Depth estimation in Integral Imaging based on a maximum voting strategy

Adolfo Martínez-Usó, Pedro Latorre-Carmona, Jose M. Sotoca, Filiberto Pla
Departament de Llenguatges i Sistemes Informàtics
Universitat Jaume I, Castelló (Spain)

Bahram Javidi
Department of Electrical and Computer Engineering
University of Connecticut, Storrs, CT 06269-4157
(USA).

February 26, 2017

Abstract

An approach that uses the scene information acquired by means of a 3D Synthetic Aperture Integral Imaging system is presented. This method generates a depth map of the scene through a voting strategy. In particular, we consider the information given by each camera of the array for each pixel, and also the local information in a neighbourhood of that pixel. The proposed method obtains consistent results for any type of object surfaces as well as very sharp boundaries.

In addition, we also contribute in this paper with a repository of a set of synthetic integral images generated by 3DS Max where the so called ground truth (real-true depth map) is available. This resource can be used as a benchmark to test any Integral Imaging based range estimation method.

1 Introduction

Three-dimensional (3D) optical image sensing and visualization technologies are currently applied in areas like TV broadcasting, 3D displays, entertainment, medical sciences and robotics [3,31,38]. An advantage of 3D in relation to traditional 2D imaging techniques is their capability to capture the spatial (structural) information of different objects that are in a scene. One promising 3D approach is based on Integral Imaging or Integral Photography, which is an autostereoscopic imaging method used to capture 3D information and to visualize it in 3D space, either optically or computationally [2,5,20,28,32,48].

Integral Imaging operates under incoherent or ambient light. This is an advantage as compared to other sensing techniques like holography or Ladar,

which require an active illumination system [6,16]. Integral Imaging can also provide the 3D profile and range of the objects in the scene, being therefore attractive for 3D object recognition [26,36]. Three-dimensional sensing with an Integral Imaging architecture has specific benefits over conventional 2D imaging as well as stereo imaging. The number of cameras, the total number of pixels and the sensing parallax which are optimum for image pick up depend on a number of parameters, including object size, distance, and depth. However, and considering these common restrictions, the use of Integral Imaging proves advantageous and may outperform 2D imaging and stereo imaging for specific applications such as segmentation of objects from heavy background, and imaging through obscuration and scattering medium (see e.g. [10–12, 18, 37] for details).

Authors in [14] propose a methodology to estimate the depth of objects in a scene using a minimum variance principle approach (hereafter *Min-Var* method). This is a well-known approach in Integral Imaging (II) literature. Spectral Radiation Pattern (SRP) in object space can be used to establish the relationship to different perspective images and thus infer depth of Lambertian surfaces. In this work, the statistical variance of a SRP function is defined on each voxel[1] from each camera. On the basis that the higher the correlation among pixels of different cameras, the most likely that information comes from an object surface point, the approach selects those z values from a range, $z \in [Z_{min} \dots Z_{max}]$, where the variance among these voxels is minimum. The main drawbacks of this methodology is that *i)* the accuracy of depth estimation is sensitive to the types of object surfaces (no Lambertian surfaces), presenting artifacts on shiny objects or light reflexes. Moreover, *ii)* this methodology is especially noisy at the boundaries of the objects [45].

In this paper, we present an approach that uses the scene information acquired by means of a 3D Integral Imaging system and generates a depth map of the scene through a voting strategy. This voting strategy is performed on the basis of the local information for each pixel, obtaining thus consistent results for any type of object surfaces and showing very sharp boundaries. In addition, we also contribute in this paper with a repository of a set of synthetic integral images generated by the 3DS Max[2] where the ground truth of the depth on each pixel is available. This resource can be used for performing Integral Imaging and Range Estimation.

The rest of this paper is organised as follows. Section 2 provides an overview of 3D Integral Imaging. Section 3 explains the new methodology proposed in this paper to do robust depth estimation. Section 4 shows the results applied on a series of real and synthetic integral imaging scenes. Section 5 discusses the results obtained and conclusions are given in Section

[1] Besides the x, y coordinates of the pixel, depth is used as the third coordinate.
[2] Autodesk 3DS Max 2015

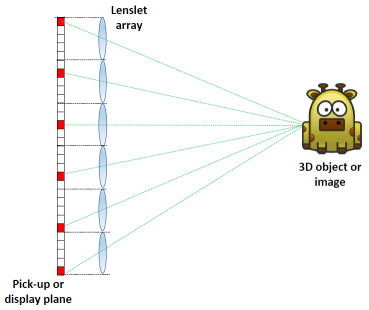

Figure 1: Principle of Integral Imaging capture and display. The object depth information is encoded into lateral relative shift between different views of the scene known as elemental images.

6.

2 Overview of 3D Integral Imaging

The historical origins of integral photography may be linked to Sir Wheatstone invention of stereoscopic viewing device which uses the principle of disparity for 3D visualization [44]. With some exceptions [8, 21, 22, 34, 39], there was no substantial activity in integral photography for most of the twentieth century due to the lack of technologies that may be able to make this acquisition and visualization strategy, a reality. With the important advances in optical detectors such as CCD and CMOS technologies, display devices such as Liquid Crystal Devices (LCDs), consumer electronics, and computer hardware, II has been significantly boosted. II has been applied to a wide variety of fields [4, 19, 24, 25, 29, 30, 33, 40, 41], and it is considered as a promising technology for 3D acquisition and visualisation [1, 46].

2.1 Capture and display stages of Integral Imaging

Acquisition and visualization of 3D objects using II can be divided into two different stages, called pickup and reconstruction. Figure 1 illustrates the principle of II for both acquisition and display purposes. In the pickup stage, multiple 2D images (referred to as elemental images) are captured through an array of small lenses (lenslet array) or an array of cameras. Each lenslet contains a unique projective transformation that maps the 3D object space onto 2D elemental images. As a result, an array of inverted real images is formed on the image sensor. In the reconstruction stage, 3D scene reconstruction can be made optically or computationally. Computational reconstruction of the 3D image can be achieved by simulating the optical back-projection of elemental images in the computer. This reconstruction

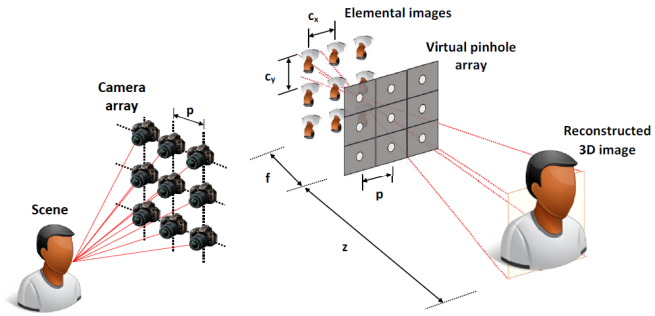

Figure 2: Synthetic aperture Integral Imaging principle. Left part of the figure shows the acquisition strategy using a camera array (or a camera moving on a grid). Right part of the figure shows 3D scene reconstruction by back projecting the elemental images through virtual pinholes. Each elemental image is back-projected through its own viewpoint and the superposition of the ray cones projected from the elemental images reconstructs the 3D scene.

method uses a computer synthesized virtual pinhole array for inverse mapping of each elemental image into the object space. All elemental images are computationally overlapped afterwards. With this process, the intensity distribution can be reconstructed at arbitrary planes inside the 3D object volume.

