Depth Estimation – An Introduction

Pablo Revuelta Sanz, Belén Ruiz Mezcua
and José M. Sánchez Pena

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/45904

1. Introduction

Depth estimation or extraction refers to the set of techniques and algorithms aiming to obtain a representation of the spatial structure of a scene. In other terms, to obtain a measure of the distance of, ideally, each point of the seen scene. We will talk, as well, about 3D vision.

In this chapter we will review the main topics, problems and proposals about depth estimation, as an introduction to the Stereo Vision research field. This review will deal with some essential and structural aspects of the image processing field, as well as with the depth perception capabilities and conditions of both computer and human based systems.

This chapter is organized as follows:

- This introduction section will present some basic concepts and problems of the depth estimation.
- The depth estimation strategies section will detail, analyze and present results of the main families of algorithms which solve the depth estimation problem, among them, the stereo vision based approaches.
- Finally, a conclusions section will summarize the pros and contras of the main paradigms seen in the chapter.

1.1. The 3D scene. Elements and transformations

We will call "3D scene" to the set of objects placed in a three dimensional space. A scene, however, is always seen from a specific point. The distorted image that is perceived in that point is the so-called projection of the scene. This projection is formed by the set of rays crossing a limited aperture arriving to the so-called projection plane (see figure 1).

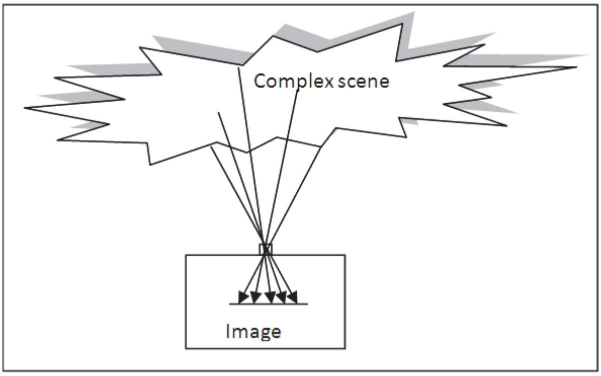


Figure 1. The 3D scene projected into a plane.

This projection presents some relevant characteristics:

- The most evident consequence of a projection is the loose of one dimension. Since in each pixel only one point of the real scene is projected, the depth information is mathematically erased during the projection process into the image plane. However, some algorithms can retrieve this information from the 2D image, as we will see.
- On the contrary, the projection of a scene presents important advantages, such a simple sampling by already well developed devices (the so-called image sensors). Moreover, dealing with 2D images is, by obvious reasons, much simpler than managing 3D sets of data, reducing computational load.

Thus, the scene is transformed into a 2D set of points, which can be described in a Cartesian plane:

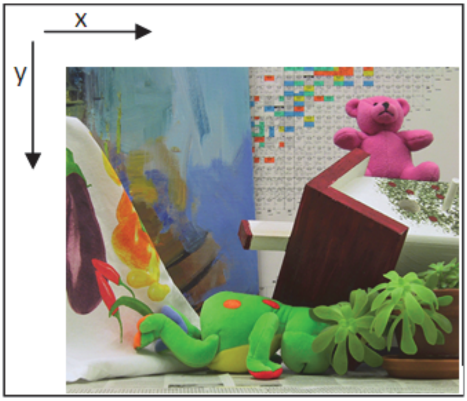


Figure 2. A 2D projection of a scene. "Teddy" image (Scharstein, 2010).

The 3D vision processes have as goal the reconstruction of this lost information, and, thus, the distances from each projected point to the image plane. An example of such reconstruction is shown in figure 3.

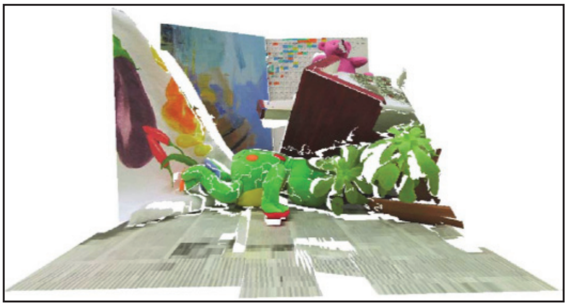


Figure 3. A 3D reconstruction of the previous image (Bleyer & Gelautz, 2005).

The reconstruction, also called depth map estimation, has to face some fundamental problems.

On the one hand, some extra information has to be obtained, for an absolute depth estimation. This aspect will be discussed in section 1.3.12.

On the other hand, there are, geometrically, infinite points in the scene that are not projected and, then, must be, in some cases, interpolated. This is the case of occluded points, shown in figure 4.

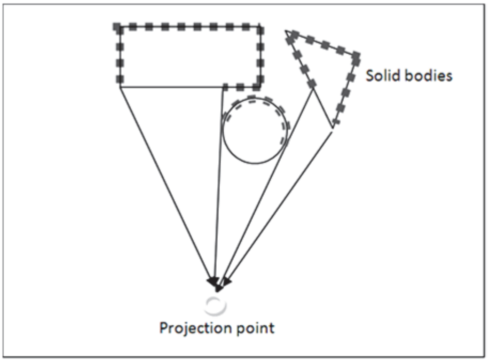


Figure 4. Occluded points, marked with squares.

1.2. Paradigms for 3D images representation over a plane

As we saw in the previous section, the projection onto a plane forces the loose of the depth dimension of the scene. However, the depth information should be able to be represented in a plane, for printing purposes, for example.

There are three widely used modes for depth representation:

- Gray scale 2.5D representation. This paradigm uses the gray scale intensity to represent the depth of each pixel in the image. Thus, the colour, texture and luminosity of the original image are lost in this representation. The name “2.5D” refers to the fact that this kind of images has the depth information directly in each pixel, while it is represented over a 2D space. In this paradigm, the gray level represents the inverse of the distance. Thus, more a pixel is bright, closer is the point represented. Vice versa, the darker is a pixel, further is the represented point. This is the most commonly used way for depth representation. Figure 5 shows an original image and its gray scale 2.5D representation.

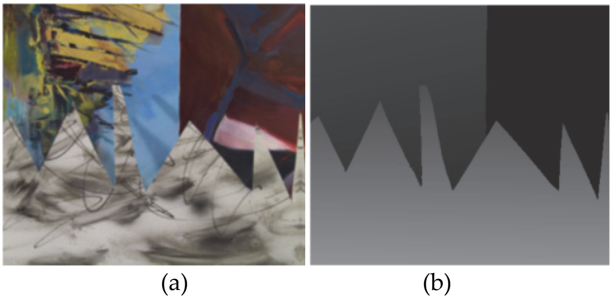


Figure 5. (a) The "Sawtooth" image and (b) its gray scale 2.5D representation (Scharstein, 2010).

- Colour 2.5D representation. This representation is similar to the previous one. The difference is the use of colours to represent the depth. In the following image, red-black colours represent closer points, and blue-dark colours the further points. However, other colour representations are available in the literature (see, for example, (Saxena, Chung, & Ng, 2008)). Figure 6 shows an example of the same image, represented in colour 2.5D.

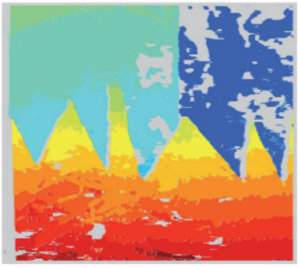


Figure 6. Colour based representation of the depth map (Kostková & Sára, 2006). In gray occluded parts.

- Pseudo-3D representation. This representation provides different points of view of the reconstructed scene. Figure 3 showed an example of this.

The main advantage of the first two methods is the possibility of implementing objective comparison among algorithms, as it is done in the Middlebury data base and test system (Scharstein, 2010).

