A Review On 3D Reconstruction Techniques From 2D Images

M. Aharchi and M. Ait Kbir

LIST Laboratory, Faculty of Sciences and Technologies, Tangier, Morocco
maharchi@uae.ac.ma, m.aitkbir@fstt.ac.ma

Abstract. In recent years, 3D model visualization techniques have made enormous progress. This evolution has not only touched the technical side but also the hardware side. It is no longer necessary to have expensive machines to see a world in 3D; a simple computer can do the trick. Historically, research has focused on the development of 3D information and acquisition techniques from scenes and objects. These acquisition methods require expertise and complex calibration procedures whenever the acquisition system was used. All this creates an important demand for flexibility in these methods of acquisition because of these different factors, many techniques have emerged. Many of them only need a camera and a computer to create a 3D world from a scene.

Keywords: 3D reconstruction, Camera, Multiple Views, 2D images, depth perception.

1 Introduction

Creating a 3D model from 2D images that is realistic as possible is one of the fundamental issues of image-based modeling and computer vision. The best choice is an automatic reconstruction of the scene with little or no user intervention. Currently, there are several methods of 3D reconstruction from 2D images; each algorithm has its own conditions of execution, its strengths as well as its weak points.

In this paper, we give a definition and the domains of uses of 3D reconstruction from 2D images. We propose a set of 3D reconstruction algorithms from these 2D images as well as a comparison between them. Afterwards, we conclude this paper with summarize the related to this study.

2 Overview On 3d Reconstruction From Images

2.1 2.1 What Is 3D Reconstruction from Images

3D reconstruction from multiple images is the creation of three-dimensional models from a set of images. It is the reverse process of obtaining 2D images from 3D scenes.

An image does not give us enough information to reconstruct a 3D scene. This is due to the nature of image forming process that consists of projecting a three-dimensional scene onto a two-dimensional image. During this process, the depth is lost. The points visible on the images are the projections of the real points on the image.

2.2 Fields of Application of 3D reconstruction from Images

3D reconstruction has applications in many fields. They are:
Medicine, Free-viewpoint video reconstruction, robotic mapping, city planning, Gaming, virtual environments and virtual tourism, landslide inventory mapping, robot navigation, archaeology, augmented reality, reverse engineering, motion capture, Gesture recognition and hand tracking …

In medicine, 3D reconstruction from 2D images can be used for both therapeutic and diagnostic purposes by using a camera to take multiple images at multiple angles. Even if there are medical imaging techniques like MRI and CT scanning, they are still expensive and can induce high radiation doses, which is a risk for patients with certain diseases, and they are not suitable for patients with ferromagnetic metallic implants. Both the methods can be done only when the patient is in lying position where the global structure of the bone changes. There are some techniques of 3D reconstruction like Stereo Corresponding Point Based Technique or Non-Stereo corresponding contour method (NCSS) which can be performed while standing and require low radiation dose by using X-ray images. [1]

In the world of robotic navigation, an autonomous robot can use the images taken by its camera to create a 3D map of its environment and use it to perform real-time processing in order to find its way or to avoid obstacles that can arise at any moment in its path, as well as making measurements on the space where it is located. [2]

The estimation of landslides dimension is a significant challenge while preparing the landslide inventory map, for which satellite aerial/data photography is required, which is very expensive. An alternative is the use of drones for such mapping. The result of 3D reconstruction from 2D images is very accurate and gives the possibility of measurements up to cm level and even small objects could be easily identified. By using images taking by a drone in combination with 3D scene reconstruction algorithms we can provide effective and flexible tools to monitor and map landslide. [3]

2.3 3D Reconstruction from Images Requirements

To obtain, as desired, the coordinates of the points of the scene, it is necessary to solve a certain number of problems:

Calibration problem.
Calibration problem or how the points of the scene are projected on the image. For this, the pinhole model is used and it is then necessary to know so-called intrinsic parameters of the camera (focal length, center of the image ...). Then, it is necessary to know the relative position of the cameras to be able to determine the coordinates of the points of the space in a reference of the scene not linked to the camera. These parameters, called extrinsic, are the position and orientation of the camera in space.

