2D TO 3D VIDEO CONVERSION

by

Aysun Çoban Aydın

B.S., Electronics Engineering, Ankara University, 2009

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering

Boğaziçi University

2016

2D TO 3D VIDEO CONVERSION

APPROVED BY:

Assoc. Prof. Burak Acar ..... (Thesis Supervisor)

Assoc. Prof. Ali Emre Pusane .....

Assoc. Prof. Engin Erzin .....

DATE OF APPROVAL: 21.06.2016

ACKNOWLEDGEMENTS

I am very grateful to my supervisor, Assoc. Prof. Burak Acar for his guidance, support and patience. He has been emboldening, caring and kind throughout the whole research period. It has been a pleasure and a chance for me to work with him.

I would like to thank Fahrettin Başarır from ARÇELİK A.Ş for his support and suggestions along this project.

I would like to thank Prof. Bülent Sankur for his valuable suggestions, comments and caring during the learning process of this thesis.

I also wish to thank Assoc. Prof. Ali Emre Pusane and Assoc. Prof. Engin Erzin for their valuable and enlightening comments which helped me to form the final version of my thesis.

I would like to thank all the people that I have known at BUSIM. In particular, my sincere thanks to Reza, Abdullah, Nihan and Erinç for their appreciation and for being there since the beginning of the way. I had very nice moments with them.

I would like to express my gratitude to my generous brother for his endless support and encouragement. Finishing this work would be so hard if he were not there for me.

I offer my greatest and sincere thanks to my better half for his support, patience and love. His understanding and encouragement gave me the enthusiasm to finish this work. Finally I would like to thank my beloved family, who supported me with their encouragement and endless patience through the entire process. I will be grateful forever for their love and support.

ABSTRACT

2D TO 3D VIDEO CONVERSION

Stereoscopic 3D visualisation is increasingly embedded into social life through the use of commercially available 3D-TV sets. In this work, a hybrid approach for 2D to 3D conversion is presented to produce stereoscopic 3D video automatically from 2D mono video frames. Each frame is synthesized to stereo pairs. Disparity/depth information required for 3D view is extracted from mono frame sequences based on motion and geometrical cues. Depth estimation of the scene is considered separately for background and foreground. Background geometry of the scene is determined by using geometrical cues such as vanishing point and straight lines in the image. According to this geometry, relevant information on the background depth field of a single image is estimated to generate a canonic disparity map of the background. For foreground depth estimation, on the other hand, two approaches are presented. First approach is based on detection of moving foreground objects. A depth value is assigned to each object based on its corresponding location in the background depth map. In the second approach, background registration is applied for consecutive frames that are captured by a moving camera. By this method, disparity in foreground regions is distinguished from background disparity that leads to a distinctive 3D effect on foreground regions. Consequently, depth/disparity information of foreground regions is combined with background canonic disparity map. According to these final disparity maps, pixels of the original frames are shifted to generate virtual frames to enable 3D views. This work is accompanied by a subjective evaluation on the basis of user test which compare our 3D results with commercially available 3D-TV sets.

ÖZET

2B-3B GÖRÜNTÜ DÖNÜŞÜMÜ

Bu çalışmada, tek bir kamera ile çekilmiş 2 boyutlu (2B) videodan 3 boyutlu (3B) stereo video elde etmek için yeni bir yöntem önerilmiştir. 3B stereo görüntü oluşturabilmek için sahnenin derinlik bilgisi gereklidir. Önerilen yöntemde, verilen 2B görüntü üzerindeki resimsel özelliklerden sahne derinliği kestirilmiştir. Bunun için, tek bir görüntü üzerindeki kaçış noktası ve birbirine yakınsayan düz çizgiler gibi geometrik özellikler çıkarılmış ve arka plan sahnesinin genel derinlik bilgisi çıkarılmıştır. Ön plandaki nesneler üzerinde derinlik etkisi oluşturmak için ise, sahne hareketine bağlı iki farklı yöntem önerilmiştir. İlk yöntemde, sahnedeki hareketli nesneler Gaussian karışım modelleri (GMM) ile çıkarılmış ve bu nesnelere arka plan derinlik haritasında bulundukları konuma bağlı bir derinlik ataması yapılmıştır. İkinci yöntemde ise, yatay kamera hareketi içeren görüntü dizilerinde, sahne üzerindeki paralaks etkisinden faydalanılarak ön plandaki nesnelere farklı bir ayrıklık kazandırılması amaçlanmıştır. Bunun için, farklı iki zamanda alınmış çerçeveler arasındaki arka plan geometrisinin hareketi kestirilmiş ve geri çatılmıştır. Böylece, ön plandaki nesneler üzerinde marjinal bir ayrıklık elde edilmiştir. Sonuç olarak elde edilen bu arka plan ve ön plan derinlik ve ayrıklık bilgileri birleştirilmiş ve tüm sahne için bir ayrıklık haritası oluşturulmuştur. Verilen 2B imgenin pikselleri, bu ayrıklık haritasına bağlı olarak kaydırılarak yeni bir imge oluşturulmuştur. Orjinal imge sağ göze, elde edilen yeni imge sol göze sunulmak üzere 3B stereo imge elde edilmiştir. Sonuçlar bir grup kullanıcı tarafından değerlendirilmiş ve performansları var olan ticari 2B-3B çeviriciler ile karşılaştırılmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

C_l	Left camera center
C_r	Right camera center
d	Disparity
f	Focal length
I	Image sequence
I_t	Image at time t
K	Number of distributions
t_c	Baseline distance
T	Threshold
W_i	Weight in HT space
x_l	Horizontal coordinate on left image
x_r	Horizontal coordinate on right image
X_t	Pixel value at time t
(x, y)	Image coordinates
Z	Real depth, distance from camera
α	Learning rate
Δ_l	Horizontal motion
η	Gaussian probability density function
θ	Angle parameter of HT space
μ	Mean
ρ	Distance parameter of HT space
σ	Standard deviation
Σ	Covariance matrix
ω	Weight of distribution

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional (Stereoscopic)
DFD	Depth from Defocus
DFF	Depth from Focus
DOF	Depth of Field
GMM	Gaussian Mixture Model
HT	Hough Transformation
LSE	Least Square Estimation
MTD	Modified Time Difference
SFM	Structure from Motion
VP	Vanishing Point