There are alternative ways to capture 3D information to the lenslet-based approach. For instance, an array of image sensors distributed in an homogeneous or (alternatively) random grid such as Synthetic Aperture Integral Imaging (SAII) [23]. The acquisition strategy used in this work (Figure 2(left)) is SAII in a homogeneous distributed case. The superposition of properly shifted elemental images provides the 3D reconstructed images as follows [17] (Figure 2(right)):

$$I(x,y,z) = \frac{1}{O(x,y)} \sum_{i=0}^{K-1} \sum_{j=0}^{L-1} E_{ij}\left[x', y'\right], \tag{1}$$

with

$$x' = x - k\frac{N_x \cdot p}{c_x \cdot M},$$

$$y' = y - l\frac{N_y \cdot p}{c_y \cdot M},$$

where $I(x,y,z)$ represents the intensity of the reconstructed $3D$ image at depth z, x and y are the indexes of the pixel, E_{ij} represents the intensity of the ith row $(0 \leq i < K)$ and jth column $(0 \leq j < L)$ elemental image,

$N_x \times N_y$ is the total number of pixels for each elemental image, M is the magnification factor $\left(M = \frac{z}{f}\right)$, $c_x \times c_y$ is the physical size of the camera sensor, $O(x, y)$ is the overlapping number matrix, p is the pitch of the cameras.

The proposed II approach is also intrinsically different from stereo vision approaches in the sense that a higher number of cameras are used, and they are usually located in a rectangular or square grid, providing also some straightforward properties for the visualisation of 3D scenes with full parallax and the capability to deal with occlusions [46]. In II the photo-consistency assumption can be better exploited due to the rather large number of views and the geometrical setup, which allow to deal with the depth discontinuities and partial occlusions.

2.2 Depth estimation using variance minimisation

A methodology to estimate the depth objects in a scene using a minimum variance criterion is proposed in [14]. Firstly, by utilizing all the elemental images and back-projection technique, 3D objects are volumetrically reconstructed by means of a stack of 3D planes with the weighted average of all the overlapped and shifted elemental images.

Let us suppose a Spectral Radiation Pattern (SRP) function described as $\mathcal{L}(\theta, \phi, \lambda)$, that is the radiation intensity for a certain point in a 3D space (x, y, z), where (θ, ϕ) determine the radiation intensity direction and λ is the wavelength. The statistical variance of this function $\mathcal{L}$ for a voxel (x, y, z) is defined as

$$V(x, y, z) = \frac{\sum_{c=1}^{3} \sum_{k=0}^{K-1} \sum_{l=0}^{L-1} \left[\mathcal{L}(\theta_{kl}, \phi_{kl}, \lambda_c) - \bar{\mathcal{L}}(\theta, \phi, \lambda_c)\right]^2_{(x,y,z)}}{3}$$

for certain image sensors located in a $K \times L$ plane. $\bar{\mathcal{L}}$ denotes the mean of the SRP functions over all the image sensors and the variance is the average over the three colour channels (c). If a 3D point belongs to an object surface, then the SRP samples are expected to be correlated with each other and the variance V should reach a minimum along all the possible depth levels (Z). Formally,

$$\hat{z}(x, y) = arg \min_{z \in Z} V(x, y, z)$$

However, depth estimation accuracy will degrade when object surfaces do not satisfy the Lambertian assumption or when concave surfaces exist. In addition, sensor position uncertainties may also contribute to the increase in the depth estimation error [45].

3 Photoconsistency-based maximum voting approach

In this section, a voting process to estimate depth information in 3D II is described. Despite state-of-the-art approaches that base this estimation on minimising the variance for each pixel at each depth level [14,45], in this approach a soft-voting procedure that takes into account the level of agreement (similarity) among the different camera views is proposed (hereafter *Max-Voting* method), more in the line of the approaches presented in [15,43]. The rationale behind this voting process is the fact that, when an object is in focus at a certain depth level z, the pixels that form part of the object also appear sharper or clearer. This means that the cameras agree on focusing that object or, in other words, the pixel intensity values from each camera view are very similar, that is, they are photoconsistent. On the contrary, when an object is not in focus, its pixels are blurred in the II reconstruction, which means that each camera is focusing on a different depth, being the pixel intensity values from each camera view very different.

Let us consider an II reconstruction process where, at a certain depth level $z \in Z$ (being Z all the possible depth levels for each pixel of the scene), the pixel at the position (i, j) of the image I and its square surrounding window W are defined as follows:

$$W_{ij} = \{\ I(i+x, j+y)\ :\quad -\tau \leq x, y \leq \tau\ \}$$

where τ directly relates to the size of the window W.

Let us suppose an odd and squared rack of cameras C, being $\|C\|$ the number of cameras whose central camera is $R \in C$. Let us also suppose I as the reconstructed image at the depth level z. For each pixel (i, j) and its neighbouring pixels (x, y) within the window W_{ij} (i.e. $\forall (x, y) \in W_{ij}$), we propose a photoconsistency criterion based on a voting procedure where each camera votes in favour of the pixel (i, j) at depth level z depending on how similar the intensities of the pixels of each camera $C^k \in C$ and the reference camera R pixel intensities are. Note that we suppose an odd and squared rack of cameras in order to assume the central camera as an obvious reference, however, it can be easily generalised to any other camera as a reference.

Therefore, a camera votes in favour of a certain depth level depending on a similarity measure and a threshold value (THR) that denotes whether this similarity is good enough. Let us measure the similarity in terms of the Euclidean distance d inferred from the intensity values of each pixel. A hard-voting procedure might be considered at this point if each camera's vote depends on whether the distance d is below the threshold or not (see Figure 3, left). However, we propose a soft-voting procedure where each camera's vote is weighted depending on this distance d, being maximal (i.e. equal to 1) when the distance is zero and decreasing exponentially until 0 (i.e. not to

vote in favour) when the distance exceeds the threshold THR (see Figure 3, right).

