JOURNAL OF SPATIAL INFORMATION SCIENCE
Number 2 (2011), pp. 59–83
doi:10.5311/JOSIS.2011.2.6

RESEARCH ARTICLE

A combined GIS and stereo vision approach to identify building pixels in images and determine appropriate color terms

Philip James Bartie[1], Femke Reitsma[1], and Steven Mills[2]

[1]Department of Geography, University of Canterbury, Christchurch, New Zealand
[2]Areograph Ltd (NZ), 90 Crawford St, Dunedin, New Zealand

Received: December 30, 2010; returned: February 26, 2011; revised: April 14, 2011; accepted: April 28, 2011.

Abstract: Color information is a useful attribute to include in a building's description to assist the listener in identifying the intended target. Often this information is only available as image data, and not readily accessible for use in constructing referring expressions for verbal communication. The method presented uses a GIS building polygon layer in conjunction with street-level captured imagery to provide a method to automatically filter foreground objects and select pixels which correspond to building façades. These selected pixels are then used to define the most appropriate color term for the building, and corresponding fuzzy color term histogram. The technique uses a single camera capturing images at a high frame rate, with the baseline distance between frames calculated from a GPS speed log. The expected distance from the camera to the building is measured from the polygon layer and refined from the calculated depth map, after which building pixels are selected. In addition significant foreground planar surfaces between the known road edge and building façade are identified as possible boundary walls and hedges. The output is a dataset of the most appropriate color terms for both the building and boundary walls. Initial trials demonstrate the usefulness of the technique in automatically capturing color terms for buildings in urban regions.

Keywords: GIS, computer vision, stereo depth mapping, color terms, referring expressions, building façade, structure from motion, wayfinding instructions, color entropy

1 Introduction

When talking about a place, people like to include descriptive words to conjure up a pictorial representation in the listener's imagination. Such feature descriptions are also often included in wayfinding instructions, such as the details of a building façade material or color, as in "We're the red brick house with the long white fence." Capturing this level of detail has received a lot of attention in recent years, including initiatives such as Google's Street View [56] and Microsoft's Street Slide [31]. However texture information is presented as imagery and not transferred into values suitable for use in cartographic symbology, or for inclusion in wayfinding feature descriptions. This paper presents a method whereby depth mapping is used to automatically filter foreground objects from images, allowing the automatic extraction of color terms to describe buildings, so that a database of building colors may be created.

Object descriptions are known in natural language research as "referring expressions" and are used, for example, to draw someone's attention to a particular building in a cityscape [14]. They include visual clues which the speaker considers to be useful aids for the listener to determine which item in view is the intended target. The most useful terms are those which the listener can identify quickly, and limit the number of candidates rapidly without leading to any confusion. In some ways the process of creating a referring expression is similar to determining landmark saliency, which focuses on ways to measure the prominence or distinctiveness of a building according to a number of factors, including visual, semantic, or structural attraction [52]. There are two main methods for extracting landmark candidates, by assigning a saliency score based on various attributes. Elias [16] uses characteristics such as the building area, number of corners, density of buildings in the district, orientation to north and so on. An alternative definition for saliency measurement was proposed by Raubal and Winter [45], later updated with Nothegger [41], scores buildings according to Sorrows and Hirtle's [52] visual, semantic, and structural characteristics. The visual factors include façade area, shape, color, and visibility, translating well to the egocentric projective view experienced by street observers. These visual variables closely reflect Bertin's [8] set of seven visual variables (position, orientation, size, color, value, texture, and form) which should be considered when displaying graphical information. While traditional GIS datasets store position, orientation, and planimetric size of buildings, they fail to show information relating to building height, color, or texture. A challenge exists therefore in how this information may be sourced and made accessible in a format suitable for use in constructing referring expressions. While LiDAR now offers a viable solution for capturing building height and form [43, 48], color and texture details are either unavailable or stored in an inaccessible form, such as street level images.

Oblique aerial imagery could provide a source for the missing color information, offering more detail on the sides of buildings than can be obtained from traditional overhead aerial imagery. However while the textures can be directly mapped on to the surfaces of building models [22, 35], foreground objects such as cars and trees are indistinguishable and are incorrectly included in the building façades. New techniques are being developed which attempt to remove the unwanted foreground elements [21], or fill in the background using images from alternative angles [60]. A sensor fusion approach has also been successful where laser ranging equipment is used in conjunction with cameras to collect textures [15, 44, 57]. While these techniques look to offer the solution to produce clean façade textures in the near future, the timescale is unknown and current coverage is sparse.

In the meantime the research presented here offers a way to capture building color details from street level using low-tech equipment available to most communities, automatically excluding foreground objects allowing the remaining building pixels to be classified with an appropriate color term. The challenges are how to automatically identify which pixels in an image correspond to a building, classify the colors in those pixels using the most appropriate color term, and then associate those values with the correct building polygon on a map.

There are a range of applications which would benefit from having access to building color details, including the ability to generate more natural wayfinding instructions [41,45], which would be especially beneficial for children [25] and people with learning difficulties [42]. The ability to create more descriptive navigational instructions that would be indistinguishable from those generated by a human has been referred to as the "spatial Turing test" [58]. In addition color can be used in forming referring expressions [14]. These expressions are particularly useful for describing objects in "vista space" [39]—that is the region currently visible to an observer. Being able to verbalize color information will be an important component in the future of speech interfaces whereby a user operating both hands-free and eyes-free may request information on a building in view, selected by its description [37]. A location based service (LBS), such as a virtual city guide application [4], may direct the user's attention to a specific building using a narrative which singles it out in the current view, simulating a natural language description. Emergency services would also benefit from access to a building color database when attempting to locate people based on a description of their surroundings [34].

The paper is arranged as follows: Section 2 discusses two appropriate computer vision techniques which can be used to calculate pixel depth values, enabling the selection of pixels in the image at distances which correspond to that of the designated target building as explained in Section 3. These pixel values may be translated into relevant color terms using a fuzzy set approach as discussed in Section 4. A trial of the proposed method is demonstrated in Section 5.

2 Stereo depth mapping for façade color retrieval

There are a number of methods which may be used to recover depth information from images, including the *structure from motion* approach [30, 54], and the *stereo vision* method [9, 28, 36, 50].

The requirement here is to determine the distance from the camera to the real-world object represented in each image pixel, such that when combined with a GIS building layer those pixels corresponding to a designated target at a known distance may be retrieved, while foreground objects are excluded. Once the depths per pixel have been recovered then those within the expected range are selected, as outlined in Figure 1.

2.1 Structure from motion approach

Through combining many views of the same real world object from different distances and angles, it is possible to reconstruct the object's structure. This is achieved through a processing pipeline which begins with matching pixels in at least three images. Pixel matching is an automated procedure, whereby points with high contrast gradients are found, such

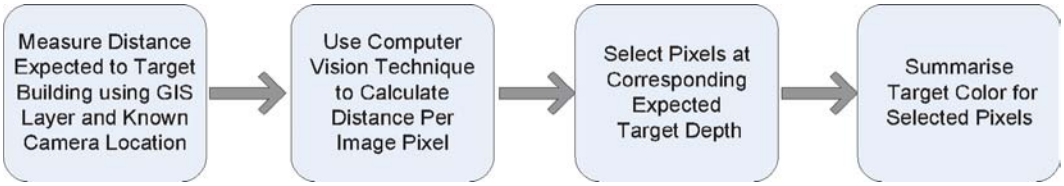

Figure 1: Overview of the process to select target pixels from captured image

as SURF-based features [5]. These features are tracked between images, known as correspondences, and the three-dimensional structure of the real world object is recovered along with camera pose estimates.

The following images (Figure 2) were rendered from a model built using this structure from motion approach, whereby 26 images collected while walking in an arc around a property were processed using PhotoScan software [3].

There are a number of considerations when using this technique in urban regions. Firstly a large number of images are required from a variety of viewing angles; a linear set of images from a single drive-by will not be sufficient to produce a detailed model. In addition to produce good results the source images need to be captured at high spatial density, meaning a new image is required each meter or so. Furthermore a robust solution requires that features need to be visible in at least three images captured from different locations, which can be rather difficult to achieve in confined streets, through foreground vegetation, or when there are many moving pedestrians and cars. Additionally processing times are fairly high, as feature matching, bundle adjustment [55], and geometry reconstruction are computationally intensive tasks. Finally low texture regions lack features for the matching process, resulting in poor depth estimates.

There have been a number of attempts to overcome these shortcomings, such as sourcing images from popular image sharing websites such as FlickR to build community volunteered virtual models [51], and improvements in depth reconstruction by imposing model restrictions such as enforcing planar surfaces [38]. However currently this approach is best suited to capturing information in the more open urban spaces where a large number of images may be captured from a multitude of angles. A good example of this exists for Cathedral Square in New Zealand, where Photosynth user Redpaw [47] used 330 images to construct a scene. A section of this that corresponds to the Christchurch Cathedral is shown in Figure 3, depicting: the point cloud (Figure 3a); and an overview map generated from the process (Figure 3b). To assist the reader, the figure also includes a sample image from the captured image set (Figure 3c) and an aerial image from the location as seen in Google Earth (Figure 3d). The points in the data cloud may be georeferenced and rectified to fit the corresponding GIS layer. As the color details from the original image pixels are maintained it is therefore possible to retrieve color information for any part of the building where a correspondence exists.

This technique shows promise for mapping open expanses, but the high number of overlapping images required to produce high density point clouds restricts its use in more confined and cluttered spaces. Therefore a less demanding approach which requires only a single stereo image pair, taken a known distance apart, is considered next.

