High-Level Structure Extraction from a Single Image

Juan Antonio de Jesús Osuna Coutiño, José Martínez Carranza

Technical Report No. CCC-17-004
July 07, 2017

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Abstract

High-Level Structures (HLS) extraction consist in recognizing/extracting 3D elements from images. There are several approaches for HLS extraction and several algorithms have been proposed. Most previous work processing two or more camera views or processing 3D data in the form of point clouds. In general, two camera views/3D point cloud approaches have good performance for certain scenes with video sequences or image sequences, but they need sufficient parallax or use thresholds in order to guarantee accuracy. Other approach, more promising in terms of scope and flexibility, is HLS from a single image. Unlike the other trends (using two views or using 3D point clouds) this approach extracts HLS without parallax constraint and without thresholds. This is useful due to in real world applications several data are limited to a single view from an unknown scene, for example, internet images, personal pictures and so on. For HLS extraction based on a single view, there is an important limitation since only planes are extracted. In practice, this limits the three-dimensional/geometrical understanding since other HLS such as spheres, cylinders and cubes provide richer 3D information than only planar structures.

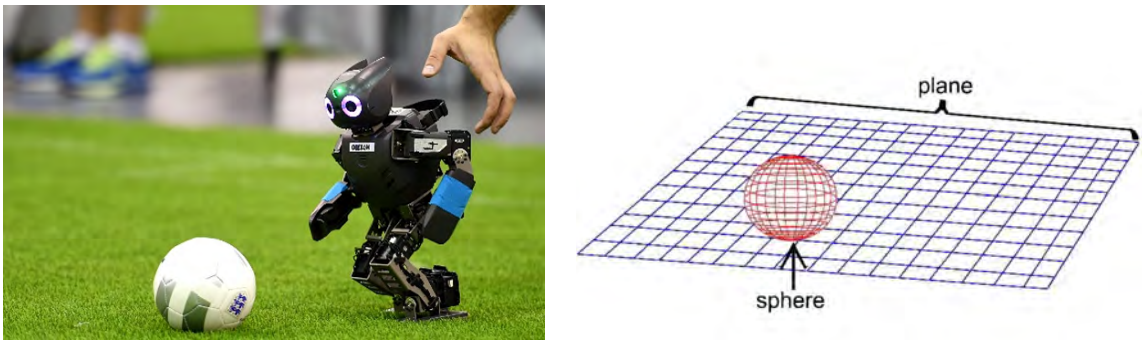
In this thesis work, we are interested in extracting HSL such as spheres, cylinders and cubes. We believe that this would be useful because a methodology of more diverse structures (spheres, cylinders, cubes, etc.) would provide more rich scene information than previous work. In addition, 3D structures such as spheres, cylinders and cubes would increase the performance in several real-world applications where HLS are used such as navigation, augmented reality, 3D model, etc. We will propose the use color, gradient and texture features as input in a learning algorithm, it will deliver a geometric classification of the scene. This classification will be used as a reference for HLS (spheres, cylinders and cubes) extraction. As work in progress, we have been proposed a new texture feature based on binary patterns which provide discriminant values for HLS recognition, a dataset of urbanized environments with light intensities variations, a method to obtain dominant structures orientation and an augmented reality application using planar structure recognition.

keywords: High-Level Structures, HLS Extraction, Single Image, 3D model, Urbanized Scenes.

Contents

Chapter 1
Introduction

In computer vision, High-Level Structures (HLS) extraction consist in recognizing/extracting 3D elements from images. There are several HLS that can be extracted (lines, planes, spheres, cylinders, cubes, etc.) and several approaches for HLS extraction have been proposed. In general, the use of HLS provides rich scene information since in man-made scenes (urbanized environments) there exist abundant HLS. In addition, HLS reduces computational processing by covering large areas with a few parameters. Due to this characteristics (rich scene information and computational processing reduction), several tasks, for example: robotics (1), augmented reality (2), navigation (3), 3D reconstruction (4) and Simultaneous Localization and Mapping (SLAM) (5), use HLS in order to performance improvements, as mathematical simplification and sped up. In **Fig.** 1.1, an example of HLS extraction is shown.



a) Input image b) 3D High-Level Structures

Fig. 1.1: High-Level Structures extraction. The 3D sphere in image (b) correspond to ball in the input image (a), while the blue plane in image (b) correspond to the scene floor (a).

There exist several approaches for HLS extraction: the first, analyzes two or more images captured from different camera views (6), (7). This approach has high performance under image sequences (collections of images related by time, such as frames in a movie or magnetic resonance imaging), unfortunately, it is necessary sufficient parallax[1], i.e., some difference (as shown in **Fig.** 1.2) between camera views in order to reach accurate results.

Other approach associates HLS with a 3D point cloud (8), (9), these methods rely on fitting algorithms, typically RANSAC and some optimization technique in order to fit HLS within 3D point clouds. Nevertheless, several thresholds and specifically set up are required in order to guarantee high performance for a specific scene. This is an important limitation because is several cases it is difficult to set appropriate thresholds set up values.

Other approach, and which we are interested in this research is the extraction of HLS from single image (10). Unlike the other trends (using two views or using 3D point clouds)

[1]Parallax is defined as the angle obtained by the objects displacement from an image sequence. i.e., closer objects have a larger displacement between images, while distant objects have small displacements, see **Fig.** 1.2.

this approach extracts HLS without parallax constraint. This is useful because in real world applications several data are limited to a single view from an unknown scene, for example historical images, internet images, personal pictures, holiday photos and so on. So, in current work, HLS from single image represent a promising solution with high performance, however, there are several challenges because there is insufficient information recorded in an image, i.e., there is not depth information from the image pixels or parallax information.

In recent work (11; 12), important progress in 3D structure interpretation have been made. This was achieved via learning algorithms that learn the relationship between visual appearance and scene structure. Motivated by the results of such techniques, and the potential benefits that single-image perception provides (HLS extraction without parallax constrains, extraction without threshold values), this thesis proposal focuses on 3D reconstruction from a single image. We believe this is a very interesting task, since despite the considerable challenges involved, some kinds of single image structure interpretation do indeed seem to be possible.

a) View 1 b) View 2

Fig. 1.2: Image parallelism. In view 2 the sewer cover has a larger displacement in axis x (28 pixels) that the flower pot (9 pixels).

1.1 Justification

Previous works demonstrated that HLS extraction deliver rich 3D information because in urbanized environments there exist abundant HLS. In addition, HLS extraction methods from a single view unlike the other trends (using two views or 3D point clouds) allow the HLS extraction without parallax constraint and thresholds in order to guarantee high performance for a specific scene. These characteristics (HLS extraction without parallax and without threshold values) are useful due to in real world applications several data are limited to a single view of an unknown scene, for example, historical images, internet images, personal pictures, holiday photos and so on. Unfortunately, there is an important limitation since HLS such as spheres, cylinders and cubes, that would deliver rich scene information

often not been extracted. In most previous work that uses a single image, only planes are extracted.

In this thesis proposal, we are interested in extract HLS that in previous work not extract. We believe that this would be useful because a methodology of more diverse structures (spheres, cylinders, cubes, etc.) would provide better 3D scene information compared with previous work. In addition, 3D structures such as spheres, cylinders and cubes would increase the performance in several real-world applications where HLS are used such as navigation, augmented reality, 3D models, etc. Finally, this method will be easy to replicate and use in others research (open several research lines and applications) because only will use as input device an RGB camera facilitating its implementation in personal devices as cell phones, personal assistants, personal computers, among other.

1.2 Problem statement

There are several single view HLS extraction approaches, unfortunately, most previous works are limited to plane 3D structure extraction. In practice, this is an important limitation since other 3D structures such as spheres, cylinders and cubes would deliver more rich scene information. This is a hard challenge because there is insufficient information recorded in an image, i.e., there is not depth information from the image pixels or parallax information.

1.3 Research questions

1.- **For a single view processing, which features allow us to obtain visual information for the recognition of 3D orientation of structures such as planes, spheres, cylinders and cubes?**

First, we will explore the visual features used in previous work (texture, gradient and color), this because they presented promising results for depth extraction and planar structures orientation, from a single image. Then, we want to extend the use of these visual features (texture, gradient, color, etc.) to obtain the orientation of spherical, cylindrical and cubic structures.

2.- **For a single view processing, what methodology allow us to extract 3D structures such as planes, spheres, cylinders and cubes?**

Previous work has shown that the use of geometric classification and a horizon estimation are sufficient to provide an automatic single-view reconstruction. For that reason, we will design a method that model the relationship between the visual information (texture, gradient and color) of a patch and key elements orientation (buildings, objects, street and grass) to obtain a geometric classification of the scene (3D orientation). Finally, to extract 3D structures we will use this geometric classification and a horizon estimation detection.

1.4 Hypothesis

Previous work has shown that visual features (texture, gradient and color) as inputs for a learning algorithm have promising results for planar structures extraction using a single image. Therefore, it should be possible to extend the use of these visual features (texture, gradient, color, etc.) as inputs for a learning algorithm to obtain 3D information that allow extraction of spherical, cylindrical and cubic structures using a single image.

1.5 Main objective

To develop a high-level structures extraction method, which provides 3D structures such as planes, spheres, cylinders and cubes from a single image under urbanized outdoor scenes.

1.5.1 Specific objectives

1.- To investigate the visual features (texture, gradient, color, etc.) that provide information of HLS extraction (planes, spheres, cylinders and cubes)

2.- To investigate the develop new visual features that provide information of HLS extraction (planes, spheres, cylinders and cubes)

3.- To design a single image processing methodology to provide 3D orientation of urbanized images

4.- To develop a 3D reconstruction method using urbanized images labeling and 3D orientation

1.6 Contributions

This research implies the exploration/development of visual features and to develop a method to deliver a scene geometric classification (3D orientation of scene elements). Both contributions, could be a promising tool under several computer applications such as augmented reality, navigation, 3D reconstruction, SLAM, among other applications. On the other hand, the proposed HLS extraction method (using a single image) has to improve the current state of the art since it will extract more elaborated HLS (spheres, cylinders and cubes) than plane structures, it is an important contribution in the area of 3D reconstruction from a single image.

1.7 Publications

At this moment, we have two articles accepted in international conferences, and another one in a journal (in process to be of submitted):

1.- *Osuna-Coutiño J. A. J., Martinez-Carranza J., Arias-Estrada M., Mayol-Cuevas W., (2016).* ***Plane Recognition in Interior Scenes from a Single Image.*** *IEEE International Conference on Pattern Recognition (ICPR), (pp. 1924-1929)*: in this

manuscript, we present a new dominant plane recognition method from a single image that provides five 3D orientation (right, left, front, top and bottom). This method combines three key elements (learning algorithm, contour detection method and segmentation technique) to obtain structures planar recognition and 3D orientation.

2.- *Osuna-Coutiño J. A. J., Cruz-Martínez C., Martinez-Carranza J., Arias-Estrada M., Mayol-Cuevas W., (2016).* ***I want to change my floor: dominant plane recognition from a single image to augment the scene.*** *IEEE International Symposium on Mixed and Augmented Reality Adjunct Proceedings (ISMAR), (pp. 135-140)*: in the second manuscript, we present a floor recognition method using a single view. We have applied our method in an augmented application. In order to infer the floor light intensities variations, we proposed a rule system that integrates three variables: texture features, blurring and superpixels-based segmentation.

3.- *Osuna-Coutiño J. A. J., Martinez-Carranza J., Mayol-Cuevas W., (2017).* ***A method to high-level structures recognition from a single image, with augmented reality.*** *tentative journal: Virtual Reality – Springer*: in the third manuscript, we will present a floor recognition method using a single image, suitable for indoor/outdoor augmented reality applications. In order to find the relation between image features (texture and color) and the floor, our method uses a supervised learning algorithm with regularized logistic regression. To validate our learning algorithm, we will use five different floor type (grass, road, smooth carpet, tile and squared carpet) with light intensities variations (we captured images to different times 9:00 am, 1:00 pm and 5:00 pm.). In addition, we will introduce a new texture feature based on binary pattern. This feature provides discriminant values for HLS recognition.

In future, we will expect to present a new HLS extraction method using a single view, key elements labeling and 3D orientation. Also, we will expect to present an augmented reality application. For both cases, we set as tentative journal/conferences: ISMAR, CVPR and Computer Vision and Image Understanding.

1.8 Research visit

Research visit at University of Bristol to the group of Prof. Walterio W. Mayol-Cuevas (01/October/2016 to 30/November/2016), funded by the Royal Society-Newton Advanced Fellowship with reference number NA140454. In this research visit, we develop a new dataset and a new texture feature. This dataset is integrated of urbanized images with light intensities variations and different scene perspectives (more details see section 5.1). The new texture feature is based on binary patterns which provide discriminant values for HLS recognition (more details see section 5.2). Using the result obtained along the visit, one article was writhing (in process to be of submitted), this article presents a floor recognition method using a single image, suitable for indoor/outdoor augmented reality applications.

1.9 Organization of the document

In order to describe our approach in more detail, this thesis proposal has been organized as follows. In chapters 2 and 3, the theoretical basis is presented, with the fundamental concepts for the development of this work and the previous works that determine the location of the research and the comparison with the results that would be obtained; each step in our methodology is discussed in chapter 4; chapter 5 describes the preliminary experiments designed to evaluate the feasibility of the proposed solution and the results achieved; finally, the conclusions and future work are indicated in the last chapter.

Chapter 2
Theoretical basis

2.1 Features

Detection of image features is an important task in computer vision because it allows abstractions of image information. Two types of features can be extracted from an image. Global features describe the image as a whole; they can be interpreted as a particular property of the image. On the other hand, local features aim to detect key points/feature points within the image. In most of the case, HLS extraction from a single image uses global features to find the relation between the image features and its HLS. In the following subsections, a description of the global features most used in HLS extraction is presented.

2.1.1 Texture

In computer vision, image texture is a set of metrics to quantify the color variations within a continuous surface.