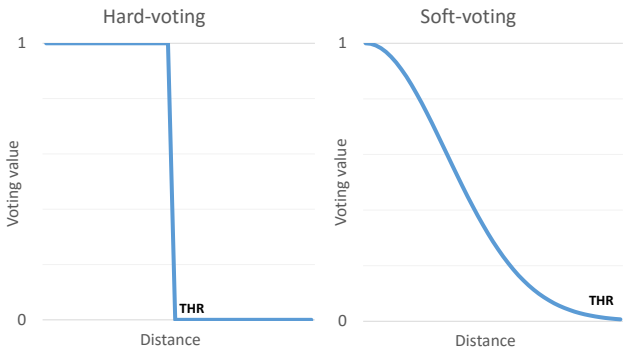

Figure 3: Hard-voting (left) and soft-voting (right) options for weighting the vote of each camera (y-axis) depending on the distance value (x-axis).

Formally, let us consider each elemental image $E(p_1, p_2, p_3)$, with p_1 and p_2 as the pixel coordinates and p_3 as the camera number. Thus, centred on the pixel position (i, j), for each neighbourhood pixel $(x, y) \in W_{ij}$ and $\forall\ C^k \in C$, distance d_{ij} is defined as the Euclidean distance among the pixel (i, j) from camera R and the pixels (x, y) from each camera C^k:

$$d_{ij}(x, y) = \sqrt{\sum_{k}^{\|C\|} \left(E(x, y, C^k) - E(i, j, R)\right)^2} \tag{2}$$

Note that the central camera R is always the same, being considered as the reference camera for the correct intensities on each pixel of the scene.

Distance d_{ij} is worked out for the pixel (i, j) at each position of the window W_{ij} and accumulated in V as follows,

$$V(i, j, z) = \frac{\sum_{(x,y)}^{W_{ij}} e^{\frac{-\left(d_{ij}(x,y)\right)^2}{THR}}}{CoK_{ij}} \tag{3}$$

Note that, in addition to the soft-voting function, the voting value is also weighted by CoK_{ij} in order to take into account only those cameras that "see" the pixel (i, j), that is, some positions of the scene in the reference camera R are only seen by certain cameras and, therefore, the number of votes is averaged by the number of cameras that are able to properly vote at this position.

After assessing all the possible $V(i, j, z)$, $\forall z \in Z$, a vector $V(i, j, Z_{min})$ $\ldots$ $V(i, j, Z_{max})$ is obtained on each pixel (i, j). The proposed *Max-Voting* method estimates the depth parameter z for each pixel position (i, j) on the

basis of the maximum number of votes obtained when the previous process is applied at each depth level z.

$$\hat{z}(i,j) = arg\max_{z \in Z} V(i,j,z)$$

Therefore, the z value where the number of votes in $V(i,j,z)$ is maximal is selected for the pixel (i,j) and this is done for all the pixels of the reference image R, creating thus a depth map of the scene.

3.1 Colour space

As a colour discriminant measure, we have used an Euclidean distance in the L*a*b* colour space in order to calculate the distance between two colours. This assumption is taken because, unlike RGB colour space, in L*a*b* the distances between two colours are approximately proportional to the human perceptual difference between them. Moreover, L*a*b* colour space appears to possess more uniform perceptual properties than other perceptual CIE spaces [9].

Only the chroma values of each pixel (a*b* planes) are taken into account when the Euclidean distance is worked out, using this distance to measure the colour difference between two pixels. It is important to emphasize that we have decided not to take into account the intensity plane L* but the chroma planes a*b* because it somehow involves a special invariance to brightness and shadows.

3.2 Post-processing

The *Max-Voting* algorithm obtains a vector of real values $V(i,Z_{min}) \dots V(i,Z_{max})$ representing the votes on pixel i for each depth z as described previously. However, there exist some points where this vector does not have a clear maximum and even ties can be found, being the 3D reconstruction on these points ambiguous. These uncertain points have a very low variance across the vector $V(i,Z_{min}) \dots V(i,Z_{max})$ and are often found in those objects (or part of the reconstructed image) whose boundaries are out of the scene. As we will see in our experimental section, this is generally consistent with those pixels that belong to the background of the image.

To overcome this problem, a heuristic post-processing stage has also been included in our approach. On those pixels (i,j) where there is no clear maximum in the number of votes for each level of depth, we take into account their variances in terms of pixel intensity at each reconstructed image $I(i,j,z)$. If these variances along all the z results is less than 1 for a certain pixel, we consider that pixel with a high level of uncertainty and it is assigned to the background ($z = \max(Z)$).

Although no better depth map can be guaranteed after applying this post-process, the results presented on this work show a significant improvement of the depth estimation.

4 Results

Two groups of experiments have been conducted to highlight the advantages of the *Max-Voting* methodology. One of them is based on images from real-world scenarios [42, 49] (Fig.4). The other group is obtained by simulating the real integral imaging pickup process in 3DS Max, using the same methodology as in [45, 47]. For these synthetic images[3], the ground truth of the depth map (*Z-buffer*) is available and a quantitative analysis of the results can be performed (Fig.5). Z-buffering or depth buffering of an image is the process of obtaining a two-dimensional array with the depth coordinates. This process also decides which elements of the rendered scene are visible, and which are hidden. The Z-buffer obtained from the 3DS Max returns the depth level of each pixel in terms of *graphical units* (GU). We have transformed these graphical units to a standard metric taking into account the actual measurements that could be expected from the objects of the scene. This transformation has resulted in the equivalence $1GU \equiv 1mm$.

The results obtained by the *Max-Voting* method have been compared with the results obtained by the *Min-Var* method [14]. Table 1 shows the image set-up used for the experiments. Second and third columns show the camera rack configuration and the depth range from Z_{min} to Z_{max} with a step size of Z_{step}. Note that the Z_{step} has been set to 1 cm for the *Head* and *Beethoven* images because higher precision for the depth estimation is needed on these cases. Fourth to fifth columns give the specifications details for the II pickup process. See Fig.2 for nomenclature details for *physical size of the camera sensor* (c_x, c_y) and *pitch of cameras* (p). A *focal length* ($f = 50mm$) has always been applied except for *PersonHand* image ($f = 8mm$).

Figure 4 shows the input images (central camera view) used for the experiments with real images. These images show a scene with unknown illumination, irregular background and in the presence of occlusions. The first two images (*CarsBrushes* and *CarsHelicopter*) have cars, an helicopter and forest plants whereas the third one (in grey) shows a person with an outstretched arm. Elemental (multiview) images for these scenes are captured as illustrated in [42, 49].

