Exploring the usage of edge gradients within images to perform coarse localisation

Keele University

MPhil Computer Science

Dean Graham Jarvis

March, 2021

Abstract

With the rise in autonomous systems being integrated into the world around us, it has become increasingly important that theses systems have functions that allow the navigation of environments. One of the key functions is the recognition of the environment in which the system resides. This thesis seeks to contribute to methods that a given system can use to recognise an environment. To do this, an omni-directional camera is used to produce images of locations which contain sharp edges that lay at certain angles. By counting the pixels on these sharp edges and putting them into histograms based on the corresponding angles, a data structure can be formed to describe the location depicted in the image. This data is taken from multiple images over two locations and then compared to one another. These comparisons show that a system can differentiate between images of locations with this data structure showing a significant difference between two locations. Knowing this, it was then analysed how the differentiating ability of this kind of system developed as the amount of locations increased. This was done by increasing the amount of locations and having the system make a decision as

to whether two images belong to the same location. This is then compared to how a human participant performed with the exact same image set. This experiment needs to be performed on a larger data set for any kind of statistical significance, however these initial results show that there is a steady decline in the ability to differentiate between images with this system. However the system had a very high false positive rate which is something that should be studied in more detail.

Contents

Definitions

Definition 1. *For the duration of this thesis, an alternative definition of position and location are used. Position will refer to an exact measure of where the robot is (this is analogous to exact coordinates x,y). Pose define the position (x,y) and orientation (theta) of the robot. Location will refer to a set of positions that share the same contextual name, for example the kitchen or the lab. This coarse localisation will attempt to differentiate between locations as opposed to positions.*

Definition 2. *Simultaneous localisation and mapping (SLAM) refers to a problem within robotic navigation wherein a robot or agent of some description must simultaneously create a map of an unknown environment whilst also estimating its own position within that environment[18].*

Definition 3. *Features in the context of this thesis refer to something distinct and prominent about the input information. This could be something like the the distribution of colour in the image or a list of corners and their positions and sizes.*

Definition 4. *Features within an image can be broadly classified into global features and local features[14]*[1]*. Local features, within the context of an image, are features that are located at specific positions within the image. Examples of a local feature may be the positions of corners their orientation and their size. Global features on the other hand describe very broadly the whole image with some metric. An example of a global feature may be the distribution of brightness across the image. Global features tend to be much faster to compute but are much more generalised and hence contain less information. Local features are much harder to find and compute as there are usually many of these features per image but the benefit is that they contain a lot more information.*

[1] Although it is worth mentioning that some researchers also break local features down into 2 more types of features which are block based local features and region based local features [43].

Assumptions

Assumption 1. *The information provided by the simple geometric shapes such as edges and corners will provide enough information to be used in the process of localisation. It is known that this information alone is not used within the neurological processes of animals to perform the task of localisation. This work extrapolates the importance of this information in order to produce a model of how it may be used in the process of localising.*

Assumption 2. *Each location will have a unique edge gradient distribution.*

Assumption 3. *A human participant would be able to determine whether two images were from the same location regardless of the perspective provided the two images are from the same perspective.*

Chapter 1

Introduction

1.1 Motivation

As technology such as sensors and high performance low power processors becomes more affordable and advanced, the usage of ever more complicated systems has become much more common in our lives. For example, the use of autonomous systems for road vehicles has started to become a point of interest for technology companies and vehicle manufactures[35]. The sensors and processors have allowed these functions have become more accessible over the years and can add to the convenience of the driver as well as making driving a much safer endeavour. These systems range from integrating low

risk manoeuvres (Level 1&2) such as cruise control and parallel parking the vehicle, all the way to automated driving on well known roads such as motorways (level 4)[35][1]. In the home, automated systems such as small vacuum robots like the Dyson 360 eye[21] and social robots such as pepper[45] are beginning to be integrated into peoples lives. This again is happening to the increase of the availability of the systems that underpin these devices and the convenience that it provides the consumer. All these autonomous systems need the ability to self navigate in their respective environments. In the case of autonomous vehicles, the use of GPS and pre-built maps allows them to navigate[35]. For indoor robots however, the accuracy required can be on the order of centimetres and although GPS can provide an accuracy up to the centimetre range, the typical accuracy in a medium density city is on average 2m[40]. In these situations, the use of systems that implement simultaneous localisation and mapping (SLAM)[20, 6] could be used to produce a map of the internal environment and navigate using this constructed map. The map is built while exploring an environment. To do this the robot must compare the relative positions of landmarks that are currently visible to the set of

[1]The levels of autonomy are a taxonomy of the differing capabilities that a vehicle may have in terms of autonomous driving. This particular taxonomy is defined by the society of automotive engineers (SAE) at the following website `https://www.sae.org/standards/content/j3016_201806/`

past locations landmarks, if a match is not found then a position is added to the map. If the robot recognises this position then it checks whether it is connected to the last visited location, if it is, nothing changes, if it is not, then a connection is made. Theoretically, this problem has been solved and we know how to perform this task, but there are issues in the practicality of the implementations[44]. These issues lay in the usage of long term information, the scaling of the map as new areas are explored, the association of data from different sensors to common origins, and robustness to adverse conditions[12, 27]. The issues of long term data usage, data association, and scalability are linked problems. These issues are bound by the quadratic scaling of required computation with the amount of landmarks in a map[44]. This limits the size of environments that SLAM systems can function in real time. The issue of robustness refers to the ability of the robot to consistently detect the correct position in differing environmental conditions and minor changes in the positions of objects in the location that the position is in. For example, if the lighting conditions are different due to the time of day. This will affect what features (such as land marks) are extracted from the robot's immediate environment, if enough of the mapped features are not identified or if to many non mapped features are extracted then the robot will not

always get the correct position.

1.2 Literature

Navigation can be broken down into three distinct tasks, localisation, mapping, and path finding[1] and there are two ways to approach robot navigation: human conceived methods and biologically inspired methods. Human conceived methods include examples such as those described in[2, 4]. These methods have come a long way in producing reliable results for navigation but they often produce a lot of data and high fidelity maps which then require optimisation before they can be used efficiently (see[2]). This means that these methods have a high demand for memory and computational power. In contrast animals seem to do the same tasks without high fidelity maps, using primarily visual stimuli.

By taking inspiration from the way nature has solved problems we are provided with a stepping stone to be able to model and solve complex problems such as, minimising distances in networks with the use of ant colony optimisation[19] and searching for close fitting parameters for problems with numerous variables for a desired result using a genetic algorithm[32]. Some of

these tasks can be completed from simple rule sets that, when used in combination can result in more complex behaviours; emergent behaviours[10]. Examples of emergent behaviours can be seem in many places in nature; a good example can be shown in the hunting strategies of wolves[41]. Here it is shown that by using two simple rules, the behaviour of a hunting pack of wolves can be reproduced. There are two categories that biologically inspired computational methods can be split into[33], 'top-down' and 'bottom up'. *Top-down* methods stem from observations of how humans and animals behave. These methods may often produce fewer more simple rules but the quality of these rules may be subject to the interpretations of the observer, who may miss or misunderstand some of the behaviours. The research investigating the behaviour of wolf packs is a good example of a *top-down* approach to modelling natural systems. *Bottom-up* methods however make use of observations about the components of a natural system produce a given behaviour. A group of neurons and how they may respond to certain stimuli. A good example of this *bottom-up* approach is provided by[33], which looks at simulating the behaviour of head direction cells within the brain that aid in modelling the orientation of an agent within an environment. Another example of the *bottom up* approach is one where the place cells of rats are

used as a basis to produce a SLAM system known as Rat-SLAM[38]. In the hippocampus of the mammalian brain, the place cells are used to model the probability of the position. Rat-SLAM models these place cells as a competitive attractor network. The research describes this network as having excitatory connections to neurons that are close and inhibitory connections to neurons that are more distant. This causes the network to converge to a stable point based on the inputs into it. The behaviours that the *bottom-up* approach produces tend represent what seems like a simplistic mechanic behind a behaviour when compared to the whole behaviour of hunting of wolves, but being able to understand and reproduce these more fundamental systems will lead to more complicated systems and behaviours that can be reproduced by building upon this work layer by layer via abstraction.

Cameras are a commonly used type of sensor for any kind of robot due to the rich source of data, low cost, and small physical footprint. A lot of the information given from a camera may not be useful for a given task. To filter out the useless information, features are extracted from the images, this is analogous to how a brain may filter information. Features can consist of simple structures such as corners and edges, to more complex structures such as whole objects. However, one of the issues that can arise, due to

the use of visual information is the different lighting conditions in an environment may change as time passes. This can affect the features that are extracted from the environment in some algorithms due to over-saturation or under-saturation of light that makes strong features such as edges harder to see. Although many feature extraction algorithms do not explicitly consider colour information, they do rely on the contrast between regions in the image. This issue can be alleviated by looking into the research area of colour constancy, this area attempts to provide methods of making images invariant to changes in lighting conditions.

SLAM systems use comparisons between recent and past key points of information. In visual SLAM these points of information are part of the images received from the camera. This then raises the question of how might a system compare two images to measure similarity. Some methods look for common landmarks between two images using local features such as corners and objects that appear to have similar spacial relations to each other. Other methods look for global features such as distributions of the colour values across the images. These comparison methods need to be performed in real time whilst also being robust to perspective changes if they are to be used as an identifier for a location. Local features take longer to compute but

can provide a more detailed description of the location. Global features on the other hand are much faster to compute but give a more generalised description of a location.