Local binary patterns

In computer vision, Local Binary Patterns (LBP) (13) is a visual descriptor type used for texture classification. LBP were used as first steps to texture classification within circular pixel neighbors. The LBP provides the texture information in a patch of the image $I(x, y)$ applying **Eq.** (2.1.1 - 2.1.4). Where, ς is the neighbor pixel number, τ is the radius, 2^p is a binomial factor, v_c is the central pixel value in grayscale and $v_{i,j}$ are the neighbor pixel values in grayscale.

$$LBP(\varsigma, \tau) = \sum_{p=0}^{\varsigma-1} S(v_c - v_{i,j})2^p \tag{2.1.1}$$

The pixel distribution within LBP circle is shown below:

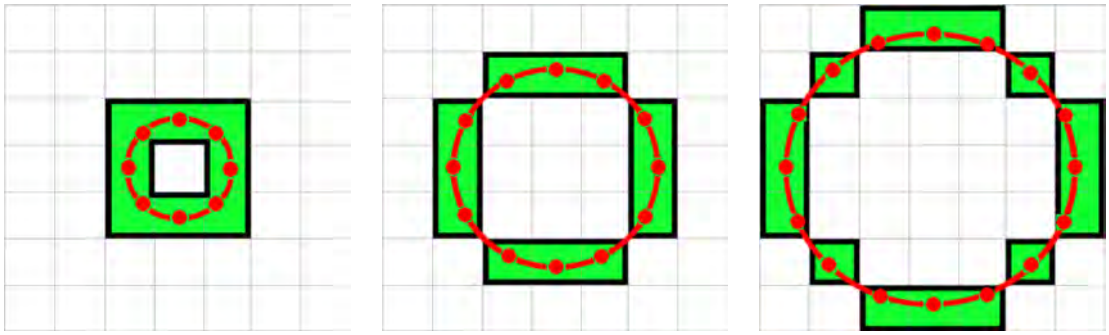
$$i = \tau \sin \frac{2\pi p}{\varsigma} \tag{2.1.2}$$

$$j = \tau \cos \frac{2\pi p}{\varsigma} \tag{2.1.3}$$

The binary values are obtained using the **Eq.** (2.1.4).

$$S(v_c - v_{i,j}) = \begin{cases} 1 & \texttt{if} \quad v_c - v_{i,j} \geq 0, \\ 0 & otherwise, \end{cases} \tag{2.1.4}$$

Fig. 2.1 shows an example of LBP circles, where each red circle corresponds to the neighbor pixel position within LBP circles, each red ring is one LBP circle, the green squares are the pixels of the LBP circles and black lines corresponding to LBP circles limit.



a) $\tau = 1, \varsigma = 8$ b) $\tau = 2, \varsigma = 12$ c) $\tau = 3, \varsigma = 16$

Fig. 2.1: The images in (a), (b) and (c) show the different radius (τ) and pixel number (ς) in LBP circles.

Co-occurrence matrix

A co-occurrence matrix (14) is a matrix which presents the distribution of co-occurring pixel values (grayscale values, or colors) in an image. In order to compute the co-occurrence matrix texture is necessary to determine five texture features in a patch, these features measure energy, entropy, contrast, homogeneity, correlation **Eq.** (2.1.5 - 2.1.9). Where, the numbers of normalized co-occurrence matrix are denoted by $c_{i,j}$, the averages of $c_{i,j}$ are denoted by μ_i, μ_j and the standard deviations of $c_{i,j}$ are denoted by σ_i, σ_j.

$$energy = \sum_{i=0}^{n}\sum_{j=0}^{n} c_{i,j}^2 \tag{2.1.5}$$

$$entropy = \sum_{i=0}^{n}\sum_{j=0}^{n} c_{i,j}(log\ c_{i,j}) \tag{2.1.6}$$

$$contrast = \sum_{i=0}^{n}\ \sum_{j=0}^{n} c_{i,j}(i-j)^2 \tag{2.1.7}$$

$$homogeneity = \sum_{i=0}^{n}\ \sum_{j=0}^{n} \frac{c_{i,j}}{1+(i-j)^2} \tag{2.1.8}$$

$$correlation = \sum_{i=0}^{n}\ \sum_{j=0}^{n} \frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i\sigma_j)}} \tag{2.1.9}$$

Law's texture energy measures

Law's texture energy measures (15) are a texture feature based on image filtering and identification of high energy points. Some of the vectors used to obtain of Laws masks are shown below:

$$L_3 = \left[\ 1,\quad 2,\quad 1\ \right]$$

$$E_3 = \begin{bmatrix} -1, & 0, & 1 \end{bmatrix}$$

$$S_3 = \begin{bmatrix} -1, & 2, & -1 \end{bmatrix}$$

From these vectors can be generated 9 different convolution masks by means of the multiplication of a vertical vector with a horizontal vector. The list of all 33 matrix is:L_3L_3, L_3E_3, L_3S_3, E_3L_3, E_3E_3, E_3S_3, S_3L_3, S_3E_3 and S_3S_3. A convolution matrix example is shown below.

$$L_3S_3 = L_3^T S_3$$

$$L_3S_3 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} -1 & 2 & -1 \end{bmatrix}$$

$$L_3S_3 = \begin{bmatrix} -1 & 2 & -1 \\ -2 & 4 & -2 \\ -1 & 2 & -1 \end{bmatrix}$$

2.1.2 Gradient

An image gradient is a change in the intensity of color in an image. In computer vision, gradient is used for identify a gradual change of color which can be considered as an even graduation from low to high values.

Nevatia-Batu gradient

Nevatia-Batu (16) is an edge detection feature in which a set of 5×5 masks are utilized to detect the edges in 30^o increments. Larger template mask would provide both a finer quantization of the edge orientation angle and a greater noise immunity, but the computational requirements increase. Examples of Nevatia-Batu masks are shown in the following

matrices.

$$0^o = \begin{bmatrix} 100 & 100 & 0 & -100 & -100 \\ 100 & 100 & 0 & -100 & -100 \\ 100 & 100 & 0 & -100 & -100 \\ 100 & 100 & 0 & -100 & -100 \\ 100 & 100 & 0 & -100 & -100 \end{bmatrix} \frac{1}{1000} \qquad 30^o = \begin{bmatrix} 100 & -32 & -100 & -100 & -100 \\ 100 & 78 & -92 & -100 & -100 \\ 100 & 100 & 0 & -100 & -100 \\ 100 & 100 & 92 & -78 & -100 \\ 100 & 100 & 100 & 32 & -100 \end{bmatrix} \frac{1}{1102}$$

Histogram of oriented gradients

The Histogram of Oriented Gradients (HOG) (17), is a feature descriptor used in computer vision and image processing to obtain the orientation each pixel. To build the HOG (17), the local orientation at each pixel is obtained by convolving the image with the mask $[1, 0, 1]$ and $[1, 0, 1]^T$ in the x and y directions separately, to approximate the first derivatives of the image. This gives the gradient values G_x and G_y, for the horizontal and vertical directions respectively, which can be used to obtain the angle **Eq.** (2.1.17) and magnitude m of the local gradient orientation **Eq.** (2.1.18).

$$\theta = tan^1 \frac{G_y}{G_x} \tag{2.1.17}$$

$$m = \sqrt{G_x^2 + G_x^2} \tag{2.1.18}$$

2.1.3 Color

A color space is an arbitrary agreed upon way to define color. There is any number of ways to visualize color. Each color space has its different advantages and disadvantages.

YCbCr color space

YCbCr (18) is a color space used as a part of the color image pipeline in video and digital photography systems, where Y is the intensity channel, and Cb and Cr are the blue-difference and red-difference respectively. YCbCr from R,G,B pixels is derived as follows:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.229000 & 0.5870 & 0.114000 \\ 0.168736 & 0.331264 & 0.500000 \\ 0.500000 & 0.418688 & 0.081312 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + 128$$

Fig. 2.2 shows an example of YCbCr color space. Where the black box represents the process for obtaining Y, Cb and Cr components from R,G,B pixels.

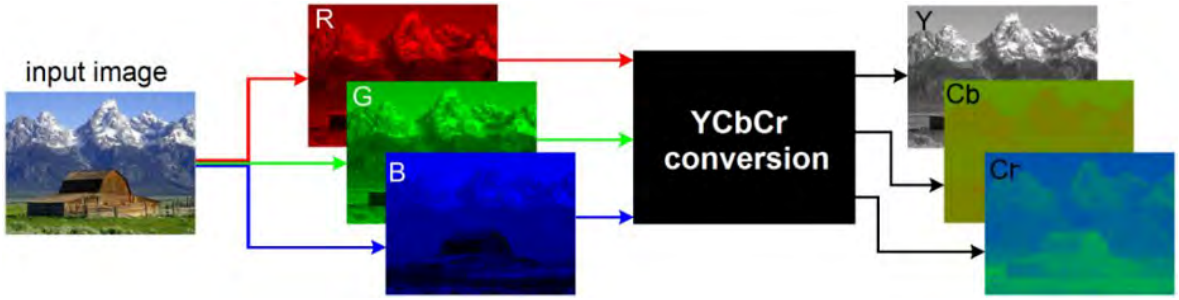


Fig. 2.2: The conversion of YCbCr color space is show in its Y, Cb and Cr components.

2.2 Learning algorithms

In previous work was demonstrated that learning algorithms are a useful tool for images interpretation. For computer vision, the use of learning algorithms has shown an important progress to learn the relationship between visual appearance and the scene structure. In most of the case, the HLS extraction from a single image uses learning algorithms to find the relation between the visual features and HLS. For this reason, in the following subsections are described the learning algorithms most used in the related works.

2.2.1 Regularized logistic regression

Regularized logistic regression (19) is a regression model where the dependent variable is binary, i.e., where it can take only two values, "0" and "1". The logistic regression hypothesis used to predict dependent variable is presented in **Eq.** (2.2.1). Where, the logistic regression classifier $h^i_\theta(x)$ for find the probability that y is equal to the classes i, i.e., $h^i_\theta(x) = P(y = i|x;\theta)$.

$$h^i_\theta(x) = g(\theta^T_j x_j) \tag{2.2.1}$$

The element θ_j is a parameter adjusted of the logistic regression, the elements x_j are the features. Where, the sigmoid function or logistic function g is expressed as $g(c) = \frac{1}{1+e^{-c}}$. The logistic regression hypothesis is defined as **Eq.** (2.2.2).

$$h^i_\theta(x) = \frac{1}{1 + e^{-\theta^T_j x_j}} \tag{2.2.2}$$

2.2.2 Markov random field

Markov random field (MRF) (20) is a set of random variables with a Markov property described by an undirected graph. Any Markov random field (with a strictly positive density) can be written as a log-linear model with feature functions f_k such that the full joint

distribution can be written as:

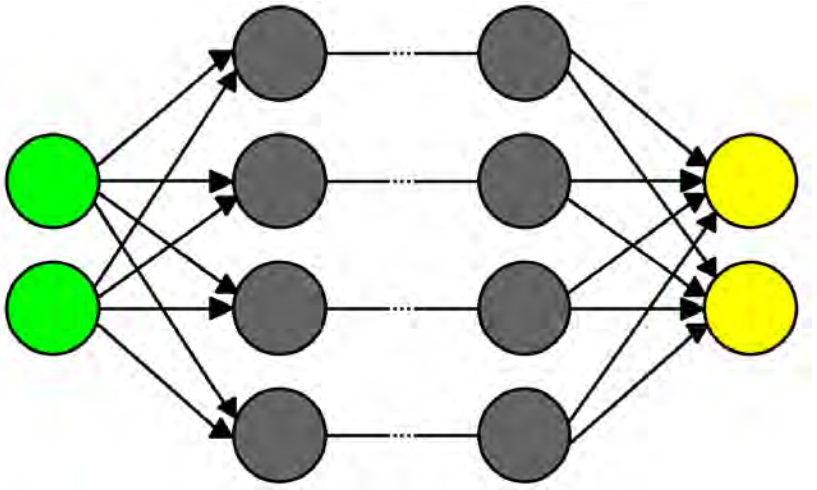
$$P(X = x) = \frac{1}{z} exp(\sum_{k}^{m} w_k^T f_k(x_k)) \tag{2.2.3}$$

The element f_k is the feature functions, z is the normalization constant for the model, m is the total number of elements to analyze, w_k is the real value feature vector of elements to analyze and X denotes the set of all possible assignments of values to all the network's random variables x.

2.2.3 Artificial Neural Networks

The Artificial Neural Networks (ANN) are distributed parallel processing structures, i.e., these parallel processing structures can do more than one task at the same time. In ANN the concept is similar the human brain that has the neuron as the elemental unit. The objective of an artificial neuron is learning from experience how a biological neuron (21). The artificial neuron operation involves the evaluation of a function from the input data and the transfer function compute. In addition, ANN synapse is the connection between two artificial neurons, this permits to share information with other neurons and establish a communication system. Each synapse has a component called weight, which is adjustable during the network training (22).

In general, an ANN architecture is structured by the input layer (the examples number given to training the ANN), one or more hidden layers, and an output layer (outputs number correspond to the classes number). **Fig.** 2.3 shows a neural network architecture example.



a) input layer b) n hidden layers c) output layer.

Fig. 2.3: Artificial Neural Networks architecture.

2.3 Shape From X

Shape From X (SFX) is a collection of methods to infers from image cues the 3D shape of one surface. These methods using features such as shading, texture, defocus, zoom, etc.

2.3.1 Shape From Shading

The Shape From Shading (SFS) (23) problem computes the 3D shape of a surface from the brightness of one black and white image. Unlike of other 3D methods (stereo, structure from motion, SLAM and photometric stereo, etc.), in the SFS problem uses a single image. The brightness equation is defined as **Eq.** (2.3.1). Where, I is the brightness image, (x_1, x_2) are the coordinates of a point x in the image I, R is the reflectance map, $L(x)$ is the light vector and $n(x)$ is the normal vector.

$$I(x_1, x_2) = R(n(x_1, x_2)) \tag{2.3.1}$$

$$R = cos(L, n) = \frac{L}{|L|}.\frac{n}{|n|} \tag{2.3.2}$$

2.3.2 Shape From Texture

Shape From Texture (SFT) (24) is a computer vision technique where a 3D object is reconstructed from the texture of one single image. SFT uses the term texture gradient in order to denote the areas that have similar texture. In order to obtain the 3D orientation of one texture gradient, it is necessary to find the tilt angle. In order to find tilt angle is use the **Eq.** (2.3.3), where D is a diagonal matrix, U gives the tilt direction of T and V are orthogonal matrixes.

$$T_{f>i} = UDU^1(UV^1) \tag{2.3.3}$$

Chapter 3
Related works

3.1 Depth perception from a single image

Saxena et al. (25) presented the first steps to depth estimation from a single image using a supervised learning algorithm with Markov Random Field (MRF). In order to obtain the feature vector, this work divides the image into small rectangular patches, and estimate a single depth value for each patch. For that, it is proposed the use two features types. The first are absolute depth features (to estimate the absolute depth at a particular patch) using texture variations, texture gradients, and color. The second are relative depth features (magnitude of the difference in depth between two or more patches). Finally, this method to learn depth information is based on the relationship between image features (absolute features and relative features) and image depth. The image depth was acquired using a custom built laser scanner unit.