Matching problem.
Is the ability to recognize and associate the points that appear on several pictures.

Reconstruction problem.
It is a question of determining the 3D coordinates of the points from the associations made and the parameters of calibration.

The density of the reconstruction.
Once the coordinates of a certain number of points in space have been obtained, it is necessary to find the surface to which these points belong to obtain a mesh, a dense model. Otherwise, in some cases, when we obtain a large number of points, the cloud of points formed is enough to visually define the shape of the object but the reconstruction is then sparse.

3 Active And Passive Reconstruction Methods

The depths restoration process of visible points on the image can be achieved by active or passive methods.

3.1 Active methods

In order to acquire a depth map, active methods actively interfere with the object to be reconstructed through radiometric or mechanical techniques (laser rangefinder, structured light and other active detection techniques). For example, a depth map can be reconstructed using a depth gauge to measure the depth relative to an object placed on a rotating plate or using radiometric methods through moving light sources, colored visible light , time-of-flight lasers, microwaves or ultrasounds that emit radiance towards the object and then measure its reflected part.

3.2 Passive methods

Passive methods of 3D reconstruction do not interfere with objects to be rebuilt; they only use a sensor to measure the luminance reflected or emitted by the surface of the object in order to deduce its 3D structure through image processing. The sensor used in the camera is an image sensor in sensitive to visible light. The input elements for this process are a set of digital images (one, two or more) or video. For this case, we are talking about image-based reconstruction and the output element is a 3D model [5] .

4 2D TO 3D CONVERTION ALGORITHMS

Depending on the number of input images, we can categorize the existing conversion algorithms into two groups: algorithms based on two or more images and algorithms based on a single still image. In the first case, the two or more input images could be taken by multiple fixed cameras located either at different viewing angles or by a single camera with moving objects in the scenes. We call the depth cues used by the first group the multi-ocular depth cues. The second group of depth cues operates on a single still image, and they are referred to as the monocular depth cues.

Table 1 summarizes the depth cues used in 2D to 3D conversion algorithms.

Table 1. Depth cue used in 2D to 3D conversion algorithms.

Number of Input Images	Depth Cues
Two or More Images	Binocular disparity Motion parallax Image blur Silhouette Structure from motion
One single image	Linear perspective Atmosphere scattering Shape from shading

4.1 Binocular disparity

By using two images of the same scene captured from slightly different points of view, we can manage to recover the depth of a point present on the two images. First, a corresponding set of points in both images are found. Then, using the method of triangulation we can get to determine the depth of a point on the images [6] .

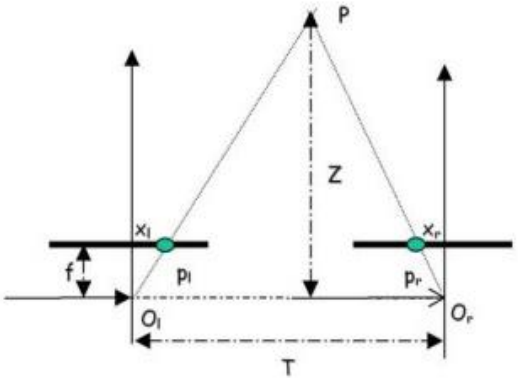

Figure 1. Binocular disparity. [15]

We assume that Pl and Pr are the two projections of the points P on the two images and O1 and Or are the origins of the coordinate systems of the two cameras. Based on the relationship between the triangles (P P1 pr) and (P O1 Or) the depth Z of the point P can be obtained where D= xr - xl.

$$Z = f\frac{T}{D}$$

4.2 Motion Parallax

The relative movement between the camera and the scene provides important clues in the perception of depth. Objects that are close to the camera move faster than the objects that are near. The extraction of 3D structures and the camera is termed as structure from motion. The motion can be seen as a form of disparity over time, represented by the concept of motion field. The motion field is the 2D velocity vectors of the image points and the observed scene. The basic assumptions for structure-from-motion are that do not deform objects and their movements are linear. This fact has been exploited in several applications, such as wiggle stereoscopy [7] where motion parallax is used as a metaphor for stereoscopic images, or parallax scrolling [8] used in games where, by moving foreground and background at different speeds, a depth sensation is evoked. The strength of this cue is relatively high when compared to other monocular cues and also when compared to binocular disparity.