1. INTRODUCTION

3D video have started to gain increasing popularity through the use of commercially available 3D TV sets which generate the 3D effect by displaying stereo frames sequentially and in synchrony with a pair of active glasses. 3D TV offers depth impression of the observed scenes and has enhanced viewing experiences in comparison to 2D TV, but there is a pertinent problem to the 3D image world that is how to ensure availability of appropriate content. Although it is possible to directly produce 3D content by using stereo cameras, a more cost efficient and faster way of generating content is to design a system converting already recorded 2D multimedia to 3D ones. Despite the advances in stereoscopic display technologies, stereoscopic content generation is still insufficient. The 3D content is generated by a stereo camera set and a large team of technicians/cameramen, that is achievable at the expense of increasing the cost significantly. Furthermore, the results are not always satisfactory due to several artifacts in stereo video capture. Conversion of 2D videos to 3D videos is an efficient and fast solution to these problems and reduces the overall cost of the system. More importantly, this makes it possible to reuse existing 2D media resources as 3D.

As in the most technological advancements, 3D imaging rely on mimicking human abilities. Human stereoscopic viewing rests on two slightly different projections of the world on left and right eyes. This slight difference which originates from the spatial displacement between eyes is referred to as binocular disparity. Human brain can combine these two images with horizontal disparity and produce 3D depth perception [1]. Therefore, 2D to 3D image conversion techniques depend on horizontal disparity estimation to build the new image from mono image.

2D-3D video conversion refers to generating the stereo twin frame of a given video frame by applying horizontal pixel shifts according to the estimated depth of the scene, followed by hole-filling to correct for disparity discontinuities. Shift amount (disparity) of each pixel is directly related to its depth. Therefore, depth map estimation in this process is the most challenging part of stereoscopic video generation from 2D frames.

In order to extract depth information from mono frames, generally image and motion-based cues are used. Various methods have been applied for depth estimation. The main approaches are depth from focus/defocus, motion based depth estimation and methods that utilize geometrical cues in the image. Consequently, given one frame and its depth map/disparity map, stereo pair of this frame can be generated. Given frame can be taken as left/right image and right/left image can be obtained by shifting its pixels according to determined disparity map.

Since mono cameras cannot capture true depth information, depth maps obtained by these methods are just approximations to true 3D geometry. It has been found that human visual perception tolerates the particular inaccuracies of estimated depth maps [2], so these depth approximations are expected to be sufficient for 3D visualization.

In this thesis, a hybrid approach is proposed for depth estimation from mono frames/images, which is based on geometrical cues and motion analysis. We deal with background and foreground depth determination issues separately and obtain final disparity map by combining them. Relative depth map of background is generated based on an improved vanishing point and major line detection in the Hough Transform space. In the first approach, motion estimation is used to detect foreground objects to which a depth is assigned using the background depth field. In the second approach, background-only registration is used to determine disparity for foreground objects without explicit depth estimation process. Final stereo frames are generated by shifting the pixels of mono frames according to disparity maps.

The proposed approach is also evaluated with real images from versatile scenes. Vanishing point detection method is experimented on a number of images for which the results are compared to natural human perception recorded by manual procedure. For every proposed algorithm, stereo image and video sets are presented to a group of subjects on 3D screens. The evaluation results yielded quite success in terms of creating a depth perception for 3D visualisation.

In conclusion, we propose methods to produce stereoscopic videos from 2D videos automatically. The aim is not to find true depth values in real world, it is to enhance end-user experience on 3D visualization by creating illusion of depth via stereopsis. The main approach pursued in this thesis is to understand the background geometry of scene and, determine the foreground objects, and assign user adjustable disparities to the background and foreground objects based on the obtained depth order information. The method is implemented on standard hardware and evaluated empirically. A group of users watched the generated 3D stereo videos and images in 3D TV sets and grade their quality based on depth impression.

Chapter 2 provides a background information on stereoscopic vision and introduces the related work in 2D to 3D conversion. In Chapter 3, proposed method is given in detail. Experiments are presented in Chapter 4 and conclusion is given in Chapter 5.

2. OVERVIEW OF 2D TO 3D VIDEO CONVERSION

Since 2D to 3D conversion techniques are based on stereoscopic viewing, various methods depend on horizontal shifting of pixels according to their depth values. In this chapter, firstly, some general information is given about stereo vision and the relationship between disparity and depth. Secondly, related work on depth estimation for 2D to 3D video conversion is presented.

2.1. Fundamentals of Stereo Vision

Binocular viewing of a scene means two slightly different images due to the locations of the eyes on the head. This slight difference between left eye image and right eye image is referred as *disparity* and it provides information about image depth. Given these two images with horizontal disparity, our brain can generate depth perception by combining them together. Stereopsis is the design and implementation of the algorithms that mimic our ability to fuse the pictures recorded by our eyes and exploit the difference between them. This difference is referred as *disparity* and used to strengthen the depth sense. 3D imaging generally relies on this concept according to which stereo camera systems are used to create stereo images. In case of 3D stereo generation from 2D, the scene is captured by only one camera and a virtual new image is created by horizontal shifting of the original image pixels. To discuss the relationship of depth and disparity, stereo geometry of a stereo camera system, which consists of two identical parallel cameras, is illustrated in Figure 2.1.