We can appreciate a difference in the definition between the image of the figure 5.b and that of the figure 6. The image shown in figure 5.b is the so-called ground truth, i.e. the exact representation of the distances (obtained by laser, projections, or directly from 3D design environments), while the image of figure 6 is a computed depth map and, hence, is not exact. The ground truth is used for quantitative comparisons in distances between the extracted image and the real ones.

1.3. Important terms and issues in depth estimation

The depth estimation world is a quite complex research field, where many techniques and setups have been proposed. The set of algorithms which solve the depth map estimation problem deals with many different mathematical concepts which should be briefly explained for a minimum overall comprehension of the matter.

In this section we will review some important points about image processing applied to depth estimation.

1.3.1. Standard Test beds

The availability of common tests and comparable results is a mandatory constraint in active and widely explored fields. Likewise, the possibility of objective comparisons make easier to classify the different proposals.

In depth estimation, and more specifically in stereo vision, one of the most important test bed is the Middlebury database and test bed (Scharstein, 2010).

The test beds provide both eyes images of a 3D scene, as well as the ground truth map.

Figure 7 shows the “Cones” test set with its ground truth.

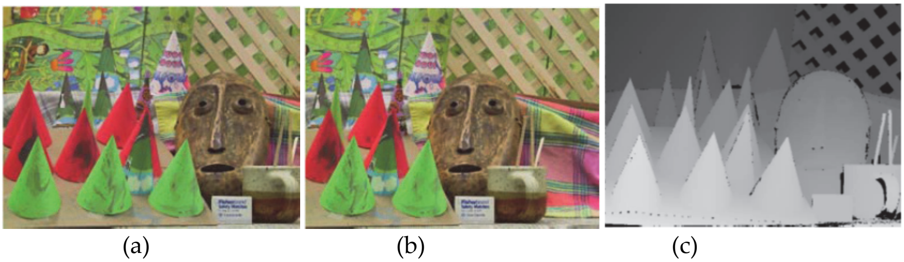


Figure 7. (a) Left eye, (b) right eye and (c) ground truth representation of the “Cones” scene (Scharstein & Szeliski, 2003).

The same test allow, as said, algorithms classification. An example of such a classification can be found in the URL http://vision.middlebury.edu/stereo/eval/

1.3.2. Colour or gray scale images?

The first point when we want to process an image, whichever is the goal, is to decide what to process. In this case colour or gray scale images.

As it can be seen in the following figure, colour images have much more information that gray scale images:

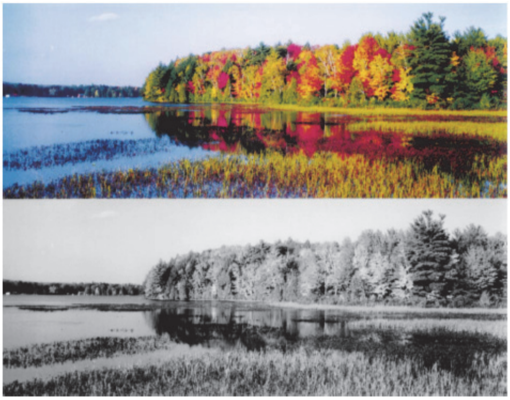


Figure 8. Advantages of colour vision (Nathans, 1999).

Colour images should, hence, be more appropriated for data extraction, among them, depth information.

However, the colour images have an important disadvantage: For a 256 level definition, they are represented by 3 bytes (24-bit representation), while gray scale images with the same level only require one single byte.

The consequence is obvious: colour image processing requires much more time and operations.

An example of the improvement of the depth estimation of colour images can be seen in the following table, where the same algorithm is run over gray scale images and a pseudo-color gray scale version of the same images sets, from (Scharstein, 2010):

Images set	Mode	Error (%)	Time
Tsukuba	Gray	55	50ms (20fps)
	Colour	46.9	77.4ms (12fps)
Teddy	Gray	79	78.9ms (12.7fps)
	Colour	60	114.2ms (8fps)
Venus	Gray	73.9	76.6ms (13fps)
	Colour	77	11.8ms (8fps)

Table 1. Comparison the colour based and gray scale processing of the same algorithm (Revuelta Sanz, Ruiz Mezcua, & Sánchez Pena, 2011).

1.3.3. The epipolar geometry

When dealing with stereo vision setups, we have to face the epipolar geometry problem.

Let C_l and C_r be the focal centres of the left and right sensors (or eyes), and L and R the left and right image planes. Finally, P will be a physical point of the scene and p_l and p_r the projections of P over L and R, respectively:

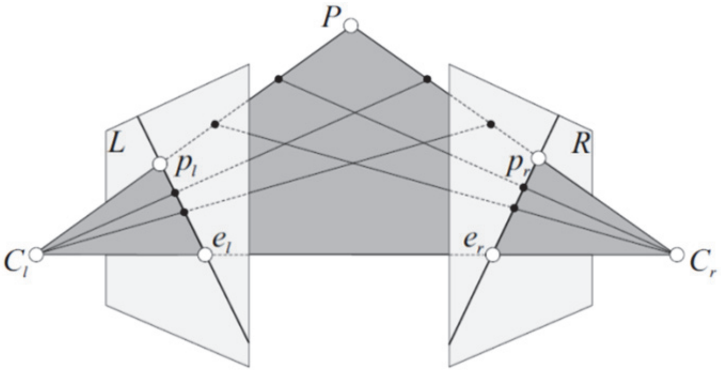


Figure 9. Epipolar geometry of a stereo vision system (Bleyer, 2006).

In this figure, we can also see both "epipoles", i.e., the points where the line connecting both focal centres intersects the image planes. They are noted as e_l and e_r.

The geometrical properties of this setup force that every point of the line Pp_l lies on the line p_re_r, which is called "epipolar line". The correspondence of a point seen in one image must be searched in the corresponding epipolar line in the other one, as shown in figure 10.

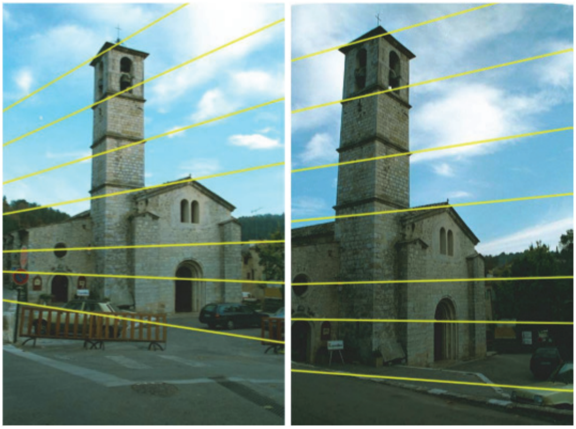


Figure 10. Epipolar lines in two different perspectives (Tuytelaars & Gool, 2004).

A simplified version of this geometry arise when the image planes are parallel. This is the base of the so-called *fronto-parallel hypothesis*.

1.3.4. The fronto-parallel hypothesis

The epipolar geometry of two sensors can be simplified, as said, positioning both planes parallel, arriving to the following setup:

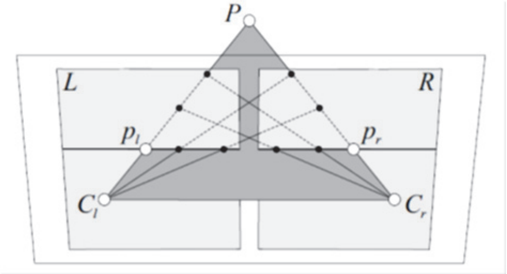


Figure 11. Epipolar geometry of a stereo vision system in a fronto-parallel configuration (Bleyer, 2006).

