Computer Vision and Image Understanding 115.1 (2011): 91-104

DOI: http://dx.doi.org/10.1016/j.cviu.2010.09.009

Shape-Based Image Segmentation Through Photometric Stereo

Carme Julià[a], Rodrigo Moreno[a], Domenec Puig[a], Miguel Angel Garcia[b]

[a] *Universitat Rovira i Virgili, Intelligent Robotics and Computer Vision Group*
Dept. of Computer Science and Mathematics,
Av.Països Catalans 26, 43007, Tarragona, Spain
[b] *Dept. of Informatics Engineering, Universidad Autónoma de Madrid*
Francisco Tomás y Valiente, 11, 28049, Madrid, Spain

Abstract

This paper describes a new algorithm for segmenting 2D images by taking into account 3D shape information. The proposed approach consists of two stages. In the first stage, the 3D surface normals of the objects present in the scene are estimated through robust photometric stereo. Then, the image is segmented by grouping its pixels according to their estimated normals through graph-based clustering. One of the advantages of the proposed approach is that, although the segmentation is based on the 3D shape of the objects, the photometric stereo stage used to estimate the 3D normals only requires a set of 2D images. This paper provides an extensive validation of the proposed approach by comparing it with several image segmentation algorithms. Particularly, it is compared with both appearance-based image segmentation algorithms and shape-based ones. Experimental results confirm that the latter are more suitable when the objective is to segment the objects or surfaces present in the scene. Moreover, results show that the proposed approach yields the best image segmentation in most of the cases.

Keywords: Photometric stereo, 3D surface normals, graph-based image

1. Introduction

Image segmentation is an important stage in computer vision as a preliminary step towards higher level analysis and recognition stages. It aims at partitioning a given image into a set of non-overlapping homogeneous regions that likely correspond to different objects or geometric structures that may be perceived in the scene. In most image segmentation algorithms, the homogeneity criterion that determines the final partition is related to visual cues such as intensity, texture or color (e.g., [1], [2], [3], [4]). This implies that the obtained regions are determined by the visual appearance of the objects present in the scene rather than by their actual 3D shapes. Although an appearance-based segmentation is useful in many applications, it usually leads to oversegmentation when working with textured objects. This oversegmetation can complicate the interpretation of the scene when the goal is to segment the objects contained in it.

Other image segmentation algorithms are based on the 3D information of the image points. This information can be encoded in a range image, whose pixels express the distance between a known reference frame and a visible point in the scene. Range images are also referred to as depth images, depth maps, xyz maps, surface profiles and 2.5D images. They are frequently obtained with laser range finders or structured light scanners. The main drawback of these sensors is their high cost, weight and size. In addition, the range image acquisition process is usually highly time consuming.

Another possibility to obtain 3D information is through a stereo camera.

Unfortunately, the obtained 3D data are very noisy in general. Furthermore, the 3D information cannot be computed for all the pixels in the 2D image. On the one hand, there are regions with poor texture in which it is difficult to find correspondences between both cameras. On the other hand, some points are only visible from one of the cameras. Hence, the obtained 3D data contain holes that correspond to points whose 3D coordinates cannot be estimated. These points constitute missing regions that cannot be segmented.

Although several approaches have been proposed, image segmentation is not a solved problem yet. One of the main difficulties is the definition of a good measure of the obtained error. This measure should depend on the final application. For instance, it could be desirable to obtain either the highest possible percentage of correctly segmented pixels with under-segmentation or any amount of correctly segmented pixels with oversegmentation. Experimental comparisons of several range image segmentation algorithms were presented in [5] and [6]. In a more recent paper, Wirjadi [7] presents a survey of 3D image segmentation methods. This survey concludes that it is not possible to have a standard segmentation method that can be expected to work equally well for all tasks. Wirjadi points out that, given a new image processing problem, the method that best solves the given task should be specifically chosen.

This paper presents an extension of the shape-based image segmentation algorithm previously introduced in [8], which consists of two stages. In a first stage, both the 3D surface normal and reflectance corresponding to every image pixel are estimated through a robust photometric stereo technique [9]. Photometric stereo [10] aims at estimating the normals and reflectance of a

3D surface from several intensity images, each obtained with different lighting conditions. The light source position for each image is also recovered. The general assumptions are that the projection is orthographic, the camera and objects are stationary and the light source is away from the objects. Thus, it can be assumed that the light illuminates every point of the scene from the same angle and with the same intensity. The computed surface normals can be used to reconstruct the surface of an object present in the scene.

Then, in the second stage of the proposed approach, the image pixels are clustered according to the orientation of their 3D surface normals through a graph-based segmentation algorithm proposed in [11]. In this particular case of image segmentation, the vertices of the graph are constituted by the pixels of any of the 2D input images, whereas the weight of every edge is some measure of dissimilarity between the two pixels linked by that edge. In particular, weights are defined by using the 3D surface normals estimated at the first stage. The goal is to group neighbouring pixels that are likely to belong to the same geometric structure.

The proposed approach is not a new 3D image segmentation scheme, but a new application of photometric stereo to image segmentation based on surface normals. One of the advantages of the proposed approach is that it only requires a conventional monocular camera and a set of inexpensive strobe lights, thus avoiding costly range finders [5] or calibrated stereo rigs. Additionally, it should be highlighted that, as shown later, only six images are necessary to recover the surface normals. Although better results could be obtained with a higher number of images, this paper aims at designing a system as portable as possible, which can be mounted on a mobile robot

together with a camera. The proposed technique is especially suited for poorly illuminated scenarios, such as those typical in search and rescue.

Photometric approaches (e.g., [12], [13], [14]) usually handle 2D intensity images containing a single object, without considering background pixels, which can contain noisy data and whose geometry and material can be very different from the ones of the object. On the contrary, the image segmentation approach proposed in this paper can cope with several objects and does consider background pixels.

Following a different approach, Koppal and Narasimhan ([15], [16]) proposed a technique for 2D image segmentation that also clusters the scene points according to their surface normals. The key point of their proposal is to use the continuity of a smoothly moving, distant light source to extract information about the scene geometry. In particular, they use the fact that the brightness measurements at each pixel form a *continuous appearance profile* and show how the derivatives of this profile are related to the geometry of the scene. One of the main drawbacks of this approach is that they need a large number of images in order to obtain a smooth, distant light source behaviour. Furthermore, an additional step is necessary to compute the normals. More recently, Shi et al. [14] proposed an approach that solves the *generalized bas-relief* (GBR) ambiguity [17] by using intensity profiles to cluster pixels with the same albedo, but different surface normals. Unfortunately, this method cannot deal with grey-level images.

A previous image segmentation approach also based on [11] was presented in [18]. However, the main difference with the proposed strategy is that the graph in [18] is created from the 3D points acquired by a binocular stereo

camera, as well as from the normals inferred from those 3D points through the tensor voting framework [19]. Unfortunately, that technique tends to oversegment the images due to the high levels of spatial noise typically present in the 3D points obtained from stereo vision.

Although the results provided in [8] showed the viability of the proposed approach, comparisons and experimental validations were not carried out in that work. In turn, the current paper performs an extensive evaluation in which the proposed approach is compared with several image segmentation algorithms based on either visual appearance or shape. One of the advantages of shape-based image segmentation with respect to appearance-based one is that the obtained regions are more related to the surfaces of the physical objects present in the scene, with independence of their visual appearance. Thus, a planar wall covered with pictures, for instance, would be segmented into a collection of patches by an appearance-based segmenter, whereas it would be segmented into a single region by a shape-based segmenter as the one proposed in this work.

The rest of the paper is organized as follows. Section 2 presents in detail the two stages of the image segmentation algorithm previously introduced in [8]. Firstly, the formulation of the photometric stereo problem is reviewed. In addition, the robust photometric stereo approach applied in the first stage of the proposed strategy is briefly outlined. Then, the graph-based segmentation algorithm applied in the second stage of the proposed scheme is described. An experimental evaluation of the proposed image segmentation approach is presented in Section 3. In particular, the proposed approach is compared with several image segmentation algorithms, both appearance-

based and shape-based. Finally, concluding remarks are summarized in Section 4.