In Figure 2.1, C_l and C_r are origin of the camera coordinate systems of left and right cameras with focal length f. Cameras are placed in parallel with a baseline distance, t_c. An object point in 3D world, P, is captured by these cameras and the projections of this 3D point on the left image and right image are denoted by p_l and p_r. The horizontal image coordinates of these projection points correspond to x_l and x_r. The depth value of P is indicated by Z. The disparity, which is the horizontal

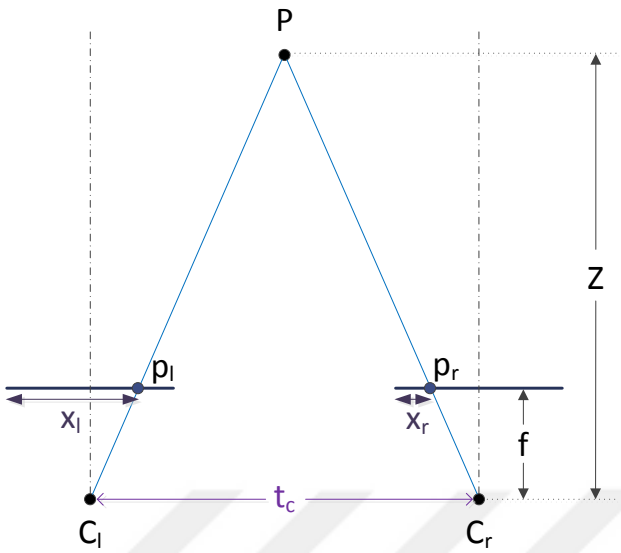


Figure 2.1. Stereo geometry for two identical parallel cameras. The difference between image coordinates of left and right projections (p_l and p_r) is referred to as disparity (d).

difference between left and right image coordinates is defined as follows:

$$d = |x_l - x_r| \tag{2.1}$$

The disparity, d, can be stated in terms of the depth value, Z, depending on the relationship between similar triangles (P, p_l, p_r) and (P, C_l, C_r) in Figure 2.1 as follows:

$$|x_l - x_r| = \frac{ft_c}{Z}$$

$$d = \frac{ft_c}{Z} \tag{2.2}$$

From the geometrical relationship in Figure 2.1, disparity, d, is found inversely proportional to depth, Z. As seen in Equation 2.2, disparity can be found if real depth (Z) is known where the distance between cameras and the focal length are known. For automatic 2D to 3D conversion, it is aimed to find relative disparity values for artificial stereo image generation without the information of camera parameters such

as t_c and f. However, since these camera parameters are constant values, the inverse proportional relationship between depth and disparity can be used to built a coarse disparity map from depth map of an image. Therefore, if the relevant information on the depth of a single image is present, its pixel-wise relative disparity values can be estimated based on Equation 2.2. As a result it can be inferred that, the main issue in stereo video generation from mono video is to estimate relative depth values of pixels in mono frames.

2.2. Previous Works on 2D to 3D Video Conversion

Main and the most difficult part of 2D to 3D image or video conversion is depth map estimation. There are many methods which use different pictorial cues of the images captured by single camera for depth estimation. In this section, three main approaches are explained in terms of their primary source of information as focus/defocus, motion, and geometrical information of the image/frame used for relative depth map estimation.

2.2.1. Depth From Motion

Motion estimation is the most common method that is used to determine depth or disparity for 2D to 3D conversion. Displacement between two images which are taken by a moving camera is more on the objects closer to the camera and less on the objects further to the camera. Owing to that parallax effect, motion vector of each pixel gives information about its depth. Therefore, motion estimation can be used to find out the shifts between two consecutive frames by determining correspondence between them. In this respect, 2D motion estimation is used for depth estimation [1]. A study reported in [3] is an example of utilization of this principle, where parallax information obtained from horizontal motion of sequential input images are used to create depth. They applied modified time difference (MTD) method which selects images that would be a stereo pair based on the estimated horizontal motion in the sequential input images. Similar ideas can be found in [4, 5]. According to the results, MTD works good for simple scenes with simple motion, but it is not successful in complicated cases.

Jung et al. proposed a 2D-3D conversion method by using depth estimates based on motion parallax in [6]. In the case of global motion, they create depth maps directly based on the motion vectors. In the case of local motions, they use depth templates by applying motion-guided depth refinement to those templates. In [7] Liu et al. used structure from motion (SFM) method for 2D to 3D conversion. Their method includes steps such as projective transformation, self-calibration and depth refinement by color segmentation. They have good depth map results but their system is too time-consuming to be able to work in real-time.

In [8], a H.264-based depth map estimation technique is used. Their method utilizes the horizontal motion information between consecutive frames to approximate the depth map of the scene. They used a H.264-based scheme that includes motion vectors which are provided accurately in the standard. In [9], optical flow is used to determine depth ordinal by using compressed videos. Motion vectors extracted from the compressed video files (MPEG4) to determine the depth map. Magnitude of motion vector is calculated by the Euclidean distance of vertical and horizontal motion estimation components for each pixel. Disparities to shift pixels to create virtual stereo image are determined by the motion vector magnitudes. In [10], a depth extraction method is proposed based on motion and geometric information for 2D to 3D conversion. They used H.264 motion estimation result and moving object detection to generate a motion-based depth map. Finally they fused it with a geometry-based depth map which is obtained by edge detection. In [2], an object based algorithm is proposed for stereoscopic 3D video generation from 2D content. They used motion information available in the scene to determine occlusion in video sequences. Therefore, they aim to segment object regions using motion inconsistency along the borders of moving objects. They used optical flow based occlusion reasoning to determine the depth ordinal and produced stereo videos. Depth from motion is the most common depth estimation method from monocular image sequences ([11]). However, these techniques have problems, such as nonlinear motion issue and the problem of separation of camera motion from object motion. In addition, the motion between two consecutive frames is very small and sensitive, and it is improbable to generate consistent and adequate results [12].

2.2.2. Depth From Focus

In the approach of obtaining depth from focus, the idea is to estimate the depth information based on the amount of blur of each region in the focused image [13], [14]. Constitutively, there are two types of this approach which are called *depth from focus* and *depth from defocus* (DFD) (also referred to as *depth from blur*). In the DFD approach, depth map is generated from the degree of blurring in the images, and it needs two or more images with different focal settings [15]. The DFF methods determine the distance of an object by comparing the sharpness of the object over a series of images of the scene with different focus distances and also with varying camera positions [16].

The distance to scene points which are in depth of field can be determined by the focal parameters of the camera. Depth of field (DOF) is the range distances through which objects are well focused. If depth of field of the camera is very limited, then only points with same distance from camera will be in perfect focus and will appear sharp. Other points in the scene will be out of focus and appear blurred. Thus, with the knowledge of the camera focal parameters such as the position of the image plane, the aperture size and the focal length and besides the degree that the object is out of focus, the depth of and object can be fairly inferred. On this basis, depth from focus method proposes that depth information of the points in a scene that is acquired with a small DOF can be calculated by modelling the effect of the camera's focal parameters on the image [17].