Saxena et al. (26) apply an MRF learning algorithm to capture some monocular cues, and incorporate them into a stereo system. This method to obtain the monocular cues uses an MRF to model the relation between the depth information of a patch and the depth of its neighboring patches. In this case, the depth of a particular patch depends on the features of the patch, but is also related to the depths of other parts of the image. For example, the depths of two adjacent patches lying in the same building will be highly correlated. This work shows that by adding monocular cues to stereo, they obtain significantly more accurate depth estimations than using monocular or stereo cues alone.

The model presented by Saxena et al. (27) used a supervised learning algorithm with MRF to find the relation between the image features and its depth. It incorporates them into a stereo system (method mentioned above), but increases the experiments in exterior and interior scenes. In addition, a simplified version of the algorithm was used to guide an autonomous vehicle over unknown terrain.

In Liu et al. (28), an algorithm for depth estimation was presented, unlike other approaches, they use scene segmentation and the semantic labels to obtain depth information. This algorithm works in two steps. The first step predicts the semantic class of each pixel (sky, tree, road, grass, water, building and so on) and the location of the horizon using MRF. The second step estimates depth. It was demonstrated that knowing the semantic class of each pixel, depth and geometry constraints can be easily obtained (e.g., "sky" is far away and "ground" is horizontal).

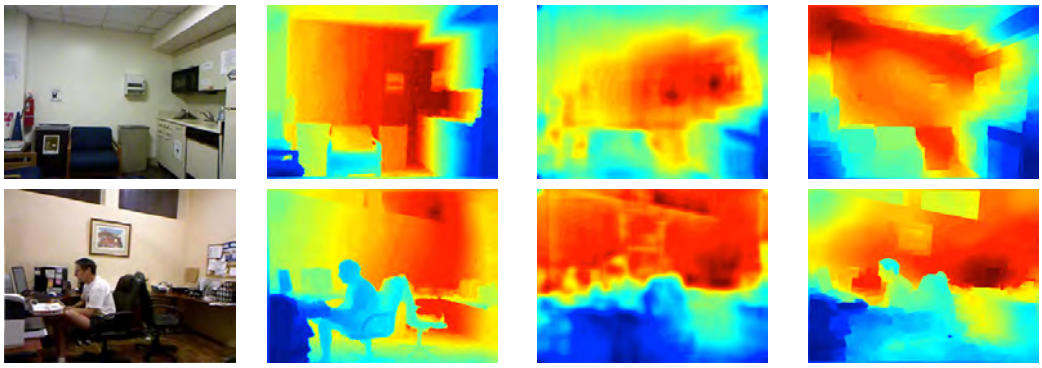
Karsch et al. (29) used a depth transfer approach that has three steps. First, given a database RGBD images, this method finds "candidate images" that are similar to the input image in RGB space. Then, the RGB and depth images (candidate images) are aligned with the input image. Finally, an optimization procedure is used to interpolate and smooth the depth values of candidate images, to obtain the depth in the input image. While this is a very interesting approach to recovering depth, this method needs building a database of reasonably sized with similar images of the input image.

Mota-Gutierrez et al. (30) provide a fast approach for monocular SLAM initialization by constructing an initial 3-D map with interest points that are susceptible to be tracked. This

work uses a depth learning algorithm based on regularized linear regression over interest points with FAST algorithm. In addition, this work used a supervised learning algorithm to find the relation between the characteristics (color and texture) of an image and its depth information. Eigen et al. (31) presented a depth estimation method from a single image through training a neural network. To obtain the depth information a neural network with two components is used: the first component estimates the absolute depth of the scene and the second refines the local depth. Both neural network components are applied to the original input, but in addition, the output of the first component is passed to the second component as an additional feature. In this way, the neural network uses the global depth prediction for feedback of local depth.

Liu et al. (32) propose a discrete-continuous Conditional Random Field (CRF) model to take into consideration the relations between adjacent superpixels and depth. More specifically, this method formulates depth estimation as a discrete-continuous optimization problem, where the continuous variables (centroid and plane normal of superpixel) encode the superpixels depth in the input image, and the discrete ones represent relationships between neighboring superpixels. Zhuo et al. (33) propose an approach to explore the global structure and the hierarchical representation of scenes for depth inference in interior images using CRF and a superpixel analysis for region extraction. To estimate depth, each superpixel is represented as a 3D plane and the method looks for the best depth parameters for each superpixel. In addition, the superpixel edges encode the depth interactions within and across the different image sessions.

One of the most current and accurate methods for estimating depth from a single image is the proposed by Liu et al. (34), this method uses a learning model from Convolutional Neural Networks (CNN) for the identification of the images depth in urban environments. For that, first, the input image is segmented in superpixels. Each superpixel is cropped as a patch centered around its centroid, then resize and this patch is used in a CNN to obtain the depth information. These networks have been trained using datasets that provide both RGB images and corresponding depth maps such as NYUv2 (35) and KITTI (36) dataset. In **Fig.** 3.1 is shown depth maps obtained from a single image.

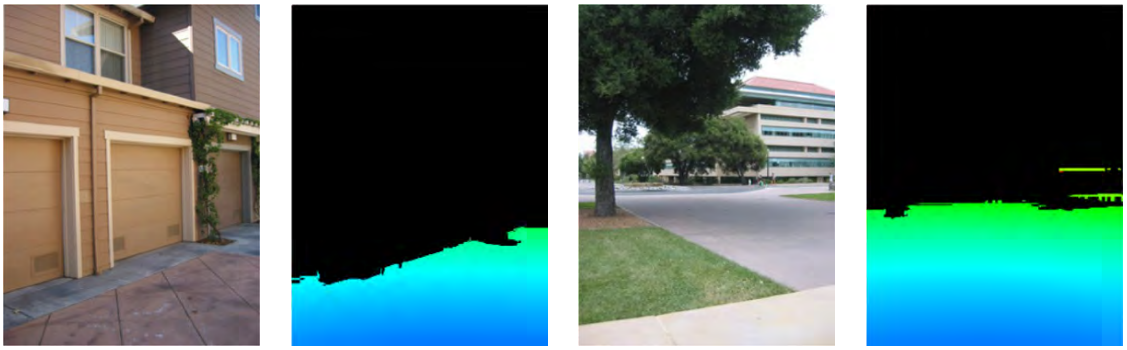


a) Input image b) Ground-truth c) Eigen et al. (2014) d) Liu et al. (2016)

Fig. 3.1: The images in (c) and (d) show the depth maps obtained from a single image, where the color indicates depths (red is far, blue is close). Images adapted from Liu et al. (2016) (34).

3.2 Related work: depth estimation for HLS extraction

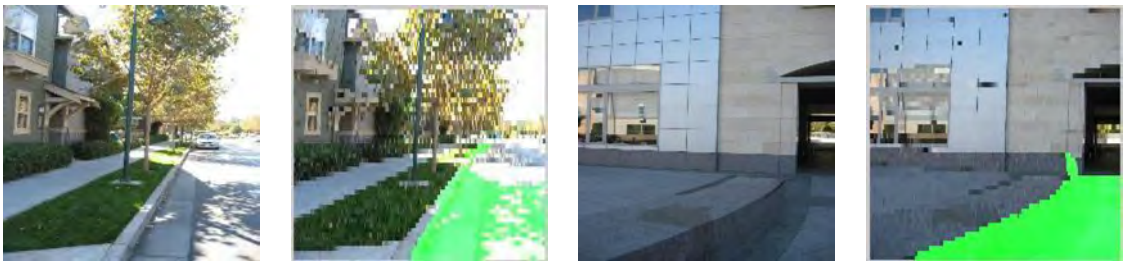
In previous work, there are several approaches for depth estimation (section 3.1), although depth estimation seems to be a different than HLS extraction. It is a promising baseline for HLS extraction. In current literature, several techniques for HLS extraction use depth estimation as keystone of their mathematical formulation Cherian et al. (37), Saxena et al. (11) and Rahimi et al. (38). In **Fig.** 3.2, as example of HLS extraction based on depth estimation is shown. Although depth estimation facilitates the planar structure extraction procedures, nevertheless, in some cases HLS extraction based on depth maps presented several challenges due to low sharpness in depth information (see **Fig.** 3.1), even in the extraction of plane structures.



a) Input image b)Planar structure c) Input image d) Planar structure

Fig. 3.2: The images in (b) and (d) show the planar structure recognition and estimated depth. Images adapted from Rahimi et al. (2013) (38).

In previous work, learning algorithms combined with depth estimation can be used to recognize high-level structures through single image analysis. The training could be performed using visual characteristics information as texture, gradient and color. However, in some cases the recognition accuracy depends on the training data quality.



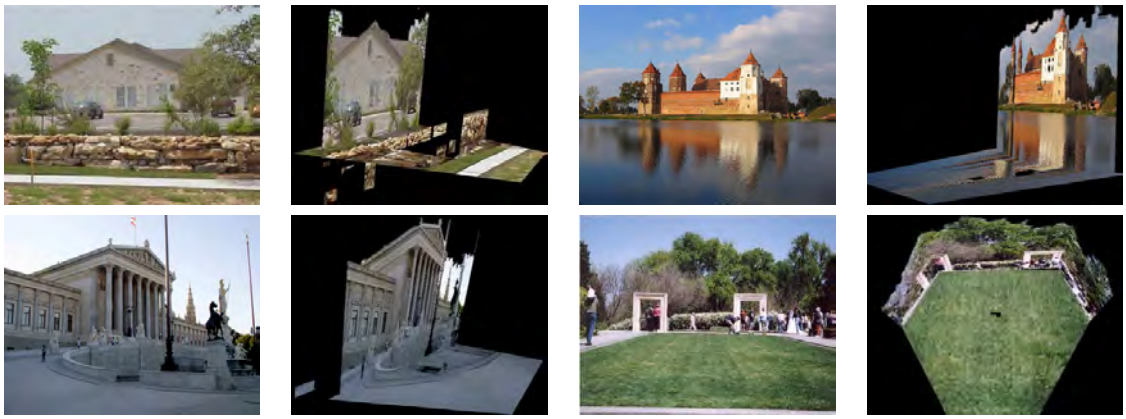
a) Input image b)Ground plane c) Input image d) Ground plane

Fig. 3.3: The images in (b) and (d) show the planar structure recognition. Images adapted from Cherian et al. (2009) (37).

Cherian et al. (37) introduce a methodology for estimating the ground plane structure and to estimate the 3D location of the landmarks from a robot using a single image.

This work used a supervised learning algorithm (MRF) to find the relation between image characteristics (texture and gradient) and its depth information. In addition, in order to differentiate the ground plane boundaries on the depth map, this method divide the original image into regions of similar textures using superpixels to feedback the depth map and locate the ground plane. In **Fig.** 3.3 an example is show, where it is presented a ground plane structure.

The method presented by Saxena et al. (11) estimates depth maps for single images of outdoor scenes for creating 3D models with planar structures. For that, this method segments the image into superpixels and compute three features (texture variations, texture gradients, and color). These features allow them to estimate both relative and absolute depth, as well as local orientation. In addition, for each superpixel and respective features, it uses an MRF to infer a set of "plane parameters" that capture both the 3D location and 3D orientation. In **Fig.** 3.4 is show an example, where is presents 3D models with planar structures. This approach has shown promising results, in images downloaded from the internet. However, it is limited 3D models with planar structures without providing information of other structures such as spheres, cylinders, etc.



a) Input images b)3D models c) Input images d) 3D models

Fig. 3.4: The images in (b) and (d) show the extraction of 3D models with planar structures. Images adapted from Saxena et al. (2009) (11).

Rahimi et al. (38), propose a method to estimate a ground plane structure and its depth information from a single static image. This methodology works in two steps. The first step estimates superpixel sections depth using a gradient boosting regression to take into consideration visual features relation (texture and gradient) with depth in the scene. In the second step, a RANSAC based plane estimator uses the superpixels depth information in order to fit with the planes in the scene. In **Fig.** 3.2 an example is show, it is presented the ground plane structure and its depth information.

Other approach is the use of a depth sensor to obtain HLS extraction. In the method presented by Firman et al. (39) it is proposed an algorithm to build a complete 3D model of a scene, given only a single depth image, i.e., they propose an algorithm that can complete the unobserved geometry using a prediction computed from the observed geometry. This is

possible because this method assumes that objects of dissimilar semantic classes often share similar 3D shape components. This work used a structured Random Forest to prediction the unobserved depth. In **Fig.** 3.5 a 3D model with the prediction of the unobserved depth is shown.

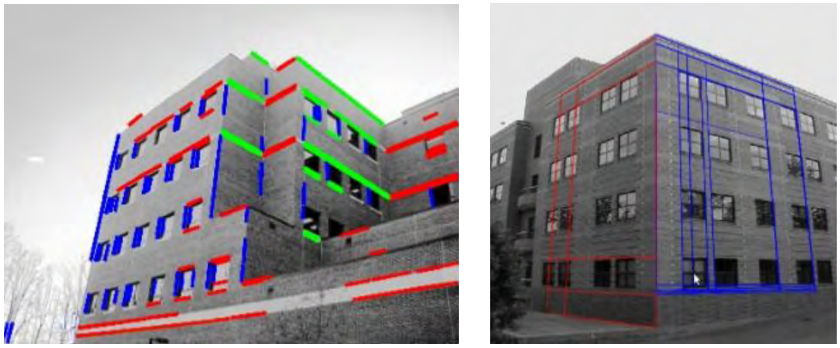
a) Input image b) 3D projection of the depth image c) Prediction the unobserved depth

Fig. 3.5: The image (a) is the Input image, image (b) show the 3D projection of the depth image and image (c) presents a 3D model with the prediction the unobserved depth. Images adapted from Firman et al. (2016) (39).