The epipoles are placed in the infinite, and the epipolar (and search) lines become horizontal. The points (except the occluded ones) are only decaled horizontally:

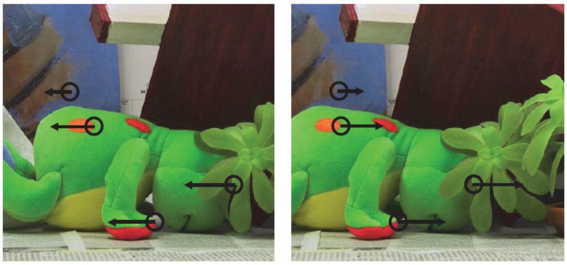


Figure 12. Corresponding points in two images, regarding the opposite image (Bleyer, 2006).

This geometrical setup can be implemented by properly orienting the sensors, or by means of mathematical transformation of the original images. If this last option is the case, the result is called "rectified image".

Other assumptions of the fronto-parallel hypothesis are described in detail in (Pons & Keriven, 2007; Radhika, Kartikeyan, Krishna, Chowdhury, & Srivastava, 2007).

The most important consequences of this geometry, regarding the Cartesian plane proposed in figure 2, can be written as follows:

- $y_l = y_r$. The height of a physical point is the same in both images.
- $x_l = x_r + \Delta d$. The abscissa of a physical point is decaled by the so-called *parallax* or *disparity*, which is inversely related to the depth.
- A point in the infinite has identical abscissa coordinates in both image planes.

1.3.5. Matching

When different viewpoints from the same scene are compared, a further problem arises that is associated with the mutual identification of images. The solution to this problem is commonly referred to as matching. The matching process consists of identifying each physical points within different images (Pons & Keriven, 2007). However, matching techniques are not only used in stereo or multivision procedures but also widely used for image retrieval (Schimd, Zisserman, & Mohr, 1999) or fingerprint identification (Wang & Gavrilova, 2005) where it is important to allow rotational and scalar distortions (He & Wang, 2009).

There are also various constraints that are generally satisfied by true matches thus simplifying the depth estimation algorithm, such as similarity, smoothness, ordering and uniqueness (Bleyer & Gelautz, 2005).

As we will see, the matching process is a conceptual approach to identify similar characteristics in different images. It is, then, subjected to errors. The matching is, hence, implemented by means of comparators allowing different identification strategies such as minimum square errors (MSE), sum of absolute differences (SAD) or sum of squared differences (SSD).

The characteristic compared through the matching process can be anything quantifiable. Thus, we will see algorithms matching points, edges, regions or other image cues.

1.3.6. The minimum distance measure constraint

It is assumed that the image planes are finite in area. Taking the fronto-parallel hypothesis into account, we can see that there is a minimum distance until which corresponding points can be found, but not below this distance. The geometrical representation of this constraint is shown in the following figure, were two image sensors with arbitrary cone of view present a blind area, which correspond to pixels out of both images:

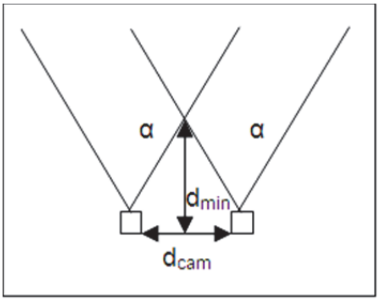


Figure 13. Minimum distance measurable in terms of the cone view angle α and the distance between sensors d_{cam}.

Some algorithms also impose an extra constraint, allowing a maximum disparity value, over which the points in the image plane are not recognized as the same physical point. This additional constraint present the advantage of reducing the number of operations: given that for one point, for example, in the left image, every pixel of the corresponding scan line in the right one must be compared to the original one, if the comparison presents a limit and, hence, not every pixel is compared, the algorithm improves its efficiency. However, some available matching will not be found.

1.3.7. The region segmentation

Region segmentation is a conceptual approach to image segmentation which is based on the similarities of adjacent pixels. The image is chopped into non-overlapping homogeneous regions which are based on a specific characteristic. In mathematical terms, let Ω be the image domain. Segmented regions can be expressed as (Pham, Xu, & Prince, 2000):

$$\Omega = \sum_{k=1}^{K} S_k \tag{1}$$

where S_k means the k_{th} region and $S_k \cap S_j = \emptyset$ for $k \neq j$.

This method is commonly applied to binary images, where the region segmentation is ambiguousless. Many different approaches have been developed regarding gray-scale medical imaging (Pham et al., 2000) and other imaging fields (Gao, Jiang, & Yang, 2006; Espindola, Camara, Reis, Bins, & Monteiro, 2006) or color images (Wang & Wang, 2008). The potential of this last option is greater than the second one, however more than three times the amount of operations are required. However, region segmentation has proven to be a very efficient method (but not the most exact) as it is capable of segmenting the image after a single analysis of the pixels contained within the image.

1.3.8. Edges and points extraction

Edges and points are important cues of the image, and are often used as descriptors. For that purpose, they must be extracted from or identified within the image.

Both edges and points are retrieved by means of different spatial operators, such as Laplacians or Laplacian-of-Gaussians (LoG). Figure 14 shows some typical operators for features extraction:

-1	-2	-1
0	0	0
1	2	1

Sobel vertical edge detector

0	1	0
1	-4	1
0	1	0

Discrete Laplace operator

-1	-1	-1
0	0	0
1	1	1

Prewitt's vertical operator

Figure 14. Three examples of image processing operators: Sobel, Laplace and Prewitt.

Figure 15 shows an original image and the results of the processing (convolution) with the previous operators:

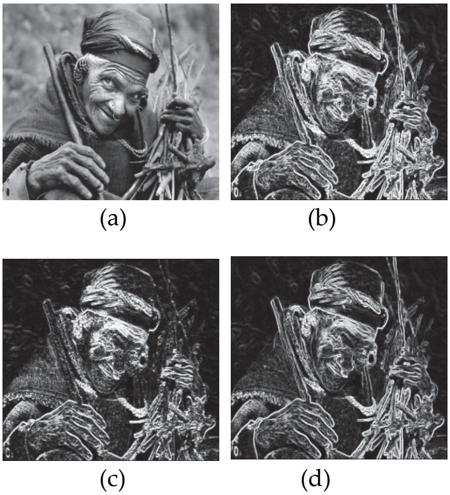


Figure 15. (a) Original image. (b) Sobel bidirectional (vertical and horizontal) filtering, (c) Prewitt's bidirectional filtering and (d) Laplacian filtering (Rangarajan, 2005).

Points are also extracted convoluting a mask, or kernel, with the whole image.

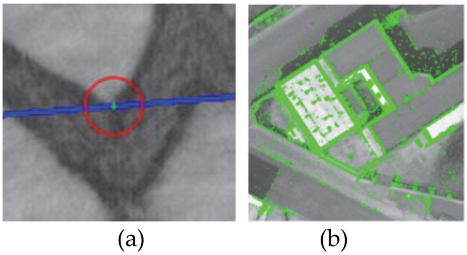
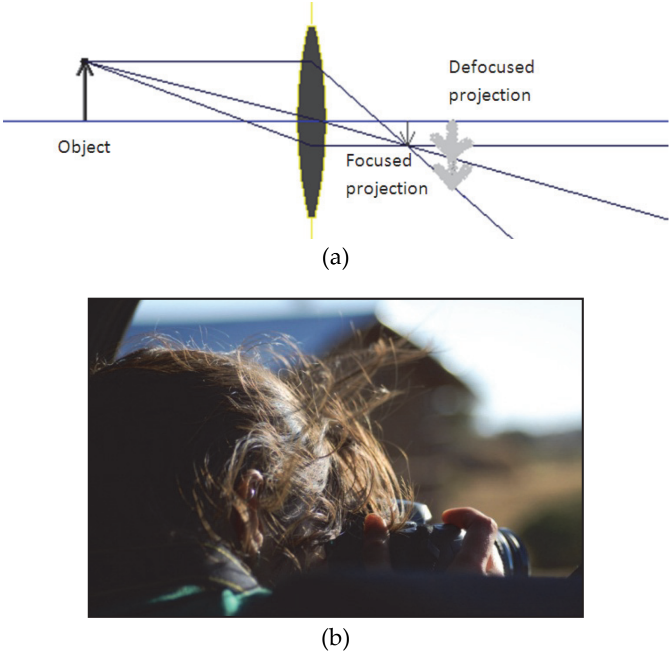


Figure 16. Relevant point retrieval. (a) Corner extraction; in blue, epipolar line. (b) The whole image already processed and the detected points in green. Both images extracted from (Yu, Weng, Tian, Wang, & Tai, 2008).