In spite of the resolution and sensitivity limitations of depth from focus methods in comparison to triangulation based depth from motion techniques; depth from focus avoids occlusion and correspondence ambiguity problems [18]. Generally, depth from focus methods depends on varying the camera focus on an object to determine its depth, that is not applicable for many cases. Depth from focus is mainly motivated by specific applications such as robotic sensing and industrial examination which have a specific imaging system for specific type of scenes. It also can be used in conjunction with other depth perception techniques as a part of a general purpose system [19].

Since depth from focus and depth from defocus methods require more than one images which are taken with different intrinsic and extrinsic parameters of camera setup for a scene, it is not useful for automatic conversion of existing 2D data to 3D. Automatic 2D to 3D conversion requires depth map determination without the knowledge about camera setup and scene

There are other depth from focus methods that can estimate blur without depending on camera parameters [20]. Considering focus/defocus effects, if a region is more blurred compared to the other one, then this region is concluded to be farther away. Valencia et al. proposed a 2D to 3D stereo conversion system based on blur in image [21]. Their depth map estimation method depends on measuring focus cues which includes a local spatial frequency measurement and a regularity estimation of significant edges. In [20] Tam et al. proposed a depth map generation method from 2D images based on blur and edge information using intensity gradients in local regions. However, reliability of this approach is questionable since the blur can also arise from other factors, such as convergence plane (lens aberration), atmospheric interference, and motion. In addition, same amount of blur can occur in both cases of the object is farther away from the focused position or closer than the focused position of the camera.

2.2.3. Depth From Geometric Perspective

Perspective distortion can be used for estimating depth information from single monocular images. In the environments with geometric elements, there are parallel lines that converge to a point called vanishing point (VP) in the image plane. Vanishing point is the eventually reached point by lines that appear to converge with distance and can be claimed to be the deepest region of the scene. Cantoni et al., proposed methods to detect vanishing point in [22]. The main idea is to detect straight lines in the image and estimate their intersection point that is considered to be the vanishing point [23]. After finding vanishing point which is an important geometrical cue of the scene, the depth layers can be determined as relative distances from the detected VP based on the geometrical structure of the scene. Generally, manmade world (e.g. buildings)

possesses strong geometric cues where this method has reliable results. However, one major drawback of this method is that there may be no vanishing points in a given image such as landscape images. Hence the success of this method heavily depends on the scene geometry. Battiato et al. used vanishing point and vanishing lines to establish depth gradient lines in [24].

In [25], color-based segmentation is applied for image classification such as outdoor, indoor, vs. For each specific class, geometrical structure is determined to extract relevant depth information. Depth map generation is based on several steps which are gradient planes generation, depth gradient assignment, consistency verification on detected region. Finally, color-based segmentation results are combined with geometrical results and depth-map is reconstructed. In [26], they utilized motion parallax and geometrical perspective to generate depth map for 2D to 3D conversion. In [27] 2D to 3D conversion method is proposed to reduce depth map flickering in videos. Their method segments objects based on color and motion history to ease depth error.

2.2.4. Other Approaches

The authors in [28] and [29] suggest machine learning approach for visual scene understanding and building rough depth map from monocular image features. They apply supervised learning to predict depth value as a function of the image using multiscale local and global image features such as texture variants, texture gradients and color. They train a set of images and their corresponding ground truth depth maps to determine the plane parameters. In [30], they improved their learning model using image segmentation and object detection methods. They also proposed a method to create 3D models using a set of images. However, the algorithm based on that approach was found less efficient for operating in real time.

Zhang et al. proposed a depth calculation method by integrating occlusion and visual attention analysis in [31]. They initialized depth by applying scene classification and they computed saliency map to weight it. By occlusion analysis they found a coarse depth map of the scene. Zhang et al. also proposed a 2D to 3D conversion

method based on saliency map that is generated based on color histogram in [32]. In their proposed method, more salient areas are assigned with closer depth values. Their depth values are based on color histogram, so depth orders are mostly not correct, although they obtained sharper and stable depth maps.

Besides fully-automatic 2D to 3D conversion techniques, manual and semi-automatic approaches are also proposed in literature. A semi-automatic approach based on depth propagation is given in [33]. The main principal of depth propagation is to create high quality depth maps at key frames manually or semi-manually and propagate these depth maps to other non-key frames. Jung et al. [33] applied a superpixel matching method for motion estimation between key image and current image. Then, they generated depth map of the current frame by conducting depth compensation based on the motion vectors. Phan et al., also proposed a semi-automatic method by using key frames for user application and propagating the strokes automatically. For the segmentation algorithms, they used Graph Cuts and Random Walks [34]. Another interactive method is proposed by Criminisi et al. in [35] to compute 3D geometry of a single perspective view of a structured scene by assuming that vanishing points and lines are known. They take some specific information from user such as reference height of an object, object segmentation, and 3D coordinates of some points. Depending on that given real measurement values and perspective geometry of the scene they reconstruct the 3D geometry of the objects in the scene. In [36] and [37], depth map generation methods which are depend on user inputs are proposed. [37], depth is assigned and propagated among over-segmented image regions.

In conclusion, the problem of 3D scene reconstruction from the images captured by a single camera has not a complete solution. All depth extraction methods have their own advantages and disadvantages. Depth from motion approach is the most preferable in terms of the applicability to more general cases [12].

3. METHOD

For the 3D stereoscopic video generation from 2D videos, we propose a hybrid model which utilizes both geometrical perspective and motion cues in single frames. Firstly, a relative background depth map is estimated on the basis of the geometrical structure of the scene. Then, foreground objects' depth estimation was done by two different methods. In the first method, moving foreground objects are detected based on motion, and depth values are assigned according to their position in the scene. For that method, foreground objects were assumed to be moving while camera was stationary. The second method is developed for scenes which are taken by a moving camera. Benefiting from the parallax effect between sequential frames, disparity for foreground object is obtained by applying background registration. After the depth and disparity estimation, the stereo pair is rendered by shifting the pixels of the original image based on its final disparity map. In Figure 3.1, the proposed hybrid method is demonstrated as a flowchart of the process for conversion of a single image to a stereo pair.