Figure 5 shows the elemental images from the central camera of the synthetic images used in this work. The depth information generated by the 3D Max is shown in the first column of the Figure 8. The synthetic images show three indoor spaces (*Livingroom*, *Bathroom* and *Toysroom*) and two

[3] In `http://www.vision.uji.es/IntegralImaging/`, synthetic images and their ground truth images are available. Parameters of how they have been created are also offered.

Image name	C-rack	Z_{min}:Z_{step}:Z_{max}	(c_x, c_y)	p
CarsBrushes	5×5	200:10:2000	(36,24)	10
CarsHelicopter	7×7	200:10:650	(36,24)	10
PersonHand	3×3	350:10:1850	(4.76,3.57)	3.0,3.5
Livingroom	7×7	370:10:900	(36,36)	5
Bathroom	7×7	220:10:830	(36,36)	5
Toysroom	7×7	220:10:750	(36,36)	5
Head	7×7	190:1:341	(36,36)	5
Beethoven	7×7	139:1:341	(36,36)	5

Table 1: Experimental set-up features. The first three images are real world scenarios whereas the following ones are synthetic images created in 3DS Max. Units for the three last columns in centimetres. *PersonHand* image has a different pitch on each axis (x-axis and y-axis respectively).

Figure 4: Real images. CarsBrushes image (left), CarsHelicopter image (center) and PersonHand image (right).

foreground images, *Head* and *Beethoven*. On the one hand, the three indoor spaces are ideally designed for evaluating the performance of the algorithms with many objects of different shapes and sizes and in different depths. On the other hand, the two foreground images are suited for evaluating how precise the algorithms could be.

Figure 5: Synthetic images. From left to right, Livingroom, Bathroom, Toysroom, Head and Beethoven images.

4.1 Studying the influence of the local parameters

As explained in Sect. 3, our approach proposes an evaluation on a pixel level basis of the photoconsistency among the cameras. Hence, both *i)* the threshold applied to measure the difference between two pixels (i.e. when two cameras agree) and *ii)* the size of the local window applied on each pixel are important parameters in our approach. Regarding the former, a very restrictive value ($THR = 1$) has been set up.

Table 2 shows how the pixel differences in terms of intensity values relate to the vote of each camera in terms of the soft-voting strategy proposed in this work (Eqs. 2 and 3). Let us suppose that the values from the first row (d) are the Euclidean distances between the intensity values of two pixels[4]. From these values, the vote value (V) of a single camera is shown in the second row, given $THR = 1$. As it can be seen, the V distribution is not linear but exponential. Moreover, there is no camera contribution when the intensity differences between two pixels is greater than 3, i.e. $d \geq 3$.

d	0	0.3	0.6	0.9	1.2	1.5	...
V	1	0.91	0.69	0.44	0.24	0.10	...
d	...	1.8	2.1	2.4	2.7	3	
V	...	0.039	0.012	0.003	0.0006	0	

Table 2: Voting values related to the intensity value differences between two pixels. For a single camera and for just one position of the local window, first row shows hypothetical values applying Eq.2 whereas the second row shows the contribution of the vote of the camera to the total vote of all cameras (Eq.3).

The window size parameter needs a more detailed explanation. Figure 6 shows the results for the *Bathroom* image, where different window sizes has been applied (for a constant $THR = 1$ value). From left to right and top to bottom, we show the generated depth map by the *Max-Voting* algorithm with the following window sizes: 3×3, 5×5, 7×7, 9×9, 11×11 and 13×13. It can be seen a smoothing behaviour whose degree increases with the window size used. However, this smoothing also makes some object details to get lost. Note how the boundaries of the easy chair or the plant are well-shaped for the 3×3 and even 5×5 windows whereas, when the size of the window increases, there exists an important loss on these details.

Using real images generally implies assuming certain level of noise. Computer vision processes are sensitive to this noise level, which is imperceptible for human eye. Therefore, on real images we have applied a restrictive 3×3 window size whereas on synthetic images, which are cleaner than the real

[4]This difference is worked out using the chroma planes (a*b*) from the L*a*b* colour space.

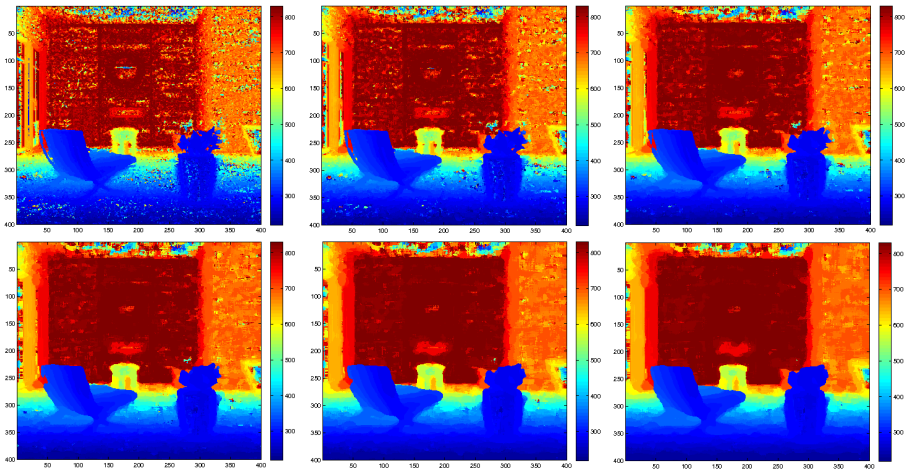

Figure 6: Results applying different window sizes on Bathroom image. From left to right and top to bottom, window sizes of 3×3, 5×5, 7×7, 9×9, 11×11 and 13×13.

ones, we have been able to relax this parameter and a 5×5 window size has been used.

4.2 Results on real images

Let us compare in this section the results obtained by the *Min-Var* method against the results obtained by the *Max-Voting* method using real images. On real images, a 3×3 window size has been applied with $THR = 1$. Figure 7 shows the results for the *CarsBrushes*, *CarsHelicopter* and *PersonHand* images. Note that *PersonHand* image is a grey image and no L*a*b* transformation has therefore been done in this case. The first column shows the results obtained by the *Min-Var* method whereas the two following ones show respectively the results obtained by the *Max-Voting* algorithm and by the *Max-Voting* algorithm when the post-processing stage is applied, respectively.