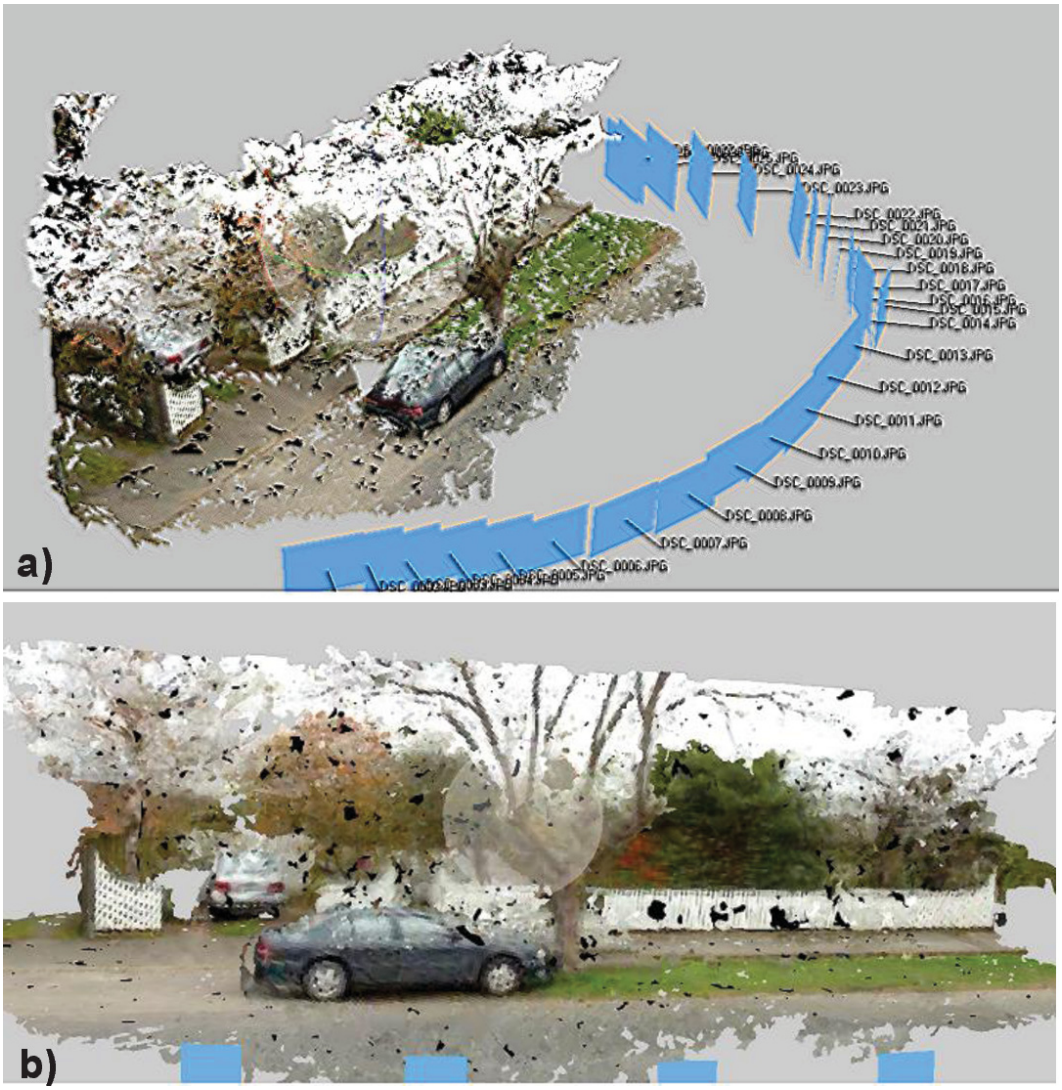

Figure 2: Images rendered from a model constructed using structure from motion approach

2.2 Stereo baseline vision

Stereo baseline vision works by comparing two images taken parallel to each other but a short distance apart, known as the baseline distance. The technique requires the same feature to be identified in both images, so that the horizontal disparity may be measured. Larger disparities indicate objects are closer to the camera, as a result of distance parallax.

Computer vision techniques are used to identify stereo correspondences by locating interesting features (e.g. corners, edges) in one image, for example using Harris [23], Förstner [20], or SURF [5] definitions, and search the paired image for the most similar matching template. The automatic process produces a disparity map, which can be trans-

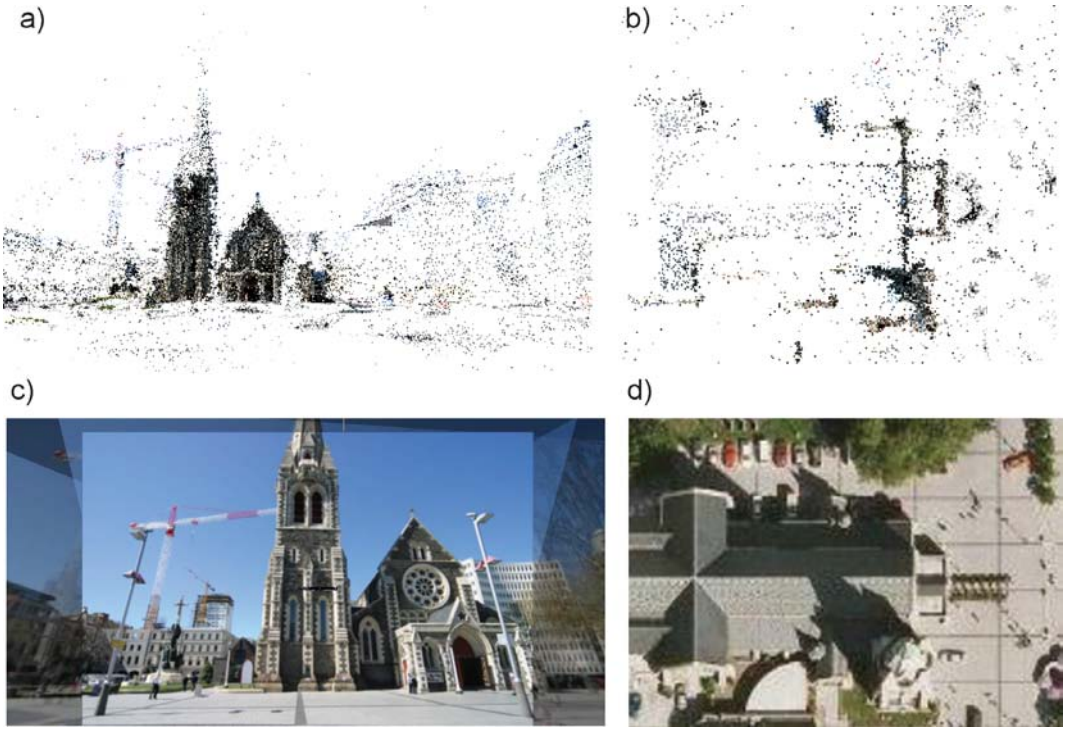

Figure 3: a. 3D point cloud; b. map view; c. sample image of Cathedral Square, Christchurch captured by RedPaw (2008); and d. equivalent region view in Google Earth

formed into a depth map with knowledge of the camera's focal length, pixel size, and the baseline distance between the image pair [11].

In order to collect a sample dataset to trial color extraction without foreground objects in a suburban region, a simple stereo rig was built from two low cost webcams, set a baseline distance of 60cm apart. It is necessary to perform a one-off calibration of the stereo camera setup to calculate various intrinsic and extrinsic details, allowing lens distortions to be removed from future images. This was done by capturing multiple views of a flat chessboard pattern using Emgu.CV [1], a C# implementation of OpenCV [2]. Initial static trials of the rig showed that depth could be recovered successfully up to a distance of around 30 meters; however the webcams suffered from various distortions when moving at 50km/h due to their shutter design, rendering the images unsuitable for depth mapping.

Instead a single high quality video camera with optical stabilization was used to capture 25 images per second with a shutter speed of $1/500^{th}$ second to ensure sharp images without motion blur. The camera was fixed perpendicular to the direction of travel. A GPS device was used to log speed, orientation, and location at 1Hz. The advantage of this setup is that it is extremely simple to implement, requiring only intrinsic details for a single lens, and uses technology available to a wide audience, potentially allowing it to be implemented on public transport vehicles such that urban color datasets could be regularly updated.

To solve for depth using the stereo vision approach requires an input of the baseline distance between image capture locations. This can be calculated by measuring the ground distance traveled between frames, which for a car traveling at 50km/h would be a baseline distance of 56cm, as shown in example A of Table 1. Rather than using GPS location information and measuring the distance between shutter releases, the relative offset between frames may be calculated more accurately using GPS speed. The main difference is that GPS speed is considered to be accurate to within 0.2m/s (0.72km/h) [59], and is calculated using Doppler shift making it more robust in multipath environments. At a frame rate of 25 images per second the speed accuracy would constitute a baseline discrepancy in the region of 0.8cm, resulting in negligible depth inaccuracies. This can be observed in example B of Table 1, where an increase in camera speed of 0.2m/s results in a baseline distance 0.8cm greater than shown in example A.