3.1. Background Depth Map Estimation

Initially, a background depth map is created using only strong geometric cues of the scene. Depth can be estimated in many real images based on perspective distortion. A group of parallel lines in the real world converges to a single point referred to as *vanishing point* (VP) due to perspective distortion when they are projected within 2D image. Vanishing point and associated lines of the scene, gives information about depth in the image to determine 3D structure of the scene. For instance, if we consider creating the depth map of a corridor, the two strongest lines leading to VP can be selected as borders between the walls, floor and ceiling as shown in Figure 3.2. After the main geometry determination, scene depth layers are created based on the main lines by taking VP as reference point for the maximum depth region. As a result, background depth map is formed as seen in Figure 3.3. Detection of the vanishing point and the main lines, and depth map formation is explained in the following subsections.

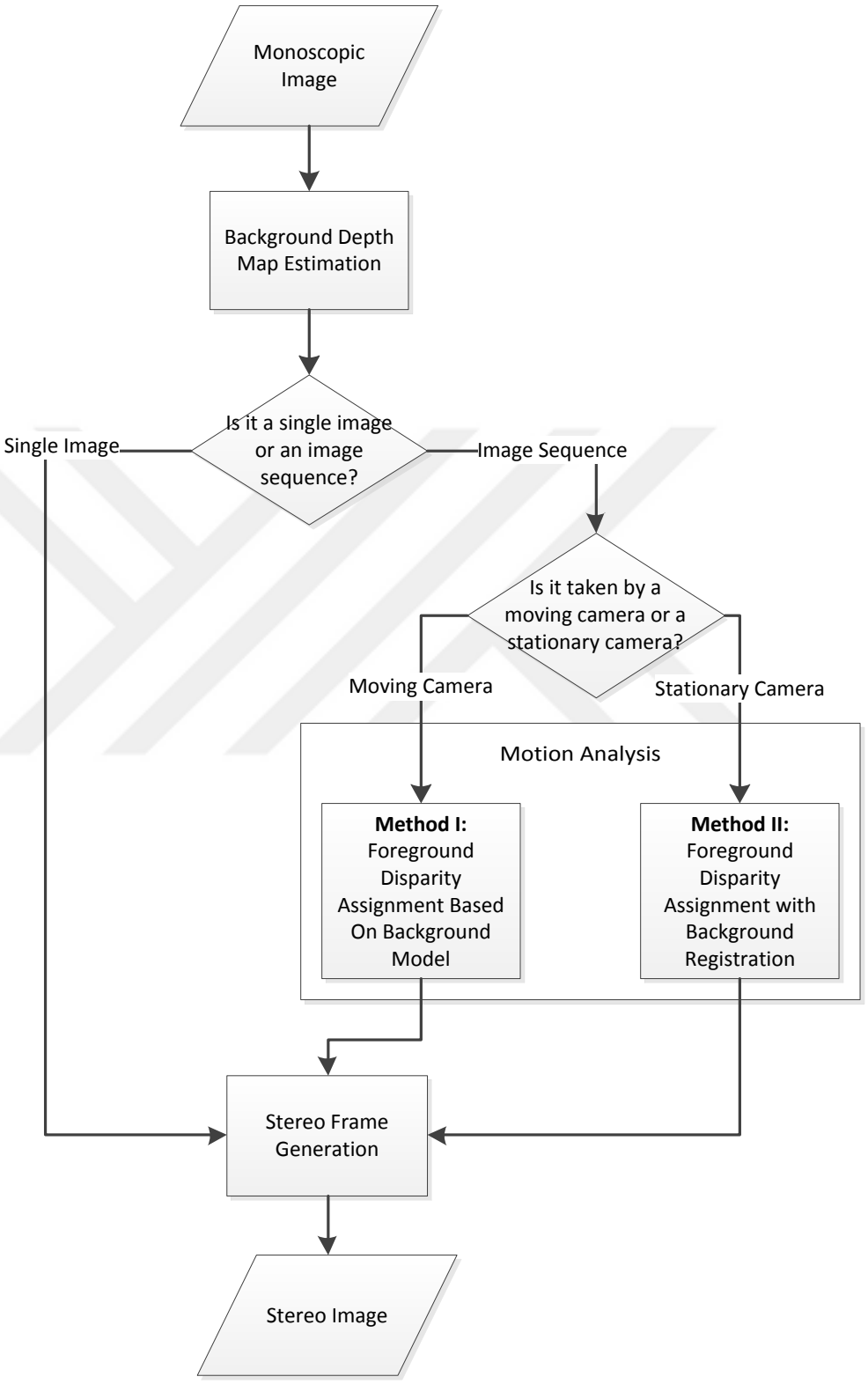


Figure 3.1. General flowchart of the proposed hybrid method of stereo image generation from a single image.

Figure 3.2. Determined geometrical cues of a background structure: vanishing point is marked by a red point and main lines are showed as blue lines

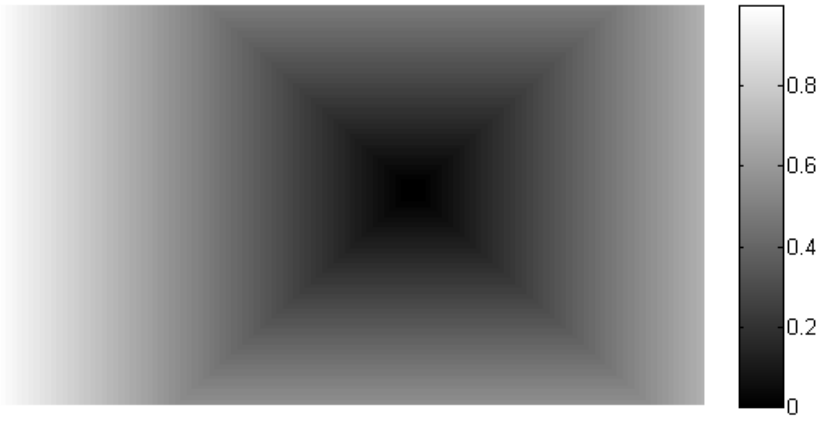


Figure 3.3. Background depth map of Figure 3.2 that is produced by proposed method

3.1.1. Vanishing Point Detection

In 2D images, parallel lines in space look like converging to an infinity point because of perspective distortion. This point, called as *vanishing point* (VP), can be claimed to be the deepest region in the scene. In order to determine VP, at least two converging straight lines is needed. For finer estimation, we need to determine the most characteristic straight lines of scene geometry and optimize VP location that best fits to intersection point of all such lines. We assume that the lines which are prominent in terms of length and density are strong lines and give cues about scene. Hough Transformation is used for the detection of the strongest straight lines of the scene geometry and estimation of VP as intersection point of these lines. Flow diagram of VP detection method proposed in this work is shown in Figure 3.4.