2. Shape-Based Image Segmentation Algorithm

The key point of the proposed algorithm is that it is exclusively based on information acquired from several 2D images in order to perform image segmentation based on 3D shapes. The proposed approach consists of two stages described below.

2.1. Estimation of 3D surface normals through photometric stereo

At this first stage, both the 3D surface normal and reflectance at every pixel is estimated through a robust photometric stereo technique [9]. The position of the light source for each image is also recovered. The formulation of the photometric stereo problem [10] is introduced below.

2.1.1. Photometric stereo

Photometric stereo aims at estimating the surface normals and reflectance of an object by processing several intensity images obtained under different lighting conditions. The general assumptions are that the projection is orthographic, the camera and objects are stationary and the light sources are far away from the objects.

The image intensity at pixel (u, v) depends on the optical properties of the surface material, the surface shape and the spatial distribution of the incident illumination. The reflectance characteristics of a given surface can be represented by a *reflectance* function ϕ of three unit vectors: surface normal $\mathbf{n} = (n_x, n_y, n_z)^t$, light source direction $\mathbf{m} = (m_x, m_y, m_z)^t$, and viewer

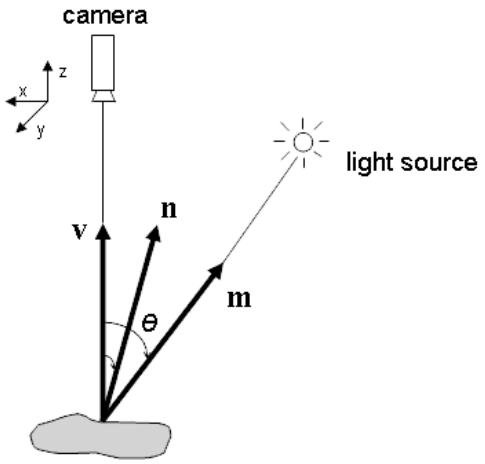


Figure 1: Viewer-oriented coordinate system.

direction $\mathbf{v} = (v_x, v_y, v_z)^t$. Using the reflectance function ϕ, the following equation describes the image-generation process:

$$i = t \cdot \phi(\mathbf{n}, \mathbf{m}, \mathbf{v}) \quad , \tag{1}$$

where t represents the light source intensity associated with each image.

Assuming that the image projection is orthographic and that there is only a distant point light source, the viewer direction and the light source direction can be considered to be constant over the image plane. The coordinate system is considered to be associated with the camera in this paper. That is, the z axis is collinear with the imaging axis of the camera, while the x and y axes are defined by the image plane coordinates, as shown in Fig. 1.

This paper assumes a *Lambertian* reflectance model, which states that materials absorb and reflect light uniformly in all directions. The following equation expresses the intensity at every pixel when this model is considered:

$$i(u, v) = t \cdot \phi(u, v) = t \cdot r(u, v)\mathbf{n}(u, v)^t\mathbf{m} \quad , \tag{2}$$

where $r(u,v)$ is the albedo at pixel (u,v), $\mathbf{n}(u,v)$ is its surface normal and $\mathbf{m}$ represents the light direction associated with each image. The albedo describes the fraction of light reflected at each point on the object.

This linear property suggests the use of *factorization techniques* to model the image formation process and to recover each of the factors that contribute to it. Thus, let I be a measurement matrix of p rows and f columns. Every column contains the gray-levels of the p pixels corresponding to a single image frame. Each of the f frames has been acquired with a different light source position. Assuming a Lambertian reflectance model, this matrix can be factorized as:

$$I = RNMT \ , \tag{3}$$

where

$$R_{p\times p} = \begin{bmatrix} r_1 & & 0 \\ & \ddots & \\ 0 & & r_p \end{bmatrix} \tag{4}$$

is the surface reflectance matrix that contains the surface reflectance at each of the p pixels,

$$N_{p\times 3} = \begin{bmatrix} \mathbf{n}_1 \dots \mathbf{n}_p \end{bmatrix}^t = \begin{bmatrix} n_{1x} & n_{1y} & n_{1z} \\ \vdots & \vdots & \vdots \\ n_{px} & n_{py} & n_{pz} \end{bmatrix} \tag{5}$$

is the surface normal matrix ($\mathbf{n}$ represents the surface normal at each of the p pixels),

$$M_{3\times f} = \begin{bmatrix} \mathbf{m}_1 \dots \mathbf{m}_f \end{bmatrix} = \begin{bmatrix} m_{x1} & \dots & m_{xf} \\ m_{y1} & \dots & m_{yf} \\ m_{z1} & \dots & m_{zf} \end{bmatrix} \tag{6}$$

is the light source direction matrix (**m** represents the light source direction at each of the f frames), and

$$T_{f\times f} = \begin{bmatrix} t_1 & & 0 \\ & \ddots & \\ 0 & & t_f \end{bmatrix} \tag{7}$$

is the light source intensity matrix that contains the light source intensity at each of the f frames. Using these definitions, the surface matrix S and the light source matrix L are defined as follows:

$$S_{p\times 3} = RN, \quad L_{3\times f} = MT \tag{8}$$

Hence, the *measurement* matrix can be decomposed as:

$$I = SL \tag{9}$$

This decomposition can be obtained by using factorization techniques. In particular, the adaptation of the Alternation technique proposed in [9] for the photometric problem is applied in order to obtain the above decomposition (9).

2.1.2. Photometric stereo with adapted Alternation

A common assumption in most photometric stereo approaches is that images do not contain shadows nor saturated regions, which correspond to points with very low and high intensity values respectively. This is due to the fact that these points do not follow a Lambertian model. The robust photometric approach introduced in [9] assumes that these points are missing entries in I. This reduces their influence in the results. That work presents

an adaptation of the Alternation technique, which has been widely studied in computer vision (e.g., [20, 21, 22]), to decompose matrix I. The algorithm is summarized below:

1. Set a lower and an upper threshold to define the shadows and saturated regions respectively, namely: σ_l and σ_u. That is, $I(i,j)$ corresponds to a shadow if $I(i,j) \leq \sigma_l$ and to a saturated pixel if $I(i,j) \geq \sigma_u$.
2. Define the following set:

$$\Omega = \{(i,j), 1 \leq i \leq p, 1 \leq j \leq f | \sigma_l < I(i,j) < \sigma_u\} \tag{10}$$

   This set contains the pixels in I that do not correspond to shadows or saturated regions. Hence, only those pixels that follow a Lambertian model are used during matrix factorization. Therefore, shadows and saturated regions, which correspond to pairs $(i,j) \notin \Omega$, are considered as missing data in I.
3. Apply the Alternation technique to I. The algorithm starts with an initial random $p \times 3$ matrix S_0 (analogously with a $3 \times f$ random matrix L_0) and repeats the next two steps until the product $S_k L_k$ converges to I:
   - Compute L: $\quad L_k = (S_{k-1}^t S_{k-1})^{-1}(S_{k-1}^t I)$
   - Compute S: $\quad S_k = I L_k^t (L_k L_k^t)^{-1}$

     These products are computed by only considering those pixels $I(i,j)$ such that $(i,j) \in \Omega$.

After convergence, every row of S consists of the 3D surface normal associated with each image pixel, every column of L consists of the 3D light

direction and intensity for each frame, and the product SL is the best rank 3 approximation of I. However, the obtained decomposition is not unique, since any invertible matrix Q with size 3×3 gives the following valid decomposition:

$$I = SL = \hat{S}QQ^{-1}\hat{L} \tag{11}$$

Therefore, at the end of the algorithm, one of the two constraints proposed in [12] is used to determinate matrix Q:

1. If the relative value of the surface reflectance is constant or known in at least six pixels, matrix Q can be computed with the following system of p equations:

   $$\hat{s}_k Q Q^t \hat{s}_k^t = \mathbf{c}_1, \quad k = 1, \cdots, p \tag{12}$$

   where $\hat{s}_k$ is the kth-vector of $\hat{S}$ and $\mathbf{c}_1$ the value of the surface reflectance.