Example	Frame rate per second	Speed (km/h)	Speed (m/s)	Baseline distance (m)
A	25	50.00	13.89	13.89 / 25 = 0.5556
B	25	50.72	14.09	14.09 / 25 = 0.5636

Table 1: Example of how to calculate baseline distance

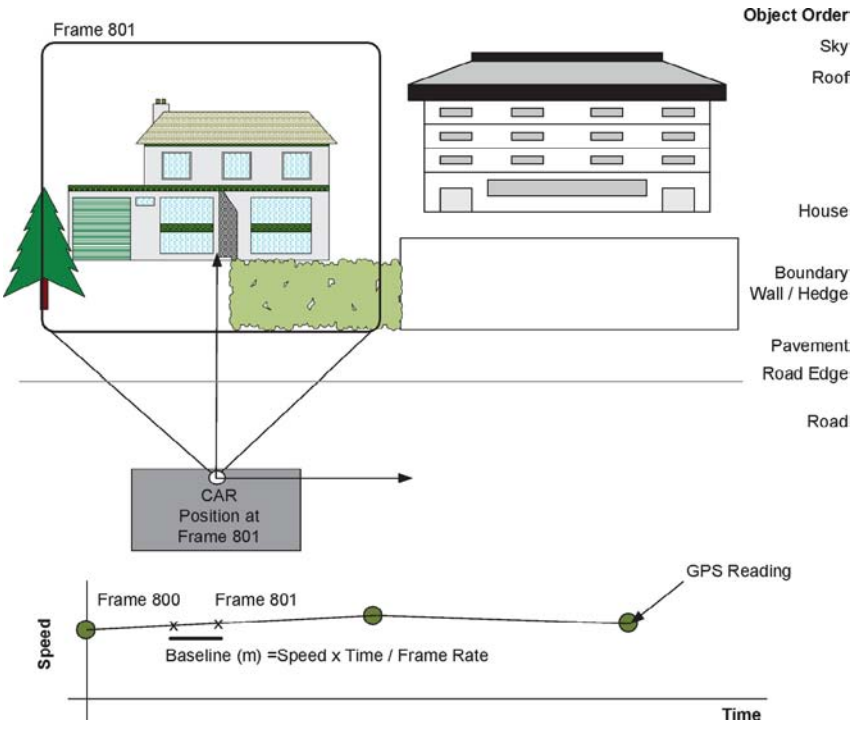

Figure 4: Calculating the baseline and expected disparity

To ensure the most accurate baseline estimate for any frame pair, the speed was interpolated from the surrounding GPS measurements, as shown in Figure 4. Here a captured

frame (801) falls between two GPS readings, resulting in an interpolated speed, from which the baseline distance may be calculated (ground distance between frames 800 and 801) so that disparity information may be translated into real world depth units (meters).

Disparity is inversely proportional to depth. Being a non-linear relationship, high depth resolution is only available for objects near the camera. The equation is given as:

$$d = fBp/Z$$

where d is the disparity expected between features in the image, f is the camera's focal length, B the baseline distance moved between frames, p the pixels per centimeter on the camera's sensor, and Z is the distance from the camera to the house [11].

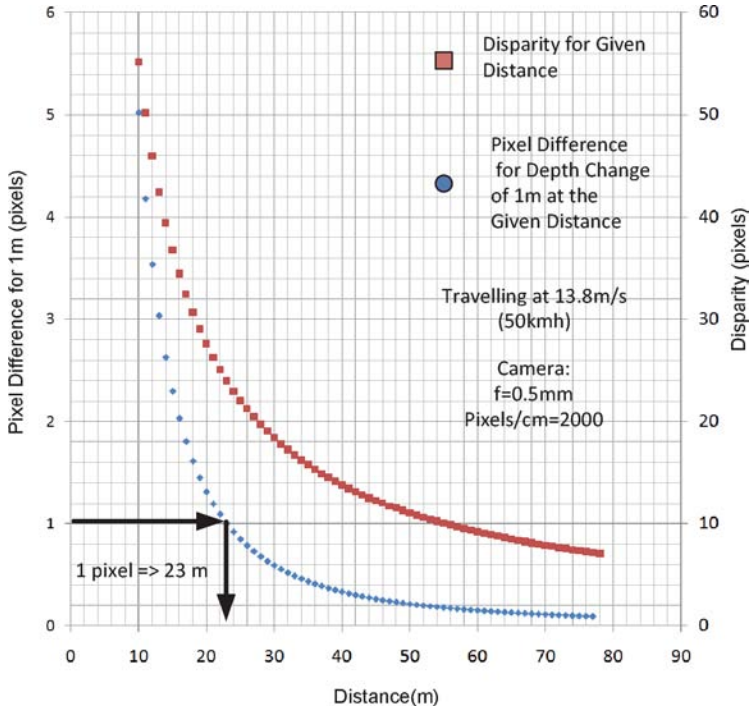

Figure 5: Example of the relationship between distance and pixel disparity

An example of this relationship for a camera traveling at 50km/h is shown in Figure 5, plotting disparity (right-side y-axis) against world object distance (x-axis). In addition the left-side y-axis displays the difference in pixel disparity which would be observed for a one meter change in object distance, considered as a reasonable margin of error for limiting the incorrect inclusion of any cars parked in front of buildings. As an example an object at a distance of 23 meters from the camera would have a disparity of 24 pixels, while an object 24 meters away would have a 23 pixel disparity. This is the limit at which a one meter change in distance can be measured at the pixel level for this camera at a speed of 50km/h, essentially defining the maximum reliable working depth. Beyond this the recovery range can be increased by using a larger baseline accommodated here by using a two frame offset between left and right images, extending the one meter depth resolution limit to a distance

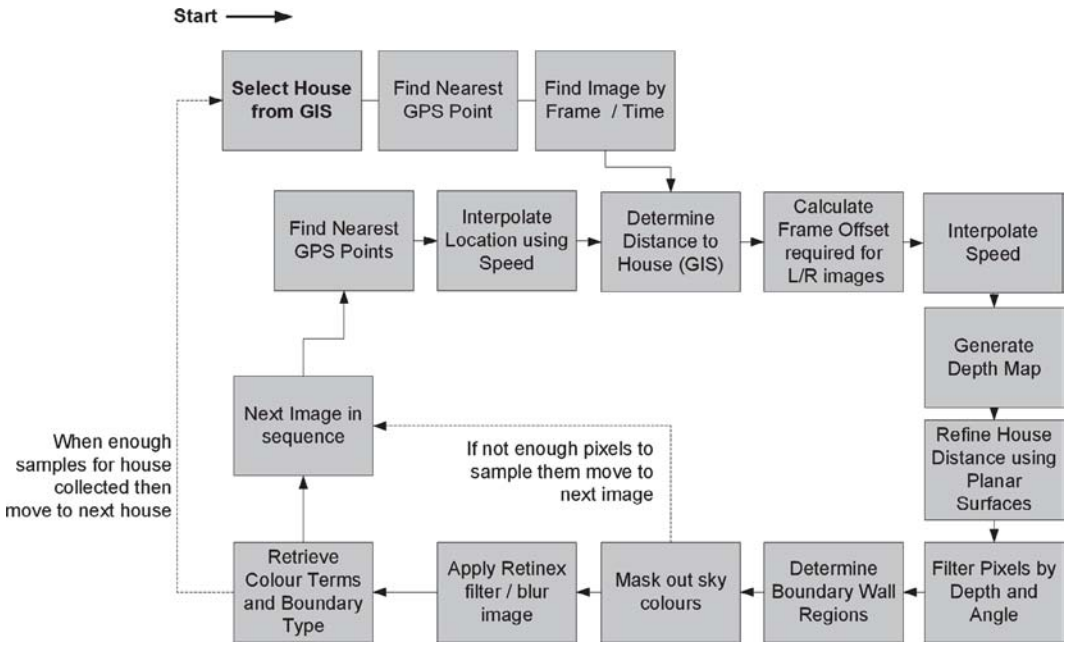

Figure 6: Detailed processing pipeline

of 32 meters. A three frame offset extends the search resolution to 40 meters from the camera. However greater offsets result in less similar foreground views, rendering it more difficult to find stereo correspondences for closer items [50].

A slower moving car will result in a shorter baseline between images, dropping depth resolution for distant items. This can be overcome automatically by calculating the current forward speed for each frame and the expected distance to the target building from a GIS layer, then choosing the lowest frame offset suitable for the required depth, thereby balancing the highest level of stereo correspondence matching and depth resolution. This approach offers both simplicity in data capture, and reduced computational requirements in data processing, with the flexibility to capture objects at a wide range of distances, and was therefore adopted for this research.

3 Processing pipeline

The processing pipeline implemented is summarized in Figure 6, followed by explanations of the components. The process begins with the identification of a target building from a GIS dataset, after which the nearest GPS data sample point is determined and the corresponding frame retrieved from the image stream. The distance from the camera to the target is measured from the GIS layer. In conjunction with the camera speed at the time of image capture, this distance is used to determine the optimum baseline distance required between the stereo image pair. In addition the baseline distance is used to transform the disparity map into a depth map (Section 2.2), refined using a depth frequency histogram approach (Section 3.2). The pixels at the corresponding target depth are selected, with those

matching the current sky hue dropped from the selection (Section 3.4). If the pixel count remains above a specified threshold the image is considered suitable for inclusion in the classification, otherwise the next image in the sequence is processed. Before classifying, a Retinex filter is applied (Section 3.5), and the image blurred and dilated slightly to remove pixel color noise. The selected pixel color values are saved as an array with the target building identification number. The process is repeated until 10 locations are collected for each target building. In some cases it was impossible to collect enough good data for a target and so these were marked as irretrievable.

At this point in the processing pipeline two pixel groups were identified, one for the most likely house pixels and the other for any likely boundary wall (Section 3.3). The final stage of the process determines the most appropriate color terms for each of these groups, as explained in Section 4.

3.1 Sampling strategy

A sampling strategy was implemented to ensure that a number of images would be captured for each target, and should the building be obscured behind foreground objects that alternative views would be used. In all cases the first sample was set at the mid-point along the building's length from the road, with subsequent samples taken radiating outwards from this point offset a frame in either direction (i.e., +1 frame, then -1 frame, then +2 frames, -2 frames etc), forming a sampling sequence as shown in Figure 7.

The camera's position for each sample was interpolated from known GPS locations, weighted according to the acceleration or deceleration experienced at that time. It proved necessary to also estimate the approximate horizontal position of the building within the image frame, so that objects at similar depths to the side of the intended target, such as neighboring houses, could be filtered out from the pixel selection process. This was achieved by projecting the building's edges into the image based on the known camera's field of view, which was calculated in the initial camera calibration stage and remained constant throughout data capture.

Each sample location was only considered valid if the number of pixels successfully placed within the expected house location was above a given threshold. For our trials this threshold was arbitrarily set to 100 pixels, deemed the minimum requirement for a fair reflection of the house color. It was also considered that 10 successful sample locations, giving 1000 pixels in total, should be used before selecting the next target feature.