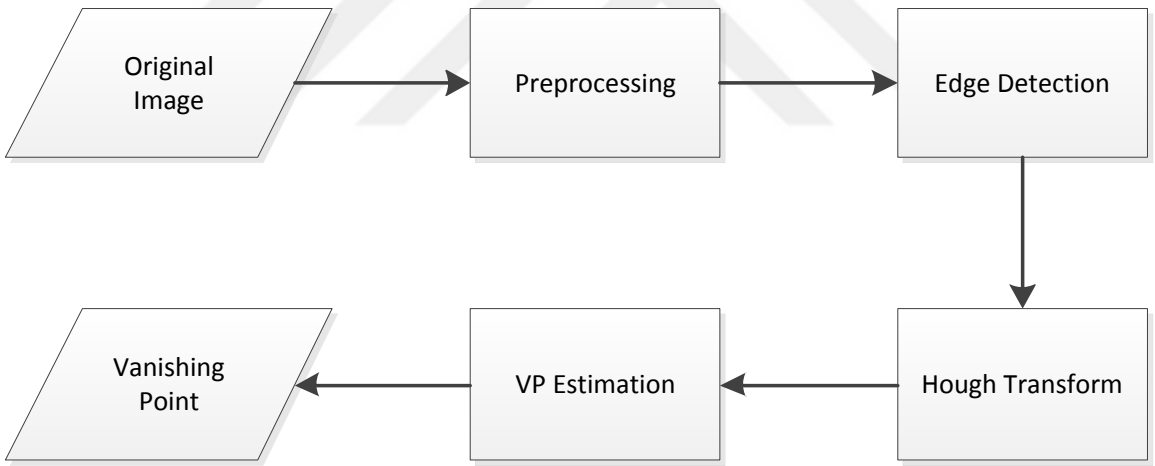


Figure 3.4. Flow diagram of vanishing point detection method

Firstly, color image is converted to gray image using the MATLAB function *rgb2gray* [38]. Then, median filtering is applied which provides very good noise reduction capabilities, with less blurring than linear smoothing filters. Median filter replaces the value of a pixel with the median of the intensity levels in the neighborhood of that

pixel. It can be expressed as follows:

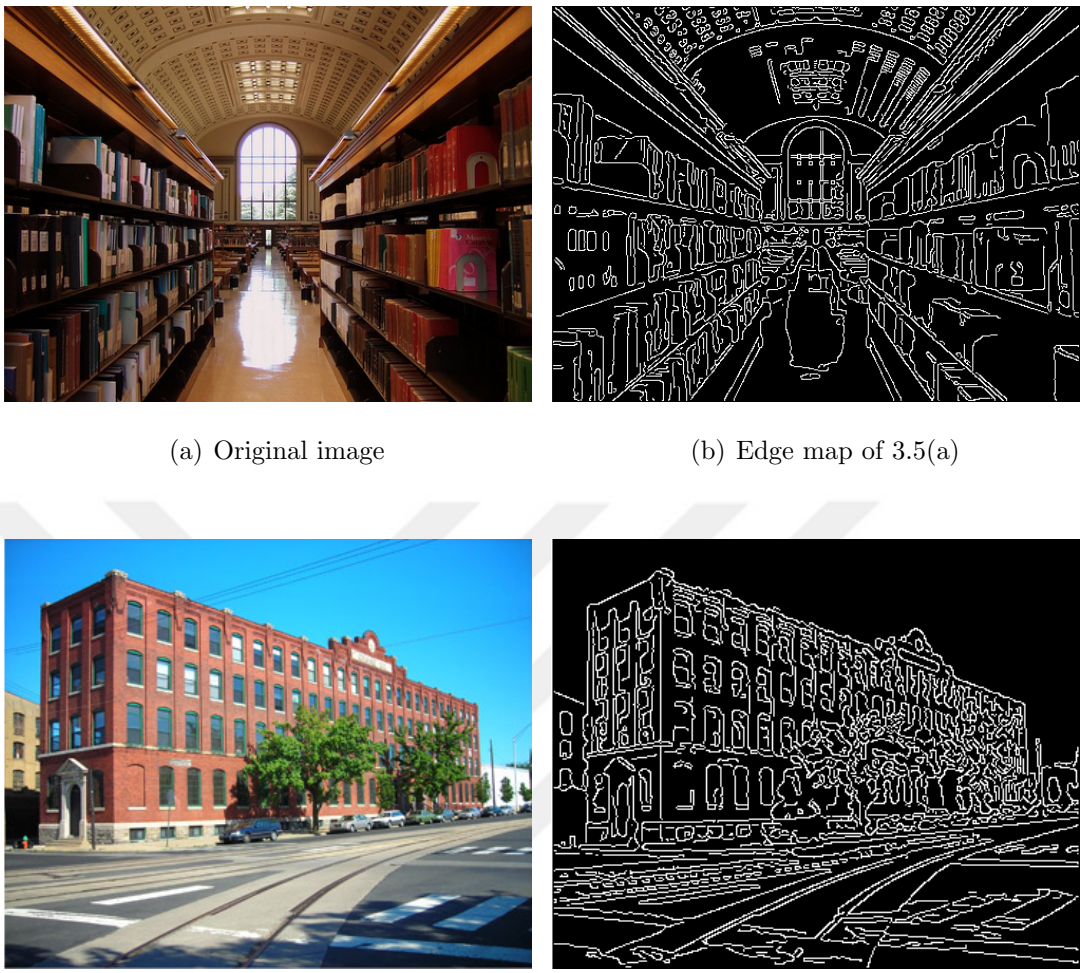
$$f(x,y) = \underset{(i,j)\in S_{ij}}{median}\{g(i,j)\} \tag{3.1}$$

where $g(i,j)$ represents a rectangular subimage window of size $m \times n$, centered at point (i,j), and S_{ij} represents the set of pixel coordinates in the neighborhood $g(i,j)$. Using median filtering, the salt and pepper noise which can give rise to the detection of wrong lines in the image is avoided.