The literature explored describes different methods of image comparison, colour constancy, localisation, and simultaneous localisation and mapping (SLAM). This literature informs decisions made throughout the experiments performed and is relevant to the use of computer vision as a primary sense for the localisation method. There are 2 experiments that were performed during the course of this thesis. The first experiment explored whether the use of a global feature that is composed of the angles of the edges in the room contains enough information to distinguish between two separate locations. A second experiment was then performed to test what the limit is on the amount of locations that can be differentiated using this feature.

The work in this thesis will be primarily based around the use of visual information as a key input into the task of localisation. There are some pieces of literature that perform these tasks without using visual information but these implementations can be limited in their feats. For example, I Ashokaraj et-al[5] use a combination of an inertial measurement unit, wheel encoders, gyroscopes, and ultrasonic sensors to estimate the robot's position on a 2

dimensional map. This process involved an extended Kalman filter[22, 29]. Other such methods are mentioned in a review of localisation systems by Deak et-al[17]. In the survey, it is mentioned that there are 2 broad categories of localisation: active and passive. Active localisation requires some form of artificial salient feature such as a tag to be added into the world for a system to be able to localise. This limits the available environments that an active localisation system can navigate. Passive localisation does not require any alterations to the external world meaning that it will need to extract salient features from the environment. Features in the context of this thesis refer to something distinct and prominent about the input information. This could be something like the the distribution of colour in the image or a list of corners and their positions and sizes.

The next few sections will be dedicated to issues faced within the use of camera systems and how you might use them to identify locations. Images are a very rich source of information[13] and so it is quite difficult to compare between images for similarities and differences without first transforming the image so that it can be represented by a collection of features. The transformation process can introduce extra salient feature such as borders and corners of shadows, these features are not stable due to the fact that

shadows can be influenced by the direction, the intensity, the position of the light source(s) and the number of light sources. Some of these issues can be tackled by looking into how an image can be altered into an illumination invariant form prior to feature extraction.

When beginning to look at image based localisation methods, it was necessary to investigate a few issues related to the how the images were to be used. These were, shall it use colour information or grey scale, and what kinds of features will be extracted from the image to help identify a location.

Initially the distribution of colours within the scene was used as a starting feature to extract and identify a location. Very quickly it became apparent that there would be an issue due to the varying lighting conditions. An alteration in lighting conditions can cause a scene to look drastically different as the objects within it are perceived as being a different colour. This is because an image recorded by a camera is dependant on three things: the content of the scene, the illumination of the content, and the camera properties[7]. Before deciding to use grey scale images a brief investigation was performed to check whether it was a problem that could be overcome. There are varying methods and assumptions that can help make a scene invariant to changes in illumination. Some methods are relatively simple and reply on statistical

techniques and assumptions; techniques like the grey world method[11]. Although praised for being simple to implement, it only performs as well as the initial parameters and it assumes a spatially uniform light source[23] making it unreliable in circumstances where there are multiple illuminates. Other methods are much more complicated to implement and use neural networks to predict the error in the hue values for small regions of images[9, 47]. The conclusion reached following the investigation of colour constancy is that, although there are methods to reduce the variance due to differences in lighting, the methods are either not reliable due to the assumptions made or are to resource intensive for real-time operation. This lead to the decision to use grey scale images for the feature extraction and the decision not to use colour distribution as the defining feature.

The next issue that is faced is determining what features to extract. There are a number of metrics that can be used to differentiate and compare between images taken of objects and locations. There are two types of features that can be extracted from an image; global features and local features[14] .Global features are pieces of information that broadly describe the whole image[34]. A good example of a global feature may be the colour distribution, which was mentioned earlier. Colour distribution is good at describing

an image whilst dealing with slight changes in the pose of the camera, but this feature also disregards a lot of other useful information such as spacial and texture information[42]. This leads to cases where two images that look very different being identified as being similar due to the similar distribution of colour. On the contrary, local features are a set of features that describe multiple points in an image. This makes them much more robust to changes to positioning of the contents and any occlusions. However, They require specialised classification techniques that can handle variable amounts of features[34]. A good example of a method used to extract local features is called SIFT (scale invariant feature transform). Sift is a popular method for feature extraction[26] as the features that are produced are scale invariant (can be recognised regardless of how much of the image it takes up). Local features tend to tackle issues such as scale, rotation, viewpoint or illumination variances, but with the increase in generalisation comes a higher demand for computational resources[26].

Both global and local features have their pros and cons and are usually used together, examples of such work can be seen here[14, 50]. Due to the time taken to implement and perform this study, only a global feature was used. The reason features extracted are extracted from an image is to re-

duce the dimensionality of the information collected (lower the amount of data that needs to be compared whilst keeping as much of the meaning as possible). Reducing the dimentionality of the data means that systems do not need to store as much data or use as much processing time searching and comparing that data. This process of reducing the dimentionality of the data also happens within nature. Take for instance reasearch by Anzai et al[3] which investigates the visual system of monkeys. This study shows how the information that is provided to the visual systems of monkeys is processed in layers. These layers feed into each other and begin to describe more and more abstract ideas. The research showed how neurons within the first two layers of the monkey visual cortex respond highly to simple geometric shapes such as lines, edges and curves. This information has been used and extrapolated in such a way as to provide an idea for the global feature that is used for this work. The feature that has been chosen for this thesis is information about the edges in the scene, specifically the orientations of the edges. It is assumed that despite the low level abstraction of the raw image that there is enough useful information to distinguish between any two given locations.

Research by Kosecka et-al[30] looks at using image comparison to localise

a device using a similar method to that used in this research. Kosecka et-al uses the gradients of edges as a feature but uses a standard limited perspective digital camera to take images. This means that one location may be given two or more separate labels. The clustering algorithm Kosecka et-al uses is known as learning vector quantisation. It is used to produce prototype histograms to represent classifications, in this case location. To compare these histograms the Chi-squared[24] statistic is used to compute a difference metric between a current histogram (from the query image) to all prototype histograms (the classes or locations). A confidence level is then produced by looking at the ratio of the smallest and second smallest Chi-squared value of this set. If confidence levels rise above 1.6 the classification is considered to be accurate. If the classification is achieved with a low confidence, it is refined by dividing the current image into sub images. These 5 images are then used in the same process to attempt to find a higher confidence match. There are 5 sub images in total. Four in the corners and 1 in the centre. The sub image comparison addition mentioned in this research is not clear about the exact method however.

1.3 Aim

There is a clear problem with the use of SLAM systems due to the large and complex data sets required for the navigation of large environments. This thesis explores a method to determine the location of a robot with a coarse localisation method. This could be as a preliminary step to limit the search space for high resource algorithms such as SLAM. Although there are implementations that refer to methods of coarse localisation, for instance work by Milford et al[39], they are usually refering to more coarse grids. In contrast to using whole locations that are defined by a system.

This thesis will explore the usage of the global feature of edge gradients/angles in a coarse localisation strategy. To do this, the feature must first be tested for robustness to small displacements in position and whether it can differentiate between two locations. If it can differenciate between two locations then how the distrimintion ability changes as the amount of locations increases.

1.4 Hypotheses

It is hypothesised that changes in position and orientation within the same location will produce no significant difference to the distribution. However it is also hypothesised that there will be a significant difference between changes due to position and changes due orientation; where changes in position will give higher differences than changes due to orientation. That is to say for any in pose within one location, the resulting change in the orientation distribution will have a lager component due to changes in position than for changes in orientation. It is also hypothesised that there will be a significant difference in the gradient distribution extracted between two locations, but that the ability for the system to differentiate locations will diminish as the amount of locations is increased. It is hypothesised that the accuracy will diminish as the amount of locations increases due to an increased chance of two locations having a similar enough descriptor that the system will not be able to separate them without the system

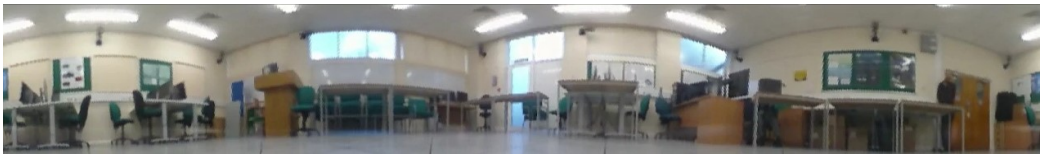
Chapter 2

Orientation of edge gradients as a global feature for image comparison.

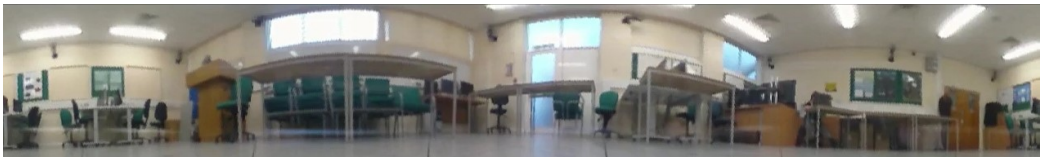
There have been studies that use edges and gradient information from images to compute image similarity[36, 28] and localise[31, 30]. However there is a lack of information about how this kind of information changes with respect to the pose of the camera and whether or not it can be used as a stand-alone method to differentiate between locations in the navigable environment. Before the edge gradients can be used as a feature for localisation, the method

must be tested to see whether it produces a useful set of data that can be used to differentiate between locations whilst being robust to small changes in position and orientation.