3.2.1 Single view HLS extraction without depth estimation

There are other approaches for HLS extraction without depth estimation. In those cases, the idea is to provide a more direct formulation. The approach presented by Kovsecká and Zhang (40), Micusik et al (41) and McClean et al. (42) show a methodology for extracting dominant planar structures by analyzing the pattern of the lines and vanishing points of an image. The method is based on the assumption that there are three orthogonal directions present in the scene. In **Fig.** 3.6 a) is shown an example, where it is presented lines colored by their assignment to the three directions of vanishing points. In addition, to find rectangular surfaces from this, two pairs of lines, corresponding to two different vanishing points, are used to localize planar structures. An example of planar structures detected is shown in **Fig.** 3.6 b). This approach has shown promising results, in both indoor and outdoor scenes, and the authors mention that it would be useful for robot navigation and Augmented Reality (AR) applications. Unfortunately, it is limited to scenes with this kind of planar structures. i.e., planes which are perpendicular to the ground, but are oriented differently from the rest of the planes, would not be easily detected.

The approach proposed by Hoiem et al. (43) and Hoiem et al. (10) interprets the geometric context from a single image using a learning algorithm. This geometric context is assigned to one of three main classes: ground, sky, and vertical, of which the latter is further subdivided into left, right, forward, porous and solid. Although this approach is not explicitly aimed at HLS detection, it has an understanding the general structure of scenes, as the image is partitioned into planar structures (ground, left, right, forward) and non-planar structures (sky, solid, porous). Some examples from their results are shown in **Fig.** 3.7. The classification of this approach is achieved using a large variety of features, including color (summary statistics and histograms), filter bank responses to represent texture, image



a) Three directions of vanishing points b) Planar structures detected image

Fig. 3.6: The images in (a) shows the lines colored by their assignment to the three vanishing points directions and image (b) presents planar structure recognition. Images adapted from Kovsecká et al. (2005) (40).

location, line intersections, shape information and vanishing point. These cues are used in the various steps of classification, using decision trees and logistic regression to select the geometric context. In addition, this approach can enable simple 3D reconstruction of a scene from a single image.

a) Input images b)Geometric context c) Input images d) Geometric context

Fig. 3.7: The images in (b) and (d) show geometric context from a single image: ground (green), sky (blue), vertical regions (red) subdivided into planar orientations (arrows) and non-planar solid ('x') and porous ('o'). Images adapted from Hoiem et al. (2007) (10).

Other approach, presented by Haines and Calway (44), Haines and Calway (12) it is the planar structures extraction and their orientation using a learning algorithm. For that, it selects a subset of salient points in the image around which to obtain features. In this approach it is obtained two features: the first is gradient orientation histograms, which consist of histograms of edge orientation. Second, the color using RGB histograms, created by histograms from the red, green and blue channels. To reduce the dimensionality of the distribution of features in an image region is uses a bag of words. Finally, a learning algorithm is used to take into consideration the relations between planar surfaces and their features (gradient orientation and color). An example of planar structures recognition is shown in **Fig.** 3.8. This approach has shown promising results, in outdoor scenes. However, it is limited to planar structures recognition without providing information of other structures such as spheres, cylinders, etc.

a) Input images b)Planar recognition c) Input images d) Planar recognition

Fig. 3.8: The images in (b) and (d) show the planar structures recognition and their 3D orientation predicted. Images adapted from Haines et al. (2015) (12).

3.2.2 Single view HLS extraction challenges and future trends

There are many HLS extraction approaches, however, one of the most promising approaches and which we are interested in this research is the extraction of HLS from a single image. For the case of single view exist three approaches. The first approach use learning algorithms to predicted depth information and optimization techniques in order to fit HLS on depth information. However, in some cases, HLS extraction based on depth maps present several challenges due to low sharpness in depth information (**Fig.** 3.1), even in the extraction of plane structures. The second approach if for algorithms without depth information such as: geometric recognition, vanishing points, planar recognition and among others. Although this approach has shown promising results, in indoor and outdoor scenes, in most of the case, it is limited to planar structures recognition without providing information of other structures such as spheres, cylinders, among other. The third approach uses a depth sensor (cameras RGB-D) to obtain HLS extraction. This approach is very interesting because recent work proposes an algorithm that can complete the unobserved geometry using a prediction computed from the observed geometry. Unfortunately, these sensors often deliver low stability under outdoor scenarios. In Table 3.1, we show a technical comparison between previous works of HLS extraction from a single image.

As can be seen, in Table 3.1 most previous work are limited to the recognition/extraction of plane structures. This is an important limitation since other 3D structures such as spheres, cylinders and cubes would deliver more rich scene information, i.e., could provide a rich 3D reconstruction of historical images, internet images, personal pictures, holiday photos and so on. An alternative to extract spheres, cylinders and cubes is the use of the depth sensor (cameras RGB-D. Unfortunately, these sensors often deliver low stability under outdoor scenarios. In addition, they are not present in personal devices (cell phones, personal assistants, personal computers, etc.). Finally, the power computation, cost, and size is higher than RGB sensors. So, an HSL extraction approach that extract lines, planes, spheres, cylinders, etc., and without depth sensor could achieve high performance (compact system design, low cost) for real world applications.

In HLS current work, there are several future trends such as: develop embedded devices (smart cameras) for HLS extraction, positioning localization from a single image, extract HLS as spheres, cylinders and cubes, among other. All these future trends are interesting and these will revolutionize the HLS extraction. In this thesis proposal, we are interested in the 3D structures extraction more rich that previous work (spherical, cylindrical and cubic)

Table 3.1: Previous works of HLS extraction from a single image

Reference	Result	Workspace	Approach	Classifier	Dataset
Hoiem D. et al. (2005) (43)	3D models (planes)	Outdoor	Geometric recognition	Logistic regression	Own images
Kovsecká J. & Zhang. W. (2005) (40)	Plane structures	Indoor and outdoor	Vanishing points	Normalized Cross Correlation (NCC)	Own images
Hoiem D. et al. (2007) (10)	3D models (planes)	Indoor and outdoor	Geometric recognition	Logistic regression	Own images and Stanford dataset
Micusik B. et al. (2008) (41)	Plane structures	Indoor and outdoor	Vanishing points	RANSAC	Own images
Cherian A. et al. (2009) (37)	3D models (planes)	Outdoor	Depth recognition	MRF and superpixels	Stanford Make3D Dataset
Saxena A. et al. (2009) (11)	3D models (planes)	Outdoor	Depth recognition	Markov Random Field (MRF)	Stanford Make3D Dataset
McClean E. et al. (2011) (42)	Plane structures	Outdoor	Vanishing points	RANSAC and Expectation Maximization (EM)	Zurich Building Dataset
Haines O. & Calway A. (2012) (44)	Plane structures	Outdoor	Planar recognition	k-Nearest Neighbors (K-NN)	Osian Haines Dataset
Rahimi A. et al. (2013) (40)	Planes and depth	Outdoor	Depth recognition	Gradient boosting regression and RANSAC	Stanford Make3D Dataset
Haines O. & Calway A. (2015) (12)	Plane structures	Outdoor	Planar recognition	Relevance Vector Machine (RVM)	Osian Haines Dataset
Firman M. et al. (2016) (39)	3D models	Indoor	Depth information (Kinect Fusion)	Structured Random Forest	NYU-Depth V2 datasets and own images
Proposed method	3D models (planes, spheres, cylinders and cubes)	Outdoor	HLS recognition	Markov Random Field (MRF)	KITTI, Make3D and proposed dataset

from a single RGB image. This is a hard challenge because there is insufficient information recorded in a single image, i.e., there is not a way of recovering depth information directly from the image pixels or parallax information (larger motion in the image for closer objects, in an image pair) to distinguish even relative depths. Nevertheless, recent work has shown important progress in 3D structure interpretation from a single image. This was achieved via learning algorithms that learn the relationship between visual appearance and scene structure.

Chapter 4
Methodology

4.1 Method

In following sub sections, we will present details about all steps of our methodology to high-level structures extraction from a single image.

4.1.1 Dataset

In this work, HLS extraction focuses on urbanized outdoor scenes. In real world applications, several data from a single view include urbanized outdoor scenes, for example, historical images, internet images, personal pictures, holiday photos and so on. Therefore, we will use dataset as "KITTI" (36) and "Make3D" (45) that have RGB images (of urbanized outdoor) and ground truth depth. Since the state of the art datasets do not have ground truth HLS, we will elaborate the ground truth HLS using depth information (HLS volume) and a manually HLS delimitation.

4.1.2 Key elements labeling

We will use a method that provides labeling of key elements in urbanized environments (buildings, street, grass, trees, water, etc.). Because if we locate key elements in the image is possible to provide a different strategy to obtain 3D orientation of each key element (see subsection 4.1.4). For example, the method presented by Domke (46) uses a CRF-based classifier to perform elements labeling in urbanized environments as buildings, objects, street, grass, water, trees, mountains and sky. In **Fig.** 4.1 b) an image with key elements labeling is shown.

4.1.3 Visual features

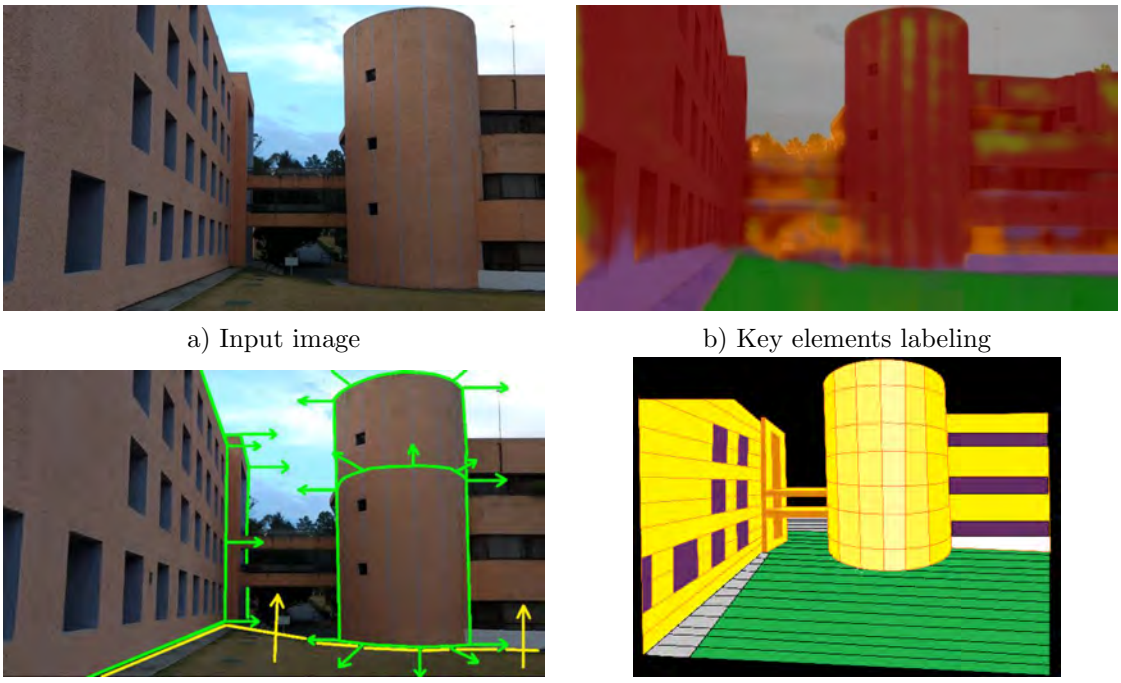
Previous work has shown that features such as texture, gradient and color have promising results for depth extraction and planar structures orientation. In this work, we want to extend the use of these visual features (texture, gradient, color, etc.) to obtain orientation of spherical, cylindrical and cubic structures.

4.1.4 Key elements orientation

We will use a learning algorithm to model the relationship between the visual information (texture, gradient, color, etc.) and its key element (buildings, street, grass, trees, water, etc.) to obtain the 3D orientation. In the learning algorithm, a different training to each key element have to be carried out. This is because each key element needs different strategies to obtain its orientation. For example, previous work has shown that the use of gradients facilitates the identification of buildings orientation (12; 44), while in the grass the use of gradients does not give good results. In **Fig.** 4.1 c) an image with key elements 3D orientation is shown.

4.1.5 HLS extraction

Previous work (10; 47) has shown that the use of geometric classification (3D orientation) and a horizon estimation are sufficient to provide an automatic single-view reconstruction of outdoor scenes. Considering the premise that is possible identifying the intersection between the floor and vertical elements (building, trees, mountain, etc.). So, identifying the floor-vertical elements intersection and geometric classification (3D orientation) in the image, we will able to present the vertical surfaces from the floor in 3D form. In **Fig.** 4.1 d) an image with HLS extraction or 3D model is shown.



a) Input image b) Key elements labeling

c) Key elements 3D orientation d) HLS extraction

Fig. 4.1: The image (a) is the Input image, image (b) show the key elements labeling (buildings in brown color, objects in orange color, street in purple color, grass in green color, trees in light green color and sky in white color) with the method presented by Domke (46), image (c) presents the key elements 3D orientation and image (d) show HLS extraction or 3D model.

4.1.6 Validation

For validation, we will compare results of proposed method with the ground truth of two databases. The first comparison will use the dataset with the ground truth HLS with depth information elaborated from the "KITTI" (36) and "Make3D" (45) dataset (subsection 4.1.1). In this comparison we will measured the 3D information computed by the proposed method compared with 3D information that is obtained by the ground truth depth ("KITTI" and "Make3D"). The second comparison we will be drawn up a dataset with ground truth of recognition HLS of a set of similar scenes to different schedules to identify

the robustness of the method to lighting changes. In the second comparison will be measured the HLS recognition computed by the proposed method compared with the ground truth HLS.

4.2 Work plan

In **Fig.** 4.2 the proposed work plan to obtain HLS extraction method is shown. The first year, we started the carried out the dataset collection, the visual feature exploration/development and the literature review (activity performed throughout the Ph.D.). In addition, a research visit was elaborated at the University of Bristol. In this research visit, we develop a new dataset and a new texture feature based on binary patterns. The second year, we will conclude the dataset collection, the visual feature exploration/development, the visual feature validation to provide discriminant values on urbanized outdoor scenes, labeling methods exploration and labeling validation on urbanized outdoor scenes. In addition, we will start to train a learning algorithm to model the relationship between the visual feature and structures 3D orientation. In the third year, we will validate 3D orientation algorithm on urbanized outdoor dataset. Furthermore, we will develop the HLS extraction method from a single image and validate our 3D reconstructions using ground truth depth "KITTI" (36) and "Make3D" (45) dataset. Finally, the fourth year, we will elaborate an AR application with HLS extraction, we will write the Ph.D. thesis and we will present the thesis dissertation.

