in the visual field.
When depth is processed in a view-based reference-frame, each depth estimate corresponds to a certain pixel of a panoramic view. In turn, each pixel corresponds to a certain angle in the visual field, and corners of the room appear as kinks in the panoramic depth picture. Any Gaussian noise in the depth estimates then blurs the corners, i.e. kinks indicating corners are smeared out to both sides of the corner leaving the corner position unchanged.
This is even true for systematic errors stemming from parabolic compression of depth estimates:

$$\boldsymbol{d}_{compressed} = \frac{\boldsymbol{d} * A}{\boldsymbol{d} + A}; \quad 0 < A$$

which is the systematic underestimation of distances, which increases with increasing distance. The equation above describes how the true distance **d** is compressed with **A** as the "distance of infinity". Here, the important point is that this compression does not change the order of the distance estimates, and hence the positions of the kinks in the picture do not change. Fig. 4.1. clearly slows that parabolic compression of depth estimates changes the perceived shape of the room but not the position of corners and with that angles defined by these corners and the position of the subject. We, therefore, conclude that models based on the perceived bearing of corners are not affected by statistical (Gaussian) and systematic errors (parabolic compression) of depth perception. Such errors reduce the contrast in the depth picture and lessen the difference between views, but they do not create a biased perception of visual angles. Models based on the smallest distance to the walls, however, are affected by the bias arising from parabolic compression.

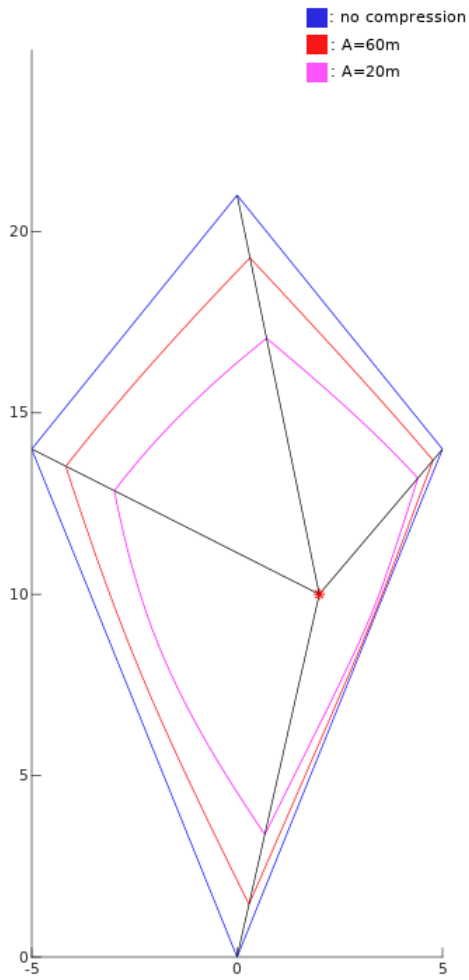


Figure 4.1: Shape of the room as perceived from position $x = (7, 10)$. With no compression (**blue**), compressed with $A = 60m$ like in the simulation presented in this work (**red**) and compressed with $A = 20m$ (**magenta**). The black lines are straight lines through $(7, 10)$ and the corners of the room.