1.3.9. Focus

Since the aperture of a sensor is finite and not null, not every point in the projection is focused. This effect, applicable to both human and synthetic visual systems, produces a Gaussian blur on the projected image, proportional to the distance of that point to the focused plane (see figure 17).

An important problem arises when using the focus to estimate the depth: the symmetry effect of defocusing. We cannot know whether an object is closer or farther to the camera from a defocusing measurement. We will discuss this later in this chapter.



(a)

(b)

Figure 17. (a) Focus and defocus scheme and (b) example.

1.3.10. Dense and interpolated depth maps

The dense depth map concept refers to those 2.5D images computed for every pixel. Oppositely, if only some relevant points' distances are computed, and the rest of them interpolated, we will talk about interpolated depth maps. Advantages and disadvantages of both strategies depend of the final application and resources.

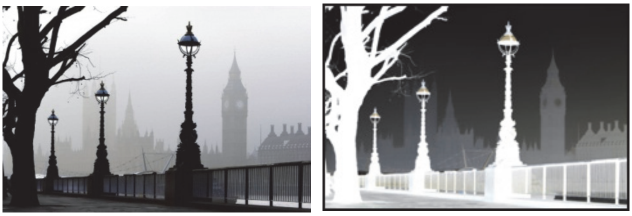
1.3.11. Relative and absolute depth measures

We will call a relative measure of the depth, when we only can know if a point is closer or farther than another one (or regarding the same point in a video sequence, when the frames go on), and an absolute measure of the depth, when we can know what is the real distance between a pixel and the camera. These results are constrained by the technology used, as we will see. Depending of the application, a relative measure, which uses to be lighter in computational load, may be enough. Likewise, we may need an absolute measure, so we will not be able to use some algorithms, technologies or setups.

1.4. The human visual perception of the depth

The human visual system is prepared for the depth perception. This perception is possible by a combination of different and complementary physiological and psychological structures and functions:

- Two eyes: the most important source of depth perception is the two eyes, sharing an important area of vision. However, the fronto-parallel hypothesis is only respected when looking at something placed in the infinity. If it is not the case, the configuration is that shown in figure 9. The angle of obliqueness (parallax) also provides information about the distance of the object.
- Focus: the crystalline is an elastic tissue which allows changing the focal distance of the eye and, hence, focusing in a wide range of distances. This information helps the brain computing the distance of the focused plane.
- Features extraction to match: many different image features extraction have been explored in the human visual system, such as shapes (Kurki & Saarinen, 2004), areas (Meese & Summers, 2009), colors (Jacobs, Williams, Cahill, & Nathans, 2007), movements (Stromeyer, Kronauer, Madsen, & et al., 1984) and other visual or psychological characteristics (Racheva & Vassilev, 2009), pattern (Georgeson, 1976) or a mixture of them (Guttman, Gilroy, & Blake, 2007).
- Differences in brightness: For constant illumination, the depth can be perceived in terms of the brightness. This method has been applied to compute the distance to stars (however, the hypothesis of constant brightness was not true), and works in daily live to help the brain knowing the distance, as perceived in figure 18.



(a) (b)

Figure 18. Depth perception through the fog. (a) original image, (b) inverse, similar to a 2.5D image.

- Finally, the structure of the perceived image can provide some depth information, although the brain can commit some errors when estimating the distance by this method, as seen in the following figure.

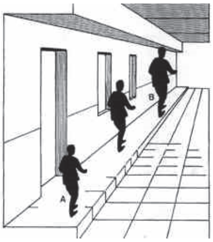


Figure 19. Visual deformation of the sizes of A and B due to structure perception of the depth.

Summarizing, we can take the human vision system as a set of functions and devices prepared to dynamically interact for a proper depth perception.

2. Depth estimation strategies

In computer vision, i.e. the set of algorithms implemented to process images or video in a complex way, the human visual system has been an important source of inspiration. Thus, we will find many algorithms trying to achieve some human capabilities, among others, the depth estimation.

However, there are other approaches to obtain the distance of a point (or a set of them). In general terms, we can divide all the methods to electronically measure the distance as *active* and *passive*.

2.1. Active methods

Active methods put some energy in the scene, projecting it in order to, in some way, illuminate the space, and processing, *passively*, the reflected energy. These methods were proposed before the passive ones, because of one main reason: the microprocessing was not even invented.

These methods present the main disadvantage, regarding the *passive* ones, in the energy needed. However, their accuracy use to be much higher, and some of them are used to obtain the ground truth.

2.1.1. Light based depth estimation

Light was the first kind of energy proposed to measure the distance. An example of this can be found in (Benjamin, 1973), working with incandescent light.

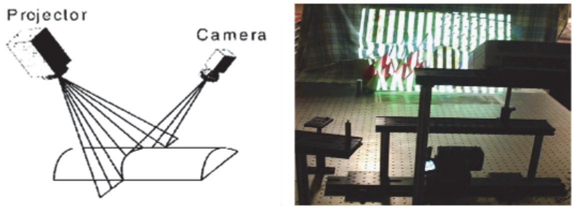
However, many light sources can be used and, hence, many different algorithms, setups and hardware are also available.

2.1.1.1. Incandescent light

Incandescent light is an uncorrelated emission of electromagnetic waves, produced by the high temperature of a coil. This is the basic setup for distance measuring and, hence, the first proposed. The information provided by such method is very rough, and only allows, under some conditions (for example, the system is very sensitive to the colour of the illuminated object), a measure in some small area, or even in a single direction. An example of this method has already been given.

2.1.1.2. Pattern projection

An improvement regarding the incandescent light (we should keep talking about incandescent light), is to produce it in a known pattern, which is projected to the scene. A camera, displaced from the light source, captures the geometrical distortion of the pattern. Figure 18 shows an example. This variant produces, with the help of a quite simple image processing, very accurate results.



(a) (b)

Figure 20. (a) pattern projection setup (Albrecht & Michaelis, 1998), and (b) figure 7 "Cones" scene from Middlebury database being processed to obtain fig.7c by structured light projection (Scharstein & Szeliski, 2003).

2.1.1.3. Time-of-Flight

The time of flight (ToF) principle uses the known speed of light to measure the time an emitted pulse of light takes to arrive to an image sensor (Schuon, Theobalt, Davis, & Thrun, 2008).

The emission can be made by IR LEDs, or Laser, the only sources to provide a short enough pulse to be useful for such measurements. Likewise, we can find different techniques inside this family, some of them moving the beam sequentially to illuminate the whole scene (as it is the case of Laser implementations, see (Saxena et al., 2008) for an example) or providing a pulse of light illuminating the whole scene in one single shot (LEDs options).

On the one hand, the main advantages of this proposal is the relatively high accuracy (in a sub-centimetre scale) and high processing rates (up to 100 fps) in the case of CMOS and LED based illumination (ODOS Imaging, 2012).This technology use to present, on the other hand, high power needs (10 W in the case of the SwissRanger (Mesa Imaging, 2011), 20 W in that used by Saxena (Saxena et al., 2008)) and cost (around $9000 for the SwissRanger).

2.1.2. Ultrasounds based methods

The ultrasounds based methods use the same ToF principle, applied to Ultrasounds. This technique has been largely applied, for example, in ultrasounds to examine foetus. As we saw in the case of light based ToF, sometimes it is necessary to perform a scanning (Douglas, Solomonidis, Sandham, & Spence, 2002).