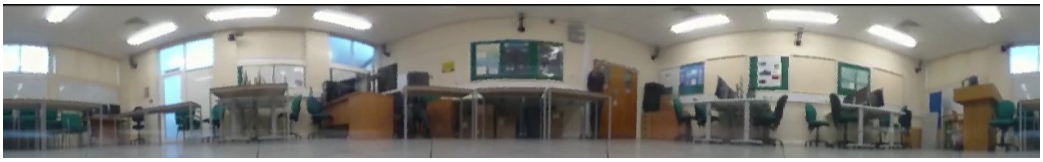
By looking at images such as those in Figure 2.1, there should be a greater difference in the gradient histograms due to changes in position than there will be due to changes in the orientation because the changing the orientation will not effect the angles of edges or the amount of edges. Changes in position seems to cause a greater distortion as this can affect how much of the image is taken up by a given edge (as it comes closer it takes up a larger portion of the image and visa versa). Also there should be a significant difference between the histograms of gradients between two different locations (between Figs. 2.1&2.4) that could be utilised as a determinant of location.



(a) Reference position image for empirical analysis.



(b) Image after moving camera forward 80cm showing an apparent distortion on the tables.



(c) Image after rotation the reference image by 60 degrees.

Figure 2.1: These photographs show the unwrapped omnidirectional images taken at two different positions in the same room. At first glance it seems that there is a greater apparent change in the images between the change in position compared to the change in orientation

2.1 Experimental setup and methodology

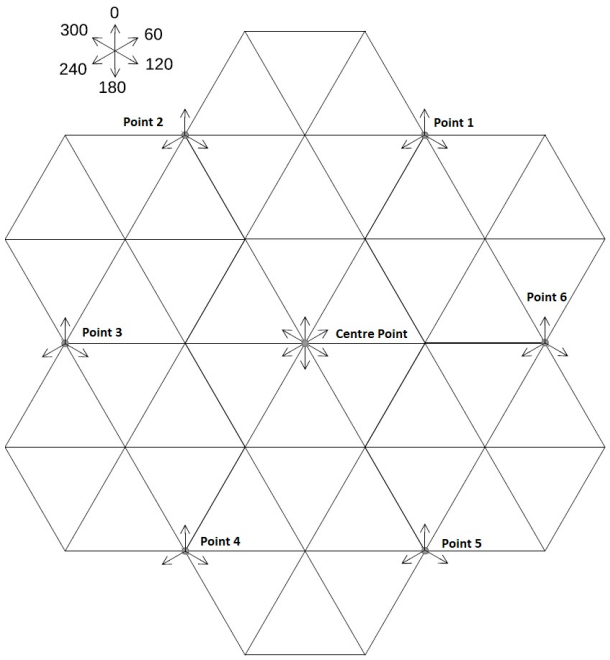
Figure 2.2: This image shows the camera with 360 degree lens attachment mounted with velcro on the back of an RC car.

A Sony Bloggie [1] camera with a 360 lens attachment, as shown in Figure 2.2, was used to capture all the images that were used in this study. A limited perspective camera was not used due to the fact that locations would require the capture of multiple views from which to be identified. Many animals benefit from wide fields of views that allow them to be aware of more of their environment without having to move around a lot to do so, this aids in hunting and detecting predators. An omnidirectional camera can handle this issue better as all possible views of a position are incorporated into a single image. The room where a majority of the images were captured (GR)

[1] https://www.sony.co.uk/electronics/support/webbie-hd-bloggie-cameras-mhspm-series/mhs-pm5 Last accessed: 1/05/2018

conveniently had a floor of tiled equilateral triangles as shown in Figures 2.3a & 2.3b. This room was chosen for this reason to ensure that distances and angles between images during tests were constant. Initially, to check if there was any inherent time dependant error due to the hardware, software, or the environment, multiple images were taken from the same pose (Figure 2.3b centre point at 0 degrees). Images taken from this one pose were used to compute the minimum difference in the distribution that could be attributed to temporal error. The positions that images were taken from are also described in Figure 2.3b, which shows the 7 positions and the various angles at which the images were taken for a total of 24 images for the GR. The centre point has smaller angle intervals as this is used primarily to test for changes due to orientation at the same position. Whereas points 1-6 are used to test for changes due to position, and changes due to different positions and orientations.

(a)



(b)

Figure 2.3: (a) Shows the GR and its convenient tiled flooring. (b) Shows a partial layout of the first room images were taken with all the positions and orientations marked. Length of triangle sides are 40cm.

For future comparison, another set of images was taken from a different

room (TR), see Figure 2.4 using the same angle increments as the centre point in the GR. The angular increments were kept constant with the use of a regular hexagon template. This is also another point of comparison for how changes in orientation alter the edge gradient distribution.

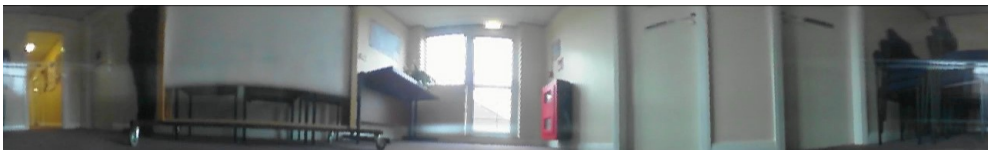


Figure 2.4: This is an image of the TR from which images were used to distinguish the viability of this method as one that would be suitable to distinguish between different locations.

2.1.1 Image Processing

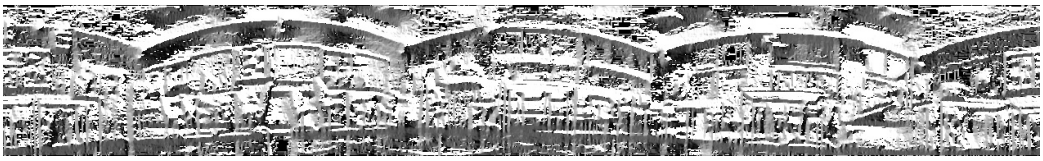
To process the images the OpenCV [2] library was used in conjunction with the Sony Playmemory Home software. Playmemory Home was used to unwrap the raw images into a panorama like image. OpenCV was then used to extract the gradient information from the image. To do this, Sobel operators [48] in the x and y direction were used to get two gradient images. These images were then input for a method that combined these images to produce two new images. One of the resulting images being the image where each pixel value represents an angular value between 0 and 360 which is also

[2]https://opencv.org/ Last accessed: 1/05/2018

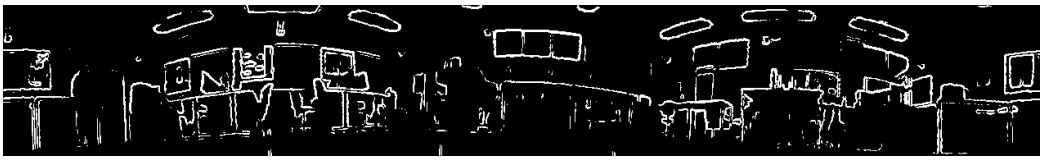
the direction of the edge as shown in Figure 2.5c. The other image being a magnitude image where each pixel value is the magnitude of the gradient as shown in Figure 2.5b. The magnitude image was altered via a binary threshold where the cut off point was one fifth the maximum magnitude. This threshold was empirically chosen. The chosen threshold appeared to be a good value where prominent and mostly continuous edges persist whilst artifacts due to low magnitude edges such as textures were removed. The resultant binary image was used as a mask to remove noisy gradients with low magnitudes that do not provide any useful information in this context. Once the gradient image had been masked the value of every non zero pixel was used to produce histograms for comparison against other image histograms.

(a)

(b)



(c)



(d)

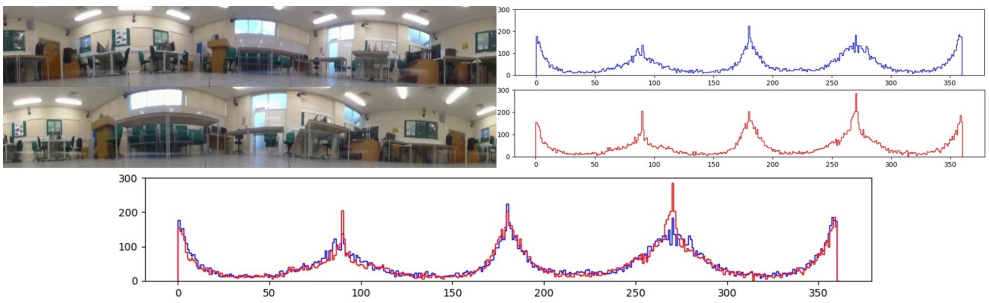
(e)

Figure 2.5: (a) Unwrapped image before processing. (b) Full magnitude image of the initial unwrapped image. This shows the magnitude in the change of intensities between pixels. (c) Full angle image of the initial unwrapped image. In this image each pixel represents an angle that corresponds to the gradient of the change of pixel intensity, no filter has been applied to remove noise information. (d) Binary threshold of magnitude image. Here is mask produced by only allowing edges with a sufficiently high difference in pixel intensity to be available. (e) Angle image after being masked with the binary threshold. This shows the location and direction of edges in degrees where each pixel value maps to a value between 0 and 360 degrees. This figure shows intermediate image processing steps. Images (b) and (c) are obtained from (a). Image (d) is obtained via a binary threshold applied to image (b). Image (d) is used to select important information from image (c). Image (e) is the final product used to produce the gradient distribution