2. If the relative value of the light source intensity is constant or known in at least six frames, matrix $P = Q^{-1}$ can be obtained by solving the following system:

   $$\hat{l}_k^t P^t P \hat{l}_k = \mathbf{c}_2, \quad k = 1, \cdots, f \tag{13}$$

   where $\hat{l}_k$ is the kth-vector of $\hat{L}$ and $\mathbf{c}_2$ the value of the light source intensity.

Notice that, at least, six pixels with constant or known reflectance or six images with constant or known light intensity are necessary in order to solve the linear systems (12) and (13), respectively. This is due to the fact that a symmetric 3×3 matrix is computed for each case ($A = QQ^t$ and $B = P^tP$).

If the values of $\mathbf{c}_1$ or $\mathbf{c}_2$ are not known a priori, they are assumed to be one. Therefore, the above constraints impose a constant reflectance at every pixel and a constant light source intensity at every image respectively. In these situations, the reflectance and light source intensity can only be recovered up to a constant. In the experiments described in the current paper, the second constraint (13) is imposed, since there is no need for selecting six pixels with equal or known reflectance.

Even after computing matrix Q, the solution is obtained up to a rotation. Notwithstanding, the recovered normals can be used for image segmentation without estimating that rotation.

2.2. Graph-based image segmentation

At the second stage, the image pixels are clustered according to the orientation of their 3D normals through a graph-based segmentation algorithm presented in [11]. The goal is to group neighbouring pixels that are likely to belong to the same geometric structure. For that purpose, the image pixels with similar 3D normals are clustered together. This second stage consists of two steps: graph creation and graph segmentation.

2.2.1. Graph Creation

In this particular problem of image segmentation, the vertices of the graph are constituted by the 2D pixels of any of the input images. In turn, an edge is defined between every pair of neighbouring pixels, p_i and p_j, with a weight w_{ij}, the latter being some measure of dissimilarity between p_i and p_j. The goal is that similar neighbouring pixels be grouped in a same cluster. In the proposed approach, the weight w_{ij} is defined according to the similarity

between the 3D surface normals computed at the previous stage, $\mathbf{n}_i$ and $\mathbf{n}_j$, respectively:

$$w_{ij} = 1 - e^{-\varphi_{ij}^2/2\sigma^2} \tag{14}$$

where $\varphi_{ij} = \arccos(\mathbf{n}_i^t \mathbf{n}_j)$ and σ is the standard deviation of the Gaussian function $e^{-\varphi_{ij}^2/2\sigma^2}$. Therefore, w_{ij} is zero when the angle φ_{ij} between $\mathbf{n}_i$ and $\mathbf{n}_j$ is zero, and close to one when φ_{ij} is larger than 3σ.

2.2.2. Graph-based segmentation

The proposed segmentation technique is based on a region-growing approach. A measure of distance MInt between every pair of neighbouring regions in the graph, C_i and C_j, is defined. Let us first introduce the *internal difference* of a component $C \in V$, which is the largest weight of the minimum spanning tree of the component, $MST(C, E)$:

$$\text{Int}(C) = \max_{e \in MST(C,E)} w(e), \tag{15}$$

For small components, $\text{Int}(C)$ is not a good estimate of the local characteristics of the data. In the extreme case, when $|C| = 1$, $\text{Int} = 0$. Therefore, Fenzelswalb et al. [11] propose the use of a threshold function based on the size of the component:

$$\tau(C) = k/|C|, \tag{16}$$

where $|C|$ denotes the size of C and k is some constant parameter. That is, for small components, a stronger evidence for a boundary is required. In practice, k defines a scale of observation, in which a larger k leads to larger components being generated. Any non-negative function for a single component can be used for τ.

Therefore, MInt is defined as follows:

$$\mathrm{MInt}(C_i, C_j) = \min(\mathrm{Int}(C_i) + \tau(C_i), \mathrm{Int}(C_j) + \tau(C_j)), \tag{17}$$

The graph-based segmentation algorithm is summarized in five different steps as defined in [11]:

The input is a graph $G = (V, E)$, with n vertices and m edges. The output is a segmentation of V into components $S = (C_1, ..., C_r)$.

1. Sort the edges of the graph, E, in ascending order of weight;
2. Start with a segmentation S^0, where each vertex v_i is in its own component;
3. Repeat step 4 for $q = 1, ..., m$;
4. Construct S^q given S^{q-1} as follows. Let v_i and v_j denote the vertices connected by the q-th edge in the ordering, denoted as e_q. Let C_i^{q-1} be the component of S^{q-1} containing v_i and C_j^{q-1} the component containing v_j.
   - If $C_i^{q-1} \neq C_j^{q-1}$ and $w(e_q) \leq \mathrm{MInt}(C_i^{q-1}, C_j^{q-1})$ then S^q is obtained from S^{q-1} by merging C_i^{q-1} and C_j^{q-1}.
   - Otherwise, $S^q = S^{q-1}$;
5. Return $S = S^m$

3. Experimental evaluation

The aim of this section is to validate the image segmentation approach detailed in Section 2 by comparing it with several image segmentation algorithms. In particular, the proposed approach is compared to both appearance-

based and shape-based image segmentation algorithms. The evaluation study is divided into three parts described next.

First of all, Section 3.1 presents experiments with noisy real data to show the viability of the proposed method. In order to show that other image segmentation algorithms can be used in the second stage of the proposed approach, results obtained with the $K - Means$ method are also reported. Unfortunately, this method requires the definition of the number of desired clusters beforehand. Additionally, results obtained with the method proposed in [16] are also shown.

Next, in Section 3.2, two appearance-based segmentation algorithms are applied to some of the images acquired in the previous experiment. Concretely, the following appearance-based segmentation algorithms have been considered:

1. The Mean Shift [1] algorithm, whose code is publicly available at: http://www.caip.rutgers.edu/riul/research/code.html. The Mean Shift image segmentation is a straightforward extension of the discontinuity preserving smoothing algorithm also proposed in [1]. Each pixel is associated with a significant mode of the joint domain density located in its neighborhood, after nearby modes are pruned.
2. The algorithm presented in [2], in which the texture features are modelled using a mixture of Gaussian distributions. It is a lossy compression clustering algorithm for segmenting degenerate Gaussian distributions. The code is publicly available at http://www.eecs.berkeley.edu/∼yang/software/lossy_segmentation/.

The aim of applying these algorithms is to show that they are not suit-

able when the objective is to segment the surfaces of the different objects contained in the scene, since textured surfaces are oversegmented in general.

Finally, Section 3.3 presents results obtained with three range image segmentation algorithms. Real range images from the OSU data set (available at http://sampl.ece.ohio-state.edu/data/3DDB/RID/index.htm) have been utilized in these experiments. In particular, the following algorithms have been studied:

1. The range image segmentation algorithm developed by the Computer Vision and Image Analysis Research Laboratory at the University of South Florida (http://marathon.csee.usf.edu/). This algorithm is described in [5] and its code is publicly available at: http://marathon.csee.usf.edu/seg-comp/SegComp.html. For simplicity, this algorithm will be denoted as USF. USF works by computing a planar fit for each pixel and then growing regions whose pixels have similar plane equations.
2. The method proposed in [23], which segments a triangular approximation of a given range image.
3. The method proposed in [18], which uses the graph-based segmentation algorithm proposed in [11], by taking the 3D normals estimated through tensor voting instead of the ones estimated with photometric stereo. In these experiments, the 3D data are not obtained with a binocular stereo camera as in [18], but with a range sensor, thus avoiding the 3D noise typical from stereo processing.