The sampling approach is depicted in Figure 7, whereby the locations A and B would be expected to give the highest pixel counts for each building. However on occasions where high fences or vegetation restricted the view the most successful samples were collected across driveways, such as at location C between houses. It is in these cases in particular that the estimation of the horizontal space occupied by the target is required, ensuring that only those pixels relating to the designated target are included in the color summary.

As neither the GPS location nor GIS building layer inputs are error-free, additional steps were taken to improve system robustness.

3.2 Improving system robustness

To accommodate system noise, from errors introduced in the GPS, GIS, or baseline calculations a further step was added to the processing pipeline. By summarizing the depth map

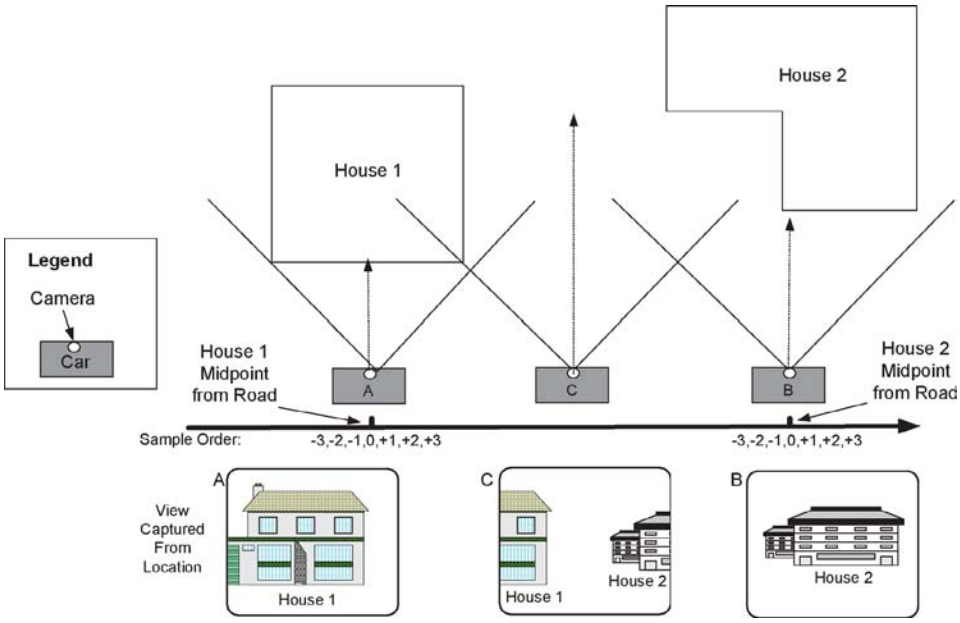

Figure 7: Estimating house position in image

vertically (against the x-axis) a frequency graph was generated which shows the most commonly occurring depths. Vertical planar surfaces result in high frequency depth counts, as a high proportion of the column exhibits the same depth value. Neighboring high frequency values across the graph (y-axis) indicate a wide planar surface exists, such as a building or wall.

Figure 8 shows an example of the depth frequency histogram superimposed and aligned onto the corresponding image, with only significant values displayed after a high pass filter has been applied to remove low frequency noise. This depth summary was used to refine the expected depth value for the target building, thereby correcting small discrepancies in the input variables. In addition a depth tolerance of 1.5 meters was permitted in all cases, judged to be shallow enough to exclude cars parked in front of target building, yet giving some flexibility to the pixel selection process.

3.3 Identifying property boundary type

In addition to searching for building color information it was also possible to retrieve information on boundary fences and wall colors by searching in the lower part of the image for the existence of planar surfaces, occurring at a distance between that of the road edge and target building (see Figure 4). Planar surfaces were identified by passing a 3×3 kernel over the depth map to compare the depth gradients between cells, identifying regions of constant gradient. Figure 9a shows a theoretical example of this where a fence would exhibit linear gradients in both x and y directions, while vegetation would result in non-uniform gradients. By using a recursive function it was possible to identify significant planar candidates with similar gradient properties, merging results to define the larger planar surfaces,

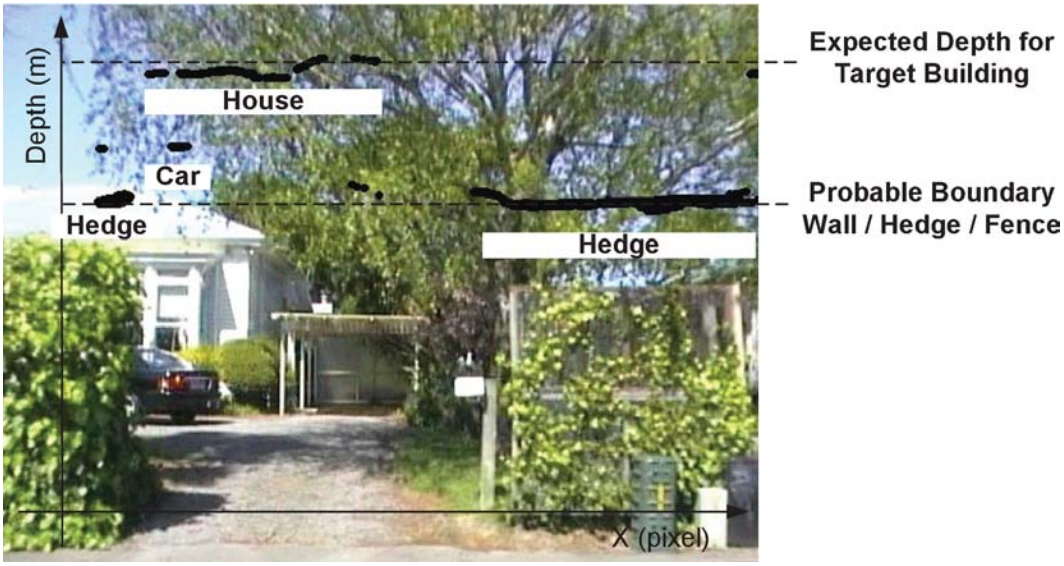

Figure 8: Using a depth frequency histogram to refine the selection of target building search

which in turn were filtered to leave only those vertical and facing the camera. Figure 9b shows the planar surfaces facing the camera which were recovered from the scene depicted in Figure 8.

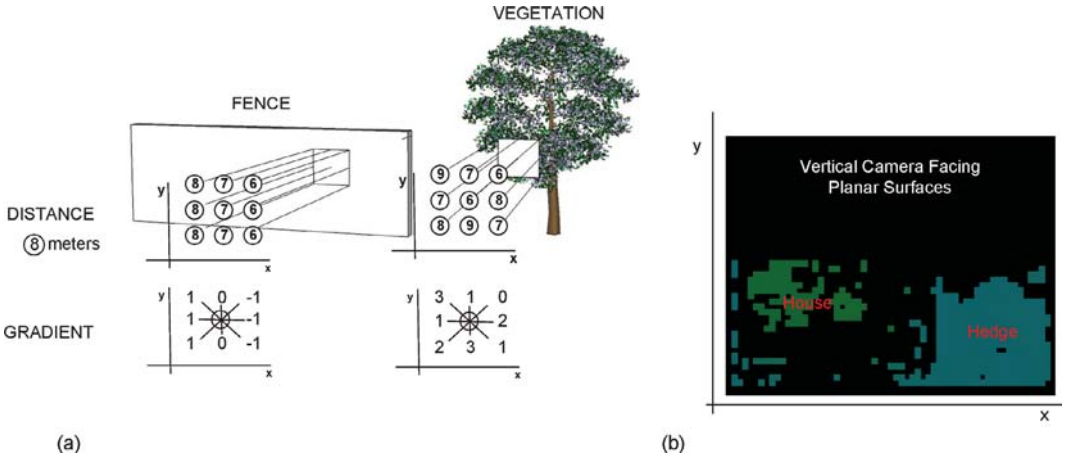

Figure 9: Detecting planar and non-planar surfaces using a 3×3 kernel

The existence of any large flat surfaces at the property boundary could be identified by using the output from the depth frequency histogram (Figure 8) in conjunction with the existence of these planar regions, compared to the known road edge distances measured from the GIS road dataset. An example of the output from each stage of the process from

stereo views, to generating the disparity map, to selecting the house pixels, and discovering planar surfaces is shown in Figure 10.

As well as discovering the boundary color an attempt was made to determine the boundary material. Material recovery is recognized to be a difficult task [24]. Therefore, to simplify matters the options were limited to a predetermined list of wall/fence, slat fence, or vegetation.

Hedges could be recognized due to the dense counts of Harris corners [23]. Slat fences can be found by applying a Gaussian blur to the image before running the Canny edge detector [12]. The part of the image corresponding to the boundary was then scanned horizontally at a number of places tallying the number of intersections with detected Canny edges. Regions exhibiting a similar number of intersections across each horizontal scan, those with a low standard deviation, were deemed to be likely slat fence candidates, as shown in Figure 11. Other borders which did not satisfy either of these conditions were labeled as wall/fence.

3.4 Windows

One of the issues encountered when automatically recovering building color histograms is the presence of windows, which have no color of their own but either reveal the interior, or reflect the surroundings. There are methods which can be used to automatically determine window locations based on gradient projection approaches [46]. However these require the extent of the façade to be pre-defined in each image. A more simplistic approach is to filter the image for sky hues in HSL (i.e., hue, saturation, lightness) color space, thereby identifying those surfaces reflecting the sky. The pixels selected with similar color may be removed from further consideration, leaving a smaller set of candidates with a higher likelihood of being part of the façade. It was found that on both bright sunny days and bright overcast days the procedure worked fairly well, although future research should look to implement more sophisticated procedures to identify windows under all lighting conditions (Figure 12).

3.5 Shadows

The color summary is also complicated by shadow regions, which are darker patches resulting from changes in lighting as a result of surrounding features [17]. A Retinex filter [33] may be applied to images to reduce the effect of illumination variation, as shown in Figure 13. Here an image is reduced to use the closest of only 11 colors before and after the Retinex filter is applied, showing an improvement in color matching for regions in shadow (such as trees) after processing.