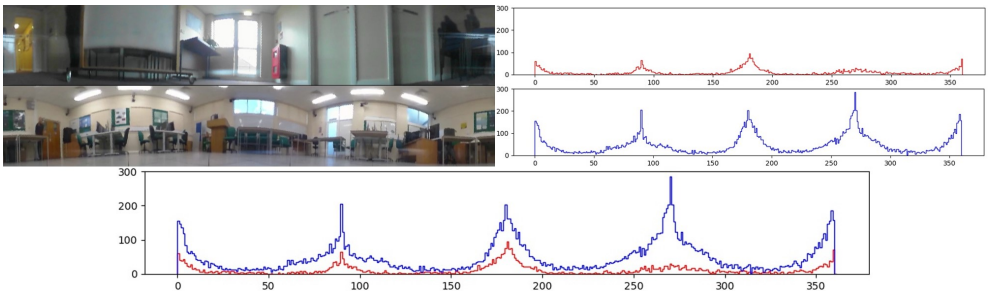
2.1.2 Data Representation

The edge information from the collected images was used to produce histograms that describe the gradient distribution at each of the sampled positions in the two locations used. These histograms were then compared using openCV's *compHist* method. The *compHist* method has four different tests it can use to compare the similarity of the histograms. The one that was chosen to compare the histograms was the chi-squared test [24] where

big values of the test statistic (chi-square) mean big differences (in terms of edge distribution) in any two images being compared, This was the prefered choice as it is more analogous to a distance. Figures. 2.6a and 2.6b illustrate how the chi-squared value represents a distance between the histograms of seemingly similar and different places.



(a)



(b)

Figure 2.6: (a) This Figure shows how two seemingly similar places look when their histograms are overlaid. This would give a very small chi-squared metric as there is a small difference between the lines at any point. (b) This Figure shows how two seemingly different places look when their histograms are overlaid. This would give a large chi-squared metric as there is a large difference between the lines at many points. The x axis of the histograms are values that gradients can take in the image. The y axis is how many times a given gradient occurs.

2.2 Results

The check for temporal errors resulted in some low level differences across different frames taken from the same pose. The mean chi-squared distance from images at the same position and orientation was 860.8 ± 98.8. This information gives an estimate of how much error there may be within any other results due purely to external factors, e.g: camera auto exposure calibration (no camera option to disable), light changes due to flickering lights, and sensor noise on the camera.

The main aim of this work was to check whether changes in pose are more dependant on position or orientation, therefore a comparison of the image data where only the position and only the orientation was changed was performed. To do this, for every image recorded the histogram generated from it was compared to that of the other images, resulting in a table of chi-squared results that could be used to easily visualise any relationships between the different poses from which the images were captured. The table in Figure 2.7 is illustrated visually (using grey-scale to represent the chi-square value) . A t-test was performed to check whether there was a significant difference between images where only position was changed and images where only orientation was changed, this was done using images from both the GR and the TR.

To do this all the data points that are the result of positional changes only were averaged and data points that are the result of changes in orientation only were also averaged. These averages were used to perform the T-test. The average chi-squared distance due to changes in position was 1947 ± 1448 whereas the average chi-squared distance due to changes in orientation only was 1851 ± 1005. A rejection of the leading hypothesis is attained showing that there is no significant difference between the means ($p = 0.393 > 0.05$).

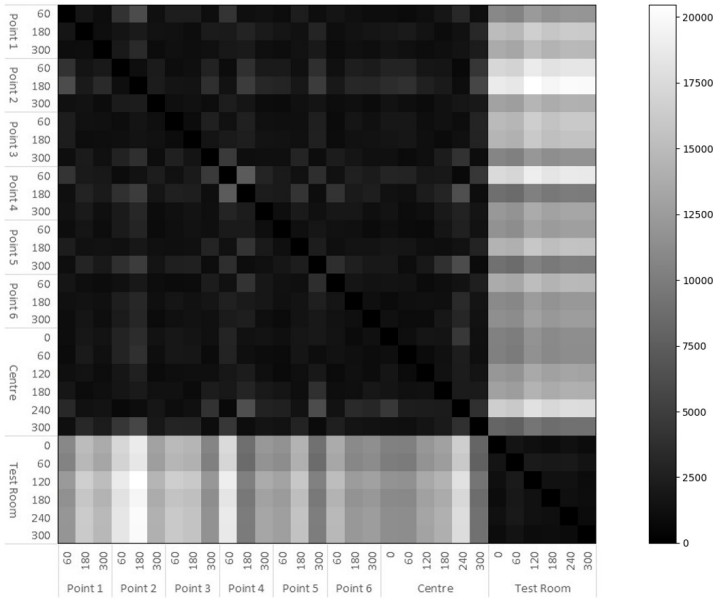


Figure 2.7: This image shows how every image was compared to all other image. Each pixel represents a chi-squared value, resulting from a comparison of two rooms which can be found on the axis. It highlights the distinction between similar and dissimilar rooms where low values-dark correspond to comparisons of images captured at different poses in the same room (GR) and high values-light correspond to comparisons of images taken in different rooms (GR&TR). The axes of this illustration refer to the position and orientation of the corresponding image; e.g Point 1 60 is the image taken from Point 1 at 60 degrees as shown in Figure 2.3b.

Again the t-test was used to compare all the data points from the Figure 2.7 that corresponded to comparisons of between different rooms (i.e the light regions). These values were averaged and compared to the average value from same room comparisons (i.e the dark regions). The mean difference between comparisons from the same room was 1917.8 ± 1062 and the mean

from comparisons of different rooms was 13574.1 ± 2920. Due to this data, the alternate hypothesis, that there will be a significant difference between the histograms of differing rooms, fails to be rejected as there is a significant difference between the means of the two sets of data ($p = 2.6 \times 10^{-95} < 0.05$).

Chapter 3

Accuracy decay Experiment

The above experiment supports the usage of the edge gradient distribution as feature that can be used to differentiate between two locations. This then raises the question of how one might expand on this and differentiate between more than two locations. How might the differentiating ability change as the number of locations stored increases. First of all there must be some system that can utilise and expand upon the work mentioned so far. As it was shown previously, comparisons from images of like locations have a lower chi squared distance than comparisons from images of unlike locations. This leads to the idea that positions from the same location would cluster together. The histogram data obtained from the images could be used as a high dimensional

data point. These data points could then be clustered using an algorithm such as K-means [25]. The reason for clustering the data points is so that images of like locations can be used to describe the actual location for future classification. This is a form of unsupervised learning, the system 'teaches' itself given some set of data which sets of images belong to the same location. This does however require that the value of K is hard coded for more simple implementations. This becomes an issue when a system is designed to be autonomous and 'make its own mind up' about what constitutes a location. What if the amount of locations presented to the system is more or less than the value of k. One slight alteration to the openCV K means implementation was made. The k-means algorithm assigns clusters a centroid value which does not have any corresponding images to act as a prototype. To allow for a real representation of the centre, the setting of the centre point of the cluster was set nearest neighbour of the centroid, this modification is often refereed to as k median.

For this experiment, it was hypothesised that as the amount of locations presented increases, the reliability of the system to differentiate between them will decrease. This hypothesis is based on the idea that as you increase the

amount of locations being actively[1] compared against, there is an increased chance that miss-classifications due to perceptual aliasing (different locations that look the same to one method of perception) will become more frequent.

3.1 Data Collection

To test how the accuracy may decay, a lager set of data was created for training and testing purposes. This data set was comprised of images from 14 locations (rooms) within the computer science department at Keele University. For each of these locations the camera setup mentioned in Figure 2.2 was placed in 10 separate positions where images were taken, with the exception of the much lager computer lab which had 25 images taken. This data set was then split so that 70 percent of the images were used to train the system and 30 percent of the images were used for testing the system.

[1]When mentioning actively comparing locations, this is in reference to the idea that in a system like this you would not necessarily have to look at every location you have ever been to, you may only need look at the N nearest neighbours.

3.2 Preparing Data

The Images taken from the 14 rooms were then duplicated in such a way as to have four sets of images that contained the contents of either 4, 7, 10 and 14 rooms. This was done so that there was a linear increase in the amount of rooms for each test. Next, the ideal number for K needed to be determined for each set. To do this, it must consider why we might not use the number of rooms as the value of K. Due to the different perspective that the robot may take in this scenario, certain items appear larger whilst other objects surfaces and other such salient features may be occluded. This can be seen in Figure 3.1. As an analogy, if you imagine what a human may consider as a room, like a small office. If you were to place a rodent on the table, the rodent may not perceive this office as a single location but multiple locations: i.e on top of the desk, under the desk etc. Another interesting point is how humans assign more complicated semantics to a location via its contents or regular usage, we as humans may put a higher precedence on the semantic use of a location than the geometric properties and relations. To this end it is not assumed that the perspective given to this robot will result in the same location labels that a person might assign. However, as mentioned in assumption 3, we would assume that when presented with images from

this perspective that a human participant would be able to agree or disagree about whether the two pictures depict the same location.

(a) (b)

Figure 3.1: Both the images in (a) and (b) are taken whilst standing in the exact same position. The only difference here is the perspective due to a change in height. In (a), it would be reasonable to consider the underneath of the desk a separate traversable location to on top of the desk. However in (b) one could not reasonably consider the underneath of the desk is a separate traversable location without first shifting to a lower perspective. This could be explained by the difference in the scale of objects and salient features relative to each other and the way that they may occlude each other.