The results obtained with these three algorithms are compared with the ones obtained with the proposed approach. However, in order to apply the photometric approach [9], at least six images taken under different lighting

conditions are needed. The problem is the availability of a public data set providing range images together with at least six images obtained under different lighting conditions. Hence, a triangulation algorithm [23] is used to generate triangulated surfaces from the tested range images. Afterwards, an OpenGL process is used to render those surfaces under different lighting positions and obtain the corresponding 2D images. Additionally, results obtained with the appearance-based image segmentation algorithms studied in Section 3.2 are also included in the comparison.

3.1. Photometric stereo segmentation

This section presents the experimental results corresponding to three sets of real noisy images: two of them correspond to indoor scenes taken in our laboratory, while the third set corresponds to an outdoor scene publicly available at: http://www1.cs.columbia.edu/CAVE/software/wild/index.php.

3.1.1. Indoor scenes

In order to capture six images of the same scene under different illumination conditions, a set of six strobelights placed at different positions has been utilized. Fig. 2 and Fig. 6 show two sets of images acquired by changing the active strobelight at a time. It can be noticed that the acquired images are very noisy. Since the strobelight positions are very close to each other, the variation of grey-level intensity at each pixel is very low. Moreover, shadows and saturated pixels can appear in the images. Nevertheless, even under these conditions, the obtained results are satisfactory as shown below.

Fig. 3 (a) and (b) show the reflectance and 3D surface normals estimated for every pixel when the photometric technique is applied to the images shown

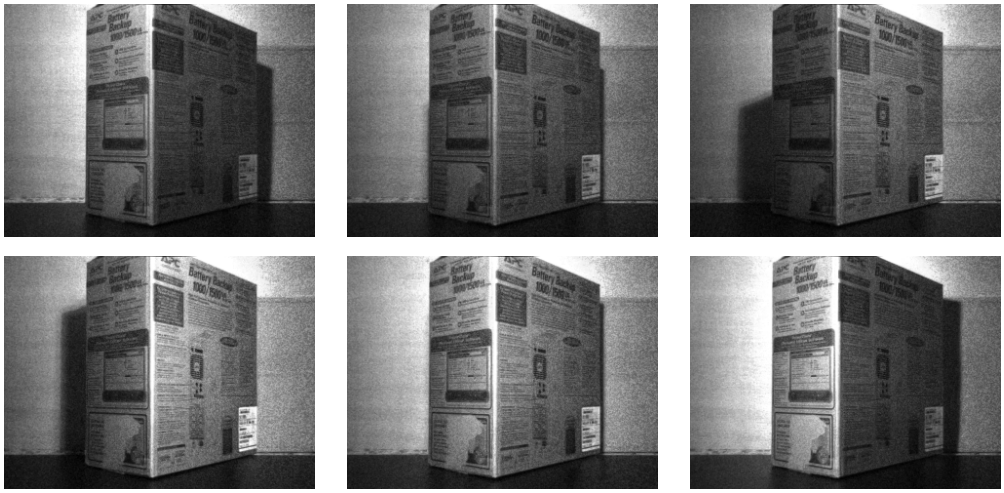


Figure 2: First set of six images obtained by alternatively switching on each of the strobelights.

in Fig. 2. In particular, Fig. 3 (b) shows the angle between each normal and the camera axis, which is perpendicular to the image plane. Dark pixels correspond to small angles, whereas bright pixels correspond to big angles. The surface normals corresponding to background pixels (and, in particular, the ones corresponding to the floor) are very noisy.

(a) (b)

Figure 3: (a) Recovered reflectance; (b) angle between the recovered normals and the camera axis, which is perpendicular to the image plane.

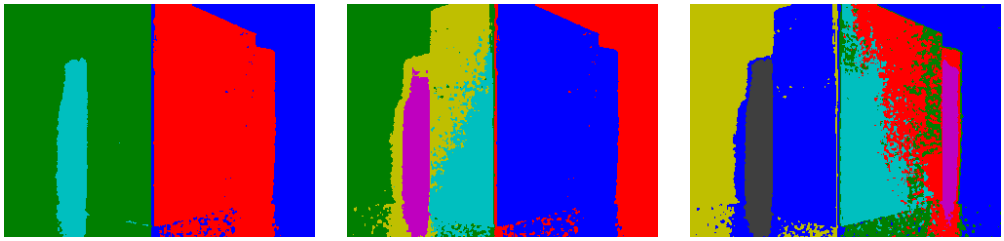
Once the normals have been computed, a graph is created as described in Section 2.2, and the method presented in [11] is applied. Fig. 4 shows the image segmentation obtained when different values for parameter k, which defines the scale of the obtained components (see eq. (16)), are considered. In particular, the results correspond to $k = 1$, 4 and 7, respectively. It can be seen that the best segmentation is obtained when $k = 4$ (Fig. 4 (b)) and that the scene is, in general, correctly segmented based on the shape of the objects contained in it. Only a small region of background pixels is wrongly segmented and joined together with a face of the object. Recall that only six images are used. With such a low number of images, the method cannot deal with shadows or saturated regions, which can be present in the background, probably leading to some inaccurate results. Those points could be considered as missing data if more images were available, as in [9]. However, this problem is out of scope of this paper. As mentioned in Section 1, the main goal of the proposed approach is to use a low number of images in order to obtain a simple technique that can be applied on a mobile robot.

(a) (b) (c)

Figure 4: Image segmentation obtained with the proposed approach for different values of k (see eq. (16)): (a) $k = 1$; (b) $k = 4$; (c) $k = 7$.

Fig. 5 shows the image segmentation obtained when $K-Means$ is used in

the second stage of the algorithm. In particular, these results correspond to the cases in which $K = 4$, 6 and 8, respectively. Notice that the graph-based method yields a better image segmentation.



(a) (b) (c)

Figure 5: Image segmentation obtained with $K-Means$ for different values of K: (a) $K = 4$; (b) $K = 6$; (c) $K = 8$.

Fig. 6 shows a second set of images obtained by considering a scene with several objects (including non-planar objects) and a variable background. Fig. 7 (b) shows that the 3D normals corresponding to background pixels are even noisier than in the previous experiment.

Some constraints should be imposed when dealing with such noisy normals. Otherwise, an oversegmentation of the scene is likely to be produced. This paper proposes a measure to decide when the normals have been properly estimated. The idea is to penalize pixels that present a low grey level intensity variation along the images. The *mean deviation* (MD)[1] is used to study the variation of the grey level intensity at every pixel over different images. The surface normals, $\mathbf{n}$, corresponding to pixels with a MD value below a given threshold (τ_{MD}) will be set to $\mathbf{n} = (0, 0, 0)$ before the segmentation

[1] $MD(X) = \frac{1}{N}\sum_{j=1}^{N}|x_j - \bar{x}|$, being $X = (x_1, ..., x_N)$ and $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

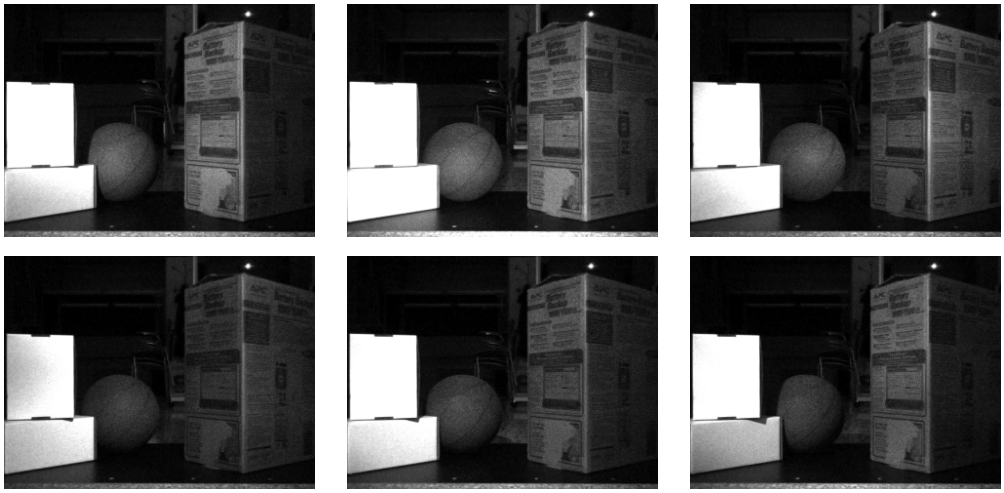


Figure 6: Second set of images obtained by alternatively switching on each of the strobelights.