Edge detection is performed by Canny edge detector [39] which is one of the most powerful edge detection methods in the literature. Canny edge detection algorithm steps are explained briefly as follows.

- Image is smoothed to reduce noise. A Gaussian filter with a standard deviation, σ, is applied.
- Gradient magnitude and angle (edge direction) is computed for each point.
- By preserving only local maxima points, non-maximum suppression is applied to the gradient image. Thus, broad edges of gradient image is thinned to sharp edges. The points with locally maximum gradient along the edge direction are defined as edge points.
- Double thresholding and connectivity analysis is applied for edge detection and edge linking. The pixels remained after non-maximum suppression are thresholded with two threshold values. The pixels that have higher strength than higher threshold is labeled as *strong* edges. The ones with lower strength than lower threshold are deleted, and the pixels which are between two thresholds are labeled as *weak* edge points. The weak pixels which are 8-connected to the strong pixels and all strong pixels are defined as edge pixels.

Canny edge detection method is executed using MATLAB's *edge* function [38]. Figure 3.5 shows resulted edge maps of two real images.



(a) Original image

(b) Edge map of 3.5(a)

(c) Original image

(d) Edge map of 3.5(c)

Figure 3.5. Two original images and their edge maps

After the edge detection, Hough Transform (HT) is used to determine strong lines in the scene which will converge at the deepest region according to the scene perspective.

In Hough Transform [40], lines and points in image plane are respectively mapped into points and curves in Hough space that is also called polar space. In Figure 3.6 the relationship between image plane and Hough plane is demonstrated.

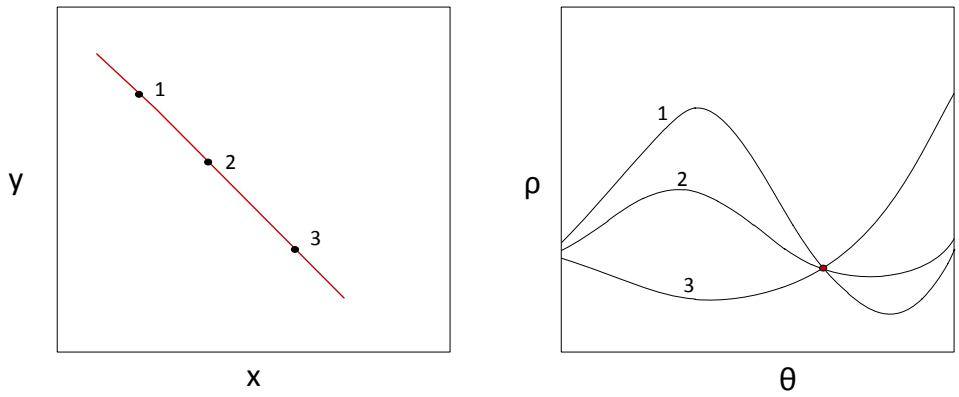


Figure 3.6. A line passing through three points on image plane is mapped a point which is the intersection of three sin curves in the Hough plane. Every point in the image plane is also mapped to a sin curve by HT.

Hough Transform maps a straight line in image plane to a point. A straight line, $y = mx + n$, in the image plane is represented in the polar space as follows:

$$\rho = x \cos\theta + y \sin\theta \tag{3.2}$$

(m, n) parameters of the straight line in image plane is transformed to (ρ, θ) parameters in the polar plane. Geometrical meanings of polar parameters are shown in Figure 3.7 where ρ represents the perpendicular distance between the line and the origin, and θ represents the angle of the vector from the origin to the closest point of the line. θ takes values in the range of $[0, \pi)$, while $\rho \in \mathbb{R}$ is limited to the size of the image.

Hough Transform also maps a point in image plane to a curve in the Hough space. If we take a point (x_0, y_0) on the image plane with some straight lines intersecting on it, its corresponding sin curve on the polar plane is presented by the following equation:

$$\rho(\theta) = x_0 \cos\theta + y_0 \sin\theta \tag{3.3}$$

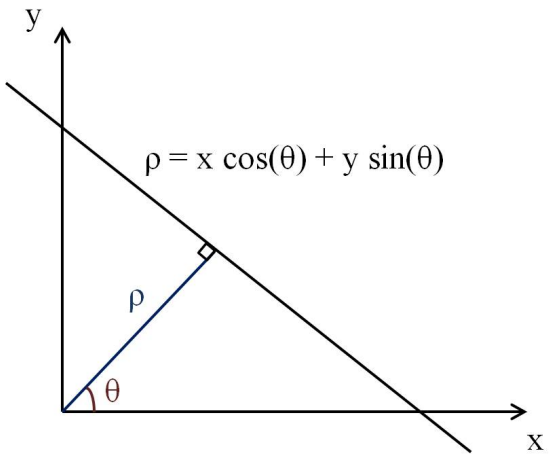


Figure 3.7. Description of ρ and θ parameters of a straight line on the image space. Line is presented by $\rho = x\cos(\theta) + y\sin(\theta)$ so it is defined by the polar parameters ρ and θ which indicate a point in the Hough space.

Each (ρ, θ) point which satisfies Equation 3.3 in the polar plane represents a straight line passing over the (x_0, y_0) in the image plane.

Vanishing point is assumed to be the most significant curve in the Hough space. Since the VP is assumed to be the intersection of the most significant lines in the image plane, it corresponds to the strongest sin curve in the polar plane.

Matessi et al. proposed an algorithm in [23] to detect the vanishing point directly on the polar plane. They use least square estimation (LSE) method, to estimate corresponding sin curve which identifies the vanishing point. LSE method is applied directly in the Hough space as given in Equation 3.4 where (x_0, y_0) stands for vanishing point coordinates in the image.

$$\min_{x_0,y_0} \sum_{i=1}^{n} W_i(\rho_i - x_0\cos\theta_i - y_0\sin\theta_i) \tag{3.4}$$

In Equation 3.4, (ρ_i, θ_i) points are weighted by W_i which comes from the number of times that the pair (ρ_i, θ_i) is observed when mapping point of a line on the image,

and n is the total number of samples. For finer estimation, they apply this statistical method iteratively. First, they measure the distances of all points in the Hough transform matrix to the first estimated curve to find a distance variance. By using this distance variance, they eliminate points farther away than this variance and make second estimation with the points left on the image. Iteratively, until the difference between last two found vanishing points is less than a threshold, they continue this point elimination and they reach the coordinates of the VP at the end of iteration.