To find the optimal value for K for each number of rooms a method known as silhouette analysis is used[46]. Silhouette analysis is a method used to rank how well a data point fits in its assigned cluster. This is done by comparing the average distance of a datum to all other points within the assigned cluster to the average distance to all other data points within the nearest neighbouring cluster. This returns a value between -1 and 1, where

-1 means that the data point fits perfectly in the second nearest cluster that it was not assigned to and 1 means that the data point fits perfectly in its currently assigned cluster. The formulae used for the silhouette analysis can be seen here.

$$a(i) = \frac{\sum_{j \in C_{i, i \neq j}} d(i,j)}{|C_i| - 1}$$

$d(i,j)$ is the distance between a point i and a point j and where C_i the cluster that point belongs to.

$$b(i) = \min_{i \neq k} \frac{\sum_{j \in C_k} d(i,j)}{|C_k|}$$

C_k is another cluster that the dissimilarity is being compared to. The cluster that provides the smallest dissimilarity will be used for this value

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)} if |C_i| > 1$$

$s(i)$ is the silhouette value for point i.

By averaging the silhouette values across all the data points over differing values of K you can find which value of K gives the largest average silhouette value. A higher value of the average silhouette value means that there was a

better categorisation of the data and hence the locations. This works well so long as the range of K isn't too close to the number of data points in the set. This is because when K = the number of data points the silhouette value becomes 0 for each data point. Once the optimal value of K has been found for each set of rooms the images are assigned to clusters using the value of K determined by the silhouette values.

3.3 Confirmation

As K-means method is an unsupervised learning strategy it is difficult to test and compare performance in any way that is meaningful. It is difficult to test due to the fact that there is no exact ground truth to compare the outputs of this system too. Assumsion 3 was required to provide some proxy for a ground truth. Using this assumption, a test was performed to compare the results from the K-means system to a humans perception of location, this will provide a measure of accuracy. To perform this test a set of images was put together for the participant to look through. This set contained pairs of images that the K-means + edge gradient system believed belonged to the same location, this made up half the set. Secondly, pairs of images that the

system believed were from different locations, this made up the other half of the set. When the participant was presented with the set there was a random order as to whether the current pair on display was grouped as the same location to the K-means system or not from the same location. The participant was asked to simply provide a yes or no answer to whether the 2 images present on the screen belonged to the same location.

3.4 Results

First we look at the values of the average silhouette values for each set of rooms for each value of K. The maximum value of K used for each set is different due to the amount of data points for each set. It was decided that to avoid getting high values of the silhouette values (due to single data points making up a cluster) that the maximum value of K would be half the total amount of data points for each set. As can be seen in figure 3.2, there is a general tendency in the silhouette values to increase as K increases (which is expected) but there is a lot of variance between subsequent values. The rest of the graphs can be viewed in appendix A.1. The peak of the silhouette values are marked on graphs along with the value for the amount of rooms

as defined by the data set. There is a debate to be had about whether or not the peak values alone are a good indicator of how well the clusters are laid out but with the information acquired at this point in time this is a good initial measure of optimal number for K.

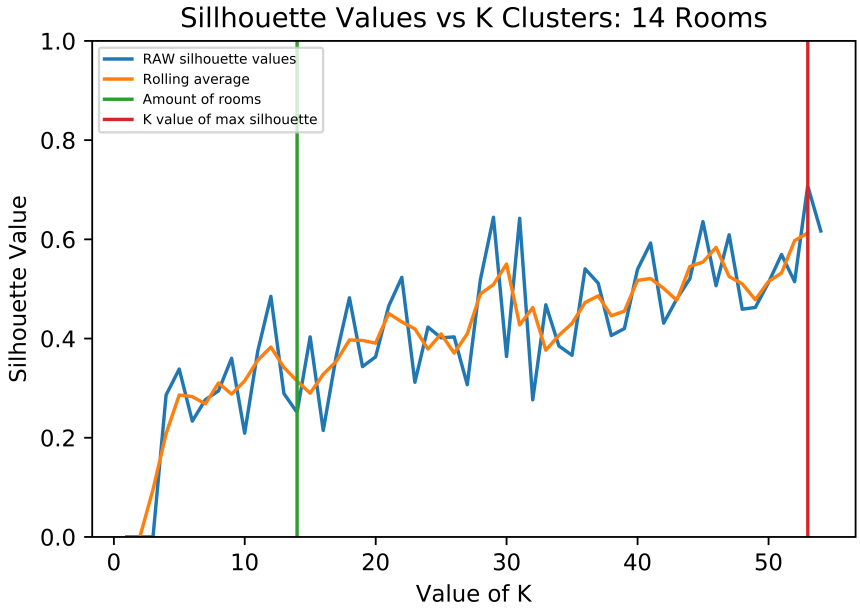


Figure 3.2: This figure shows how the silhouette value changes with respect to the values of K set for the K-means algorithm. The blue line is the raw data, the orange line is the rolling average and the green and red line represent the amount of rooms as a person may describe it and the amount of locations as the system describes it respectively.

The silhouette graphs were used to inform a decision of the optimal value of K. Once this had been decided then the clusters were formed using this K value. These clusters were then used by the system to provide a location

to the images from the testing set. The images from the test set which have now been assigned to a cluster (which can be thought of as a location), were provided to the human participant as mentioned in the method for this experiment. The results from this experiment were displayed in confusion matrices. These matrices can be found in Figure 3.3 and appendix A.2.

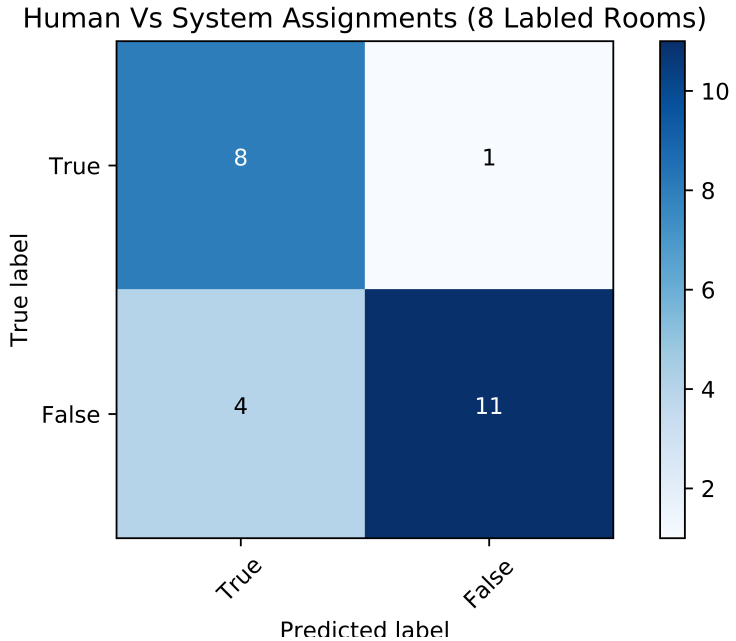


Figure 3.3: This Graph shows the comparison between the human participants answers and the computer systems answers to whether two images are from the same location. The participants answers are the true label and the computers answers are the predicted label.

The matrices were analysed using common statistics such as accuracy, recall, prevalence, and false positive rate. These statistics are important as

they give a base line for how to judge the success or failure of the system. Figure 3.4 is a graph that shows how accuracy, recall, false positive rate, and prevalence change with of K. It can be seen that the recall rate stays high throught the values of K, this may seem good alone but coupled with the rapidly decreasing prevalence rate means that there are increasingly fewer overall positives that could be agreed upon. The false positive rate sees a sharp increase with k showing that there is overall a disagreement between the human participant and the system as to whether any two images belong to the same location. These statistics show that there is overall an issue with the underlaying assumptions that the human could be used as a ground truth observer, that the system is performing poorly or that the data set provided was not large and diverse enough to provide any useful analysis. It is still however interesting that the accuracy stays as high as it does. The system and the human participant seem to regularly agree on the assignment of two images not belonging to the same location.

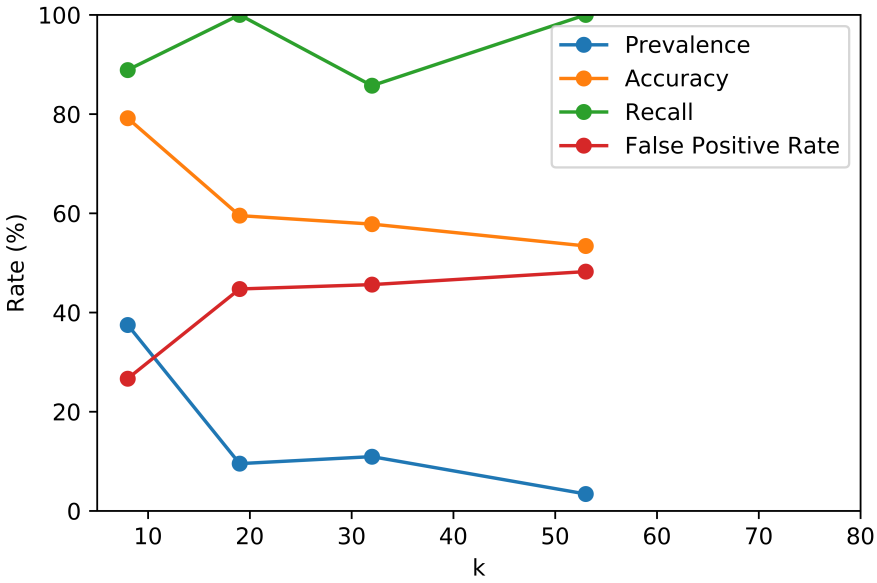


Figure 3.4: This graph depicts how the accuracy, recall, prevalence, and false positive value change with respect to K.