2.2. Passive methods

We call passive methods for depth estimation to those techniques working with natural light in the ambient, and the optical information of the captured image. These techniques capture the images with image sensors, being the problem solved in a computational way. Thus, we will mostly talk about algorithms.

In this family of algorithms we can appreciate two former groups: monocular and multiview solutions.

2.2.1. Monocular solutions for the depth estimation

The first one uses one single image (or a video sequence of them) to obtain the depth map. The main limitation of this approach, as we will see, is the intrinsic limitation of the depth characteristics lost during the projection of the scene into the image plane. An advantage of this approach uses to be the relatively low amount of operations needed to process one single image, instead of two or more.

2.2.1.1. Image structure

Structures within the image can be analyzed to obtain approximation to the volume, as it is proposed in (François & Medioni, 2001). In this approximation, some basic structures are assumed, producing a relative volume computation of objects represented in an image.

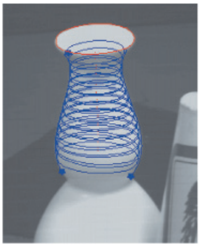


Figure 21. Structure estimation from a single image (François & Medioni, 2001).

Another related option is to compute the depth of well-known structures, such as human hands or faces (Nagai, Naruse, Ikehara, & Kurematsu, 2002), or indoor floors and walls (Delage, Lee, & Ng, 2005).

The measurement of distances in this proposal is relative. We cannot know the exact distance to each point of the image but just the relative distance among them. Moreover, some other disadvantages of these algorithms arise from the intrinsic limitation in terms of expected forms and geometries of figures appearing in the image. Perspective can trick this kind of algorithms producing uncontrolled results.

2.2.1.2 Points tracking or Optical flow

Tracking points in a set of images, which change with the time, supposed solid bodies, may drive to a structure of the space in which the video sequence has been recorded.

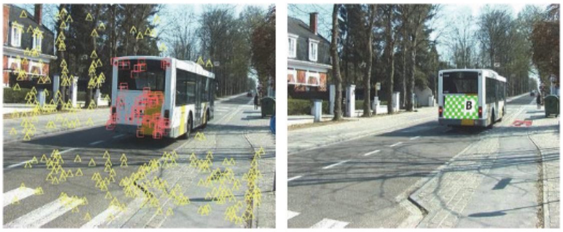


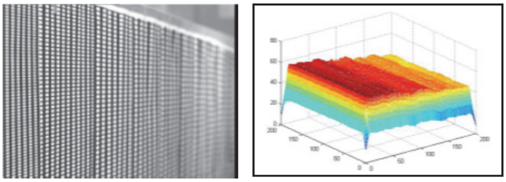
Figure 22. Augmented reality and 3D estimation through points relative movements in (Ozden, Schindler, & van Gool, 2007)

This approach provides, as in the previous case, a relative measure of the distances, tracking only relative variation in the positions of some relevant pixels.

2.2.1.3. Depth-on-defocus

The only approach that provides an absolute measurement of distance with monocular information is based in the focus properties of the image. This approach estimates the distance of every point in the image by computing the defocusing level of such points, following the human visual focusing system. This defocusing measurement is mainly done with Laplacian operators, which computes the second spatial derivative for every point in a neighbourhood of N pixels in each direction. Many other operators have been proposed, and a review of them can be found in (Helmi & Scherer, 2001).

Focused pixels provide an exact measurement of the distance, if the camera optical properties are known.



(a) (b)

Figure 23. Planar object distance estimation by focus (Malik & Choi, 2008).

This approximation has important errors when defocusing is high, and is very sensitive to texture features of the image and other noise distortions.

2.2.2. Multiview solutions for the depth estimation

In this group, we find algorithms dealing with two or more images to compute the depth map. Stereo vision is a particular case of this set, using two images. For clearness purposes, we will talk about stereo vision when two images are involved, and multiview for more than two images.

Some reasons explain why this new approach was proposed and, finally, widely used:

- Computation power available for civil and academic projects grew very fast for the last 20 years. This allows some algorithms to run in real time for the first time.
- Absolute measures may be needed in some environments, and the depth-on-focus only provides an accurate measure of the depth in a quite narrow field.
- Multiview systems, in some specific configurations, allow parallel computation, which can be a huge advantage when implementing them over GPUs or FPGAs or other parallel processing hardware.

Before presenting the most important approaches to solve the depth problem with multiview setups, we will discuss about the matching problem, which appears in this family for the first time.

2.2.2.1. The matching problem

This problem is posed for every stereo or multiview system (but not restricted to computer vision).

The matching problem can be solved with four main strategies: local, cooperative, dynamic programming and global approximations.

The first option takes into account only disparities within a finite window or neighborhood which presents similar intensities in both images (Islam & Kitchen, 2004; Williams & Bennamoun, 1998). The value of a matching criterion (sum-of-absolute-differences (SAD), sum-of-squared-differences (SSD) or any other characterization of the neighborhood of a pixel) for every windows positions is compared with the value for any other position. These windows are k×k pixel size. Then, this sum is optimized and the best match pixel is found. Finally, the disparity is computed from the abscissa difference of matched windows:

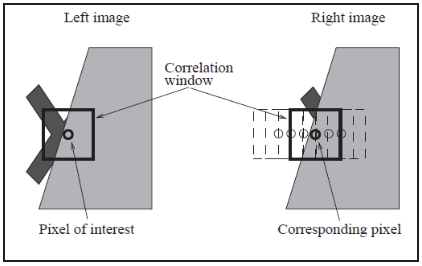


Figure 24. Moving window finding an edge. Graph taken from (Hirchsmüller, Innocent, & Garibaldi, 2002)

The main disadvantage can be clearly seen: the number of operations needed gives a global order of the algorithm of $o(n)=N^3\bullet k^4$ for a N×N image with windows of k×k pixels. This order is very high and these algorithms are not so fast, around 1 and 5 fps (Hirchsmüller et al., 2002) the fastest one. Another possibility for local matching is implemented by means of point matching. The basic idea consists on identifying important points (information relevant) in both images. After this process, all relevant points are identified and their disparity computed. These algorithms are neither too fast, achieving processing times of few seconds (Kim, Kogure, & Sohn, 2006). In the case of Lui (Liu, Gao, & Zhang, 2006), he gives time measures to obtain these results with a Pentium IV (@2.4GHz): 11.1 seconds and 4.4 seconds for the Venus and the Tsukuba pairs respectively. The main drawback is the necessity of interpolation. Only matched points are measured. After that, an interpolation of the non identified points is mandatory, increasing slightly the processing time. Another important disadvantage is the disparity computation on untextured surfaces, where the real depth reference is easily lost.

Cooperative algorithms were firstly proposed by Marr & Poggio (Marr & Poggio, 1976) and they were implemented trying to simulate how the human brain works. A two dimensional neural network iterates with inhibitory and excitatory connections until a stable state is reached. Later, some other proposals in this group have been proposed (Mayer, 2003; Zitnick & Kanade, 2000).

Dynamic programming strategy consists on assuming the ordering constraint as always true (Käck, 2004). The matching is done line by line, although the independent match of horizontal lines produces horizontal "streaks". The problem with the noise sensitivity of this proposal is smoothed with vertical edges (Ohta & Kanade, 1985) or ground control points (Bobick & Intille, 1999). These are some of the fastest proposals, achieving around 50 fps in a 3 GHz CPU (Kamiya & Kanazawa, 2008)

Global algorithms make explicit smoothness assumptions converting the problem in an optimization one. They seek a disparity assignment that minimizes a global cost or energy function that combines data and smoothness terms (Scharstein & Szeliski, 2002; Käck, 2004):

$$E(d)=E_{data}(d)+\lambda \bullet E_{smooth}(d) \tag{2}$$

Some of the best results with global strategies have been achieved with the so called graph cuts matching. Graph cuts extends the 1D formulation of dynamic programming approach to 2D, assuming a local coherence constraint, i.e. for each pixel, neighbourhoods have similar disparity. Each match is taken as a node and forced to fit in a disparity plane, connected to their neighbours by disparity edges and occlusion edges, adding a source node (with lower disparity) and a sink node (highest disparity) connected to all nodes. Costs are assigned to matches, and mean values of such costs to edges. Finally, we compute a minimum cut on the graph, separating nodes in two groups and the largest disparity that connect a node to the source is assigned to each pixel (Käck, 2004).