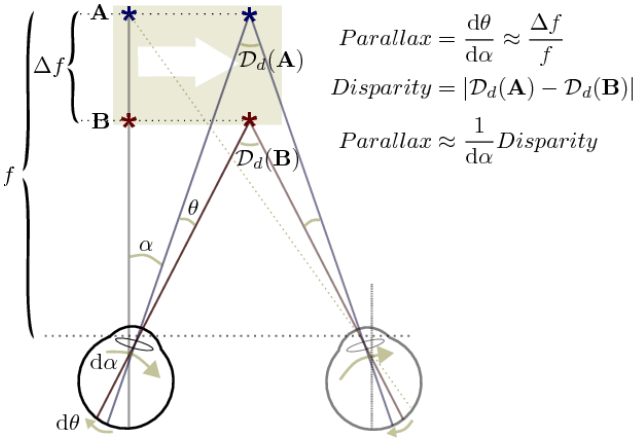

Figure 2. Motion Parallax Mechanics.

4.3 Image Blur

Evidence for the use of image blur in depth perception has been reported by Mather [9] and by Marshall et al [4] . Their papers described experiments on ambiguous figure-ground stimuli, containing two regions of texture separated by a wavy boundary. Objects that are in-focus are clearly pictured whilst objects at other distances are defocused.

The following general expression relates the distance d of a point from a lens to the radius s of its blurred image (Pentland 1987):

$$d = Frv/(rv- F(r + s)) \quad (1)$$

Where F is focal length, r is lens aperture radius, and v is the distance of the image plane from the lens. If we know the values of F, r, and v and a measure of image blur s is available, then absolute distance can be calculated. Eq. (1) can be used to predict retinal blur as a function of distance, on assuming typical values for the optical parameters of the human eye (r = 1.5 mm, v = 16 mm).

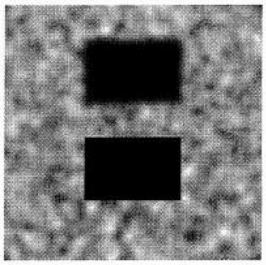

Figure 3. Blur as a depth cue in random patterns.

The upper rectangle is more far than the lower black rectangle because the upper rectangle has sharply defined edges.

4.4 Silhouette

A silhouette of an object in an image refers to the contour separating the object from the background. Shape-from-silhouette methods require multiple views of the scene taken by cameras from different viewpoints. Such a process together with correct texturing generates a full 3D model of the objects in the scene, allowing viewers to observe a live scene from an arbitrary viewpoint. Shape-from-silhouette requires accurate camera calibration.

For each image, the silhouette of the target objects is segmented using background subtraction. The retrieved silhouettes are back projected to a common 3D space with projection centers equal to the camera locations. Back-projecting a silhouette produces a cone-like volume. The intersection of all the cones forms the visual hull of the target 3D object, which is often processed in the voxel representation. This 3D reconstruction procedure is referred to as shape-from-silhouette [10] .

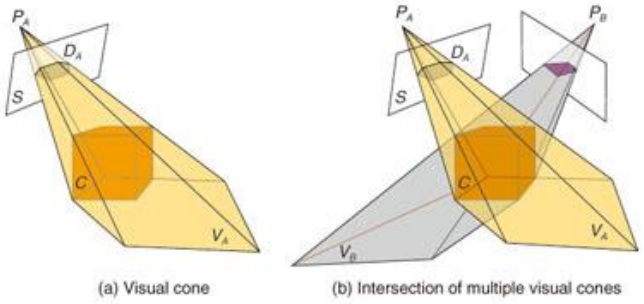

Figure 4. Silhouette volume intersection.