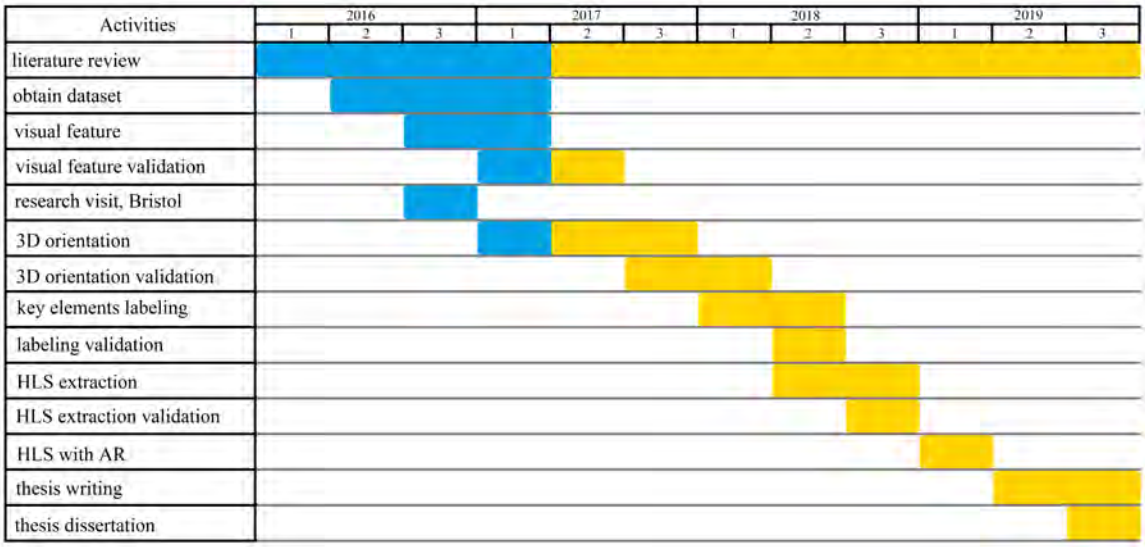


Fig. 4.2: Work plan.

Chapter 5

Preliminary results

In this section, we present preliminary results for 3D structures extraction. These results are integrated of a proposed dataset, a new texture feature based on binary patterns, a method to obtain dominant structures orientation and an Augmented Reality application using planar structure recognition.

5.1 Proposed dataset

To validate the HLS extraction features (texture, gradient, color, etc.), a dataset using urbanized environments with light intensities variations and different scene perspectives is proposed. This dataset is important because in HLS extraction is common to reconstruct elements under different lighting conditions and perspectives. This dataset consists on urbanized scenes with 1,500 images (720×1280 pixels), five different classes (grass, road, smooth carpet, tile and square carpet). This 1,500 images have floor labeled and light intensities variations (we captured images to different times 9:00 am, 1:00 pm and 5:00 pm). We chose the floor because in most cases it is an element that can be captured in large quantity in the image, it is easy to observe in different perspectives and it has similar texture (similar to most of urbanized structures). In **Fig.** 5.1 a) and **Fig.** 5.1 b) some images from the proposed dataset are shown.

There are several datasets that provide labeled for the image content. However, in most of the case these datasets are focus on a specific labeled such as road (36). This is a limitation because urbanized environments are composed of a scenarios variety. There are other datasets, more rich in terms of information about light intensities variations, but these are not from urbanized environments (50). Finally, exist other datasets, that provide large set for different scenes, unfortunately, they provide information for a single perspective (51). This is a limitation because HLS in urbanized environments are located in different perspectives of the scene.

Fig. 5.1: Images of proposed dataset, section a).

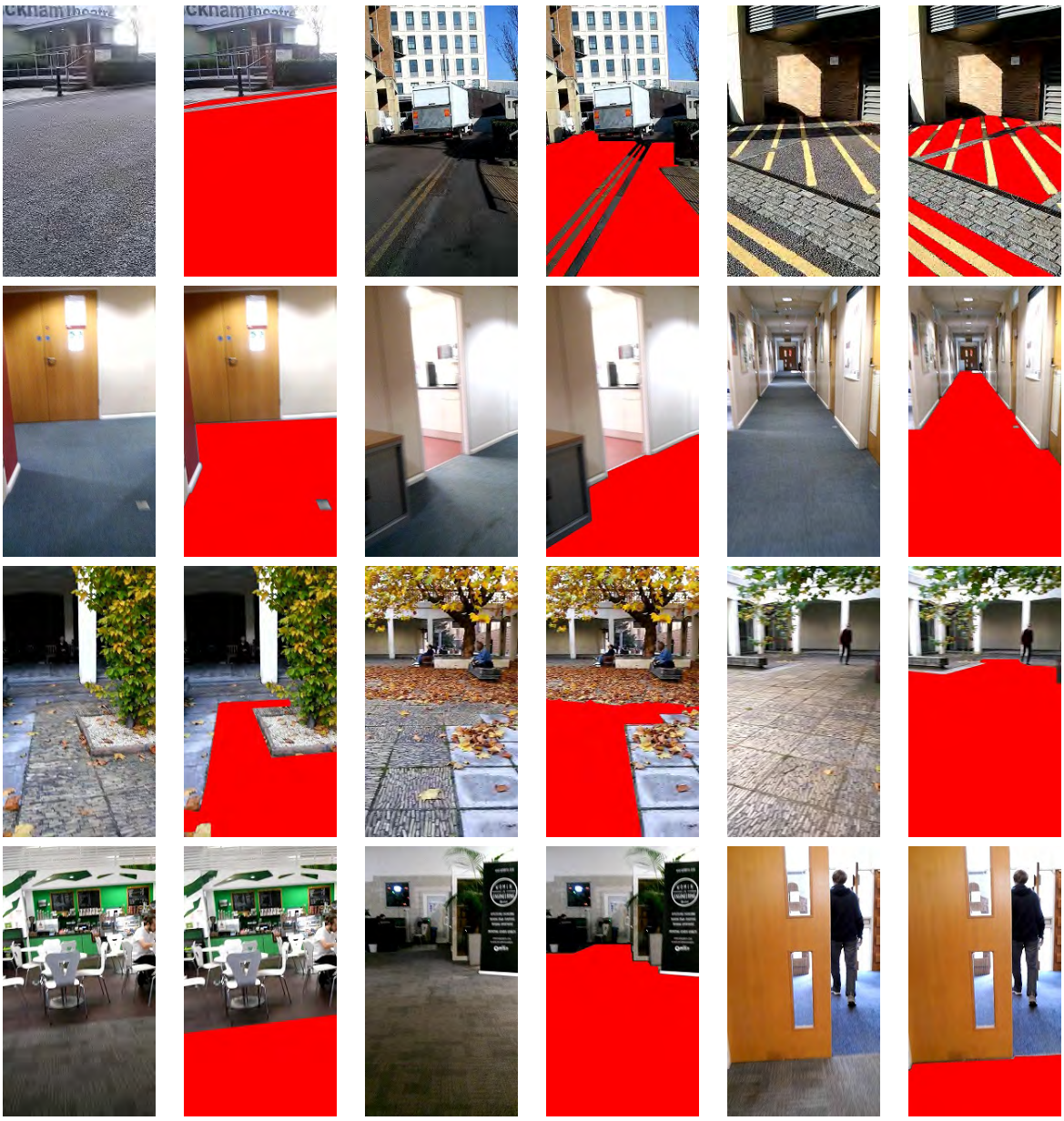
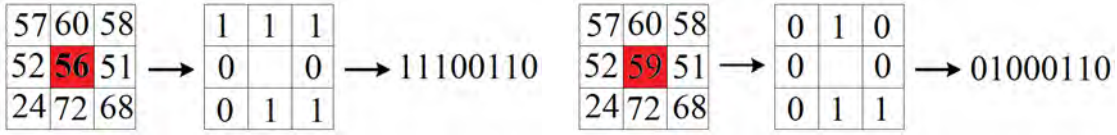


Fig.5.1: Images of proposed dataset, section b).

5.2 The proposed feature

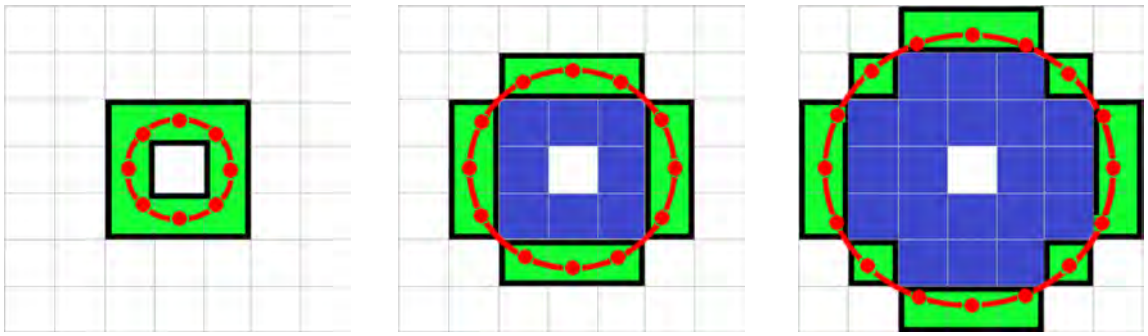
In the HLS extraction procedures, there are several challenging issues that affect the performance, for example, illumination changes, different scene perspectives, dynamic environment, etc. In order to provide HLS extraction information, one alternative is the use of LBP (see subsection 2.1.1) feature because it is simple, fast to compute and robust to illumination changes. However, the LBP feature is sensitive to noise (see Fig. 5.2), i.e., a low central pixel change affects the feature robustness.



a) LBP feature with 56 in central pixel b) LBP feature with 59 in central pixel

Fig. 5.2: Example of LBP sensitivity to noise, where a low central pixel change affects the feature robustness.

In addition, the use of some few pixels within a patch limits the information that can be used for feature description (see Fig. 5.3). To solve this inconvenience, we propose a new texture feature based on binary patterns. This new feature is robust to noise, robust to illumination changes, invariant to rotation and it considers a larger number of pixels than the LBP and LBP variants.



a) radius= 1, pixel number= 8 b) radius= 2, pixel number= 12 c) radius= 3, pixel number= 16

Fig. 5.3: Information that can be used for LBP feature. The use of some few pixels within a patch limits the performance in LBP feature. The pixels considered in the LBP computation are proportional to the used radios, where, green pixels are used in LBP feature and blue pixels are not used in LBP feature.

5.2.1 Input image

The input image is denoted as $I_{(x,y)}$. The image $I_{(x,y)}$ is used to obtain the texture features. We use an image partition of $I_{(x,y)}$ in a grid Θ to obtain a faster processing. For that, the grid Θ consists of sections Θ_w. Section Θ_w is a finite set of pixels $\Theta_w = \{x_1, ..., x_m\}$, $\Theta_w \in \Theta$, where, m is the pixels number within one section $\Theta_w \iff m$ is an odd number. A section Θ_w has a patch $\vartheta_{\varphi,\omega}$, where the pixels number from patch $\vartheta_{\varphi,\omega}$ are proportional to pixels number in the section Θ_w. Patch $\vartheta_{\varphi,\omega}$ is a finite set of pixels $\vartheta_{\varphi,\omega} = \{x_1, ..., x_u\}$, $\vartheta_{\varphi,\omega} \in \Theta$, where, u is the pixels number within one patch $\vartheta_{\varphi,\omega} \iff u$ is an odd number. Where, w denotes the w-th section in Θ, φ is the abscissa from grid Θ and ω is the ordinate from grid Θ. **Fig.** 5.4 shows a grid example Θ of 3×2.

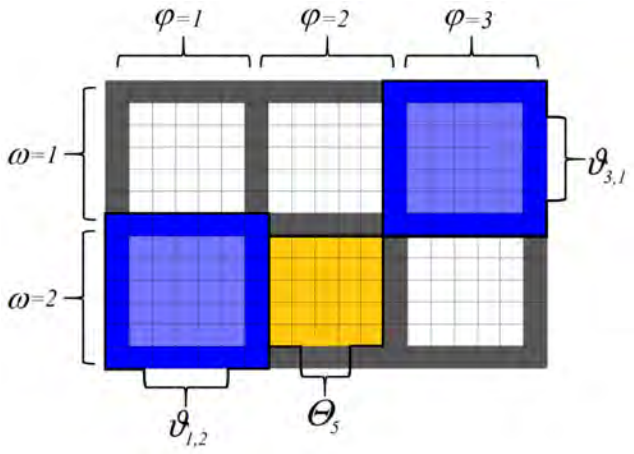


Fig. 5.4: Grid of 3×2. Blue squares are the patches $\vartheta_{\varphi,\omega}$ of 7×7 , orange square is one section Θ_w and the gray lines are the sections limits Θ_w.

5.2.2 Proposed texture feature

We propose a new feature to obtain texture based on binary patterns: BIRRN (Binary feature: Invariant to Rotation and Robust to Noise). The BIRRN feature considers a set of neighbor pixels within circular distributions with binary values, where binary values are added in each circular distribution. We refer to Δ_j as the set of neighbor pixels in circular distributions or BIRRN circles. The BIRRN provides the texture information in a patch $\vartheta_{\varphi,\omega}$. **Fig.** 5.5 shows an example of n BIRRN circles Δ_j. Where, each red circle corresponds to one neighbor pixel within BIRRN circles Δ_j, each red ring is one BIRRN circle Δ_j, the green squares are the pixels of the BIRRN circles Δ_j and black lines corresponding to BIRRN circles limit.