We can find also a group using some specific features of the image, like edges, shapes and curves (Schimd et al., 1999; Szumilas, Wildenauer, & Hanbury, 2009; Xia, Tung, & Ji, 2001). In this family, a differential operator must be used (typically Laplacian or Laplacian of Gaussian, as in (Pajares, Cruz, & López-Orozco, 2000; Jia et al., 2003)). This task requires a convolution of 3×3, 5×5 or even bigger windows; as a result, the computing load increases with the size of the operator (for separable implementations). However, these algorithms allow real-time implementations.

Another possibility of global algorithms are those of Belief propagation (Sun, Shum, & Zheng, 2002), modelling smoothness, discontinuities and occlusions with three Markov Random Fields and itinerates finding the best solution of a "Maximum A Posteriori" (MAP).

A final family of global algorithms to be referred in this study is the segment-based algorithms. This group of algorithms chops the image as explained in equation 1 to match regions. An initial pair of images is smoothed and segmented in regions. The aim of this family of algorithms addresses the problem of untextured regions. After forcing pixels to fit in a disparity plane, the depth map estimation is obtained.

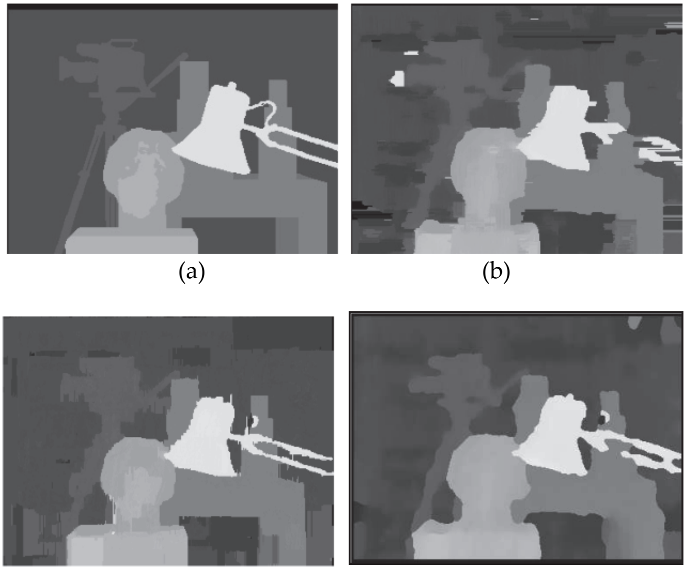
These algorithms have the advantage of producing a dense depth map, disparity estimated at each pixel (Scharstein & Szeliski, 2002), hence, avoiding interpolation. Some algorithms also perform a k×k window pre-match, and a plane fitting, producing a high computational load (and computation time of tens of seconds), and avoiding its use in real-time applications (Bleyer & Gelautz, 2005).

Combinations of segment-based and graph cuts algorithms have also been implemented (Hong & Chen, 2004).

A further group of global algorithms are based on wavelets, as described in (Xia et al., 2001). These algorithms present important problems in terms of time performance, around hours in 3 GHz CPU for two images matching (Radhika et al., 2007).

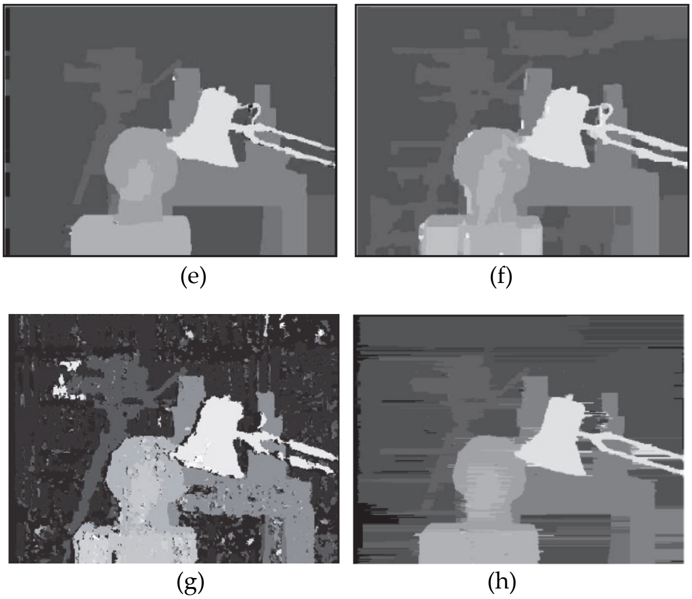
Summarizing, each of the previously described approaches to the matching problem presents several computing problems. In the case of edges, curves and shapes, differential operators increase the order linearly with the size (for separable implementations). This problem gets worse when using area-based matching algorithms, following the computational load an exponential law. The use of a window to analyze and compare different regions is seen to perform satisfactorily (Bleyer & Gelautz, 2005) however this technique requires many computational resources. Even most of segment-based matching algorithms perform a N×N local windowing matching as a step of the final depth map computation (Hong & Chen, 2004; Scharstein & Szeliski, 2002). It is important to notice that this step is not dimensional separable. Most of these algorithms, however, obtain very accurate results, with the counterpart of interpolating optimized planes that forces to solve linear systems (Hong & Chen, 2004; Klaus, Sormann, & Kraner, 2006). The calculations required for depth mapping of images is very high. It has been studied in detail, and a complete review of algorithms performing this task by means of stereovision can be found at (Scharstein & Szeliski, 2002).

Figure 25 shows some results of the presented algorithms.



(a) (b)

(c) (d)



(e)

(f)

(g)

(h)

Figure 25. (a) Ground truth of the Tsukuba scene (Scharstein & Szeliski, 2002), (b) Window 9x9 SAD matching (Hirchsmüller, 2001), (c) points matching (Liu, Gao, & Zhang, 2006), (d) cooperative algorithm (Zitnick & Kanade, 2000), (e) graph cuts depth estimation (Kolmogorov & Zabih, 2010), (f) Belief propagation (Sun et al., 2002), (g) segment regions and plane fitting (Bleyer & Gelautz, 2005), (h) dynamic programming (Scharstein & Szeliski, 2002).

In (Scharstein & Szeliski, 2002) a detailed stereo matching taxonomy can be found.

2.2.2.2. Stereo vision structure

The set of images used to compute the depth can be taken in many different ways, attending to their spatial organization. The first group being analyzed will be the stereo vision. This setup requires two cameras, closely placed and pointing to the scene. The figure 9 shows the general structure of a stereo vision images acquisition.

However, the stereo setup structure presents some free parameters, which may change the way the images should be analyzed. We have already seen some constraints, which allow some simplifications and, thus, fast algorithms, to extract the depth map, such as the fronto-parallel hypothesis (figure 11).

Stereo vision, as defined, allows obtaining a 2.5D image (or a 3D fragmented reconstruction, as it is shown in figure 3). Depending on how much are the image sensors are separated, we will be able to reconstruct more or less points of the volume analyzed. Following (Seitz & Kim, 2002), we can talk about central perspective stereo (when the displacement between both images is done in one single axis) and multiperspective stereo (otherwise). Regarding this last case, (Ishuguro, Yamamoto, & Tsuji, 1992) demonstrated how any perspective can be transformed to a stereo scene, under some geometrical and optical restrictions. In such case, the image rectification and dewrapping is mandatory.