Although the Retinex process improves the color classification, some illumination artifacts remain in the image (e.g., strong shadows show up as sky blue), and further shadow reduction techniques may prove beneficial [19, 27, 49]. The issue of strong shadows is reduced by limiting data capture to bright overcast days, when the lighting source is more diffuse.

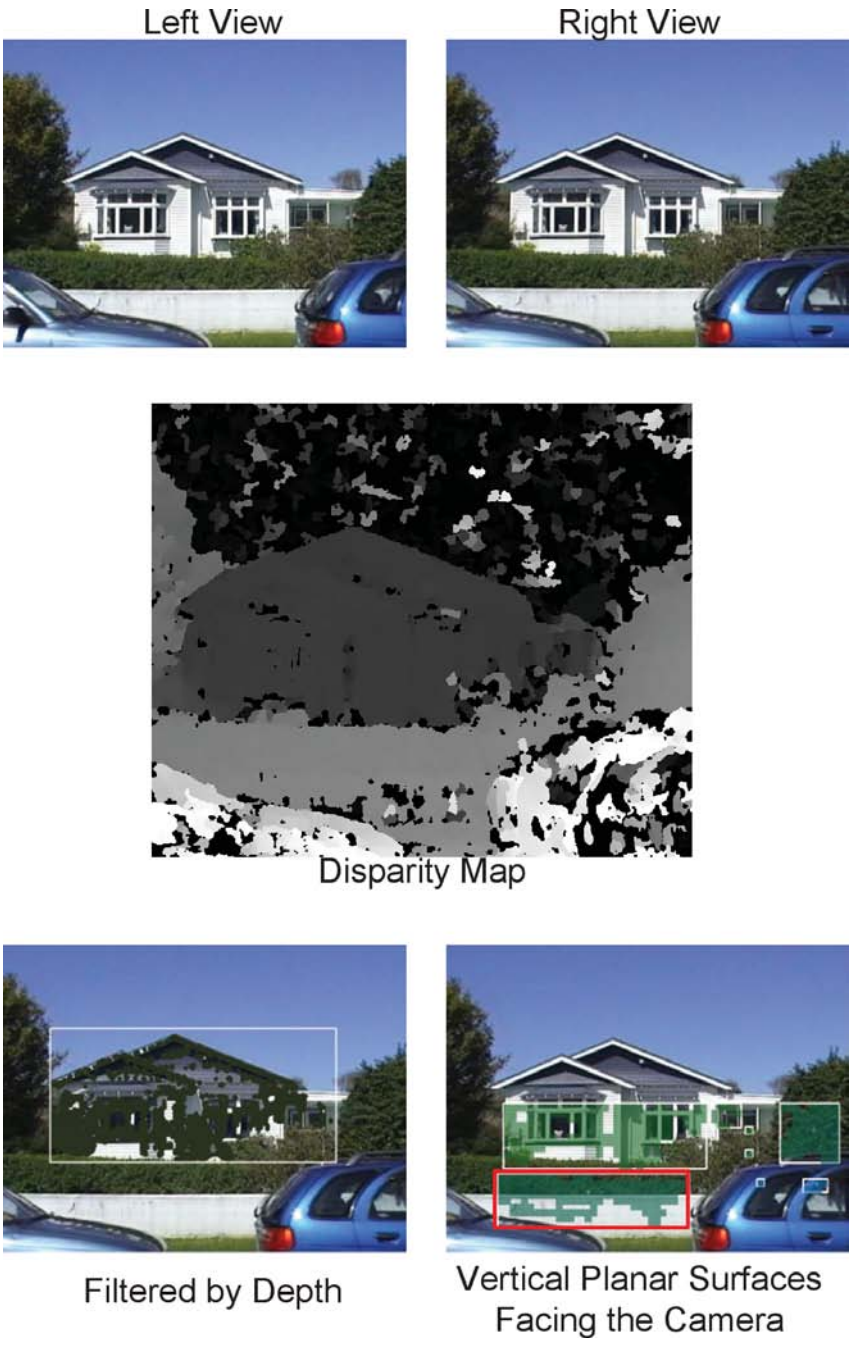

Figure 10: Stereo vision used to build a disparity map, and turned into a depth map used to identify building pixels. The most likely boundary wall/fence region highlighted in red.

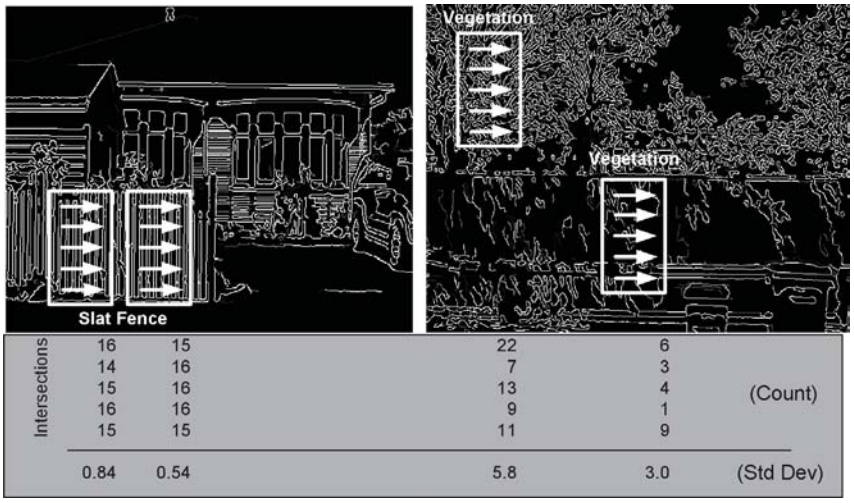

Figure 11: Slat fence detection using Canny edges and showing intersection counts and standard deviation

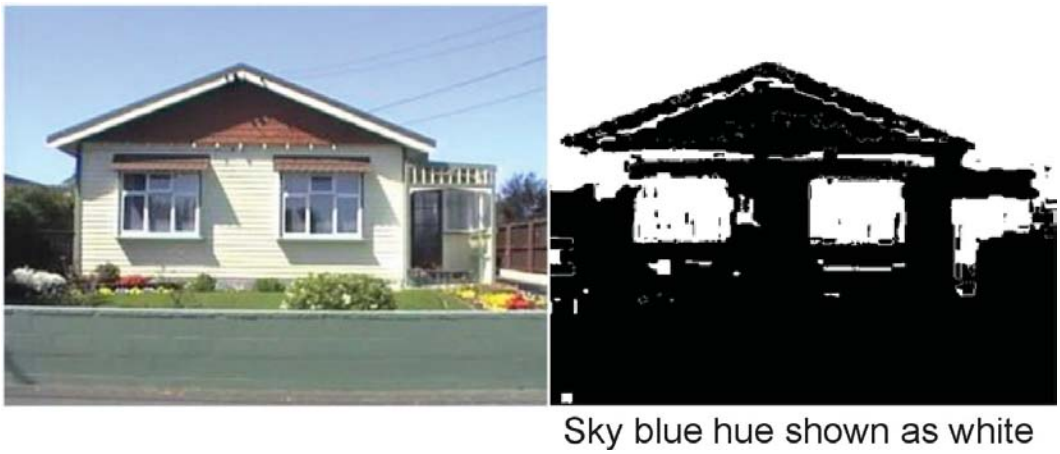

Figure 12: Avoiding window regions from selection using sky hue filter

4 Color terms

The mapping of pixel values to color terms is rather complex, as the perception of color is related to many factors [32]. There are issues of chromatic induction [26], which is when similar hues are judged to be different due to the contrast with surrounding colors, and effects of lighting where similar colors appear very differently depending on shadows cast onto the surface. Furthermore color terms describe regions in color space which are only vaguely defined, and vary depending upon the viewer.

There are many color terms which are highly specialized and not in general use by the public (e.g., chartreuse). It was therefore necessary to first determine a list of the most pop-

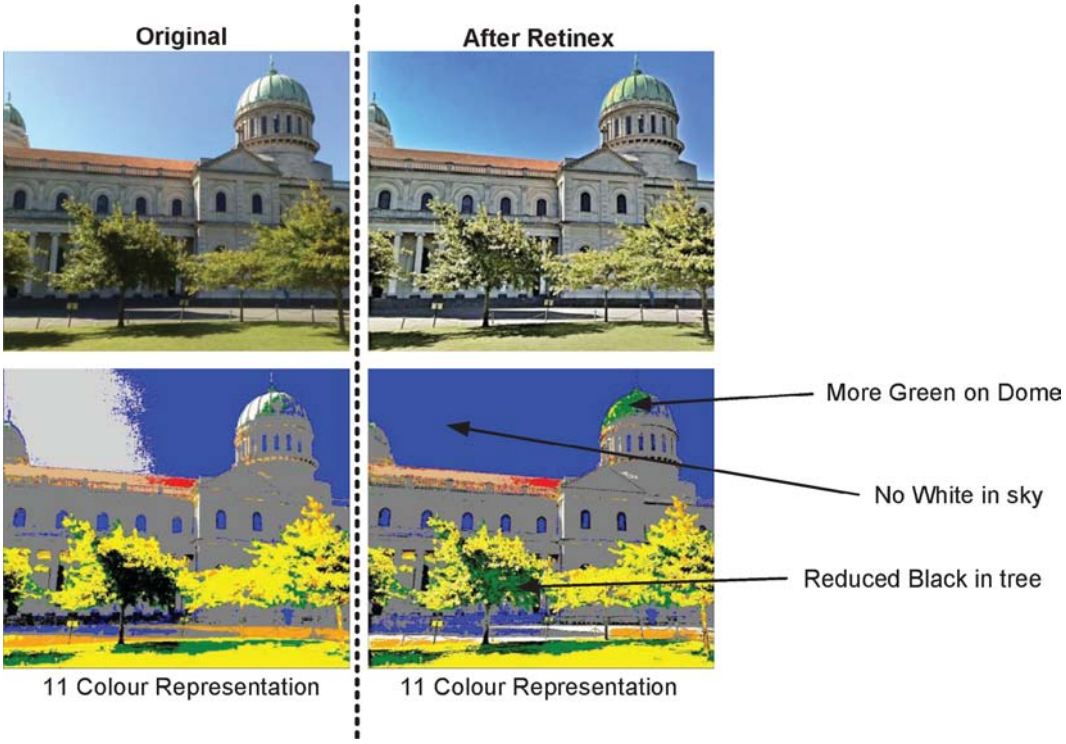

Figure 13: Using a Retinex filter to reduce illumination variations

ular terms in use and rate each building against these terms, generating a fuzzy color classification [6]. This approach means that a target could be identified with varying strengths in a number of different color groups, offering a degree of flexibility in any future system using the dataset, such that one user may refer to a building as "orange" while another calls it "red."