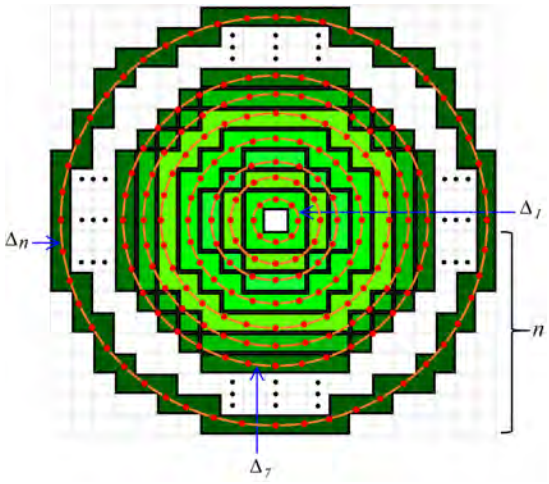
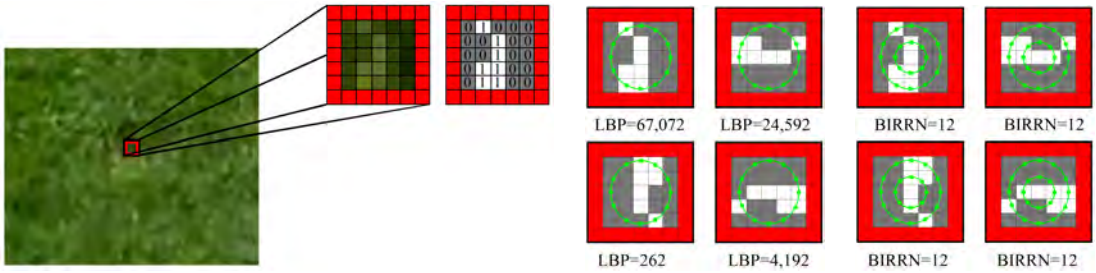


Fig. 5.5: Shows an example of BIRRN circles

5.2.3 Rotation and sensitivity to noise comparison

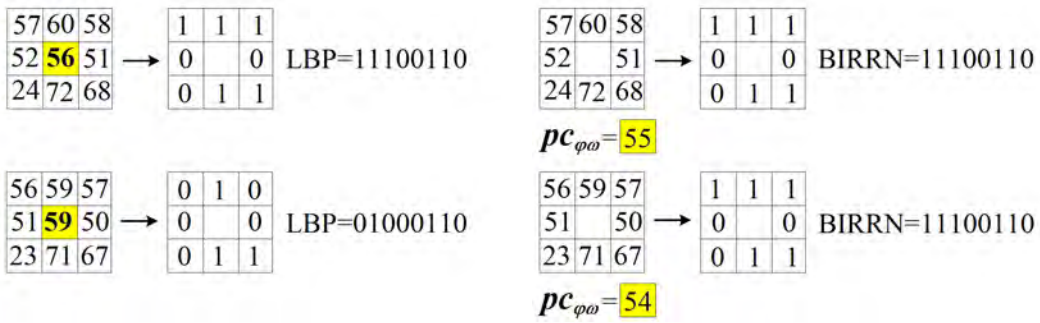
Fig. 5.6 shows a results comparison between LBP and proposed feature (BIRRN) using a patch 5×5, and considering four test cases (90^o, 180^o, 270^o and 360^o image rotations). Where, the LBP feature provides different results under the four tests, while the proposed feature delivers equal results for all cases. This is an advantage because in several urbanized elements (buildings, floors, monuments, etc.) textures values remain constant under different rotation changes. Therefore, a texture feature invariant to the rotation could be useful considering learning algorithms since would be possible to decrease its training to detect urbanized elements and it removes redundant features in detection.



a) Extraction of a patch 5×5 b) LBP features c) BIRRN features

Fig. 5.6: Rotation comparison between LBP and proposed feature. Image a) presents one patch extraction of grass and its binarization. Image b) presents the LBP features under different rotations 90^o, 180^o, 270^o and 360^o respectively. Image c) presents the BIRRN features under different rotations 90^o, 180^o, 270^o and 360^o respectively.

Fig. 5.7 shows a noise sensitivity comparison between LBP and proposed feature. For the LBP, a low change of the pixels value affects the LBP performance. For the proposed method, it provides the same results. This is possible because the proposed feature replaces the central pixel value used in LBP feature by the grayscale average of BIRRN circles.



a) Examples using LBP feature b) Examples using BIRRN feature

Fig. 5.7: Sensitivity to noise comparison. Image a) presents the binary result of LBP feature in two patches with a low change of the pixels value, where, the yellow pixels are the central pixel in LBP feature. Image b) presents the binary result of BIRRN feature in the two patches aforementioned, where, the yellow pixels are the grayscale average of BIRRN circle (neighbor pixels to central pixel).

5.2.4 BIRRN, LBP and LBP variants comparison

In this subsection, we presented a comparison between different binary features LBP (13), CsLBP (52), SLBP (53), XcsLBP (54) and the proposed feature. These 5 binary features were compared in five different scenes with 5 different classes to detect (grass, road, smooth carpet, tile and square carpet), the dataset consists of 1,500 images with different light intensities variations (see subsection 5.1).

For measurement procedures, we defined training features percentage as following: percentage of features that has more apparition in a particular element than the other elements of the image. For example, if we want to detect the grass, and the "feature (1011010) "appears 7 times in the grass and 5 times in the other elements of the image, this is considered as training feature. Otherwise, if the "feature (0001101) "appears 5 times in the grass and 7 in the other elements of the image, this is not considered as training feature. We defined confusion percentage as following: percentage in which training features detect elements different that the element being detecting. For example, if we want to detect the grass using LBP feature, and the LBP training features detect 5 features in the sky, 12 features in the trees and 10 feature in the buildings, the LBP feature has 27 features of confusion. Finally, we defined recurrence percentage as following: percentage of the training features variation, considering all images analyzed. For example, if using LBP feature and its training features are 97 in all images, the LBP feature has 100% recurrence. On the other hand, if we considering only 2 images and the first image has 100 training features and the second image has 50 training features, the LBP feature has 75% recurrence.

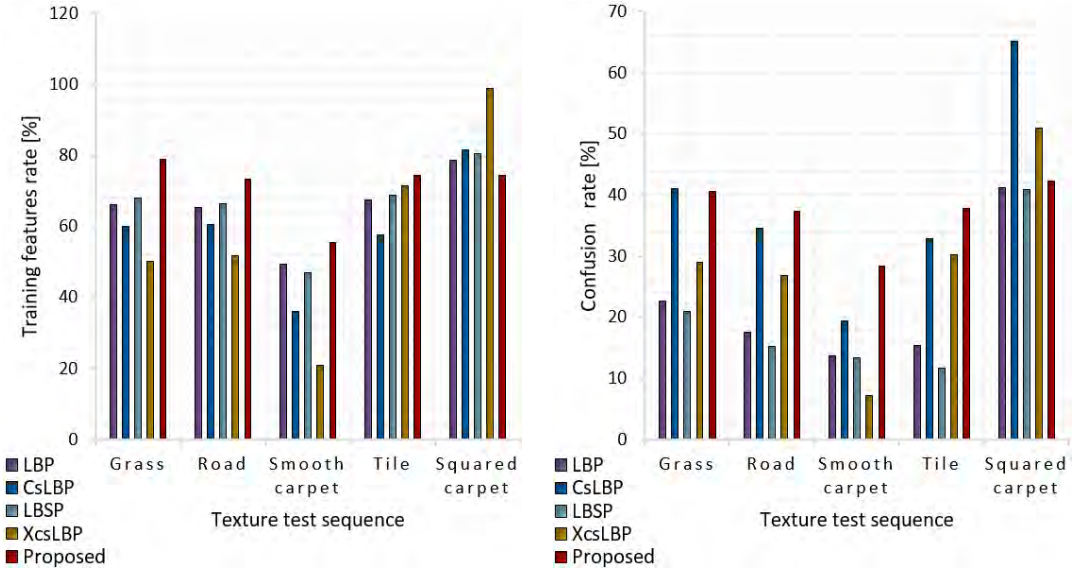
Experimental results

Table 5.1 shows the average percentage of training features, confusion and recurrence. As can be seen, the proposed feature provides a greater number of training features and more recurrence percentage than the other binary features. In addition, although it does not provide the best result in confusion this may be improved considering a second variable as presented in following subsection.

Table 5.1: Average percentage of training features, confusion and recurrence

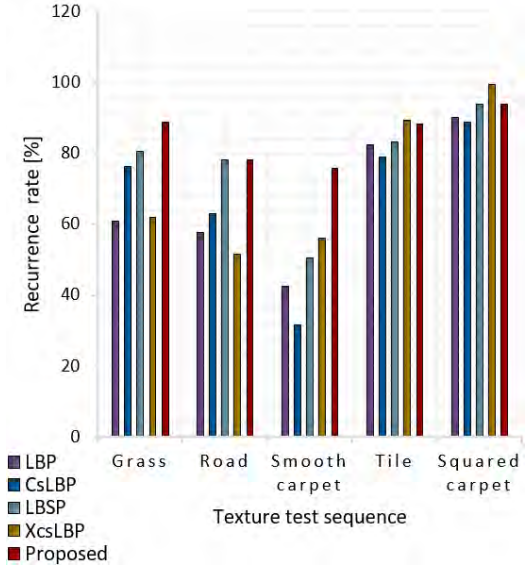
feature	training features percentage	confusion percentage	recurrence percentage
LBP (2002) (13)	65.45%	22.07%	66.69%
CsLBP (2009) (52)	59.24%	38.64%	67.72%
SLBP (2013) (53)	66.24%	**20.47%**	77.22%
XcsLBP (2015) (54)	58.75%	28.87%	71.64%
Proposed feature	**72.44%**	37.35%	**85.02%**

Fig. 5.8 and **Fig.** 5.9 show graphics that compare training features, confusion and recurrence percentage, considering the 5 binary features (LBP (13), CsLBP (52), SLBP (53), XcsLBP (54) and the proposed feature) in 5 different scenes (grass, road, smooth carpet, tile and square carpet).



a) Training features percentage by scene b) Confusion percentage by scene

Fig. 5.8: BIRRN, LBP and LBP variants comparison, section a). Images a) and b) compare 5 binary features using the metrics: number of training features and confusion. These metrics were obtained in 5 different scenes (grass, road, smooth carpet, tile and square carpet).



c) Recurrence percentage by scene

Fig. 5.9: BIRRN, LBP and LBP variants comparison, section b). Image c) compare 5 binary features using the metric: recurrence. This metric was obtained in 5 different scenes (grass, road, smooth carpet, tile and square carpet).

5.2.5 Proposed binary feature to train a learning algorithm

In order to analyze the scope of the proposed texture feature, we train a learning algorithm (logistic regression) to floor recognition. This learning algorithm was trained using color (RGB channels) and the training features obtained in subsection 5.2.4 (proposed texture feature). **Fig.** 5.10 shows some images of the floor recognition results of a learning algorithm trained with proposed texture feature and color (RGB channels).

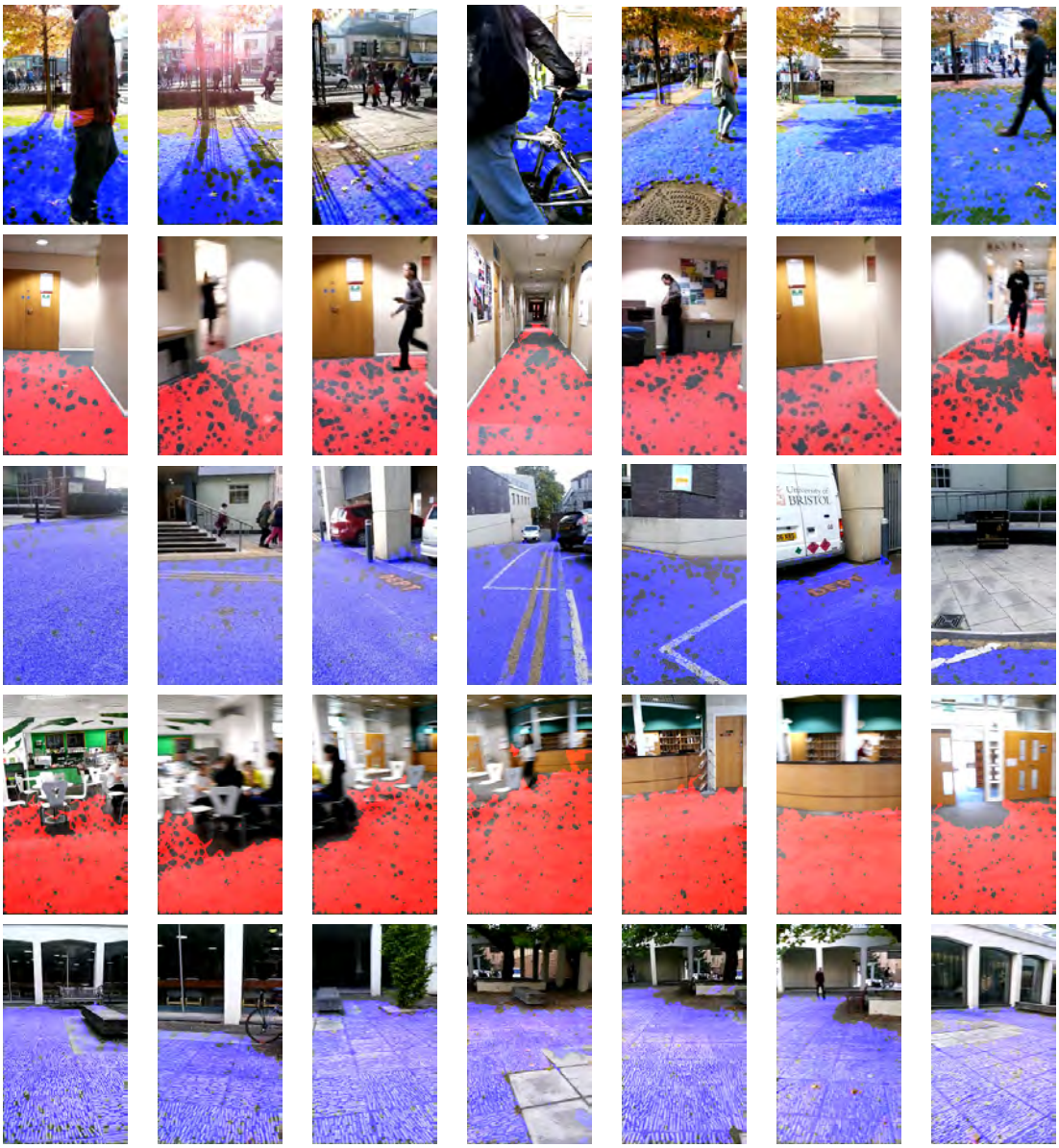


Fig. 5.10: Floor recognition of grass, smooth carpet, road, square carpet and tile respectively. Where, the blue and red areas are the floor recognition using the proposed binary feature and RGB channels how input of a learning algorithm.