2.2.2.3. 2 Multiview structure

The final case that we will present is the multiview setup. In this option, several cameras are placed around the scene, which is captured from different points of view. See figure 26 for an example.

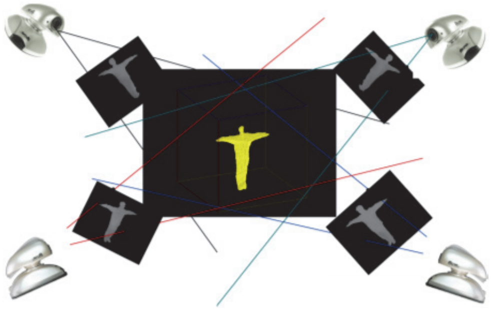


Figure 26. Multiview scheme (Kim, Kogure, & Sohn, 2006).

The algorithms dealing with this scheme need to perform a high number of matches, obtaining, however, a full 3D model, which is not restricted to a single perspective.

3. Conclusions

The depth is an important cue of a scene, which is lost in standard image acquisition systems. For that reason, and given that many applications need this information, several strategies have been proposed to extract the depth.

We have seen active methods, which project some energy onto the scene to process the reflections, and passive methods, only dealing with the natural received energy from the scene. Among this last option, we found monocular systems, working with a single perspective, and stereo or multiview systems, which work with more than one single perspective.

We have shown why these last algorithms have to solve the matching problem, or finding the same physical points in two or more images. Several strategies, again, are available in this category.

The analysis has revealed advantages and disadvantages in every system, regarding energy needs, computational load and, hence, speed, complexity, accuracy, range, hardware implementation or price, among others. Thus, there is not a concluding winner among all the analyzed solutions. Instead of that, we will have to think about the final application of our algorithm, to make the correct choice.

Author details

Pablo Revuelta Sanz, Belén Ruiz Mezcua and José M. Sánchez Pena
Carlos III University of Madrid, Spain

Acknowledgement

We would like to acknowledge the student grant offered by the Carlos III University of Madrid and the Spanish Center for Subtitling and Audiodescription (CESyA) which has allowed this research work to be performed.

4. References

Albrecht, P. and Michaelis, B. (1998). Improvement of the Spatial Resolution of an Optical 3-D Measurement Procedure. *IEEE Transactions on Instrumentation and Measurement,* Vol.47, pp. 158-162.

Benjamin, J. M. Jr. (1973). The Laser Cane. *Bulletin of Prosthetics Research,* pp. 443-450.

Bleyer, M. (2006). Thesis: "Segmentation-based Stereo and Motion with Occlusions", Institute for Software Technology and Interactive Systems, Vienna University of Technology.

Bleyer, M. and Gelautz, M. (2005). A layered stereo matching algorithm using image segmentation and global visibility constraints. *Journal of Photogrammetry & Remote Sensing,* Vol.59, pp. 128-150.

Bobick, A. & Intille, S. (1999). Large occlusion stereo. *International Journal of Computer Vision,*Vol. 33, pp. 181-200.

Delage, E., Lee, H., & Ng, A. Y. (2005). Automatic single-image 3D reconstructions of indoor Manhattan world scenes. In: *12th International Symposium of Robotics Research (ISRR),* pp.305-321.

Douglas, T. S., Solomonidis, S. E., Sandham, W. A., and Spence, W. D. (2002). Ultrasound image matching using genetic algorithms. *Medical and Biological Engineering and Computing,* Vol.40, pp. 168-172.

Espindola, G. M., Camara, G., Reis, I. A., Bins, L. S., and Monteiro, A. M. (2006). Parameter selection for region-growing image segmentation algorithms using spatial autocorrelation. *International Journal of Remote Sensing,* Vol.27, pp. 3035-3040.

François, A. R. J. and Medioni, G. G. (2001). Interactive 3D model extraction from a single image. *Image and Vision Computing,* Vol.19, pp. 317-328.

Gao, L., Jiang, J., and Yang, S. Y. (2006). Constrained Region-Growing and Edge Enhancement Towards Automated Semantic Video Object Segmentation. *Lecture Notes in Computer Science, Advanced Concepts for Intelligent Vision Systems,* Vol.4179, pp. 323-331.

Georgeson, M. (1976). Antagonism between Channels for Pattern and Movement in Human Vision. *Nature,* Vol.259, pp. 412-415.

Guttman, S., Gilroy, L. A., and Blake, R. (2007). Spatial grouping in human vision: Temporal structure trumps temporal synchrony. *Vision Research,*Vol. 47, pp. 219-230.

He, Z. & Wang, Q. (2009). A Fast and Effective Dichotomy Based Hash Algorithm for Image Matching. *Lecture Notes in Computer Science, Advances in Visual Computing,* Vol. 5358, pp. 328-337.

Helmi, F. S. & Scherer, S. (2001). Adaptive Shape from Focus with an Error Estimation in Light Microscopy. *2nd Int'l Symposium on Image and Signal Processing and Analysis,* pp. 188-193.

Hirchsmüller, H. (2001). Improvements in real-time correlation-based stereo vision. In: *IEEE Workshop on Stereo and Multi-Baseline Vision at IEEE Conference on Computer Vision and Pattern Recognition*, December 2001, Kauai, Hawaii, pp. 141-148.

Hirchsmüller, H., Innocent, P. R., and Garibaldi, J. (2002). Real-Time Correlation-Based Stereo Vision with Reduced Border Errors. *Journal of Computer Vision,* Vol. 47, pp. 229-246.

Hong, L. & Chen, G. (2004). Segment-based Stereo Matching Using Graph Cuts. *Computer Vision and Pattern Recognition (CVPR) 2004. Proceedings of the 2004 IEEE Computer Society Conference on,* Vol. 1, pp. I-74-I-81.

Ishuguro, H., Yamamoto, M., & Tsuji, S. (1992). Omni-directional stereo. *PAMI,* Vol.14, pp. 257-262.

Islam, M. S. & Kitchen, L. (2004). Nonlinear Similarity Based Image Matching. *International Federation for Information Processing,* Vol.228, pp. 401-410.

Jacobs, G. H., Williams, G. A., Cahill, H., and Nathans, J. (2007). Emergence of Novel Color Vision in Mice Engineered to Express a Human Cone Photopigment. *Science,* Vol.315, pp. 1723-1727.

Jia, Y., Xu, Y., Liu, W., Yang, C., Zhu, Y., Zhang, X. et al. (2003). A Miniature Stereo Vision Machine for Real-Time Dense Depth Mapping. *Lecture Notes in Computer Science, Computer Vision Systems,* Vol.2626, pp. 268-277.

Käck, J. (2004). *Robust Stereo Correspondence using Graph Cuts.* Master Thesis, Royal Institute of Technology. Available from:
www.nada.kth.se/utbildning/grukth/exjobb/rapportlistor/-2004/rapporter04/kack per-jonny 04019.pdf

Kamiya, S. & Kanazawa, Y. (2008). Accurate Image Matching in Scenes Including Repetitive Patterns. *Lecture Notes in Computer Science, Robot Vision,* Vol.4931, pp. 165-176.

Kim, H. K. I., Kogure, K., & Sohn, K. (2006). A Real-Time 3D Modeling System Using Multiple Stereo Cameras for Free-Viewpoint Video Generation. *Lecture Notes in Computer Science, Image Analysis Recognition*, Vol.4142, pp. 237-249.

Kim, J. Ch., Lee, K. M., Choi, B. T., & Lee, S. U. (2005). A dense stereo matching using two-pass dynamic programming with generalized ground control points. In: *Computer Vision and Pattern Recognition (CVPR) 2005. IEEE Computer Society Conference on,* pp. 1075-1082

Klaus, A., Sormann, M., and Kraner, K. (2006). Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. *Pattern Recognition (ICPR) 2006. 18th International Conference on,* pp. 15-18.