Previous research highlights 11 main perceptual color focuses [7, 53]: black, white, red, green, yellow, blue, brown, purple, pink, orange, and gray. Boynton and Olson [10] found that 424 subjects could repeatedly consistently identify the colors in this group without any confusion, therefore these 11 focuses were adopted for this research.

Before assigning color terms to an image, each color requires a definition in color space. The RGB values defined for each term were taken from the HP color thesaurus, an online databank of defined color centers constructed from people around the world [40], for example the most widely used red definition uses RGB values (216, 35, 44). The selected building pixels were then compared to the color terms by translating the values into HSL color space, before calculating their two-dimensional Euclidean distance using the hue and saturation values. The process was repeated for all 11 colors giving a fuzzy classification for the building sample against the 11 color terms. This was repeated for each sample for each target building, and the classifications were summed to produce a single building fuzzy color set.

4.1 Color entropy

Color entropy is a measure of the ambiguity of the assigned color [13], which may be measured by counting the number of fuzzy classes with significant values (Figure 14). In cases where a building has a single strong classification the color entropy is low, and the generated natural description may include a single color term. However where color entropy is high a number of color terms may be required, such as "the red-ish orange building." This fuzzy classification accommodates variation in user-formed descriptions, such that inclusion of "red" or "orange" would give the same search results. Where many buildings in view match a given selection criteria other classifications may be required to narrow the results, such as building size, or roof color.

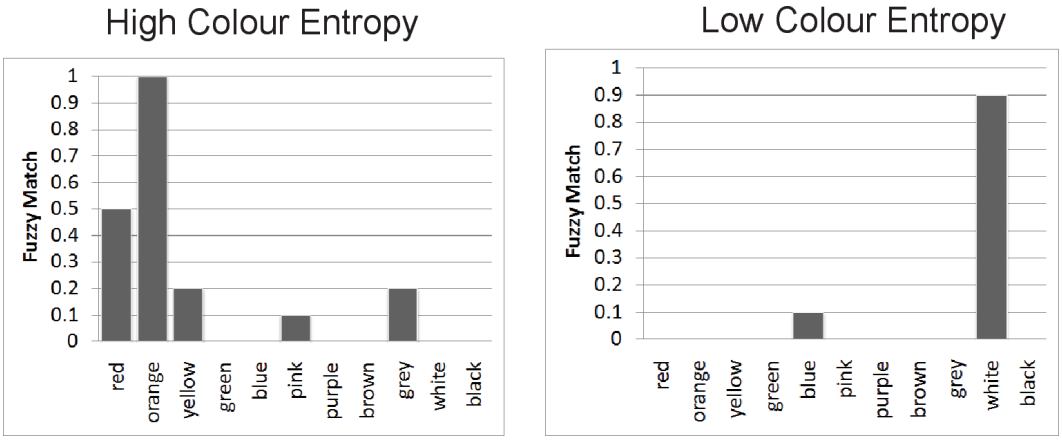

Figure 14: High and low color entropy examples

5 Trials and results

The system was trialled in a number of streets within Christchurch, New Zealand, during the summertime. The central city is grid based with a large suburban expanse, much vegetation and many gardens, providing a suitable test environment for identification and filtering of house façades from behind foreground objects such as parked cars, trees, bushes, and people.

Figure 15 shows a selection of the identified façades, with corresponding fuzzy color classifications. Notice that despite fairly limited views house color could still be recovered from the views across driveways (A), and also where foreground vegetation was present (B, C). Roofs were not included in selections as a result of being non-vertical planar surfaces, and could more easily be captured from aerial imagery. Garage doors proved to be an issue however (C, D) and were generally ignored for the classification as stereo depth mapping failed to produce stable results on the very similar textured surfaces. Despite efforts to reduce selections on window surfaces there were occasions where pixels were selected (E). However on most occasions the non-window façade pixels dominated the selection, rendering the window pixels less significant in the color classification.

To evaluate the system's performance a comparison was carried out with a sample collected from a walk along a number of streets noting house color information. It was found that the automated procedure was able to correctly identify the most prominent color of 33 out of 43 houses (77%) when compared to these manual values. The occasions it failed were mostly a result of incorrect color term classification due to shadows, rather than incorrect pixel selection.

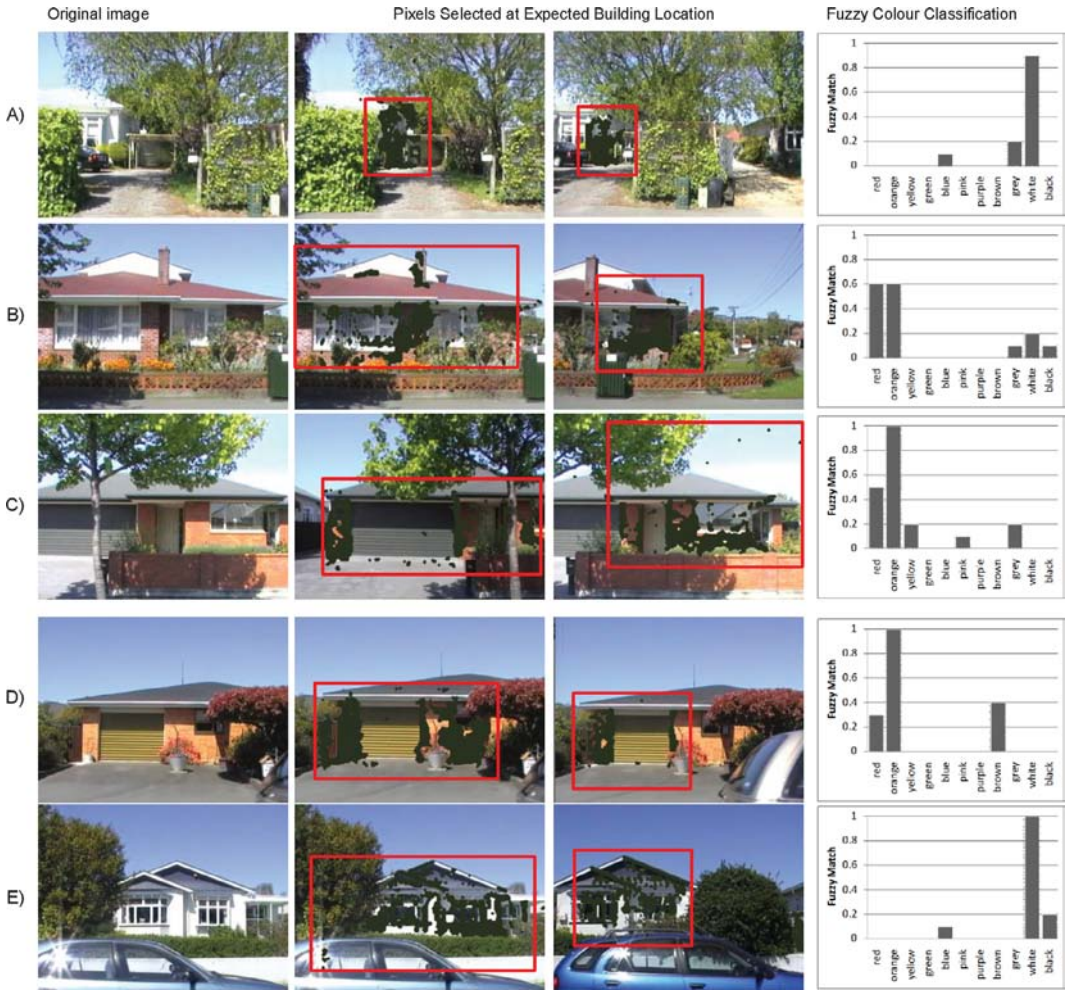

Figure 15: Example façade determination and color classification. The minimum bounding box for building pixels selected shown in red.

In addition boundary wall information was gathered by selecting linear planar features. Boundary wall locations were identified correctly in 36 of the 43 targets (84%). Material recovery proved to be a more difficult task, and during trials vegetation shadows sometimes caused incorrect classifications, particularly when cast onto plain painted fences. Slat fence detection was the most reliable classification with no false positives, and few false negatives. A summary of the target houses processed are shown in Figure 16, including the

graph of depth used to first identify the presence and location of vertical linear features in front of the house, the main planar regions, and the results of the Harris corner detection and Canny filter process.