C denotes a cube, which is an example of a 3D object, S denotes a 2D screen, PA denotes a viewpoint in 3D space, DA denotes a 2D polygon on the screen, which is the silhouette of the cube, and VA denotes the visual cone backprojected from the viewpoint PA. PB, DB, and VB denote the corresponding meanings to PA, DA, and VA. [18]

4.5 Linear Perspective

Linear perspective refers to parallel lines such as roads or pathways that converge with distance. The points of line of these lines are less visible than those of the nearest ones. The approach proposed by Battiato, Curti et al. [12] works for images containing surfaces with rigid geometry. The intersection with the most intersection points in the neighborhood is considered to be the vanishing point. The major lines close to the vanishing point are marked as the vanishing lines. Between each pair of neighboring

vanishing lines, a set of gradient planes is assigned, each corresponding to a single depth level. The pixels closer to the vanishing points are assigned a larger depth value and the density of the gradient planes is higher [11] .

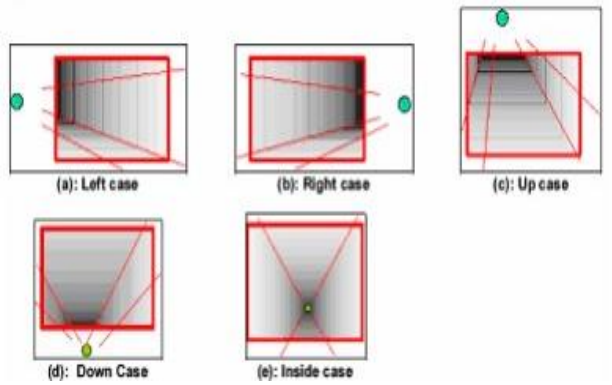

Figure 5. Blur as a depth cue in random patterns. [16]

4.6 Atmosphere Scattering

Atmosphere scattering approach is based on the fact that the power and direction of light are changed when the light passes through the atmosphere because of small particles present in it. The objects that are close to the camera appear clearer while those near are more blurred. [12] .In 1997, Krotkov and Cozman and [13] presented an analysis of this conversion. It was based on Lord Rayleigh's 1871 physical scattering model. Their algorithm is suitable for estimating the depth of outdoor images containing a portion of sky.

Figure 6. Depth map from atmosphere scattering. [18]

4.7 Shape From Shading

Shape from shading is a technique that allows knowing the surface normal of an object by observing the reflectance of the light on this object. The amount of light reflected by the surface of the object depends on the orientation of the object. Woodham introduced this technique in 1980. When the data is a single image, we call it shape from shading, and it was analyzed by B. K. Horn P. In 1989. Photometric stereo has since been generalized to many other situations, like non-Lambertian surface finishes and extended light sources.

Multiple images of an object under different lighting are analyzed to produce an estimated normal direction at each pixel [14] .

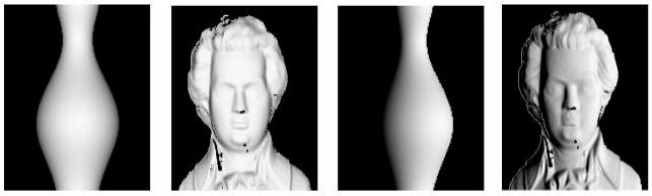

Figure 7. Explanatory figure of shape from shading [17].

4.8 Structure From Motion

Structure from Motion (SfM) is a technique that uses a series of two-dimensional images of a scene or object to reconstruct its three-dimensional structure. SfM can produce 3D models based on high-resolution point clouds.

SfM is based on the same principles as stereoscopic photogrammetry. In stereophogrammetry triangulation is used to calculate the relative 3-D positions (x,y,z,) of objects from stereo pairs. Traditionally these techniques require expensive specialized equipment and software.