As we can see, *Max-Voting* results are generally less noisy than the ones produced by the *Min-Var* method. Also, note how the post-processing step has improved a lot the estimation of the depth of those pixels that belong to the background in the two first images. In the third image (*PersonHand*), this is not so obvious but almost a 10% of the pixels have been modified in the post-processing stage, especially at the limits of the image.

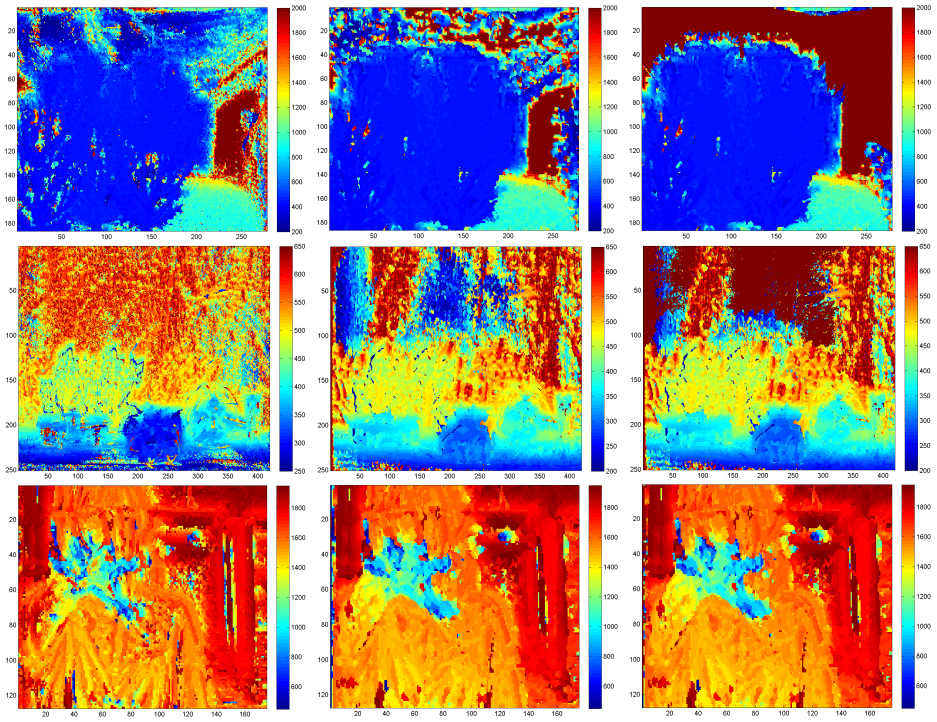

Figure 7: Results for CarsBrushes image (top), CarsHelicopter image (center) and PersonHand image (bottom) in pseudo-color. From left to right columns, we show the depth map using approach in [14], our proposal with no post-processing and our proposal with post-processing, respectively.

4.3 Results on synthetic images with ground truth

The results from Sect.4.2 only allow us to visually compare the *Min-Var* and *Max-Voting* results. In this section, we compare the results obtained by both methodologies using synthetic images where the ground truth of the depth level is available and a quantitative evaluation of the results can therefore be done. On synthetic images, a 5×5 window size has been applied with $THR = 1$. In addition, in order to ensure a fair comparison among both methodologies, no post-processing stage has been applied to our results for the case of synthetic images.

Figure 8 shows the results for the *Livingroom*, *Bathroom*, *Toysroom*, *Head* and *Beethoven* images. The first column shows the ground truth for the depth information of each scene. The second column shows the results obtained by the *Min-Var* method and the third column shows the results obtained by the *Max-Voting* method.

In order to quantitatively evaluate the results, the Root Mean Square

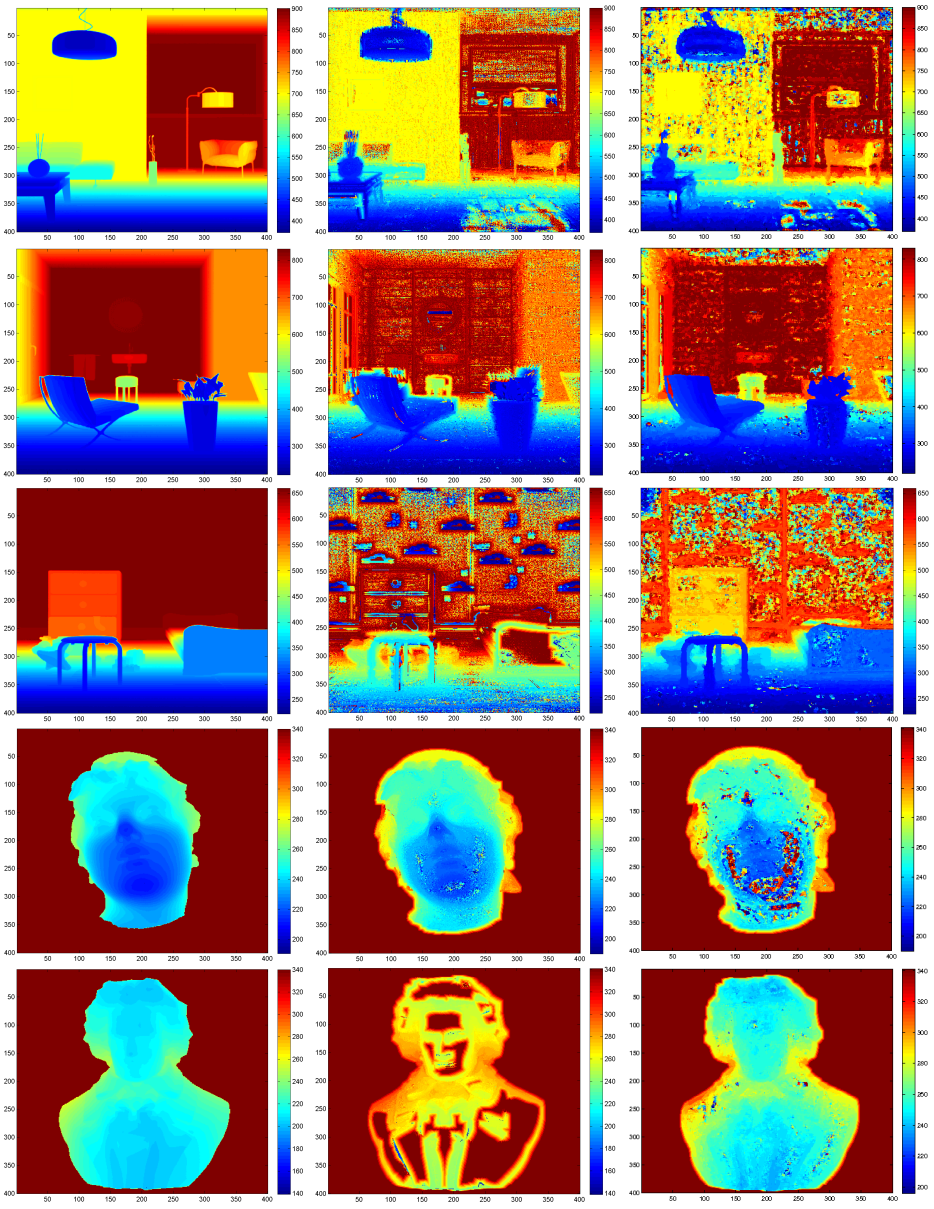

Figure 8: Depth maps results. From left to right columns, ground-truth of the depth map, results obtained by the *Min-Var* method and results obtained by the *Max-Voting* method.