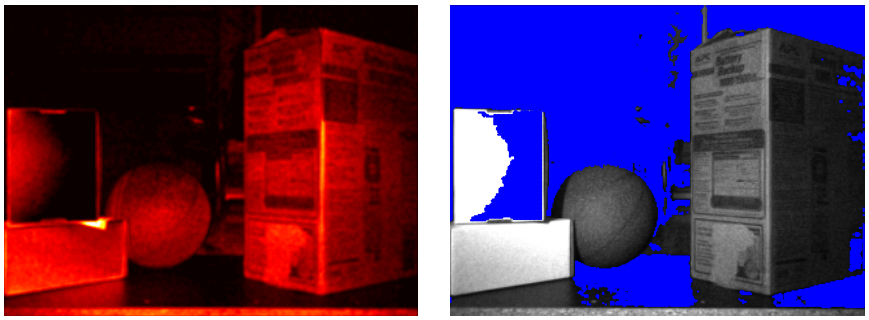
(a) (b)

Figure 7: (a) Recovered reflectance; (b) angle between the recovered normals and the camera axis, which is perpendicular to the image plane.

stage.

Fig. 8 (a) shows the MD value of every pixel for the images shown in Fig. 6. Darker pixels correspond to those with lower MD values or, which is the same, to pixels with a low grey-level intensity variation over the images of the sequence. Threshold τ_{MD} is experimentally set to 4 in this particu-

lar example. Those pixels whose MD is below τ_{MD} are marked in blue in Fig. 8 (b).



(a) (b)

Figure 8: (a) MD value of every pixel; (b) pixels with a MD value below $\tau_{MD} = 4$ are marked in blue.

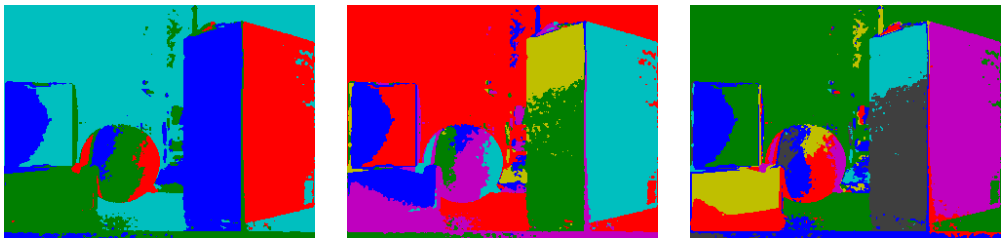
Fig. 9 shows the final segmented image when the graph-based image segmentation algorithm is used in the second stage of the proposed approach, by taking different values of parameter k. Again, the best segmentation is obtained when $k = 4$ (Fig. 9 (b)). Notice that pixels corresponding to a same face are segmented as a single region in general.

(a) (b) (c)

Figure 9: Image segmentation obtained with the proposed approach, for different values of k (see eq. (16)): (a) $k = 1$; (b) $k = 4$; (c) $k = 7$.

The image segmentation obtained when $K - Means$ is applied in the

second stage of the proposed approach is shown in Fig. 10. The results correspond to some of the tested K values: $K = 4$, $K = 6$ and $K = 8$, respectively. The results obtained with the graph-based segmentation algorithm are always better than the ones obtained with the $K - Means$ method.



(a) (b) (c)

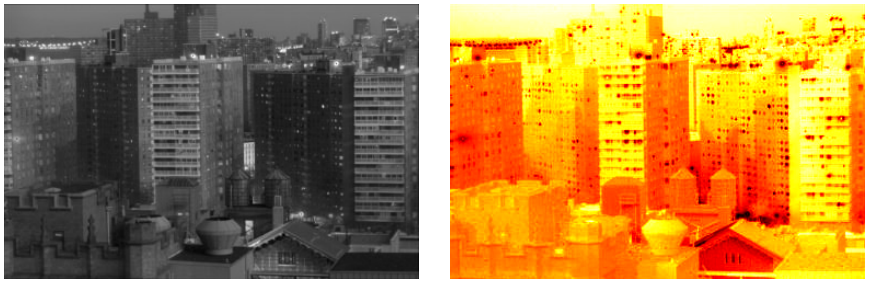
Figure 10: Image segmentation obtained with $K-Means$ for different values of K: (a) $K = 4$; (b) $K = 6$; (c) $K = 8$.

3.1.2. Outdoor scene

Fig. 11 shows 6 of the 24 images used in this experiment. They were extracted from the WILD Dataset and correspond to the particular set *"January, number 25"*. This is a very difficult sequence: on the one hand, notice that there is a high amount of both shadows and saturated points in the images. On the other hand, some of the images are very dark. The recovered reflectance and 3D surface normals estimated at every pixel are shown in Fig. 12 (a) and (b) respectively.

Fig. 13 shows the image segmentation obtained with the proposed approach when different values of k are considered. The results obtained when $K-Means$ is used in the second stage of the proposed approach are shown in Fig. 14. Finally, the approach proposed in [16], whose code is publicly available at: http://sites.google.com/site/koppaldev/code, is also applied, with

Figure 11: Images from the WILD Dataset.



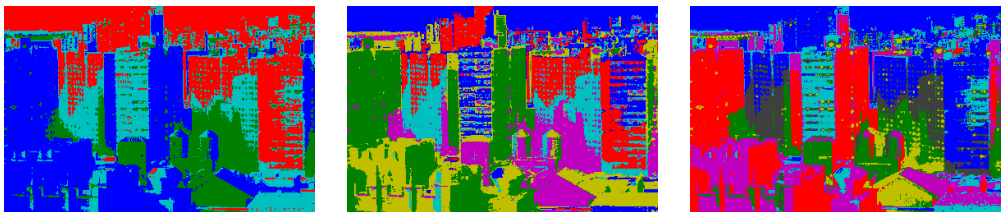
(a) (b)

Figure 12: (a) Recovered reflectance; (b) angle between the recovered normals and the camera axis, which is perpendicular to the image plane.

the obtained segmentation being shown in Fig. 15. Different values of parameter K are considered in both methods. Notice that the results obtained with $K - Means$ and with the method proposed in [16] slightly differ from the ones shown in [16]. This can be due to the fact that a different set of images may have been utilized. Koppal et al. [16] do not provide an accurate information about the particular images they consider, nor the value of K used in $K - Means$.

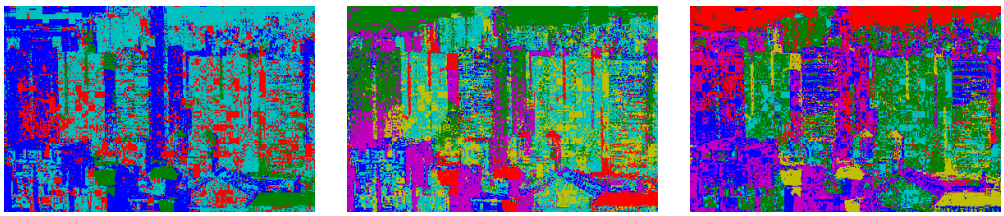
(a) (b) (c)

Figure 13: Image segmentation obtained with the proposed approach for different values of k (see eq. (16)): (a) $k = 1$; (b) $k = 4$; (c) $k = 7$.