To create a 3D reconstruction one simply needs many images of an area or an object with a high degree overlap, taken from different angles. The camera doesn't need to be specialized, standard consumer-grade cameras work well for SfM methods. The images are often be taken from a moving sensor or by a one or multiple people at different locations and angles. SfM involves the three main stages:

Step 1:

Match corresponding features and measure distances between them on the camera image plane d, d'. Scale Invariant Feature Transform (SIFT) (Lowe, 1999) [30] allows corresponding features to be matched even with large variations in scale and viewpoint and under conditions of partial occlusion and changing illumination.

Step 2:

When we have the matching locations of multiple points on two or more photos, there is usually just one mathematical solution for where the photos were taken. Therefore, we can calculate individual camera positions (x, y, z), (x', y', z'), orientations i, i', focal lengths f, f', and relative positions of corresponding features b, h, in a single step known as "bundle adjustment". This is where the term Structure from

motion comes from. Scene structure refers to all these parameters; motion refers to movement of the camera.
Step 3:

Next, a dense point cloud and 3D surface is determined using the camera parameters and using the SfM points. This step is called “multiviewstereo matching” (MVS)

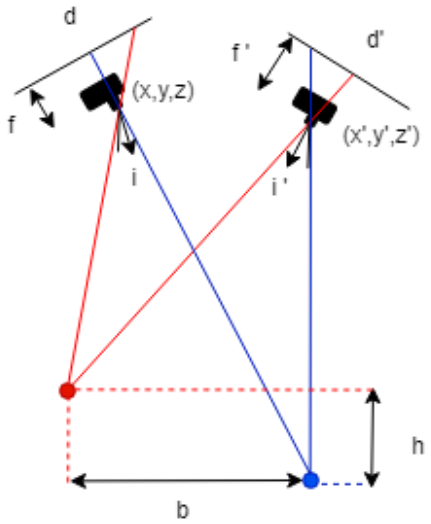

Figure 8. Explanatory figure of structure from motion.

5 Comparison

To perform an effective evaluation of the performances of the different algorithms of 2D to 3D conversion requires careful design of criteria and sets of data. Many articles do not provide a quantitative performance analysis explicit, which complicates the evaluation process. It is therefore unfair to say that a method represents a lower error rate or a slower execution time than others methods. Another complication is that for each individual algorithms, each having different characteristics and performances. Therefore, when we discuss issues such as accuracy or speed of algorithms of each depth cues, we rely, where appropriate, on the experimental results presented in the articles of these representative algorithms. The comparison is based on 8 qualitative aspects. The results are presented in the tables 2.

5.1 Relative or Absolute Depth

Most algorithms that rely on camera settings can recover the actual depth. Some algorithms provide a real (absolute) distance between the viewing camera and objects, and they are able to estimate the actual size of the object; other algorithms measure

relative depth by analyzing shading, edges and junction, etc., providing a relative depth but not the actual values. On the other hand, monocular depth cannot be used to estimate the actual depth. Except in the case where correct the values obtained using machine learning techniques.

5.2 Depth Range

This aspect describes the effective depth range a human can perceive based on each individual depth cue. For example, linear perspective works in all ranges; and atmospheric scattering works only at large distance.

5.3 Real Time Processing

Some of the articles provide explicit run-time and environment parameters, others simply state that the algorithm is suitable for a real-time application or does not mention speed at all.

The speed of algorithms is related to their accuracy. Greater accuracy requires more processing time. In order to obtain a real-time execution speed, it is simply necessary to reduce the accuracy to an acceptable limit. Traditional depth-of-focus methods [15] are fast but less accurate. Several researchers have published different techniques to improve accuracy, but they require very high computing costs. One of the examples of these suggestions for improving this algorithm is that proposed by Ahmad and Choi [16] . There the optimization technique of dynamic programming. Shape-from-Silhouette is memory intensive and normally computationally. Thanks to various techniques such as parallel PC processing or 2D intersection, it is possible to meet the real-time criterion. The hardware accelerated visual hulls algorithm developed by Li et al. [17] linked to several consumer PCs following a server-client architecture and makes arbitrary views of the visual hull directly from the input silhouette.

5.4 Image Acquisition

This aspect describes whether the method used is active or passive. In other words, parameters of the image acquisition system. Almost all multi-ocular depth cues require a special camera set-up, and most monocular depth cues do not.