Error (RMSE) figure of merit has been chosen, since it is a commonly-used error measure, and in fact, RMSE has widely been used to quantitatively compare the performance of different depth map operators [35].

Let us consider two depth maps, the ground truth G and the estimated

depth map Z, both obtained from a $N_x \times N_y$ image. The RMSE metric is defined as,

$$RMSE = \sqrt{\frac{1}{N_x \cdot N_y} \sum_{(i,j)} (G(i,j) - Z(i,j))^2} \tag{4}$$

In Sect. 4.1, we show how the different window sizes affect to the results obtained by our proposal. For synthetic images where the ground truth is available, a quantitative analysis about the influence of this window size parameter can be performed. Thus, it is worth saying that, in terms of RMSE, the results are progressively getting better while the size of the window is getting larger (see Table 3).

Window size	W3	W5	W7	W9	W11	W13
Bathroom (RMSE)	77.98	64.45	59.09	56.32	54.89	54.02

Table 3: RMSE results for the Bathroom image while the window size increases. From left to right, the window size is increasing from 3×3 to 13×13. As second row shows, the larger the window, the better the RMSE value.

Tables 4 and 5 show the depth estimation error results obtained using the *Min-Var* method and the *Max-Voting* method. Table 4 shows the percentage of pixels that have high errors when the z value has been estimated. The threshold value for considering high errors has been set to 100cm for all images but for the images marked with an asterisk, where the threshold is equal to 50cm.

Table 5 shows the RMSE values (expressed in centimetres or millimetres depending on the image) and the RMSE obtained if those pixels with high errors are not taken into account. The *Hi-Error* column from Table 4 shows the percentage of the image pixels where the estimation can be considered unacceptable. By the $RMSE^*$ column, we expect to demonstrate that most of the RMSE error made by the algorithms is concentrated on few pixels. Thus, if these few pixels are removed and the RMSE is re-calculated, the real performance of the methods substantially improves.

Finally, it is important to point out that in the foreground images (marked with an asterisk) only those pixels that belong to the object have been taken into account, both for the RMSE value and for the percentage of high errors.

5 Discussion

As the previous section shows, the proposed maximum voting approach has shown a significant improvement with regard to the *Min-Var* method both in terms of a qualitative and a quantitative assessment. We have shown the results obtained using real images acquired in a laboratory set-up, where our

Image name	*Min-Var* Hi-Error	*Max-Voting* Hi-error
Livingroom(cm)	**13.48**%	13.97%
Bathroom(cm)	12.60%	**8.32**%
Toysroom(cm)	54.13%	**16.12**%
Head*(mm)	**3.42**%	11.47%
Beethoven*(mm)	49.01%	**6.29**%

Table 4: Quantitative results on synthetic images (I). From left to right, in blocks, images, results for the *Min-Var* approach and results for the *Max-Voting* approach. *Hi-Error* column shows the percentage of pixels that have experimented high errors when the z value has been estimated.

Image name	*Min-Var*		*Max-Voting*	
	RMSE	RMSE*	RMSE	RMSE*
Livingroom(cm)	87.35	**26.40**	**83.48**	26.53
Bathroom(cm)	85.14	28.10	**64.45**	**27.16**
Toysroom(cm)	172.24	44.15	**92.95**	**33.84**
Head*(mm)	**10.94**	**10.78**	29.15	15.79
Beethoven*(mm)	81.85	45.66	**43.63**	**43.25**

Table 5: Quantitative results on synthetic images (II). From left to right, in blocks, images, results for the *Min-Var* approach and results for the *Max-Voting* approach. Second and third blocks show the RMSE values obtained on each image (*RMSE* column) and the RMSE obtained if pixels with high errors are not taken into account (*RMSE** column).

proposal with the post-processing stage has found fine details at boundaries, less pixels with high errors in their depth estimation and a significant enhancement in the background. Note also that the level of improvement is quite more noticeable in the colour images than in the grey one (*PersonHand* image) where no L*a*b* transformation has been done.

Results on artificial images where the true depth map is available have also been shown. The results obtained by our proposal show higher precision on boundaries. This fact is recognisable in the indoor scenes, where the objects that are closer to the camera set-up show very sharp edges (this is especially clear in the *Bathroom* image). In addition, there are less high-depth errors in our proposal in terms of percentage of pixels when compared with the *Min-Var* method. Note how the *Min-Var* method shows problems in the indoor images in areas such as the window at the background of the *Livingroom* image, the mirror of the *Bathroom* image or the foot board of the bed and the wallpaper clouds in the *Toysroom* image.

Figure 9 supports these facts by showing the resulting images of sub-

tracting the ground truth values from the results obtained by the *Min-Var* method (center) and the *Max-Voting* method (right). It is shown how the pixels around most of the important object boundaries in *Min-Var* method present a higher number of artifacts than in *Max-Voting* method with respect to the depth estimation. In addition, *Min-Var* method also presents some problems with reflections as it can be seen with the mirror at the background of the scene.

Regarding the foreground images, on the one hand, the *Min-Var* algorithm has obtained a poor result for the depth estimation of the *Beethoven* image. Indeed, a high percentage of the bust has been considered background by the algorithm. On the other hand, the *Max-Voting* method has obtained a worse result for the *Head* image.

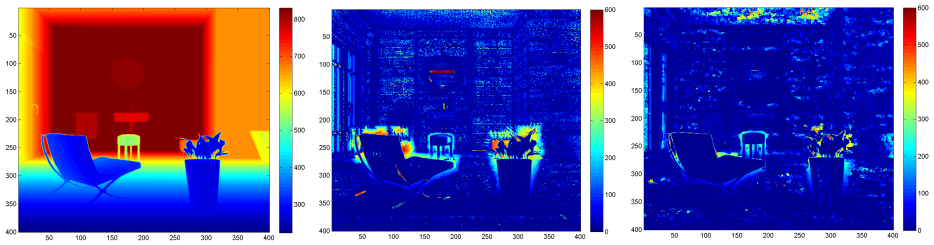

Figure 9: Synthetic images comparison by ground truth subtraction. From left to right, the ground truth image and the results of the *Min-Var* and *Max-Voting* methods, respectively.