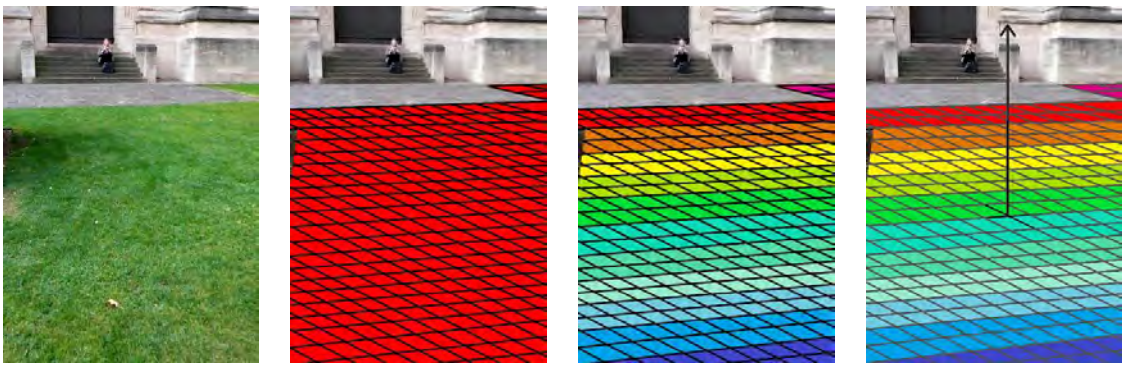
Table 5.2 presents the floor recognition percentage (accuracy) and the confusion percentage, i.e., the recognition of other elements that is not the floor. In addition, it also also provides the standard deviation of accuracy and confusion. As can be seen, the floor recognition using a learning algorithm trained with proposed texture feature and color (RGB channels) has high accuracy and low confusion using urbanized images with light intensities variations and different scene perspectives. This is possible because the proposed texture feature has robustness to illumination changes and invariance to rotation.

Table 5.2: The accuracy and confusion

	Percentage	**Standard deviation**
Accuracy	90.37%	3.27
Confusion	4.72%	1.34

Work in progress

As work in progress, we considering that the proposed binary feature can be used to obtain 3D orientation of HLS (similar to the method presented in section 5.3). We expect that using this proposed feature we will achieve more degrees of freedom in 3D orientation. Because the proposed feature can work with patches smaller than the method presented in section 5.3. These small patches can be used to obtain a blurring pattern on the HLS that provides its 3D orientation, i.e., these small patches can be used to obtain a different blurring pattern in each HLS view that provides 3D orientation information (see **Fig.** 5.11). In addition, faster processing is expected since the proposed descriptor can process 14 frames HD (640×480 pixels) by seconds, while, the method presented in section 5.3 can process 1 frame 480p (640×480 pixels) by seconds.



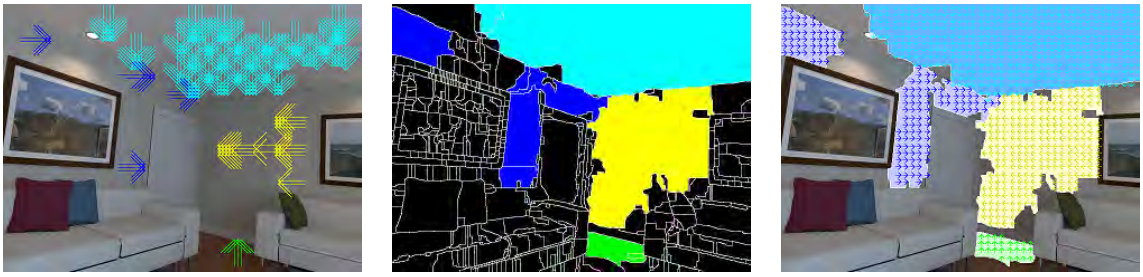
a) Input images b)Patches c) Blurring pattern d) 3D orientation

Fig. 5.11: These small patches can be used to obtain a blurring pattern on the HLS that provides its 3D orientation.

5.3 3D orientation

Our first steps for to obtain 3D orientation from a single image are presented in the manuscript (55). In this manuscript, we present a new dominant plane recognition method from a single image that provides five 3D orientation (right, left, front, top and bottom) of dominant planar structures (floor, wall and ceiling) in interior scenes. To obtain the 3D orientation of planar structures, we assume that in each planar structure view has a different texture information by the factors that affect the image as light variations, blurring and other factors. Then, we train a learning algorithm (logistic regression) with texture features to predict the 3D orientation in a planar structure. **Fig.** 5.12 a) shows an image with 3D orientation (right, left, top and bottom).

To increase the 3D orientation information provided by learning algorithm, we use the 3D orientation as growing points in a contour image. We use the method proposed in (48) to obtain the contour image. **Fig.** 5.12 b) shows growth of the 3D orientation information in contour image. Finally, all 3D orientation that belongs to the same dominant planar structure (floor, wall or ceiling) are integrated into a single element. **Fig.** 5.12 c) shows the dominant planar structures recognition with their 3D orientation.



a) 3D orientation b)Growth the 3D orientation information c) Dominant planar structures

Fig. 5.12: 3D orientation of dominant planar structures (floor, wall and ceiling) in interior.

Our method is different from others in that we do not aim at classifying every single pixel of the image in order to recognise dominant planes. In contrast, we use only those pixels for which a label could be assigned by our learning algorithm (3D orientation: right, left, top and bottom) as seeds in a connected component segmentation algorithm, which grows regions of connected pixels and whose stop criteria is set by the edges found by the contour algorithm. We intentionally chose interior scenes in order to propose a set of visual descriptors that could capture the appearance of a dominant plane in such scenes.

5.3.1 Dataset

We use the RGB images and depth images from interior scenes of the ICL-NUIM dataset (56). This dataset consists of two different scenes (living room and office room) with images of 640 $\times$ 480 pixels. Our training set and testing set consisted of 428 and 220 images, respectively. **Fig.** 5.13 shows ICL-NUIM dataset images.

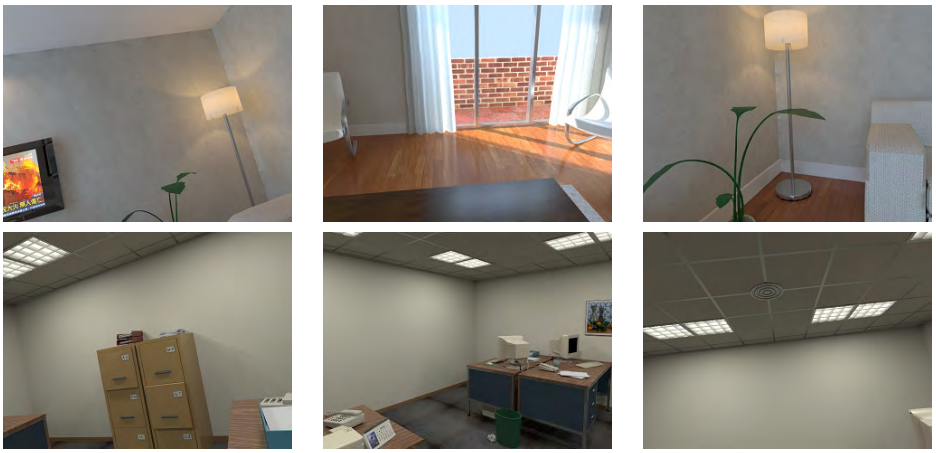


Fig. 5.13: ICL-NUIM dataset (56)

5.3.2 Experimental results

Table 5.3 is a confusion matrix of 3D orientation classification of proposed method. The correct guesses percentage are located in the main diagonal of the confusion matrix. In this main diagonal, the average of correct guesses percentage (3D orientation accuracy of proposed method) is 60.17%.

Table 5.3: Confusion matrix of 3D orientation classification

	Front orientation	Right orientation	Left orientation	Top orientation	Bottom orientation
Front orient.	**71.47%**	2.50%	1.75%	12.50%	0.00%
Right orient.	0.00%	**55.33%**	9.37%	3.12%	0.00%
Left orient.	0.00%	1.92%	**69.63%**	1.82%	0.00%
Top orient.	3.57%	2.50%	2.70%	**58.86%**	0.00%
Bottom orient.	11.07%	0.00%	0.00%	0.00%	**45.57%**

Table 5.4 presents the confusion of each 3D orientation. We defined confusion as following: the percentage that learning algorithm recognize 3D orientation within elements that are not dominant planar structures (floor, wall and ceiling). The confusion average of proposed method is 3.14%.

Table 5.4: 3D orientation confusion

	Front orientation	Right orientation	Left orientation	Top orientation	Bottom orientation
Confusion	4.72%	2.92%	2.37%	3.84%	1.87%

Fig. 5.14 shows some images of 3D orientation (right, left, front, top and bottom) of dominant planar structures (floor, wall and ceiling) in interior scenes of proposed method.

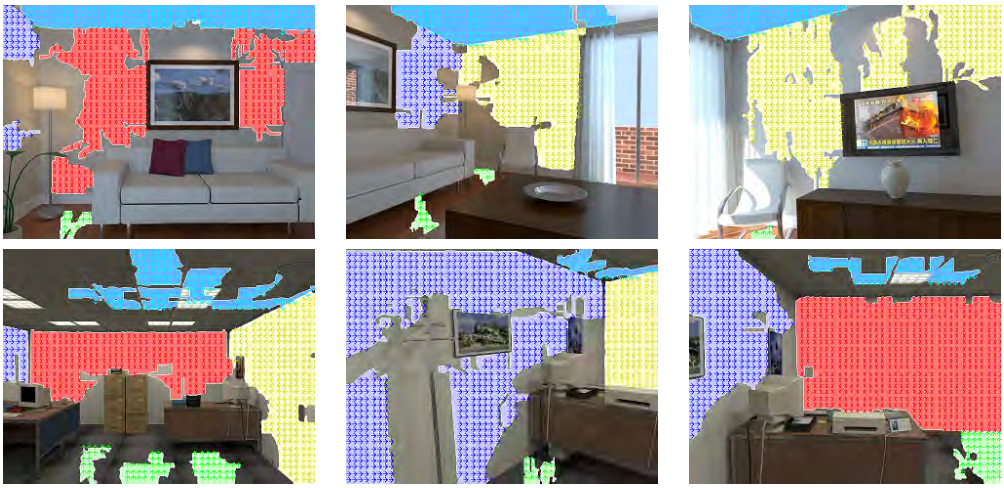
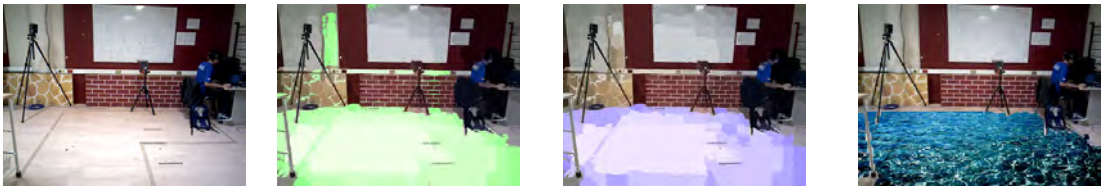


Fig. 5.14: Images of 3D orientation of proposed method.

5.4 Augmented reality

Our first steps for to obtain Augmented Reality applications from a single image are presented in manuscript (57). In this manuscript, we present a floor recognition method with virtual information as water, lava and grass (see **Fig.** 5.17). In order to detect the floor light variations, we proposed a rule system that integrates three variables: texture features, blurring and superpixels-based segmentation (49). **Fig.** 5.15 b) shows an image with floor light variations detection (green color in **Fig.** 5.15 b). In order to remove noise (floor light variations misrecognition), we proposed a technique presented in manuscript (57). **Fig.** 5.15 c) shows an image without noise using the proposed technique. Finally, the image without noise provides the area to integrate the virtual information. **Fig.** 5.15 d) shows an image with augmented reality.



a) Input image b) Floor detection c) Image without noise d) Augmented reality

Fig. 5.15: Augmented reality method. Image a) presents the input image. Image b) presents the floor recognized. Image c) presents the floor recognition (blue area) without noise (removing the bad recognition). Finally, the floor detected in image c) is used how the area to integrate the virtual information (water) to obtain the image d).

Our method recognizes the floor in an interior scene and replaces it with an augmented texture, for which a coarse light model, generated with our approach, it is applied in order to generate a more realistic augmentation of this virtual texture. Contrary to other methods, our approach does not aim at classifying every single pixel of the image in order to recognize the floor, but we exploit the assumption that homogeneous regions (similar in appearance) are likely to correspond to the same plane (the floor).

5.4.1 Dataset

We use the RGB images and depth images from interior scenes of a proposed dataset. The dataset was obtained using a sensor Asus Xtion (58). This dataset consists of a laboratory scene within our campus with images of 640×480 pixels. The proposed dataset has different floor light intensities. Our training set and testing set consisted of 80 and 250 images, respectively. **Fig.** 5.16 shows proposed dataset images.

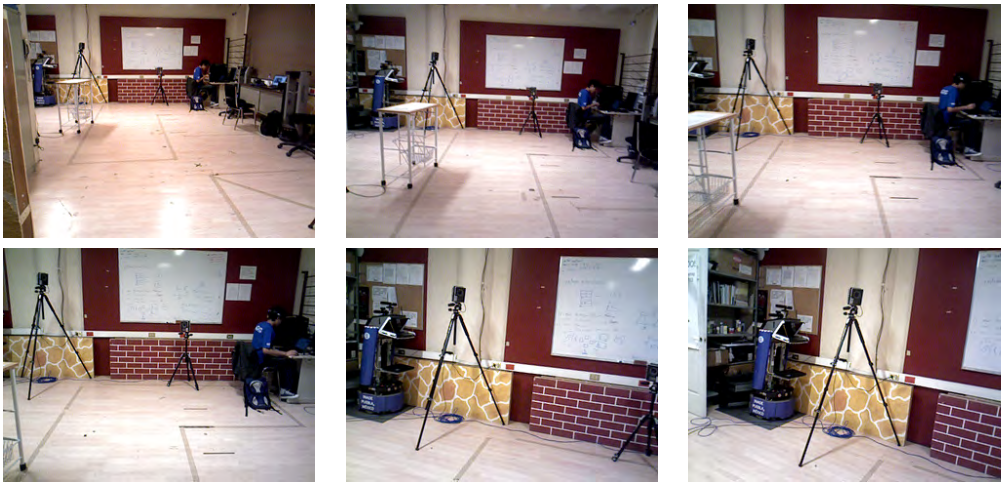


Fig. 5.16: Laboratory dataset.

5.4.2 Experimental results

Table 5.5 presents the floor recognition percentage (accuracy) and the floor misrecognition percentage (confusion), i.e., the percentage that proposed method detect elements that are not floor. Furthermore, we provide the standard deviation of the accuracy and confusion.