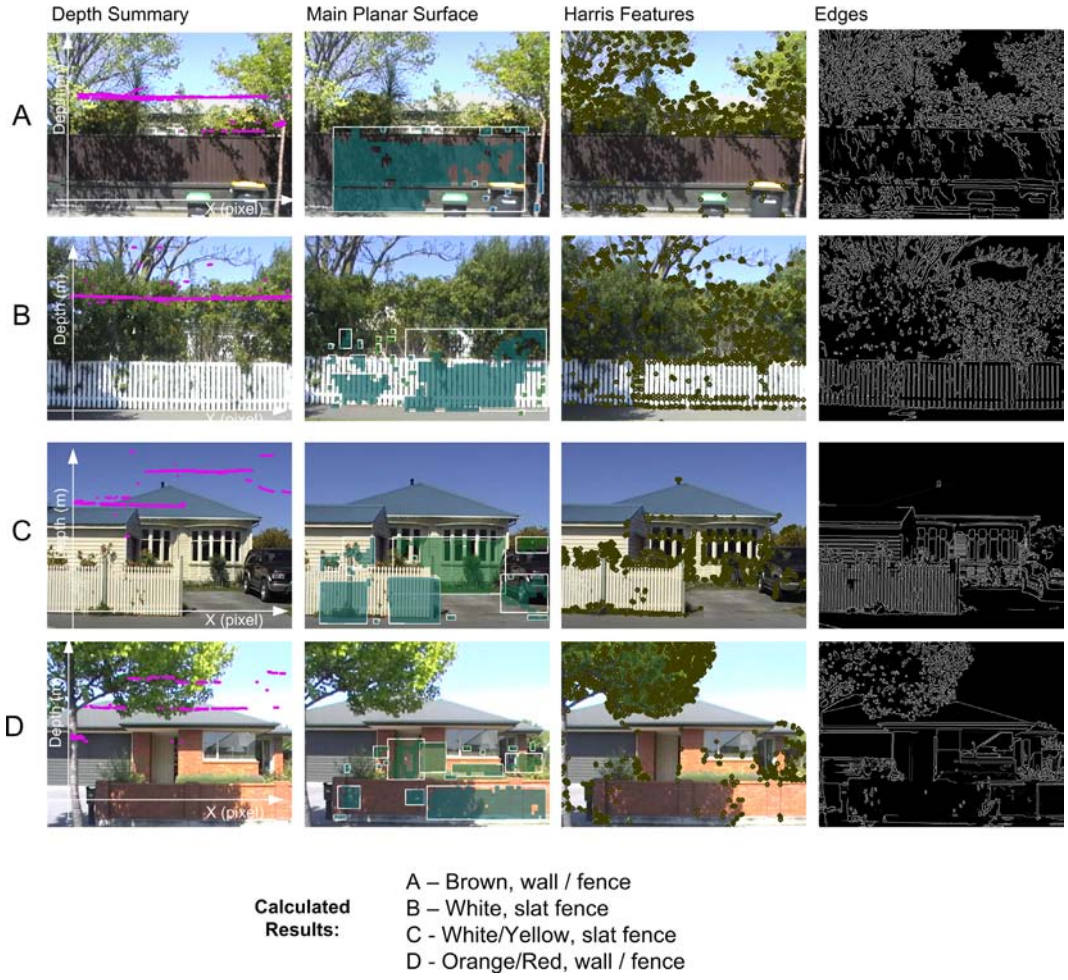

Figure 16: Detecting candidate boundary walls and their material

6 Conclusion and future work

When people form descriptions of buildings they often include references to attributes which are lacking in existing spatial datasets, such as façade color and material type. In recent times there has been a dramatic rise in efforts to capture this texture and color information from street level. However the information is locked in images restricting its usefulness beyond visualization. To be able to extract building color terms for use in forming natural language descriptions, and for verbalization, requires an ability to identify which

pixels correspond to buildings by filtering out foreground objects. These pixels may then be summarized to generate the most appropriate color term for use in forming a building description.

The work presented here demonstrates how this may be achieved using a low-tech hardware solution, available to most communities, in conjunction with spatial analysis. The processing pipeline begins by determining the expected building distance from the camera to the building by analyzing the GIS dataset. Computer vision techniques are used to build a disparity map from images captured from a moving camera. The disparity map is converted to a depth map using GPS speed information to determine the baseline distance between frames. From these inputs it is possible to filter the image for pixels at the expected building distance and generate the color summary. Although a full evaluation of color perception was beyond the scope of this work, the method of using a single moving camera and GPS unit to build stereo views was able to generate meaningful color descriptions, and also to identify boundary walls and fences in images. There were however a number of issues which require attention in future versions of the system.

The use of a single camera limits the system's operation to straight road sections, as when navigating corners the rotation of the camera means parallel image views are not available. This was only a minor problem for our tests in a grid based city, but could be overcome by using the structure from motion approach to recover an estimate of the camera pose and orientation from which pixel depth information may then be calculated. Alternatively a multi-baseline stereo camera, consisting of a set of cameras with fixed baselines, could be used [29]. This would offer the advantage of an increase in operational depth recovery compared to a single stereo camera, and an ability to validate correspondences from a number of image pairs improving depth recovery robustness.

Color term retrieval was influenced by the presence of strong shadows in images. Although the Retinex filter stage of the processing pipeline improved the situation it became evident that more sophisticated approaches would be beneficial in some cases. New techniques such as entropy minimization [18] may be worth considering for future versions of the system.

The ability to retrieve material types for both boundary walls and house facades would be beneficial when generating natural descriptions. However this proved to be difficult. For example although bricks are uniform shapes they occur in a range of colors. In theory shape detection methods would be suitable to identify them. Unfortunately Canny edge detection proved unreliable at the operating distances required. Higher quality cameras and lenses would improve this, but shadows are often cast on to the solid surfaces and may still impede these edge detection methods. Slat fences were one of the most easily recognized boundary types, as the edges were clearly defined even under strong shadow.

Currently only a single color definition is collected per building façade or fence. However some buildings have different colored side walls, or a multiple boundary types (eg low wall with a fence above). By exploiting the planar surface details it would be possible to divide target buildings into sections based on the direction the wall faces, and map these color values to the GIS layer, thereby creating color summaries for subsections of the target building. In addition the vertical boundary regions could be subdivided into those sections which appear most wall-like, and most vegetation-like, to derive more complete descriptions.

Garage doors are often different colors to the main house, and would be useful additions to the descriptors. Currently they tended to be regions ignored from the summary,

due to difficulties in retrieving stable depth value for such similar textured regions. In addition window frames and doors may be segmented in the image for separate color analysis, giving rise to very detailed house description possibilities. It is possible that windows may be identifiable as regions which change appearance with direction across views, therefore tracking a lack of consistency could be used to detect their locations.

While this study has focused on suburban regions there would be value in extending the trials to different cityscapes, including industrial and commercial regions. Color recovery in wet conditions should also be tested to determine the effect that more reflective surfaces have on the depth and color capture process.

The automation of building color capture lends itself to a possible future whereby centralized color databases are maintained from systems installed on public transport (e.g. buses, taxis), or volunteered domestic cars. In so doing color information could be updated regularly and included in a wide range of disciplines from urban cartographic maps, to emergency response, to speech based LBS applications.

Finally, it must be acknowledged that referring expressions may not always use color as a descriptor, and that a process is required to evaluate the relative merits of a range of descriptive attributes. For example if many buildings in a scene have very similar colors yet only one has a balcony, then the balcony feature would be a more suitable descriptor. The challenge of determining the most suitable attributes for visible features should form the basis of the next stage of this work.

7 Acknowledgments

The authors would like to thank Dr Cynthia Brewer and Dr William Mackaness for their contributions during the concept phase of this paper, also Associate Professor Simon Kingham and the anonymous reviewers for their comments on the original submission.

References

[1] Emgu.CV. http://www.emgu.com/wiki/index.php/Main_Page. Retrieved 15 September 2010.

[2] OpenCV. http://opencv.willowgarage.com/wiki/. Retrieved 5 September 2010.

[3] AGISOFT. PhotoScan software. http://www.agisoft.ru/products/photoscan/. Retrieved 11 October 2010.

[4] BARTIE, P. J., AND MACKANESS, W. A. Development of a speech-based augmented reality system to support exploration of cityscape. *Transactions in GIS 10*, 1 (2006), 63–86. doi:10.1111/j.1467-9671.2006.00244.x.

[5] BAY, H., ESS, A., TUYTELAARS, T., AND GOOL, L. V. Speeded-up robust features (SURF). *Computer Vision and Image Understanding 110*, 3 (2008), 346–359. doi:10.1016/j.cviu.2007.09.014.

[6] BENAVENTE, R., VANRELL, M., AND BALDRICH, R. A data set for fuzzy colour naming. *Color Research & Application 31*, 1 (2006), 48–56. doi:10.1002/col.20172.

[7] BERLIN, B., AND KAY, P. *Basic color terms: Their universality and evolution*. University of California Press, Berkeley, 1969.

[8] BERTIN, J. *Semiology of graphics*. University of Wisconsin Press, Madison, 1983.

[9] BIRCHFIELD, S., AND TOMASI, C. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision 35*, 3 (1999), 269–293. doi:10.1023/A:1008160311296.

[10] BOYNTON, R. M., AND OLSON, C. X. Salience of chromatic basic color terms confirmed by three measures. *Vision Research 30*, 9 (1990), 1311–1317.

[11] BRADSKI, G., AND KAEHLER, A. *Learning OpenCV*. O'Reilly, Sebastopol, CA, 2008.

[12] CANNY, J. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 8*, 6 (1986), 679–698. doi:10.1109/TPAMI.1986.4767851.

[13] CHUANG, J., STONE, M., AND HANRAHAN, P. A probabilistic model of the categorical association between colors. In *Color Imaging Conference*. Retrieved 17 July 2010.

[14] DALE, R., AND REITER, E. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science 19*, 2 (1995), 233–263. doi:10.1207/s15516709cog1902_3.

[15] EL-HAKIM, S. F., BRENNER, C., AND ROTH, G. A multi-sensor approach to creating accurate virtual environments. *Journal of Photogrammetry and Remote Sensing 53* (1998), 379–391.

[16] ELIAS, B. Determination of landmarks and reliability criteria for landmarks. In *5th Workshop on Progress in Automated Map Generalization* (Paris, France, 2003).

[17] FINLAYSON, G., HORDLEY, S., AND DREW, M. Removing shadows from images. In *European Conference on Computer Vision* (2002), A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds., pp. 823–836. doi:10.1007/3-540-47979-1_55.

[18] FINLAYSON, G. D., DREW, M. S., AND LU, C. Entropy minimization for shadow removal. *International Journal of Computer Vision 85*, 1 (2009), 35–57. doi:10.1007/s11263-009-0243-z.