5.5 Image content

Indicates the characteristic that the image must have in order to be processed by algorithms and work properly.

5.6 Motion Presence

This aspect describes whether the points on the image are moving on the different images or not. It is only applicable for multi-ocular depth cues. Since monocular depths operate on a single picture.

5.7 Dense or Sparse Depth Map

This aspect deals with the density level of the depth map. It can be either dense or sparse. Some depth cues can generate dense and sparse depth maps, depending on whether the specific algorithm uses local feature points or global structures. A dense depth map is constructed using the features of the overall image. Each level of depth is assigned to each pixel of the image. A sparse depth map provides only depth values for feature points. It is more suitable for the extraction of 3D shapes.

5.8 State of Each Depth Cue

This aspect indicates when a certain algorithm has been published compared to other type of algorithms in the field of computer vision.

Table 2. Depth cue comparison

Depth Cues	Relative /Absolute Depth	Depth Range	Real Time Processing	Image Acquisition	Image content	Motion Presence	Dense or Sparse Depth Map	State of Depth Cue
Binocular disparity	Absolute	< 30 m	yes	Active	All	yes	Dense/sparse	1976, Marr & Poggio [23]
Motion parallax	Absolute	< 30 m	yes	Active/passive	All	yes	Dense/sparse	1979, Ullman [24]
Image blur	Absolute	N/A	yes	Active	Objects with complex surface characteristic	no	Dense	1987, Pentland [25]
Silhouette	Absolute	Indoor size	yes	Active	Foreground objects must be distinguishable from background	yes	Sparse	1983, Martin & Aggarwal [26]
Linear perspective	Relative	All ranges	yes	Passive	Image contains geometric appearance	no	dense	1980, Haralick [27]
Atmosphere scattering	Relative	900 - 8000 m	N/A	Passive	Scene contains haze	no	dense	1997, Cozman and Krotkov [13]
Shape from shading	Relative	All ranges	no	Passive	Image must not be too dark.	no	Dense on surface	1975, Horn [28]
Structure from motion	Absolute	< 30 m	yes	Active/passive	All	yes	yes	2006,Snavely,Seitz,Szeliski [29]

6 CONCLUSION

A large number of 2D to 3D conversion algorithms are dedicated to the recovery of 3D shape from object in a scene. Each of these algorithms has it own requirements. These algorithms can be better used in different domains such as monitoring, robot navigation, etc. No depth cue is better or indispensable than other depth cue. Each cue has its own advantages and disadvantages.

It is necessary to combine multiple depth cues in order to achieve a robust conversion algorithm. Some depth cues produce less detailed surface information due to reasons such as smoothness constraints other depth cues offers a better detailed surface, combing them may leads to a better result.

References

1. 3D reconstruction from multiple images,Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/3D_reconstruction_from_multiple_images. (Accessed 6 August 2019)
2. Tamas Fazakas, Róbert Tamás Fekete, “3D reconstruction system for autonomous robot navigation”, 2010 11th International Symposium on Computational Intelligence and Informatics, Budapest (November 2010)
3. Gupta, Sharad & Shukla, Dericks. “Application of drone for landslide mapping, dimension estimation and its 3D reconstruction.” Journal of the Indian Society of Remote Sensing. . (January 2018).
4. Marshall J. A., Burbeck C. A., Ariely D., Rolland J. P., Martin K. E. “Occlusion edge blur: A cue to relative visual depth” , Journal of the Optical Society of America A, (1996). 13, 681–688.
5. Steffen Herbort and Christian Wöhler “An introduction to image-based 3D surface reconstruction and a survey of photometric stereo methods”. (September 2011) 23.
6. Joseph S Lappin, “What is Binocular Disparity. (August 2014) 1-4.
7. StereoKinetic phenomenon from Michael Bach’s "Optical Illusions & Visual Phenomena" (January 2013)
8. Parallax scrolling, Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Parallax scrolling. (Accessed 8 may 2019)
9. George Mather, The use of image blur as a depth cue: (February 1997)
10. Pilar Merchán, Antonio Adan, Santiago Salamanca, “Depth Gradient Image Based On Silhouette: A Solution for Reconstruction Of Scenes in 3D Environments”. (January 2006).
11. Qingqing Wei, “Converting 2D to 3D: A Survey”. (Dec. 2005) 11-12.
12. Qingqing Wei, “Converting 2D to 3D: A Survey”. (Dec. 2005) 12-13.
13. Cozman, F.; Krotkov, E. “Depth from scattering”, IEEE Computer society, conference on Computer Vision and Pattern Recognition, Proceedings. (1997) 801–806