It is also important to point out that a small proportion of the RMSE value obtained on each image (especially for the indoor images) comes from the size of the Z_{step} used in our experimentation. A ground truth value between two consecutive z values in our depth range, that is between z_j and z_k where $z_k - z_j = Z_{step}$, can not be exactly estimated neither the *Max-Voting* method nor the *Min-Var* method.

Quantitatively, in terms of RMSE and percentage of pixels that show high error on their estimation, our proposal has substantially improved the results obtained by the *Min-Var* method in three out of the five images (*Bathroom*, *Toysroom* and *Beethoven*), being slightly better in the *Livingroom* image and worse in the *Head* image.

As it has been shown in Table 3, larger windows sizes generate a lose of detail on some object boundaries. However, in terms of RMSE, this fact is compensated by the noise removal produced by these larger windows. Regarding this balance between noise removal and detailed boundaries, we have decided to apply a quite restrictive 3×3 and 5×5 window sizes to keep clear edges. As explained in our introduction, the *Min-Var* method is especially noisy at boundaries. Thus, in order to improve this point in our algorithm, those very restrictive window sizes has been applied since, in our opinion,

the loss of detail in boundaries is too significant from a window size of 7×7 or higher, although the RMSE value for our proposal would get better results using larger window sizes.

Regarding the *Head* and *Beethoven* images, the fact of focusing only on foreground pixels allows us to obtain a more accurate estimation of the error made by each methodology since it avoids misrepresenting the given values adding a large amount of pixels from the background where the depth has been estimated correctly.

Finally, it important to emphasise how the proposed algorithm behaves against occlusions and noise. On the one hand, the use of local information (neighbourhood of a pixel) together with the majority voting mechanism, allows the method to cope with occlusions and depth discontinuities at a local level. See for instance the result on real images shown in Figure 7, where the depth discontinuities (and therefore occlusion areas) are well-defined with sharper object boundaries in the result of the post-processed outcome of our approach with respect to the minimum variance approach, particularly in the *CarsBrushes* and *CarsHelicopter* images. On the other hand, the visual quality of the input images can be improved by noise filtering techniques [13]. However, the goal of the presented technique is to be an input for a higher level processing for a global approach that can regularize the result at this level, using for instance a Random Markov Field regularization [27] or Energy Minimization with regularization [7]. At a local level, regularization, and therefore noise reduction, can be obtained by varying the window (pixel neighbourhood) size. In this sense, as already discussed, note the experimental result with respect to RMSE error for different windows sizes in Table 3, which shows the effect when increasing the neighbourhood size.

6 Conclusions

We have presented an approach that uses the scene information acquired by means of a 3D Integral Imaging system and generates a depth map of the scene through a voting strategy based on a photoconsistency criterion. This voting strategy is performed on the basis of the image local information obtained from all the cameras in the array. The results obtained are generally consistent for any type of object surfaces, also defining accurate enough boundaries. We also contribute with a repository (`http://www.vision.uji.es/IntegralImaging/`) with the materials presented in this work. Synthetic images and ground truth images (generated by the 3DS Max), set-up parameters and the source code for the here proposed maximum voting approach are available in this repository. This resource can be used for performing Integral Imaging and Range Estimation.

Acknowledgments

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) under the projects SEOSAT (ESP2013-48458-C4-3-P) and MTM2013-48371-C2-2-P, by the Generalitat Valenciana through the project PROMETEO-II-2014-062, and by the University Jaume I through the project UJI-P11B2014-09. B. Javidi would like to acknowledge support under NSF/IIS-1422179.

References

[1] LYTRO cameras. `https://www.lytro.com/`, Last accessed: April 12th, 2016.

[2] J. Arai, F.Okano, H. Hoshino, and I. Yuyama. Gradient-index lens-array method based on real-time integral photography for three-dimensional images. *Applied Optics*, 37:2034–2045, 1998.

[3] J. Arai, F. Okano, M. Kawakita, M. Okui, Y. Haino, M. Yoshimura, M. Furuya, and M. Sato. Integral three-dimensional television using a 33-megapixel imaging system. *Journal of Display Technology*, 6(10):422–430, 2010.

[4] J. Arai, F. Okano, M. Kawakita, M. Okui, Y. Haino, M. Yoshimura, M. Furuya, and M. Sato. Integral three-dimensional television using a 33-megapixel imaging system. *IEEE Journal of Display Technology*, 6(10):422–430, 2010.

[5] H. Arimoto and Bahram Javidi. Integral three-dimensional imaging with digital reconstruction. *Opt. Lett.*, (26):157–159, 2001.

[6] S. A. Benton and V. M. Bove. *Holographic Imaging*. Wiley-Interscience, 2008.

[7] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

[8] C. B. Burckhardt. Optimum parameters and resolution limitation of integral photography. *Journal of the Optical Society of America*, 58:71–76, 1968.

[9] Heng-Da Cheng and Ying Sun. A hierarchical approach to color image segmentation using homogeneity. *IEEE Trans on Image Processing*, 9(12):2071–2082, 2000.

[10] M. Cho, M. Dansehpanah, I. Moon, and B. Javidi. Three-dimensional optical sensing and visualization using integral imaging. *Proceedings of the IEEE*, 99(4):556–575, 2011.

[11] Myungjin Cho and B. Javidi. Three-dimensional visualization of objects in turbid water using integral imaging. *Journal of Display Technology*, 6(10):544–547, October 2010.

[12] Myungjin Cho, Abhijit Mahalanobis, and Bahram Javidi. 3D passive photon counting automatic target recognition using advanced correlation filters. *Optics Letters*, 36(6):861–863, March 2011.

[13] M. Daneshpanah, B. Javidi, and E.A. Watson. Three dimensional object recognition with photon counting imagery in the presence of noise. *Opt. Express*, (18):26450–26460, 2010.

[14] Mehdi DaneshPanah and Bahram Javidi. Profilometry and optical slicing by passive three-dimensional imaging. *Opt. Lett.*, 34(7):1105–1107, 2009.

[15] A. W. Fitzgibbon, Y. Wexler, and A. Zisserman. Image-based rendering using image-based priors. *International Journal of Computer Vision*, 63(2):141–151, 2005.