Chapter 4

Discussion & Conclusion

4.1 Discussion

The first experiment conducted shows that there is enough information within the edge gradient feature that was extracted from the images of rooms to feasibly differentiate between 2 locations. The issue with this experiment is that there was no follow up to make sure the results were repeatable in different pairs of locations. Although there is a statistical significance between the two chosen locations in the feature, this only shows that this is the case for only these two locations. A reasonable follow up would be to have repeated the experiment within many differing pairs of locations to confirm

whether this statistical significance is observable across many kinds of locations. This is not to say that this invalidates the results of experiment 1. The very low p value provided by the t test shows that there is definitely circumstances where this feature can discriminate well even if it is just between these two rooms.

The second experiment has a few short failings that should be looked at in a bit more detail. As well as completing the repeats of experiment 1 there should be some further though into whether or not an unsupervised method such as k-means is appropriate for this global feature. The feature itself is very high dimensional meaning that any distances calculated will appear very large. The large dimensionality paired with the small data set used for this experiment have meant that the data was likley overfitted to the corresponding clusters. For the prior reasons there could be no statistically analysis of experiment two due to the lack of data. However that isnt to say nothing was gained from this experiment. It has shown that the underlaying assumption about a human participant being able to identifying a location based on an image taken from a different perspective may not have been correct. It has also shown that there is an issue with trying to use the silhouette analysis to automatically determine the ideal value for the number

of clusters may not be appropriate either. The results may have been more stable if instead a supervised method was used to cluster the data. If the data clustering had been supervised then salient parts of the feature could be identified which could improve results.

The dimentionality of this feature could easily be reduced with two methods. One method would be to look at the symmetry of the angles themselves. As this system looks at the difference in brightness values to get the angle of an edge there could be two possible values for the edge. Lets take a flat surface and say that its value is 90 degrees when the light is coming from the top. If the light source were to be moved below, the recorded value would become 270 degrees. This is something that could pose a real issue of the system were to ever be used out doors over a long time as the sun moves from the east the west. To account for this, all values could be modulated by 180 degrees and have the remainder be the value that is used to form the feature, in a way negating the difference due to light directionality. The second method to reduce the dimensionality would be to bin the angles into larger regions of 5 degrees rather than in 1 degree bins.

Another interesting point that was brought up during this work was what actually defines the boundaries of a location. Can a location be defined purely

by its geometric properties or are they defined by the contextual information that is provided by the objects in the location and their functions? Do all locations have hard boundaries like we may expect from a building passing between doors. Can boundaries between locations in more open environments be more fuzzy or could a system still argue that they were in location one but give reference to their sub location within in that location ie "I'm in location A to the right of landmark 1". Work by Zeil, Jochen and Hofmann[49] look at the concept of catchment areas. These are areas where as you move towards or away from a reference point the difference between the reference point and the current point, decreases or increases respectively. These catchment areas could prove to be useful in the definition of the difference between adjacent or open locations. This could be used over time to make certain locations more probable of being entered and hence possibly further limiting a search area. In locations where there may be a more clearly defined change between locations, such as at a door, one might wonder whether it would alternate between 2 locations depending on which one is more prominent. This is something that must be tested in the future. However, it may be hypothesised that the doorway, which is a transition between two locations, may be seen as a separate location in and of it self. Treating typical boundaries in such a

way may allow smoother transitions between all locations similar to the idea of catchment areas.

A system that uses this feature could also benefit from the use of video footage that could provide multiple images per second, and hence a degree of temporal information, to compare to the reference images stored with the K-means. This would allow for a better confidence of the suggested location as there will be more data and information for comparisons. An error metric could be used to filter out any erroneous data that comes in. Allowing this system to access video footage and therefore a history of the places visited in the recent past would also provide some other benefits. These include being able to ignore any transient features such as a new object coming into and out of the location and being able to update the K-means clusters, the reference image and future beliefs of the system.

Neither experiment looks at the posibility of objects or structures in the location being moved or rotated. For instance, if you were in an office and the desk and chair had been moved to lay against another wall. This scenario should not pose a problem to this system as the edge gradient feature used does not rely on the topological features of the location that it is in. That is to say an angle on the left hand side of the location is indifferent to an

angle with the same value on the right hand side of the room. Moving things within a location would still be a significant to test. Translating an object within the location will likely yield insignificant changes in the edge gradient distribution, but rotations of objects in the room may yield a noticeable difference, especially on objects with rotational asymmetry, such as elongated structures. It could be hypothesised that rotational differences of the objects would be small unless a large proportion of the objects were elongated and rotated. Although this could also look sufficiently different for a human participant also. Occlusion of prominent structures could also be an issue with this feature but this would be much harder to test. This would require knowing ahead of time due to a lack of topological information

A system using the edge gradient from images could be shadows in a scene. Hard shadows may be picked up by this system as an edge. In well lit environments like indoors where there is ample lighting in most directions, any shadows that might be cast wont be dark enough to be picked up by this system. In outdoor conditions shadows will not only be picked up by the system due to only having one primary light source, but as the position of the sun changes, so does the direction, scale, and skew of the shadows change. There is a very specific kind of colour constancy that looks at the problem

of making images shadow invariant[15]. This could be used in an outdoor setting to aid in the recall of locations by removing the shadows from the images. Though this process does change the image into a grayscale image, this wouldn't be an issue for the edge detection algorithm.

A similar feature that looks at the gradient distributions in images exists, however it is a local/grid feature. This is known as the histogram of orientated gradients (HOG)[37, 16]. This feature uses small chunks of the image and produces histograms of the gradients within each chunk. These histograms all together form the feature as a set of vectors that describes the image. This method is good at finding local features such as objects and classifying them but would not suit an entire panoramic view of a room. Rotations of the view, which causes the rolling of the image, will cause the HOG feature set to be different in such a way as to cause miss-classification; although this is something that should be tested also. In a preliminary test of speed, the method used through out this research (i.e the edge gradient distribution) was briefly compared to the HOG feature in the speed of creating the feature. The edge gradient performed faster in this small test but this would require more robust and thorough testing. If this result were to be reproducible, then it could be explained by the overhead required for HOG

to divide the image up into a regular grid and calculate the histogram of gradients for each cell.

Another avenue that could be explored within this coarse localisation is the use of contextual information about the location via object detection. An interesting paper by Betancourt etal[8] uses egocentric video footage and an unsupervised learning method to identify types of location. Their research is designed to give contextual actions to other systems but if the type of location can be discerned from the video footage it could prove to be a useful way to potentially differentiate between locations that may have otherwise been classified as the same location using the edge gradient feature. This is also an example where other individual systems could be used in tandem with the edge gradient feature where each system can function individually but pool their results for greater accuracy. This would require a large amount of testing to discover what combinations of systems would be both quick to function using existing data but also provide useful outputs.

4.2 Conclusion

In conclusion this study has shown that the edge gradient histogram can be used to differentiate between two separate locations whilst being robust to small changes of pose within said locations. The hypothesis that there will be no significant difference between the Chi-Squared distances between images in the same location was confirmed. However the hypothesis that there would be a significant difference between Chi-Squared distances due to position and Chi-Squared distances due to orientation was rejected. This was a favourable result. The hypothesis that the differentiating ability of the system will decay as the amount of locations increases was neither confirmed or rejected because there is not enough data to suggest whether this is definitely the case. However, the initial study seems to imply that this is the case. More tests with more locations and images are required to show whether there is any statistical significance to the drop in accuracy. This research has also brought to light how one might use such a feature to differentiate between more than two locations. More work could be done to look at how this method is performed such as looking at how the value of K is set. This system could also benefit from exploring other clustering techniques and perhaps dimensionality reductions techniques to reduce the size of the features. More

work needs to be done to look at the useful applications of this edge gradient feature as well as looking at what conditions it works better in.

Own Publications

Jarvis, D. and Kyriacou, T. [2018], The effect of pose on the distribution of edge gradients in omnidirectional images, in 'Annual Conference Towards Autonomous Robotic Systems', Springer, pp. 234-244.

Bibliography

[1] Afifi, M. [2018], 'Semantic white balance: Semantic color constancy using convolutional neural network', *arXiv preprint arXiv:1802.00153* .

[2] Andreasson, H., Treptow, A. and Duckett, T. [2007], 'Self-localization in non-stationary environments using omni-directional vision', *Robotics and Autonomous Systems* **55**(7), 541–551.

[3] Anzai, A., Peng, X. and Van Essen, D. C. [2007], 'Neurons in monkey visual area v2 encode combinations of orientations', *Nature neuroscience* **10**(10), 1313.

[4] Araújo, P., Miranda, R., Carmo, D., Alves, R. and Oliveira, L. [2017], 'Air-sslam: A visual stereo indoor slam for aerial quadrotors', *IEEE Geoscience and Remote Sensing Letters* **14**(9), 1643–1647.

[5] Ashokaraj, I., Tsourdos, A., Silson, P. and White, B. [2004], Sensor

based robot localisation and navigation: Using interval analysis and extended kalman filter, *in* 'Control Conference, 2004. 5th Asian', Vol. 2, IEEE, pp. 1086–1093.