14. Michael W. Tao, Pratul P. Srinivasan, Jitendra Malik, Szymon Rusinkiewicz and Ravi Ramamoorthi ,“Depth from Shading, Defocus, and Correspondence Using Light-Field Angular Coherence”, (june 2015) 1-2.
15. Nayar, S.K.; Nakagawa, Y. “Shape from Focus”, Pattern Analysis and Machine Intelligence, IEEE Transactions on Volume 16, Issue 8. (1994) 824 – 831.
16. Ahmad, M.B.; Tae-Sun Choi “Fast and accurate 3D shape from focus using dynamic programming optimization technique”, Proc. (ICASSP '05), IEEE International Conference on Acoustics, Speech, and Signal Processing Vol. 2. (2005) 969 – 972.
17. Li, M.; Magnor, M.; Seidel, H. P. “Hardware-Accelerated Visual Hull Reconstruction and Rendering”, Proceedings of Graphics Interface 2003, Halifax, Canada (2003).
18. Xiaojun Wu, High-quality Software Development Through Collaborations with Major Universities in China, https://www.nttreview.jp/archive/ntttechnical.php?contents=ntr201110fa6.pdf&mode=show_pdf. (Accessed 11 may 2019)
19. Qingqing Wei, “Converting 2D to 3D: A Survey”. (Dec. 2005) 4.
20. Battiato, S.; Curti, S.; La Cascia, M.; Tortora, M.; Scordato, E. “Depth map generation by image classification”, SPIE Proc. Vol 5302, EI2004 conference ‘Threedimensional image capture and applications VI”. (2004).
21. Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah: shape from shading: a survey. (August 1999) 697.
22. Huihui Xu Mingyan Jiang, Comprehensive depth estimation algorithm for efficient stereoscopic content creation in three-dimensional video systems. (July 2015) 8.
23. Maar, D.; Poggio, T. “Cooperative Computation of Stereo Disparity”, Science, Volume 194. (1976) 282-287.
24. Ullman, S. “The Interpretation of Visual Motion”, MIT Press, 1979.
25. Pentland, A. P. “Depth of Scene from Depth of Field”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 9, No.4, pp. 523-531, 1987.
26. Amartin, W. N.; Aggarwal, J. K. “Volumetric Descriptions of Objects from Multiple Views”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 5(2), pp. 150-158, 1983.
27. Haralick, R. M. “Using Perspective Transformation in Scene Analysis”, Computer Graphics and Image Processing (CGIP 13), Volume 13, Issue 3, pp. 191- 221, 1980.
28. Horn, B.K.P., “Obtaining Shape from Shading Information”, P.H. Winston (e.d.), the Psychology of Computer Vision, McGraw-Hill, New York, pp. 115-155, 1975.
29. Noah Snavely, Steven M. Seitz, Richard Szeliski, “Photo Tourism: Exploring Photo Collections in 3D” (2006).
30. David G. Lowe, “Object Recognition from Local Scale-Invariant Features”, Proc. of the International Conference on Computer Vision, Corfu (Sept. 1999).

31. Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz and Richard Szeliski, "Building Rome in a Day", Communications of the ACM, Vol. 54, No. 10, Pages 105-112, (October 2011).
32. Danilina, Nina & Slepnev, Mihail & Chebotarev, Spartak. "Smart city: automatic reconstruction of 3D building models to support urban development and planning". MATEC Web of Conferences, (January 2018).