(a) (b) (c)

Figure 14: Image segmentation obtained with $K-Means$ for different values of K: (a) $K = 4$; (b) $K = 6$; (c) $K = 8$.



(a) (b) (c)

Figure 15: Image segmentation obtained with the method proposed in [16] for different values of K: (a) $K = 4$; (b) $K = 6$; (c) $K = 8$.

3.2. Appearance-based segmentation

This section presents the results obtained when the image segmentation is based on the visual appearance of the depicted objects. Two different

algorithms are tested: the Mean Shift [1] and the lossy compression-based clustering algorithm proposed in [2].

3.2.1. Mean Shift

Fig. 16 (b) and (c) show the image segmentation obtained when Mean Shift is applied to the image shown in Fig. 16 (a), by using different parameter values. In particular, the image segmentation obtained with $(h_s, h_r, M) = (4, 1, 400)$ is shown in Fig. 16 (b), while the image segmentation obtained with $(h_s, h_r, M) = (3, 2, 100)$ is shown in Fig. 16 (b). Recall that h_s and h_r determine the spatial and range domain resolution, whereas M is the minimum number of pixels contained in each region. The image shown in Fig. 16 (a) corresponds to one of the 2D images used in the previous experiment, Fig. 2 (bottom-center). It can be seen that an oversegmentation is obtained in both cases due to the textured nature of the box and to shadows and highlights in the image, which can cause the oversegmentation of the same surface.

(a) (b) (c)

Figure 16: (a) Input image; (b) image segmentation with Mean Shift [1], $(h_s, h_r, M) = (4, 1, 400)$; (c) $(h_s, h_r, M) = (3, 2, 100)$.

Fig. 17 (b) and (c) show the segmentation obtained when Mean Shift [1] is applied to the image shown in Fig. 17 (a), which corresponds to the image

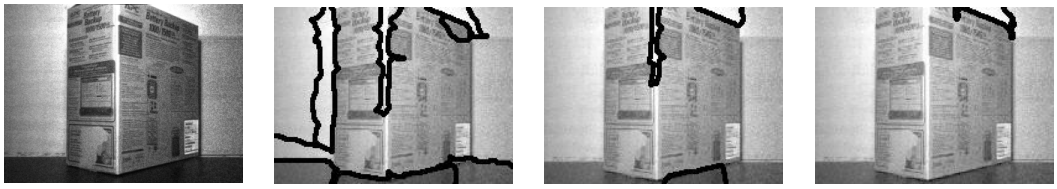
in Fig. 6 (top-left). Similarly to the previous experiment, the image segmentation obtained with $(h_s, h_r, M) = (4, 1, 400)$ is shown in Fig. 17 (b), while the image segmentation obtained with $(h_s, h_r, M) = (3, 2, 100)$ is shown in Fig. 17 (b). Again, an oversegmentation is obtained for both tested parameter values. The problem is that the obtained regions do not correspond to the faces of the depicted objects, which is the main goal of the current work.

(a) (b) (c)

Figure 17: (a) Input image; (b) image segmentation with Mean Shift [1], $(h_s, h_r, M) = (4, 1, 400)$; (c) $(h_s, h_r, M) = (3, 2, 100)$.

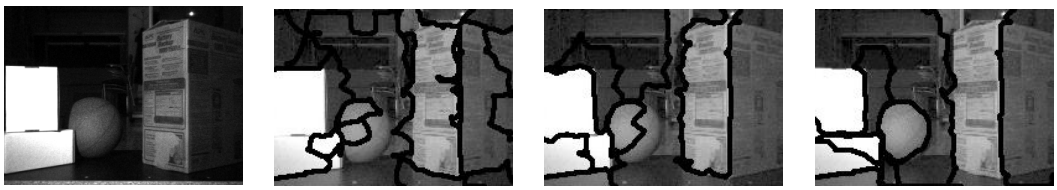
3.2.2. Lossy compression-based clustering algorithm

Fig. 18 and Fig. 19 show the results obtained when the algorithm proposed in [2] is applied to the images shown in Fig. 18 (a) and Fig. 19 (a) respectively. Different segmentations are obtained depending on the value of parameter ε. It can be seen that the obtained regions do not correspond to faces of the objects in general. Actually, the segmented regions contain background points together with faces of the objects, which is not useful when the goal is to segment the objects or surfaces present in the scene. From the results obtained in this section, it can be concluded that appearance-based segmentation is not suitable when the goal is to segment the surfaces of the objects contained in the scene.



(a) (b) (c) (d)

Figure 18: (a) Input image; (b-d) segmentation with the lossy compression-based clustering algorithm [2] under different ε's ($\varepsilon = 0.1$, $\varepsilon = 0.2$, $\varepsilon = 0.3$, respectively).



(a) (b) (c) (d)

Figure 19: (a) Input image; (b-d) segmentation with the lossy compression-based clustering algorithm [2] under different ε's ($\varepsilon = 0.1$, $\varepsilon = 0.2$, $\varepsilon = 0.3$, respectively).

3.3. Range image segmentation

Range image segmentation is the alternative to appearance-based image segmentation. In this case, the segmentation is based on the shape of the depicted objects. Real range images from the OSU data set (http://sampl.ece.ohio-state.edu/data/3DDB/RID/index.htm) are used in this section. In particular, the range images shown in Fig. 20 (a) and Fig. 24 (a) are studied. This section summarizes the results obtained with all the image segmentation algorithms studied in the current paper.

Fig. 20 shows the segmentation obtained when the three studied range image segmentation algorithms are applied to the range image data depicted in Fig. 20 (a). USF does not yield good results for non-planar regions, as can

be seen in Fig. 20 (b). Furthermore, there are some regions (marked in white) where the algorithm cannot be applied (namely *missed classification*). This algorithm tends to oversegment the image. The image segmentation obtained with the Garcia and Basañez's [23] algorithm is shown in Fig. 20 (c). The problem with this algorithm is that it segments the triangular approximation of the range image instead of that image. Therefore, the segmentation is less realistic. Finally, the segmentation obtained when the method proposed in [18] is applied is shown in Fig. 20 (d).

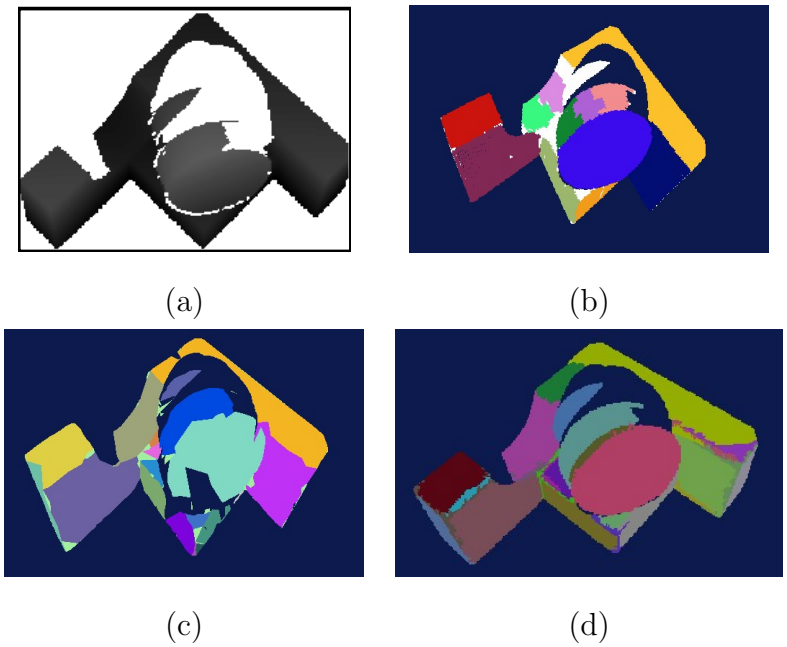


Figure 20: (a) Input range image; (b) segmentation with USF [5]; (c) segmentation with Garcia et al. [23]; (d) segmentation with [18].