[6] Bailey, T. and Durrant-Whyte, H. [2006], 'Simultaneous localization and mapping (slam): Part ii', *IEEE Robotics & Automation Magazine* **13**(3), 108–117.

[7] Barnard, K., Cardei, V. and Funt, B. [2002], 'A comparison of computational color constancy algorithms. i: Methodology and experiments with synthesized data', *IEEE transactions on Image Processing* **11**(9), 972–984.

[8] Betancourt, A., Díaz-Rodríguez, N., Barakova, E., Marcenaro, L., Rauterberg, M. and Regazzoni, C. [2017], 'Unsupervised understanding of location and illumination changes in egocentric videos', *Pervasive and Mobile Computing* **40**, 414–429.

[9] Bianco, S., Cusano, C. and Schettini, R. [2017], 'Single and multiple illuminant estimation using convolutional neural networks', *IEEE Transactions on Image Processing* **26**(9), 4347–4362.

[10] Bonabeau, E., Dessalles, J.-L. and Grumbach, A. [1995], 'Characterizing

emergent phenomena (1): A critical review', *Revue internationale de systémique* **9**(3), 327–346.

[11] Buchsbaum, G. [1980], 'A spatial processor model for object colour perception', *Journal of the Franklin institute* **310**(1), 1–26.

[12] Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I. and Leonard, J. J. [2016], 'Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age', *IEEE Transactions on Robotics* **32**(6), 1309–1332.

[13] Choras, R. S. [2007], 'Image feature extraction techniques and their applications for cbir and biometrics systems', *International journal of biology and biomedical engineering* **1**(1), 6–16.

[14] Chow, T. W. and Rahman, M. [2007], 'A new image classification technique using tree-structured regional features', *Neurocomputing* **70**(4-6), 1040–1050.

[15] Corke, P., Paul, R., Churchill, W. and Newman, P. [2013], Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation, *in* 'Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on', IEEE, pp. 2085–2092.

[16] Dalal, N. and Triggs, B. [2005], Histograms of oriented gradients for human detection, *in* 'Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on', Vol. 1, IEEE, pp. 886–893.

[17] Deak, G., Curran, K. and Condell, J. [2012], 'A survey of active and passive indoor localisation systems', *Computer Communications* **35**(16), 1939–1954.

[18] Dissanayake, G., Huang, S., Wang, Z. and Ranasinghe, R. [2011], A review of recent developments in simultaneous localization and mapping, *in* '2011 6th International Conference on Industrial and Information Systems', IEEE, pp. 477–482.

[19] Dorigo, M. and Blum, C. [2005], 'Ant colony optimization theory: A survey', *Theoretical computer science* **344**(2-3), 243–278.

[20] Durrant-Whyte, H. and Bailey, T. [2006], 'Simultaneous localization and mapping: part i', *IEEE robotics & automation magazine* **13**(2), 99–110.

[21] Dyson, U. [n.d.], 'See the new dyson 360 eye robot vacuum cleaner in action# dysonrobot'.

[22] Fujii, K. [n.d.], 'Extended kalman filter'.

[23] Gijsenij, A., Gevers, T. and Van De Weijer, J. [2011], 'Computational color constancy: Survey and experiments', *IEEE Transactions on Image Processing* **20**(9), 2475–2489.

[24] Greenwood, P. E. and Nikulin, M. S. [1996], *A guide to chi-squared testing*, Vol. 280, John Wiley & Sons.

[25] Hartigan, J. A. [1975], 'Clustering algorithms'.

[26] Heinly, J., Dunn, E. and Frahm, J.-M. [2012], Comparative evaluation of binary features, *in* 'Computer Vision–ECCV 2012', Springer, pp. 759–773.

[27] Hidalgo, F. and Bräunl, T. [2015], Review of underwater slam techniques, *in* '2015 6th International Conference on Automation, Robotics and Applications (ICARA)', IEEE, pp. 306–311.

[28] Hu, R. and Collomosse, J. [2013], 'A performance evaluation of gradient field hog descriptor for sketch based image retrieval', *Computer Vision and Image Understanding* **117**(7), 790–806.

[29] Kalman, R. E. and Bucy, R. S. [1961], 'New results in linear filtering and prediction theory', *Journal of basic engineering* **83**(1), 95–108.

[30] Kosecka, J., Zhou, L., Barber, P. and Duric, Z. [2003], Qualitative image based localization in indoors environments, *in* 'Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on', Vol. 2, IEEE, pp. II–II.

[31] Kröse, B. J., Booij, O., Zivkovic, Z. et al. [2007], A geometrically constrained image similarity measure for visual mapping, localization and navigation., *in* 'EMCR'.

[32] Kumar, M., Husian, M., Upreti, N. and Gupta, D. [2010], 'Genetic algorithm: Review and application', *International Journal of Information Technology and Knowledge Management* **2**(2), 451–454.

[33] Kyriacou, T. [2011], An implementation of a biologically inspired model of head direction cells on a robot, *in* 'Conference Towards Autonomous Robotic Systems', Springer, pp. 66–77.

[34] Lisin, D. A., Mattar, M. A., Blaschko, M. B., Learned-Miller, E. G. and Benfield, M. C. [2005], Combining local and global image features for

object class recognition, *in* 'Computer vision and pattern recognition-workshops, 2005. CVPR workshops. IEEE Computer society conference on', IEEE, pp. 47–47.

[35] Litman, T. [2017], *Autonomous vehicle implementation predictions*, Victoria Transport Policy Institute Victoria, Canada.

[36] Liu, A., Lin, W. and Narwaria, M. [2012], 'Image quality assessment based on gradient similarity', *IEEE Transactions on Image Processing* **21**(4), 1500–1512.

[37] McConnell, R. K. [1986], 'Method of and apparatus for pattern recognition'. US Patent 4,567,610.

[38] Milford, M. J., Wyeth, G. F. and Prasser, D. [2004], Ratslam: a hippocampal model for simultaneous localization and mapping, *in* 'IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004', Vol. 1, IEEE, pp. 403–408.

[39] Milford, M. and Wyeth, G. [2003], Hippocampal models for simultaneous localisation and mapping on an autonomous robot, *in* 'Proceedings of the Australasian Conference on Robotics and Automation, 2003', Australian Robotics and Automation Association Inc.

[40] Modsching, M., Kramer, R. and ten Hagen, K. [2006], Field trial on gps accuracy in a medium size city: The influence of built-up, *in* '3rd workshop on positioning, navigation and communication', Vol. 2006, pp. 209–218.

[41] Muro, C., Escobedo, R., Spector, L. and Coppinger, R. [2011], 'Wolf-pack (canis lupus) hunting strategies emerge from simple rules in computational simulations', *Behavioural processes* **88**(3), 192–197.

[42] Pass, G. and Zabih, R. [1996], Histogram refinement for content-based image retrieval, *in* 'Proceedings Third IEEE Workshop on Applications of Computer Vision. WACV'96', IEEE, pp. 96–102.

[43] ping Tian, D. et al. [2013], 'A review on image feature extraction and representation techniques', *International Journal of Multimedia and Ubiquitous Engineering* **8**(4), 385–396.

[44] Pirahansiah, F., Abdullah, S. N. H. S. and Sahran, S. [2013], 'Simultaneous localization and mapping trends and humanoid robot linkages', *Asia-Pacific Journal of Information Technology and Multimedia* **2**(2).

[45] Qureshi, A. H., Nakamura, Y., Yoshikawa, Y. and Ishiguro, H. [2016], Robot gains social intelligence through multimodal deep reinforcement

learning, *in* 'Humanoid Robots (Humanoids), 2016 IEEE-RAS 16th International Conference on', IEEE, pp. 745–751.

[46] Rousseeuw, P. J. [1987], 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

[47] Shi, W., Loy, C. C. and Tang, X. [2016], Deep specialized network for illuminant estimation, *in* 'European Conference on Computer Vision', Springer, pp. 371–387.

[48] Sobel, I. [1990], 'An isotropic 3× 3 image gradient operator', *Machine vision for three-dimensional scenes* pp. 376–379.

[49] Zeil, J., Hofmann, M. I. and Chahl, J. S. [2003], 'Catchment areas of panoramic snapshots in outdoor scenes', *JOSA A* **20**(3), 450–469.

[50] Zou, J., Li, W., Chen, C. and Du, Q. [2016], 'Scene classification using local and global features with collaborative representation fusion', *Information Sciences* **348**, 209–226.