[19] FINLAYSON, G. D., HORDLEY, S. D., AND DREW, M. S. Removing shadows from images using retinex. In *Color Imaging Conference Proceedings* (Arizona, USA, 2002), pp. 73–79.

[20] FÖRSTNER, W., AND GÜLCH, E. A fast operator for detection and precise location of distinct points, corners and centers of circular features. In *ISPRS Conference on Fast Processing of Photogrammetric Data* (Interlaken, 1987), pp. 281–305.

[21] FORSYTH, D., TORR, P., AND ZISSERMAN, A. Analysis of building textures for reconstructing partially occluded facades. In *Computer Vision*, T. Korah and C. Rasmussen, Eds. Springer, 2008, pp. 359–372. 10.1007/978-3-540-88682-2_28.

[22] FRUEH, C., SAMMON, R., AND ZAKHOR, A. Automated texture mapping of 3D city models with oblique aerial imagery. In *Second international symposium on 3D data processing, visualization and transmission* (Greece, 2004), pp. 396–403. doi:10.1109/TDPVT.2004.1335266.

[23] HARRIS, C., AND STEPHENS, M. A combined corner and edge detector. In *Proc. 4th Alvey Vision Conference* (Manchester, 1988), pp. 147–151.

[24] HAYMAN, E., CAPUTO, B., FRITZ, M., AND EKLUNDH, J.-O. On the significance of real-world conditions for material classification. In *Computer Vision*, T. Pajdla and J. Matas, Eds. Springer, 2004, pp. 253–266.

[25] HELVACIOG LU, E., AND OLGUNTÜRK, N. Color contribution to children's wayfinding in school environments. *Optics and Laser Technology 43*, 2 (2009), 410–419. doi:10.1016/j.optlastec.2009.06.012.

[26] JAMESON, D., AND HURVICH, L. M. Essay concerning color constancy. *Annual Review of Psychology 40*, 1 (1989), 1–24. doi:10.1146/annurev.ps.40.020189.000245.

[27] JIANHONG, G., LU, L., AND PENG, G. Removing shadows from Google Earth images. *International Journal of Remote Sensing 31*, 6 (2010), 1379–1389. doi:10.1080/01431160903475316.

[28] KANADE, T., YOSHIDA, A., ODA, K., KANO, H., AND TANAKA, M. A stereo machine for video-rate dense depth mapping and its new applications. 196–202. doi:10.1109/CVPR.1996.517074.

[29] KANG, S. B., AND SZELISKI, R. 3D scene data recovery using omnidirectional multibaseline stereo. *International Journal of Computer Vision 25*, 2 (1997), 167–183.

[30] KOENDERINK, J. J., AND VAN DOORN, A. J. Affine structure from motion. *Journal of the Optical Society of America A 8*, 2 (1991), 377–385. doi:10.1364/JOSAA.8.000377.

[31] KOPF, J., CHEN, B., SZELISKI, R., AND COHEN, M. F. Street slide: Browsing street level imagery. In *ACM Transactions on Graphics* (2010), pp. 1–8. doi:10.1145/1833351.1778833.

[32] LAMMENS, J. M., AND SHAPIRO, S. C. Learning symbolic names for perceived colors. In *Machine Learning in Computer Vision: What, Why and How?* AAAI Press, 1993.

[33] LAND, E. H. The retinex theory of colour vision. *Scientific American 237*, 6 (1977), 108–129.

[34] LE YAOUANC, J.-M., SAUX, E., AND CLARAMUNT, C. A visibility and spatial constraint-based approach for geopositioning. In *Geographic Information Science*, S. Fabrikant, T. Reichenbacher, M. van Kreveld, and C. Schlieder, Eds. Springer, 2010, pp. 145–159. doi:10.1007/978-3-642-15300-6_11.

[35] LENSCH, H. P. A., HEIDRICH, W., AND SEIDEL, H. P. Automated texture registration and stitching for real world models. In *Proc. 8th Pacific Conference on Computer Graphics and Applications* (2000), IEEE Computer Society, pp. 317–452. doi:10.1.1.18.4621.

[36] LUCAS, B. D., AND KANADE, T. An iterative image registration technique with an application to stereo vision. In *Proc. 7th International Joint Conference on Artificial Intelligence* (Vancouver, Canada, 1981), Morgan Kaufmann Publishers Inc., pp. 674–679. doi:10.1.1.49.2019.

[37] MICHELIS, D., RESATSCH, F., NICOLAI, T., AND SCHILDHAUER, T. The disappearing screen: Scenarios for audible interfaces. *Personal and Ubiquitous Computing 12*, 1 (2008), 33.

[38] MICUSIK, B., AND KOSECKA, J. Piecewise planar city 3D modeling from street view panoramic sequences. In *Proc. Computer Vision and Pattern Recognition (CVPR)* (2009), pp. 2906–2912. doi:10.1109/CVPR.2009.5206535.

[39] MONTELLO, D. Scale and multiple psychologies of space. *Spatial Information Theory A Theoretical Basis for GIS* (1993), 312–321.

[40] MORONEY, N. HP's online color thesaurus. http://www.hpl.hp.com/personal/Nathan_Moroney/color-thesaurus.html. Retrieved 11 October 2010.

[41] NOTHEGGER, C., WINTER, S., AND RAUBAL, M. Computation of the salience of features. *Spatial Cognition and Computation 4*, 2 (2004), 113–136. doi:10.1207/s15427633scc0402_1.

[42] OFFICE OF THE DEPUTY PRIME MINISTER. Final report for signage and wayfinding for people with learning difficulties. Tech. Rep. 6/2005, UK Government, 2006. http://www.communities.gov.uk/documents/planningandbuilding/pdf/144248.pdf.

[43] PALMER, T. C., AND SHAN, J. A. Comparative study on urban visualization using LIDAR data in GIS. *URISA Journal 14*, 2 (2002), 19–25.

[44] PYLVÄNÄINEN, T., ROIMELA, K., VEDANTHAM, R., ITÄRANTA, J., WANG, R., AND GRZESZCZUK, R. Automatic alignment and multi-view segmentation of street view data using 3D shape priors. In *3D Data Processing, Visualization and Transmission 2010* (Paris, France, 2010).

[45] RAUBAL, M., AND WINTER, S. Enriching wayfinding instructions with local landmarks. In *Second International Conference GIScience*, M. J. Egenhofer and D. M. Mark, Eds. Springer, Boulder, USA, 2002, pp. 243–259. doi:10.1007/3-540-45799-2_17.

[46] RECKY, M., AND LEBERL, F. Windows detection using K-means in CIE-lab color space. In *Proc. International Conference on Pattern Recognition (ICPR)* (Istanbul, Turkey, 2010), pp. 356–360. doi:10.1109/ICPR.2010.96.

[47] REDPAW. Cathedral square images (12mm). http://photosynth.net/view.aspx?cid=a3db4c7f-f638-4860-a790-9b3b9f66814d. Retrieved 8 August 2010.

[48] ROTTENSTEINER, F., AND BRIESE, C. A new method for building extraction in urban areas from high-resolution LiDAR data. *International Archives of Photogrammetry and Remote Sensing and Spatial Information Sciences 34*, 3A (2002), 295–301.

[49] SCANLAN, J. M., CHABRIES, D. M., AND CHRISTIANSEN, R. W. A shadow detection and removal algorithm for 2d images. In *International Conference on Acoustics, Speech, and Signal Processing* (Albuquerque, USA, 1990), IEEE, pp. 2057–2060. doi:10.1109/ICASSP.1990.115931.

[50] SCHARSTEIN, D., AND SZELISKI, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision 47*, 1 (2002), 7–42. doi:10.1023/A:1014573219977.

[51] SNAVELY, N., SEITZ, S. M., AND SZELISKI, R. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics 25*, 3 (2006), 835–846. doi:10.1145/1141911.1141964.

[52] SORROWS, M., AND HIRTLE, S. The nature of landmarks for real and electronic spaces. In *Spatial information theory*, C. Freksa and D. Mark, Eds. Springer, 1999, pp. 37–50.

[53] STURGES, J., AND WHITFIELD, T. W. A. Locating basic colours in the Munsell space. *Color Research and Application 20*, 6 (1995), 364–376. doi:10.1016/S0042-6989(96)00170-8.

[54] STURM, P., AND TRIGGS, B. A factorization based algorithm for multi-image projective structure and motion. In *Computer Vision ECCV96* (Cambridge, UK, 1996), Springer, pp. 709–720. doi:10.1007/3-540-61123-1_183.

[55] TRIGGS, B., MCLAUCHLAN, P., HARTLEY, R., AND FITZGIBBON, A. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice*, B. Triggs, A. Zisserman, and R. Szeliski, Eds. Springer, Corfu, Greece, 2000, pp. 153–177. doi:10.1007/3-540-44480-7_21.

[56] VINCENT, L. Taking online maps down to street level. *Computer 40*, 12 (2007), 118–120. doi:10.1109/MC.2007.442.

[57] WANG, Q., AND YOU, S. Automatic registration of large-scale multi-sensor datasets. In *11th European Conference on Computer Vision (ECCV)* (Greece, 2010).

[58] WINTER, S., AND WU, Y. The "spatial Turing test". In *Colloquium for Andrew Frank's 60th Birthday Geoinfo Series, Vienna, Austria, Department for Geoinformation and Cartography* (2008), Technical University Vienna, pp. 109–116.

[59] WITTE, T. H., AND WILSON, A. M. Accuracy of non-differential GPS for the determination of speed over ground. *Journal of Biomechanics 37*, 12 (2004), 1891–1898. doi:10.1016/j.jbiomech.2004.0.

[60] YI-LEH, W., CHENG-YUAN, T., MAW-KAE, H., AND CHI-TSUNG, L. Automatic image interpolation using homography. *EURASIP Journal on Advances in Signal Processing 2010* (2010), 1–12. doi:10.1155/2010/307546.