Fig. 21 shows six 2D images obtained with the aforementioned OpenGL process, which renders the surface shown in Fig. 20 (a) under different lighting conditions. It can be seen that only six images (the minimum necessary)

are generated considering light source positions very close to each other. Recall that the knowledge of these light source positions is not used by the proposed algorithm. Hence, the variation of grey-level intensity at each pixel is very low. Nevertheless, the results obtained with the proposed approach are satisfactory, as shown in Fig. 22 (a). This segmentation corresponds to the case in which $k = 4$ (see eq. (16)). Black points in the segmented images correspond to points whose depth is not known (the same points are marked in white in the corresponding range image). Recall that the obtained segmentation is similar to the one obtained with [18] (see Fig. 20 (d)). However, it should be highlighted that the 3D data are necessary in [18], whereas only 2D images are required in the proposed approach. Additionally, the results obtained when $K - Means$ is applied in the second stage of the proposed approach are shown in Fig. 22 (b). Only the best obtained image segmentation is shown ($K = 4$).

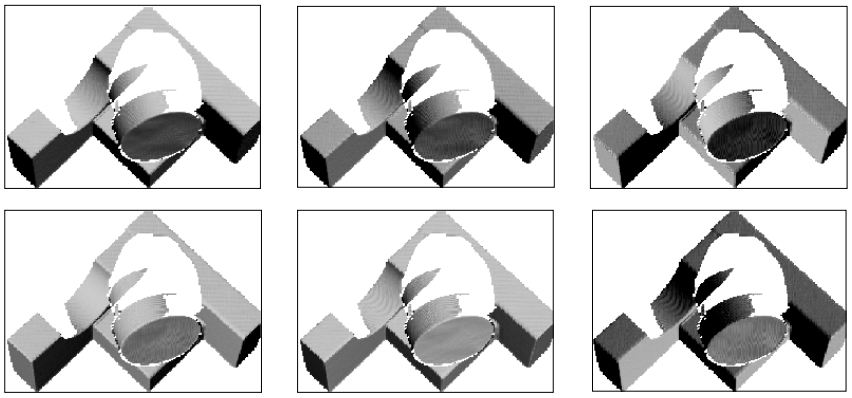
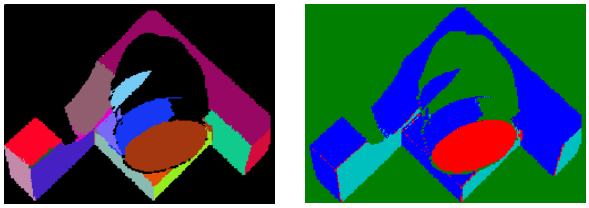


Figure 21: Input 2D images for the photometric stereo approach.

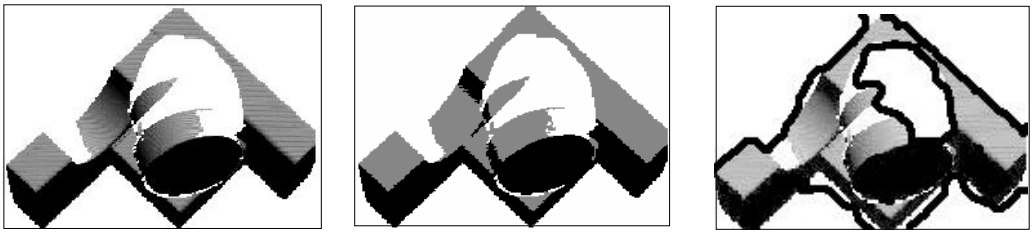
Fig. 23 shows the results obtained when the two appearance-based algorithms studied in the previous section are applied to the 2D image shown



(a) (b)

Figure 22: Image segmentation obtained: (a) with the proposed approach, $k = 4$; (b) with K-means, $K = 4$.

in Fig. 23 (a). Mean Shift yields a quite good image segmentation, as can be seen in Fig. 23 (b). Recall that the object has a single texture over all its surface. The segmentation obtained with the algorithm proposed in [2] is shown in Fig. 23 (c). This algorithm yields regions that contain pixels of the object's surface along with pixels belonging to the background (the best segmentation result is obtained with $\varepsilon = 0.4$).



(a) (b) (c)

Figure 23: (a) Input 2D image; (b) segmentation with Mean Shift [1]; (c) segmentation with [2].

The second studied image is depicted in Fig. 24 (a). Fig. 24 (b) shows the segmentation obtained with the USF algorithm, while the segmentation obtained with the Garcia and Basañez's [23] algorithm is shown in Fig. 24 (c). The approach proposed in [18] yields the segmentation shown in Fig. 24 (d).

It can be seen that an oversegmentation is obtained with the three studied algorithms.

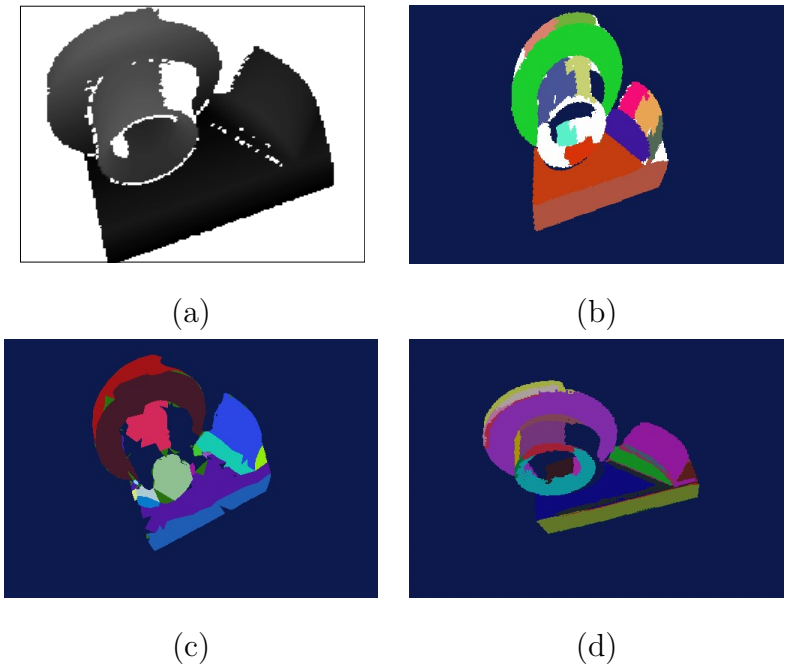


Figure 24: (a) Input range image; (b) segmentation with USF [5]; (c) segmentation with Garcia et al. [23]; (d) segmentation with [18].

As in the experiment with the first range image, an OpenGL process is used to render the surface shown in Fig. 24 (a) under different lighting conditions in order to obtain the six 2D images depicted in Fig. 25. The segmentation obtained with the proposed approach ($k = 4$) is shown in Fig. 26 (a). Fig. 26 (b) shows the segmentation obtained when $K - Means$ is applied in the second stage of the proposed approach. In this case, the best result is obtained when $K = 6$.