Table 5.5: The accuracy and confusion

	Percentage	**Standard deviation**
Accuracy	90.18%	2.24
Confusion	3.72%	1.12

Fig. 5.17 shows some images of augmented reality using our proposed method. The virtual elements are integrated of water, lava and grass, respectively.

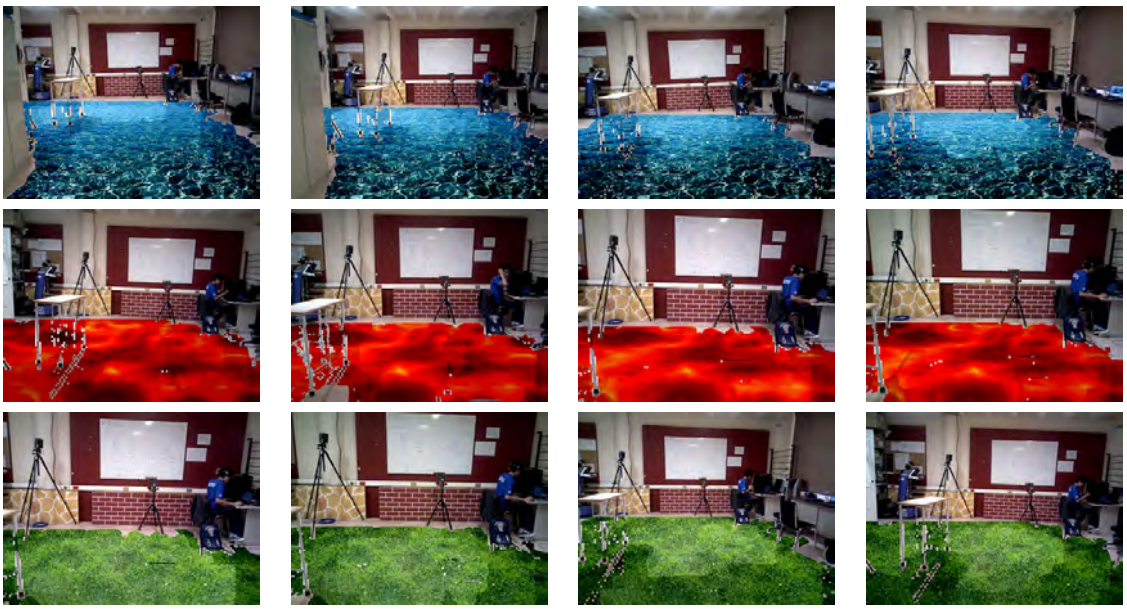


Fig. 5.17: Augmented reality of water, lava and grass.

Chapter 6
Conclusions and future work

6.1 Conclusions

HLS extractions from a single image is useful because in real world applications several data are limited to a single view, for example historical images, internet images, personal pictures, holiday photos and so on. Unfortunately, in most previous work (using a single image), only planes are extracted. This is an important limitation since other 3D structures such as spheres, cylinders and cubes would deliver more scene information. In this thesis proposal, we are interested in extract HLS that in previous work not extract. This would be useful because an extraction methodology that deliver several different structures (spheres, cylinders, cubes, etc.) would provide more 3D scene information. In addition, 3D structures such as spheres, cylinders and cubes would increase the performance in several real-world applications where HLS are used such as navigation, augmented reality, 3D reconstruction etc. Preliminary results are encouraging and show the feasibility of our HLS extraction methodology. As preliminary results in 3D structures extraction, we proposed a new dataset with light intensities variations, a new texture feature based on binary patterns which provide discriminant values for HLS recognition, a method to obtain dominant structures orientation and an augmented reality application using planar structure recognition. As work in progress, we will use the proposed binary features to obtain 3D orientation of HLS and analyze the use of other visual features to obtain the HLS extraction.

6.2 Work in progress

As work in progress, we will use the proposed binary feature (see section 5.2) to obtain 3D orientation of HLS. In addition, we will present a new HLS extraction method using a single view, key elements labeling and 3D orientation. Also, we will expect to present an augmented reality application. For that, we set as tentative journal/conferences: ISMAR, CVPR and Computer Vision and Image Understanding.

Bibliography

[1] Dani, A., Panahandeh, G., Chung, S. J., Hutchinson, S., (2013). Image moments for higher-level feature based navigation. In IEEE International Conference on Intelligent Robots and Systems (IROS), (pp. 602-609).

[2] Ventura, J., Hollerer, T., (2010). Real-time Planar World Modeling for Augmented Reality. In IEEE International Symposium on Mixed and Augmented Reality (ISMAR), (pp. 1-4).

[3] Li, W., Song, D., (2014). Toward featureless visual navigation: Simultaneous localization and planar surface extraction using motion vectors in video streams. In IEEE International Conference on Robotics and Automation (ICRA), (pp. 9-14).

[4] Vosselman, G., Dijkman, S. (2001). 3D building model reconstruction from point clouds and ground plans. International archives of photogrammetry remote sensing and spatial information sciences, (pp. 37-44).

[5] Gee, A., D. Chekhlov, A., Calway, A., Mayol-Cuevas, W., (2008). Discovering higher level structure in visual SLAM. In Robotics, IEEE Transactions, (pp. 980-990).

[6] Bartoli, A., Sturm, P., (2003). Constrained structure and motion from multiple uncalibrated views of a piecewise planar scene. In International Journal of Computer Vision IJCV, (pp. 6-122).

[7] Silveira, G. F., Malis, E., Rives, P., (2006). Real-time robust detection of planar regions in a pair of images. In International Conference on Intelligent Robots and Systems (IROS), (pp. 49-54).

[8] Gee, A., Chekhlov, D., Mayol, W., Calway, A., (2007). Discovering planes and collapsing the state space in visual slam. In British Machine Vision Conference (BMVC), (pp. 6-122).

[9] Wangsiripitak, S., Murray, D., (2010). Reducing mismatching under time-pressure by reasoning about visibility and occlusion. In British Machine Vision Conference BMVC, (pp. 6, 72).

[10] Hoiem, D., Efros, A. A., Hebert, M. (2007). Recovering surface layout from an image. International Journal of Computer Vision, (pp. 151-172).

[11] Saxena, A., Sun, M., Ng, A. Y. (2009). Make3d: Learning 3d scene structure from a single still image. IEEE transactions on pattern analysis and machine intelligence, (pp. 824-840).

[12] Haines, O., Calway, A. (2015). Recognising planes in a single image. IEEE transactions on pattern analysis and machine intelligence, (pp. 1849-1861).

[13] Ojala, T., Pietikainen, M., Maenpaa, T., (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. In IEEE Transactions on Pattern Analysis and Machine Intelligence, (pp. 971-987).

[14] Conners, R. W., Harlow, C. A. (1980). A theoretical comparison of texture algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, (pp. 204-222).

[15] Laws, K. I. (1980). Rapid texture identification. In 24th annual technical symposium. International Society for Optics and Photonics, (pp. 376-381).

[16] Nevatia, R., Babu, K. R. (1980). Linear feature extraction and description. Computer Graphics and Image Processing, (pp. 257-269).

[17] Tsai, G. (2010). Histogram of oriented gradients. University of Michigan.

[18] Kekre, H. B., Thepade, S. D. (2009). Color based image retrieval using amendment of block truncation coding with YCbCr color space. International Journal of Imaging and Robotics, (pp. 2-14).

[19] Andrew Y. Ng. Coursera, Stanford, Machine Learning. In online, Website: https://class.coursera.org/ml-003/lecture.

[20] Li, S. Z. (2009). Markov random field modeling in image analysis. Springer Science & Business Media.

[21] Vadivambal, R., Digvir, S. J. (2016). Chapter 3. Classification, statistical and neural network. In R. Vadivambal, & S. J. Digvir, Bio-Imaging. Principles, Techniques, and Applications. (pp. 47-48). New York: Taylor & Francis Group.

[22] Huang, Y., Kangas, L. J., Rasco, B. A. (2007). Applications of artificial neural networks (ANN) in food science and nutrition. Critical Reviews in Food Science and Nutrition, (pp. 113-126).

[23] Prados, E., Faugeras, O. (2005). Shape from shading: a well-posed problem? In Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society Conference on (Vol. 2, pp. 870-877).

[24] Aloimonos, J. (1988). Shape from texture. Biological cybernetics, 58(5), 345-360.

[25] Saxena, A., Chung, S. H., Ng, A. Y., (2005) Learning depth from single monocular images. In: Advances in Neural Information Processing Systems, (pp. 1161-1168).

[26] Saxena, A., Schulte, J., Ng, A. Y. (2007). Depth Estimation Using Monocular and Stereo Cues. In IJCAI, (pp. 2197-2203).

[27] Saxena, A., Chung, S. H., Ng, A. Y., (2008) 3-D depth reconstruction from a single still image. In International journal of computer vision, (pp. 53-69).

[28] Liu, B., Gould, S., Koller, D., (2010) Single image depth estimation from predicted semantic labels. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 1253-1260).

[29] Karsch, K., Liu, C., Kang, S. B. (2012). Depth extraction from video using non-parametric sampling. In European Conference on Computer Vision. Springer Berlin Heidelberg, (pp. 775-788).

[30] Mota-Gutierrez, S. A., Hayet, J. B., Ruiz-Correa, S., Hasimoto-Beltran, R., Zubieta-Rico, C. E., (2013) Learning depth from appearance for fast one-shot 3-D map initialization in VSLAM systems. IEEE International Conference on Robotics and Automation (ICRA), (pp. 2291-2296).

[31] Eigen, D., Puhrsch, C., Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In Advances in neural information processing systems, (pp. 2366-2374).

[32] Liu, M., Salzmann, M., He, X. (2014). Discrete-continuous depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (pp. 716-723).

[33] Zhuo, W., Salzmann, M., He, X., Liu, M., (2015) Indoor Scene Structure Analysis for Single Image Depth Estimation. IEEE Conference on Computer Vision and Pattern Recognition, (pp. 614-622).

[34] Liu, F., Shen, C., Lin, G., Reid, I., (2016) Learning depth from single monocular images using deep convolutional neural fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, (pp. 1-16).

[35] Silberman, N., Hoiem, D., Kohli, P., Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. In European Conference on Computer Vision. Springer Berlin Heidelberg, (pp. 746-760).

[36] Fritsch, J., Kuhnl, T., & Geiger, A. (2013). A new performance measure and evaluation benchmark for road detection algorithms. International Conference on Intelligent Transportation Systems (ITSC), (pp. 1693-1700).

[37] Cherian, A., Morellas, V., Papanikolopoulos, N. (2009). Accurate 3D ground plane estimation from a single image. IEEE International Conference on Robotics and Automation (ICRA), (pp. 2243-2249).

[38] Rahimi, A., Moradi, H., Zoroofi, R. A. (2013). Single image ground plane estimation. IEEE International Conference on Image Processing (ICIP), (pp. 2149-2153).

[39] Firman, M., Mac Aodha, O., Julier, S., Brostow, G. J. (2016). Structured prediction of unobserved voxels from a single depth image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (pp. 5431-5440).

[40] Kovsecká, J., Zhang, W. (2005). Extraction, matching, and pose recovery based on dominant rectangular structures. Computer Vision and Image Understanding, (pp. 274-293).

[41] Micusik, B., Wildenauer, H., Kosecka, J. (2008). Detection and matching of rectilinear structures. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 1-7).

[42] McClean, E., Cao, Y., McDonald, J. (2011). Single image augmented reality using planar structures in urban environments. In Machine Vision and Image Processing Conference (IMVIP), (pp. 1-6).

[43] Hoiem, D., Efros, A. A., Hebert, M. (2005). Geometric context from a single image. IEEE International Conference on Computer Vision (ICCV), (pp. 654-661).

[44] Haines, O., Calway, A. (2012). Estimating Planar Structure in Single Images by Learning from Examples. In ICPRAM (pp. 289-294).

[45] Saxena, A., Make3D Range Image Data, Website: http://make3d.cs.cornell.edu/data.html.

[46] Domke, J. (2013). Learning graphical model parameters with approximate marginal inference. IEEE transactions on pattern analysis and machine intelligence, (pp. 2454-2467).

[47] Hoiem, D., Efros, A. A., Hebert, M. (2005). Automatic photo pop-up. ACM transactions on graphics (TOG), (pp. 577-584).

[48] Arbelaez, P., Maire, M., Fowlkes, C., Malik, J., (2011). Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence, (pp. 898-916).

[49] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S., (2010). SLIC superpixels. No (p. 15).

[50] Cusano, C., Napoletano, P., Schettini, R., (2016). Evaluating color texture descriptors under large variations of controlled lighting conditions. JOSA A, (pp. 17-30).

[51] Gould, S., Fulton, R., Koller, D., (2009). Decomposing a scene into geometric and semantically consistent regions. In International Conference on Computer Vision (pp. 1-8).

[52] Heikkilä, M., Pietikäinen, M., Schmid, C. (2009). Description of interest regions with local binary patterns. Pattern recognition, (pp. 425-436).

[53] Yin, H., Yang, H., Su, H., Zhang, C. (2013). Dynamic background subtraction based on appearance and motion pattern. IEEE International Conference on Multimedia and Expo Workshops, (pp. 1-6).

[54] Silva, C., Bouwmans, T., Frélicot, C. (2015). An eXtended center-symmetric local binary pattern for background modeling and subtraction in videos. In Computer Vision, Imaging and Computer Graphics Theory and Applications, VISAPP 2015.

[55] Osuna-Coutiño J. A. J., Martinez-Carranza J., Arias-Estrada M., Mayol-Cuevas W., (2016). Plane Recognition in Interior Scenes from a Single Image. IEEE International Conference on Pattern Recognition (ICPR), (pp. 1924-1929).

[56] Imperial College London, the ICL-NUIM dataset, in: online, Website: http://www.doc.ic.ac.uk/ ahanda/VaFRIC/iclnuim.html

[57] Osuna-Coutiño J. A. J., Cruz-Martínez C., Martinez-Carranza J., Arias-Estrada M., Mayol-Cuevas W., (2016). I want to change my floor: dominant plane recognition from a single image to augment the scene. IEEE International Symposium on Mixed and Augmented Reality Adjunct Proceedings (ISMAR), (pp. 135-140).

[58] H. Gonzalez-Jorge, B. Riveiro, E. Vazquez-Fernandez, J. Martinez-Sanchez, and P. Arias, "Metrological evaluation of microsoft Kinect and asus xtion sensors," Measurement, vol. 46, no. 6, pp. 1800–1806, Jul. 2013.