Appendix A

A.1 Silhouette vs K Graphs

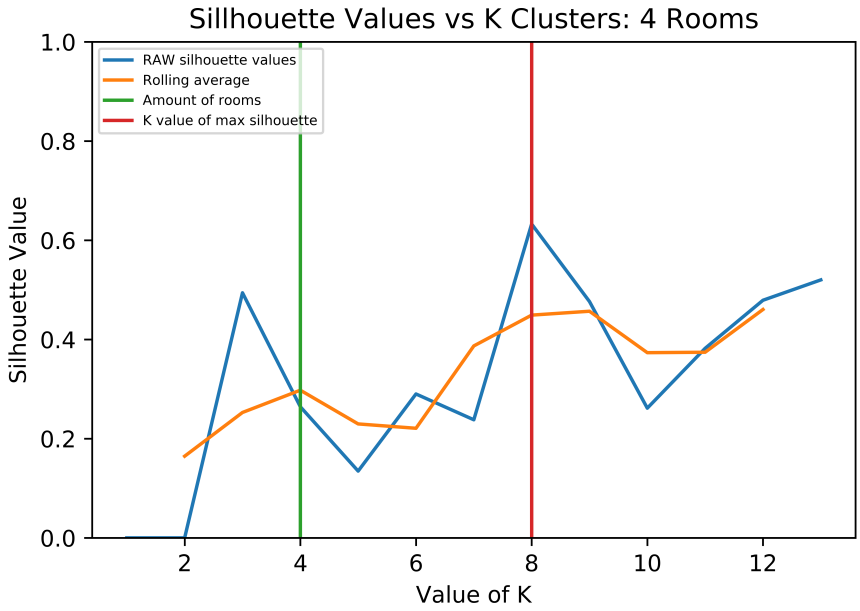


Figure A.1

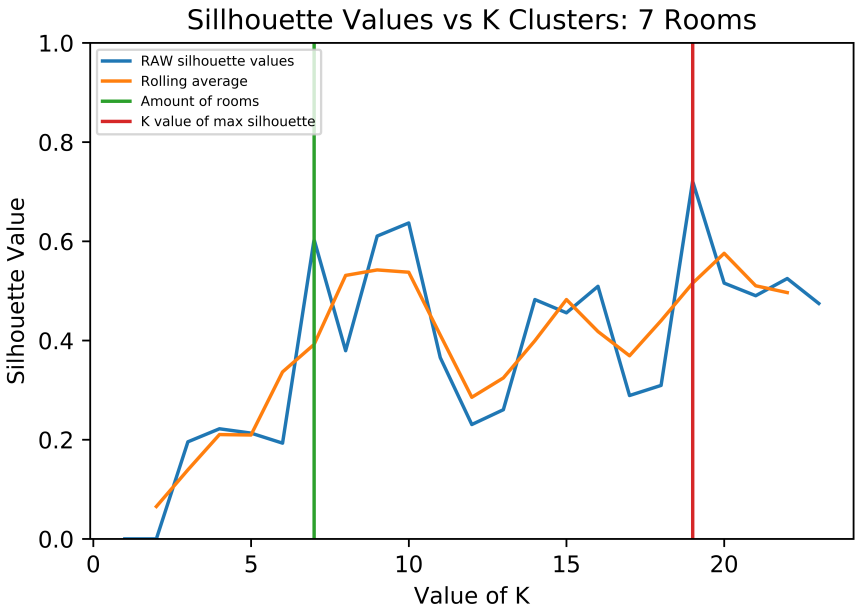

Sillhouette Values vs K Clusters: 7 Rooms
RAW silhouette values
Rolling average
Amount of rooms
K value of max silhouette
Silhouette Value
Value of K
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20

Figure A.2

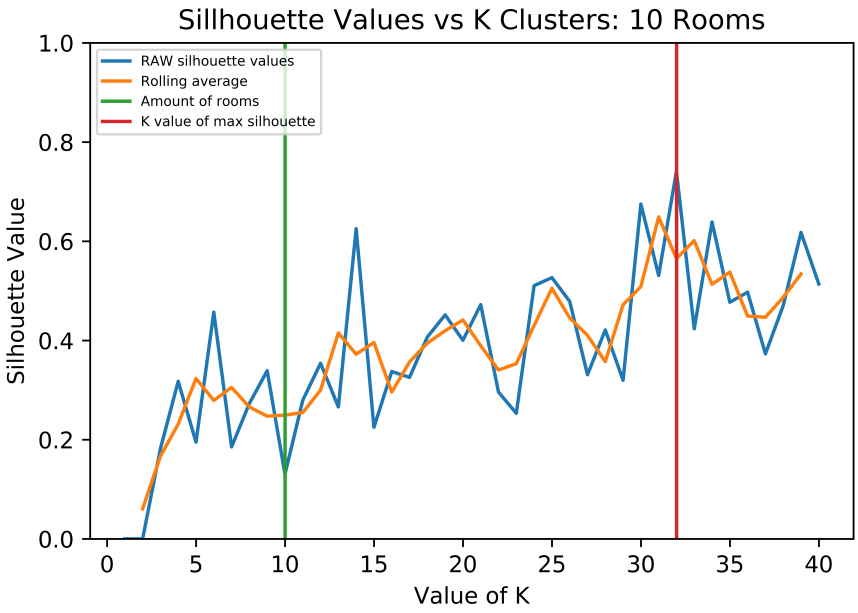

Sillhouette Values vs K Clusters: 10 Rooms
RAW silhouette values
Rolling average
Amount of rooms
K value of max silhouette
Silhouette Value
Value of K
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
35
40

Figure A.3

A.2 Confusion Matrices

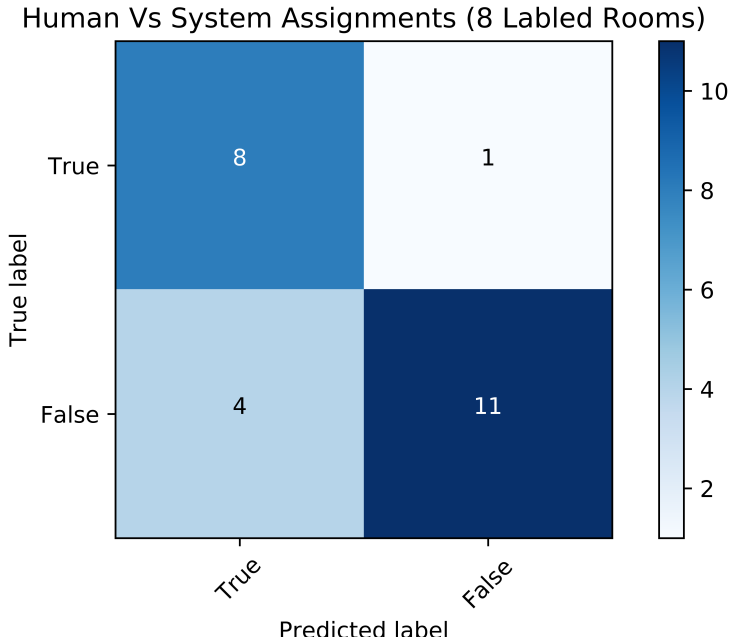


Figure A.4

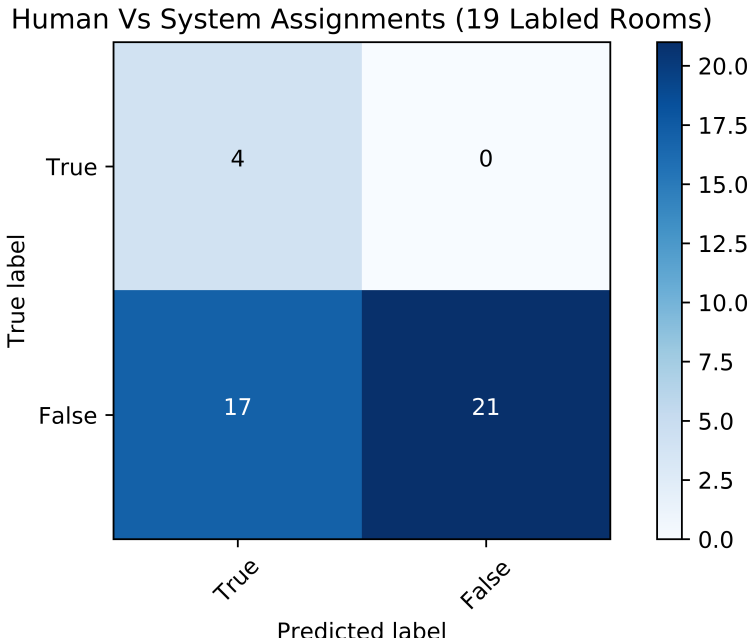

Human Vs System Assignments (19 Labled Rooms)
True
False
True label
4
0
17
21
True
False
Predicted label
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

Figure A.5

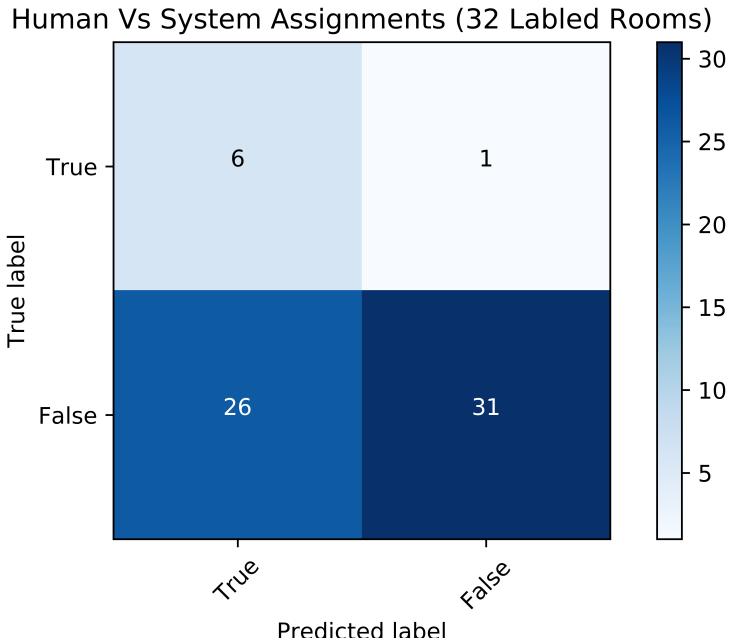

Human Vs System Assignments (32 Labled Rooms)
True
False
True label
6
1
26
31
True
False
Predicted label
30
25
20
15
10
5

Figure A.6