[16] J. W. Goodman. *Introduction to Fourier Optics*. McGraw-Hill, 1996.

[17] S.-H. Hong, J.-S. Jang, and B. Javidi. Three-dimensional volumetric object reconstruction using computational integral imaging. *Optics Express*, 3(3):483–491, 2004.

[18] Seung-Hyun Hong and Bahram Javidi. Distortion-tolerant 3D recognition of occluded object using computational integral imaging. *Optics Express*, 14(25):12085–12095, December 2006.

[19] L. Hongen, N. Hata, S. Nakajima, M. Iwahara, I. Sakuma, and T. Dohi. Surgical navigation by autostereoscopic image overlay of integral videography. *IEEE Trans. Inf. Technol. Biomed.*, 8:114–121, 2004.

[20] H. Hoshino, F.Okano, H. Isono, and I. Yuyama. Analysis of resolution limitation of integral photography. *Journal of the Optical Society of America A*, 15:2059–2065, 1998.

[21] Yutaka Igarashi, Hiroshi Murata, and Mitsuhiro Ueda. 3-D display system using a computer generated integral photograph. *Japanese Journal of Applied Physics*, 17(9):1683–1684, 1978.

[22] H. E. Ives. Optical properties of a lippmann lenticuled sheet. *Journal of the Optical Society of America*, 21:171–176, 1931.

[23] J.-S. Jang and B. Javidi. Three-dimensional synthetic aperture integral imaging. *Optics Letters*, 27:1144–1146, 2002.

[24] B. Javidi, S.-H. Hong, and O. Matoba. Multidimensional optical sensor and imaging system. *Applied Optics*, 45:2986–2994, 2006.

[25] B. Javidi, F. Okano, and J.-Y. Son. *Three-dimensional Imaging, Visualization, and Display Technology*. Springer, 2008.

[26] S. Kishk and B. Javidi. Improved resolution 3-D object sensing and recognition using time multiplexed computational integral imaging. *Optics Express*, 11:3528–3541, 2003.

[27] Stan Z. Li. *Markov Random Field Modeling in Image Analysis*. Springer Publishing Company, Incorporated, 3rd edition, 2009.

[28] G. Lippmann. La photographie integrale. *Comptes-Rendus Academie des Sciences*, 146:446–451, 1908.

[29] M. Martínez-Corral, B. Javidi, R. Martínez-Cuenca, and G. Saavedra. Integral imaging with improved depth of field by use of amplitude modulated microlens array. *Applied Optics*, 43:5806–5813, 2004.

[30] R. Martínez-Cuenca, G. Saavedra, M. Martínez-Corral, and B. Javidi. Progress in 3-D multiperspective display by integral imaging. *Proceedings of the IEEE*, 97:1067–1077, 2009.

[31] H. Navarro, R. Martínez-Cuenca, G. Saavedra, M. Martínez-Corral, and B. Javidi. 3D integral imaging display by smart pseudoscopic-to-orthoscopic conversion (SPOC). *Opt. Express*, 18(25):25573–25583, 2010.

[32] F. Okano, J. Arai, K. Mitani, and M. Okui. Real-time integral imaging based on extremely high resolution video system. *Proceedings of the IEEE*, 94(3):490–501, March 2006.

[33] Fumio Okano, Haruo Hoshino, Jun Arai, and Ichiro Yuyama. Real-time pickup method for a three-dimensional image based on integral photography. *Appl. Opt.*, 36(7):1598–1603, 1997.

[34] T. Okoshi. Three-dimensional displays. *Proceedings of the IEEE*, 68(5):548–564, 1980.

[35] Said Pertuz, Domenec Puig, and Miguel Angel Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415 – 1432, 2013.

[36] R. Schulein, C. Do, and B. Javidi. Distortion-tolerant 3D recognition of underwater objects using neural networks. *Journal of the Optical Society of America A*, 27:461–468, 2010.

[37] D. Shin, M. Daneshpanah, and B. Javidi. Generalization of three-dimensional n-ocular imaging systems under fixed resource constraints. *Optics Letters*, 37(1):19–21, 2012.

[38] S. Sinha, D. Steedly, R. Szeliski, M. Agrawala, and M. Pollefeys. Interactive 3D architectural modeling from unordered photo collections. *ACM Trans. Graph.*, (27):1–10, 2008.

[39] A. P. Sokolov. *Autostereoscopy and Integral Photography by Professor Lippmann's Method*. Moscow State University Press, 1911.

[40] J.-Y. Son, W.-H. Son, S.-K. Kim, K.-H. Lee, and B. Javidi. Three-dimensional imaging for creating real-world-like environments. *Proceedings of the IEEE*, 101(1):190–205, 2013.

[41] A. Stern and B. Javidi. Three-dimensional image sensing, visualization and processing using integral imaging. *Proceedings of the IEEE*, 94:591–607, 2006.

[42] V. Javier Traver, Pedro Latorre-Carmona, Eva Salvador-Balaguer, Filiberto Pla, and Bahram Javidi. Human gesture recognition using three-dimensional integral imaging. *J. Opt. Soc. Am. A*, 31(10):2312–2320, 2014.

[43] George Vogiatzis, Carlos Hernández Esteban, Philip H. S. Torr, and Roberto Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Transactions on Pattern Analysis Machine Inteligence (PAMI)*, 29(12):2241–2246, 2007.

[44] C. Wheatstone. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London*, 128:371–394, 1838.

[45] Xiao Xiao, Bahram Javidi, and Dipak K. Dey. Bayesian estimation of depth information in three-dimensional integral imaging. In *Three-Dimensional Imaging, Visualization, and Display (SPIE) Conference Series*, volume 9177, pages 911714–911714–8, 2014.

[46] Xiao Xiao, Bahram Javidi, Manuel Martinez-Corral, and Adrian Stern. Advances in three-dimensional integral imaging: sensing, display, and applications. *Appl. Opt.*, 52(4):546–560, 2013.

[47] Xiao Xiao, Xin Shen, M. Martinez-Corral, and B. Javidi. Multiple-planes pseudoscopic-to-orthoscopic conversion for 3d integral imaging display. *Journal of Display Technology*, 11(11):921–926, 2015.

[48] L. Yang, M. McCornick, and N. Davies. Discussion of the optics of a new 3-D imaging system. *Applied Optics*, 27:4529–4534, 1988.

[49] Yige Zhao, Xiao Xiao, Myungjin Cho, and Bahram Javidi. Tracking of multiple objects in unknown background using bayesian estimation in 3d space. *J. Opt. Soc. Am. A*, 28(9):1935–1940, Sep 2011.