Kolmogorov, V. & Zabih, R. (2010). *Computing visual correspondence with occlusions via graph cuts*. Rep. No. Technical Report CUCS-TR-2001-1838, Cornell Computer Science Department.

Kostková, J. & Sára, R. (2006). *Fast Disparity Components Tracing Algorithm for Stratified Dense Matching Approach*. Rep. No. Research Reports of CMP, Czech Technical University, No. 28.

Kurki, I. & Saarinen, J. (2004). Shape perception in human vision: specialized detectors for concentric spatial structures?. *Neuroscience Letters,* Vol.360, pp. 100-102.

Liu, L., Gao, H.-B. & Zhang, Q. (2006). Research of Correspondence Points Matching on Binocular Stereo Vision Measurement System Based on Wavelet. *CORD Conference Proceedings,* pp. 3687-3691. Available from: http://pubget.com/paper/pgtmp_a32b66de88e2f5adb012c343cb5f2bf4

Malik, A. S. & Choi, T.-S. (2008). Depth Estimation by Finding Best Focused Points Using Line Fitting. *Lecture Notes in Computer Science, Image and Signal Processing*, Vol.5099, pp. 120-127.

Marr, H. & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science,* Vol.194, pp. 283-287.

Mayer, H. (2003). Analysis of means to improve cooperative disparity estimation. *ISPRS Conference on Photogrammetric Image Analysis,* JSPRS Archives, Vol.XXXIV, Part 3/W8.

Meese, T. S. & Summers, R. J. (2009). Area summation in human vision at and above detection threshold. In: *Proceedings of the Royal Society B: Biological Sciences,* Vol.274, pp. 2891-2900.

Mesa Imaging. (2011). SR4000 Data Sheet. 20-1-2012. Avaliable from: http://www.mesa-imaging.ch/pdf/SR4000_Data_Sheet.pdf

Nagai, T., Naruse, T., Ikehara, M., & Kurematsu, A. (2002). Hmm-based surface reconstruction from single images. In. *Image Processing. 2002. Proceedings. 2002 International Conference on,* Vol.2, pp. II-561 - II-564

Nathans, J. (1999). The Evolution and Physiology of Human Color Vision: Insights from Molecular Genetic Studies of Visual Pigments. *Neuron,* Vol.24, pp. 299-312.

ODOS Imaging. (2012). 2+3D™ - real world in real time. 20-1-2012. Available from: http://odos-imaging.com/

Ohta, Y. & Kanade, T. (1985). Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.7, pp. 139-154.

Ozden, K. E., Schindler, K., and van Gool, L. (2007). Simultaneous Segmentation and 3D Reconstruction of Monocular Image Sequences. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on,* pp. 1-8.

Pajares, G., Cruz, J. M., and López-Orozco, J. A. (2000). Relaxation labeling in stereo image matching. *Pattern recognition,* Vol.33, pp. 53-68.

Pham, D. L., Xu, C., and Prince, J. L. (2000). Current Methods in Medical Image Segmentation. *Annual Review of Biomedical Engineering,* Vol.2, pp. 315-337.

Pons, J.-P. & Keriven, R (2007). Multi-View Stereo Reconstruction and Scene Flow Estimation with a Global Image-Based Matching Score. *International Journal of Computer Vision,* Vol.72, pp. 179-193.

Racheva, K. & Vassilev, A. (2009). Human S-Cone Vision Effect of Stiumuls Duration in the Increment and Decrement Thresholds. *Comptes rendus de l'Academie bulgare des Sciences,* Vol.62, pp. 63-68.

Radhika, V. N, Kartikeyan, B., Krishna, G., Chowdhury, S., and Srivastava, P. K. (2007). Robust Stereo Image Matching for Spaceborne Imagery. *IEEE Transactions on Geoscience and Remote Sensing,* Vol.45, pp. 2993-3000.

Rangarajan, S. (2005), *Algorithms for Edge Detection*, Stony Brook University. Available from: www.uweb.ucsb.edu/~shahnam/AfED.doc

Revuelta Sanz, P., Ruiz Mezcua, B., Sánchez Pena, J. M., & Thiran, J.-P. Stereo Vision Matching over Single-channel Color-based Segmentation, In: *International Conference on Signal Processing and Multimedia Applications (SIGMAP) 2011 Proceedings*, pp. 126-130.

Saxena, A., Chung, S. H., and Ng, A. Y. (2008). 3-D Depth Reconstruction from a Single Still Image. *International Journal of Computer Vision,* Vol.76, pp. 53-69.

Scharstein, D. & Szeliski, R. (2002). A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision, 47,* 7-42.

Scharstein, D. & Szeliski, R. (2003). High-Accuracy Stereo Depth Maps Using Structured Light. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) 2003,* Vol.1, pp. 195-202, Madison, WI.

Scharstein, D. Middlebury Database. (2010). www.middlebury.edu/stereo

Schimd, C., Zisserman, A., & Mohr, R. (1999). Integrating Geometric and Photometric Information for Image Retrieval. *Lecture Notes in Computer Science, Shape, Contour and Grouping in Computer Vision,* Vol.1681, pp. 217-233.

Schuon, S., Theobalt, Ch., Davis, J., & Thrun, S. (2008). High-quality scanning using time-of-flight depth superresolution. In: *IEEE CVPR Workshop on Time-Of-Flight Computer Vision 2008*, pp. 1-7

Seitz, S. M. & Kim, J. (2002). The Space of All Stereo Images. *International Journal of Computer Vision,* Vol.48, pp. 21-38.

Stromeyer, C. F., Kronauer, R. E., Madsen, J. C., and et al. (1984). Opponent-Movement Mechanisms in Human-Vision. *Journal of the Optical Society of America A-Optics Image Science and Vision,* Vol.1, pp. 876-884.

Sun, J., Shum, H.-Y., and Zheng, N.-N. (2002). Stereo matching using belief propagation. In: *European Conference on Computer Vision,* pp. 510-524.

Szumilas, L., Wildenauer, H., & Hanbury, A. (2009). Invariant Shape Matching for Detection of Semi-local Image Structures. *Lecture Notes in Computer Science, Image Analysis Recognition*, Vol.5627, pp. 551-562.

Tuytelaars, T. & Gool, L.V. (2004). Matching Widely Separated Views Based on Affine Invariant Regions. *International Journal of Computer Vision,* Vol.59, pp. 61-85.

Wang, Ch. & Gavrilova, M. L. (2005). A Novel Topology-Based Matching Algorithm for Fingerprint Recognition in the Presence of Elastic Distortions. *Lecture Notes in Computer Science, Computational Science and Its Applications ICCSA*, Vol.3480, pp. 748-757.

Wang, X. L. & Wang, L. J. (2008). Color image segmentation based on Bayesian framework and level set. *Proceeding of 2008 International Conference on Machine Learning and Cybernetics,* Vol.1, No.7, pp. 3484-3489.

Williams, J. & Bennamoun, M. (1998). A Non-linear Filtering Approach to Image Matching. In: *Proceedings of the 14th International Conference on Pattern Recognition,* Vol.1, No.1, p. 3.

Xia, Y., Tung, A., and Ji, Y. W. (2001). A Novel Wavelet Stereo Matching Method to Improve DEM Accuracy Generated from SPOT Stereo Image Pairs. *International Geoscience and Remote Sensing Symposium*, Vol.7, pp. 3277-3279.

Yu, J., Weng, L., Tian, Y., Wang, Y., and Tai, X. (2008). A Novel Image Matching Method in Camera-calibrated System. In: *Cybernetics and Intelligent Systems, 2008 IEEE Conference on*, pp. 48-51.

Zitnick, L. & Kanade, T. (2000). A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transactions on Pattern Analysis and Machine Inteligence*, Vol.22, pp. 675-684.