Finally, Fig. 27 shows the results obtained when the two appearance-based algorithms studied in the previous section are applied to the 2D image

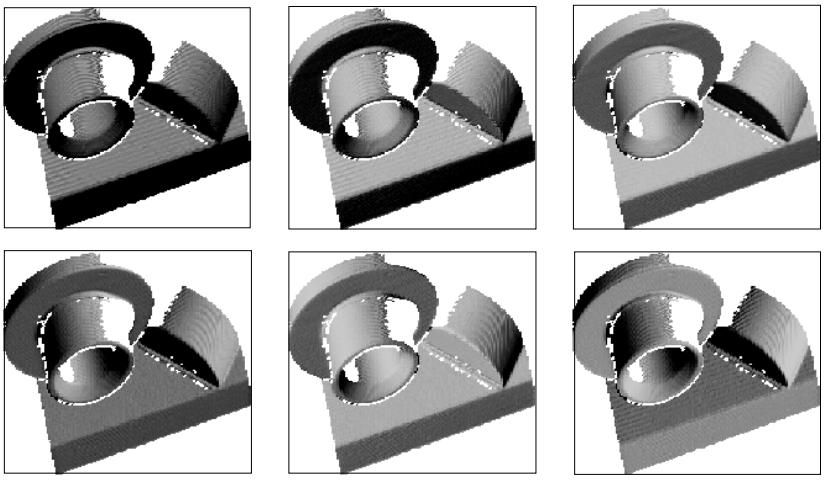
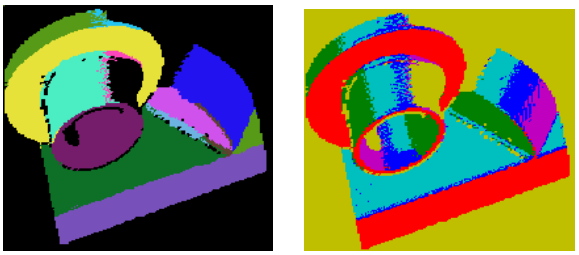


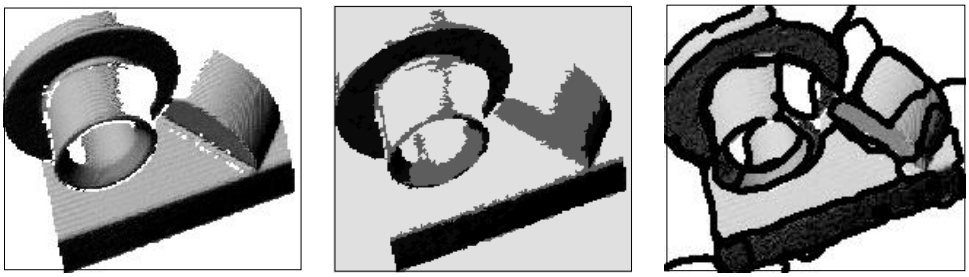
Figure 25: Input 2D images for the photometric stereo approach.



(a) (b)

Figure 26: Image segmentation obtained: (a) with the proposed approach, $k = 4$; (b) with K-means, $K = 6$.

shown in Fig. 27 (a). Particularly, the segmentation obtained with the Mean Shift algorithm is shown in Fig. 27 (b), while the segmentation obtained with the algorithm proposed in [2] can be seen in Fig. 27 (c). Again, the best segmentation is obtained with $\varepsilon = 0.4$. Notice that the proposed approach yields the best image segmentation in these experiments, as can be appreciated in Fig. 22 and Fig. 26.



(a) (b) (c)

Figure 27: (a) Input 2D image; (b) segmentation with Mean Shift [1]; (c) segmentation with [2].

4. Conclusion

This paper presents a detailed description and extensive experimental validation of a shape-based image segmentation algorithm previously introduced by the authors. The proposed algorithm consists of computing the surface normals of the objects present in the scene by applying a robust photometric stereo approach. Afterwards, the image pixels with similar normals are clustered together by using a graph-based image segmentation algorithm.

The proposed image segmentation approach is compared with several existing image segmentation algorithms, both appearance-based and shape-based ones. Experimental results show that the shape-based image segmentation algorithms are more suitable than the appearance-based ones when the objective is to segment the objects or surfaces present in the scene. In particular, the proposed algorithm yields the best image segmentation results in most cases.

Future lines of research include the study of other weight definitions in the graph generation step. In addition, other image segmentation algorithms and

strategies can be tested. Finally, once the image segmentation is obtained, the surfaces of the objects present in the scene can be reconstructed by using the estimated surface normals.

Acknowledgment

This work has partially been supported by the Spanish Government under project DPI2007-66556-C03-03, by the Commissioner for Universities and Research of the Department of Innovation, Universities and Companies of the Catalonian Government and by the European Social Fund.

References

[1] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 603–619.

[2] A. Y. Yang, J. Wright, Y. Ma, S. S. Sastry, Unsupervised segmentation of natural images via lossy data compression, Computer Vision and Image Understanding 110 (2008) 212–225.

[3] L. Bertelli, B. Sumengen, B. S. Manjunath, F. Gibou, A variational framework for multiregion pairwise-similarity-based image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 1400–1414.

[4] P. Arbeláez, M. Maire, C. Fowlkes, J. Malik, Fom contours to regions: an empirical evaluation, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 2294–2301.

[5] A. Hoover, G. Jean-Baptiste, J. X., P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, R. Fisher, An experimental comparison of range image segmentation algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (1996) 673–689.

[6] X. Jiang, K. Bowyer, Y. Morioka, S. Hiura, K. Sato, S. Inokuchi, M. Bock, C. Guerra, R. Loke, J. du Buf, Some further results of experimental comparison of range image segmentation algorithms, in: International Conference on Pattern Recognition (ICPR), 2000, pp. 877–881.

[7] O. Wirjadi, Survey of 3d image segmentation methods, Tech. rep., Institut Techno-und Wirtschaftsmathematik (2007).

[8] C. Julià, R. Moreno, D. Puig, M. Garcia, Image segmentation through graph-based clustering from surface normals estimated by photometric stereo, Electronics Letters 46 (2010) 134–135.

[9] C. Julià, A. Sappa, F. Lumbreras, J. Serrat, A. López, Photometric stereo through an adapted alternation approach, in: IEEE International Conference on Image Processing, 2008, pp. 1500–1503.

[10] J. Woodham, Photometric method for determining surface orientation from multiple images, Optical Engineering 19 (1980) 139–144.

[11] P. Fenzelswalb, D. Huttenlocher, Efficient graph-based image segmentation, International Journal of Computer Vision (IJCV) 59 (2004) 167 –181.

[12] H. Hayakawa, Photometric stereo under a light source with arbitrary motion, Optical Society of America 11 (1994) 3079–3089.

[13] R. Basri, D. Jacobs, I. Kemelmacher, Photometric stereo with general, unknown lighting, International Journal of Computer Vision 72 (2007) 239–257.

[14] B. Shi, Y. Matsushita, Y. Wei, C. Xu, P. Tan, Self-calibrating photometric stereo, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[15] S. Koppal, S. Narasihmhan, Clustering appearance for scene analysis, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2006, pp. 1323–1330.

[16] S. Koppal, S. Narasihmhan, Appearance derivatives for isonormal clustering of scenes, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 1375–1385.

[17] A. Yuille, D. Snow, R. Epstein, P. Belhumeur, Determining generative models of objects under varying illumination: shape and albedo from multiple images using SVD and integrability, International Journal of Computer Vision 35 (1999) 203–222.

[18] R. Moreno, M. Garcia, D. Puig, Graph-based perceptual segmentation of stereo vision 3d images at multiple abstraction levels, in: International Workshop on Graph-based Representations in Pattern Recognition (GbRPR), LNCS 4538, 2007, pp. 148–157.

[19] G. Medioni, M. Lee, C. Tang, A computational framework for feature extraction and segmentation, Elsevier Science.

[20] R. Guerreiro, P. Aguiar, Estimation of rank deficient matrices from partial observations: two-step iterative algorithms, in: Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR), 2003, pp. 450–466.

[21] R. Hartley, F. Schaffalitzky, Powerfactorization: 3D reconstruction with missing or uncertain data, in: Australian-Japan advanced workshop on Computer Vision, 2003.

[22] A. Buchanan, A. Fitzgibbon, Damped Newton algorithms for matrix factorization with missing data, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2, University of Oxford, 2005, pp. 316–322.

[23] M. Garcia, L. Basañez, Fast extraction of surface primitives from range images, in: International Conference on Pattern Recognition, Applications and Robotic Ssytems, 1996, pp. 568–572.