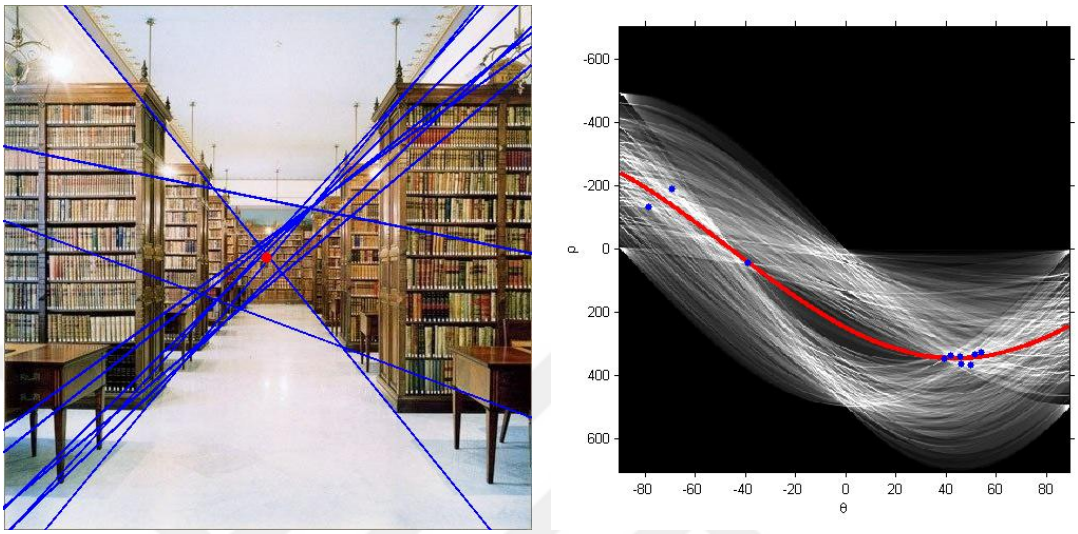
This approach has two problems: *(i)* Even if it is aimed to eliminate the effect of very small noisy lines by using weights, they affect the determination of the scene geometry negatively, and create noise. *(ii)* Vertical and horizontal lines do not converge to the vanishing point, and also distort the output because they are strong in most of scenes.

The proposed modification to Matessi's algorithm to overcome the above mentioned weaknesses involves: *(i)* Noisy lines are eliminated by detecting some number of peak points on the Hough transform plane. The results are much more improved when fitting the curve to these points. Using the strongest lines in the image gives clearer solutions since very small lines do not contribute to the geometry of the scene, but create noise. *(ii)* Vertical and horizontal are removed from the Hough transform plane.

According to these specifications, the selection of some straight lines and the calculation of VP as the estimation of the intersection point of these lines is intended. Therefore, firstly vertical and horizontal lines are removed by deleting the points around the θ values of 0° and 90° in the HT plane. After that, a number of peak values is picked which is applied as ten peak values of the remaining HT matrix. To find the coordinates of VP, (x_0, y_0), its curve, $\rho(\theta) = x_0 \cos\theta + y_0 \sin\theta$, is fitted to these peak points by using least square minimization method which is given in Equation 3.4.

In Figure 3.8(a), ten strongest lines -except vertical and horizontal ones- and estimated vanishing point is illustrated on an image. Hough Transform matrix of

this image is also given in Figure 3.8(b) with corresponding points of these lines and corresponding sin curve of VP.



(a) (b)

Figure 3.8. (a) Image with ten strongest straight lines and vanishing point estimated as intersection point of these lines. VP is marked by a red dot and lines are showed with blue lines. (b) Hough Transform of (a). Corresponding sine curve of VP is marked by a red curve and ten peak points correspond to ten strongest lines is marked by blue points.

The vanishing point is obtained by minimizing Equation 3.4. Let $\cos\theta_i = a_i$ and $\sin\theta_i = b_i$. Then we differentiate with respect to x_0 and y_0, and equate to zero. Following linear system is obtained:

$$\begin{aligned} \sum W_i a_i(\rho_i - a_i x_0 - b_i y_0) = 0 \\ \sum W_i b_i(\rho_i - a_i x_0 - b_i y_0) = 0 \end{aligned} \tag{3.5}$$

If we sum the left hand side of these two equations and make the substitutions of $A = \sum W_i {a_i}^2$, $B = \sum W_i {b_i}^2$, $C = \sum W_i a_i b_i$, $D = \sum W_i a_i \rho_i$, $E = \sum W_i b_i \rho_i$, we obtain

the following linear system:

$$\begin{aligned} Ax_0 + Cy_0 &= D \\ Cx_0 + By_0 &= E \end{aligned} \tag{3.6}$$

Thus, the coordinates of VP, (x_0, y_0), is estimated by solving this simple linear system:

$$\begin{bmatrix} A & C \\ C & B \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} D \\ E \end{bmatrix}$$

$$\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} D \\ E \end{bmatrix} \begin{bmatrix} A & C \\ C & B \end{bmatrix}^{-1} \tag{3.7}$$

3.1.2. Main Lines Detection

By taking vanishing point as the reference point for the deepest region in the scene, we determine relevant information on depth according to main lines of the scene. We select one negative slope line and one positive slope line to form depth layers of the background in the shape of nested tetragons as seen in Figure 3.3. In this obtained depth map darker regions refers to deepest regions while lighter regions represents closest regions of the scene.

Two main lines which presents general scene structure are selected as a left and a right straight lines which are closest to VP among strongest straight lines in the image. We separated the straight lines into two groups in terms of right lines and left lines. Lines with $0 < \theta < 90$ defined as right lines and lines with $-90 < \theta < 0$ defined as left lines. In order to determine scene structure we need one left line and one right line as main lines.

Initially, if ten peak points in Hough space includes both left and right sided lines, one left line point and one right line point which are closest to vanishing point curve

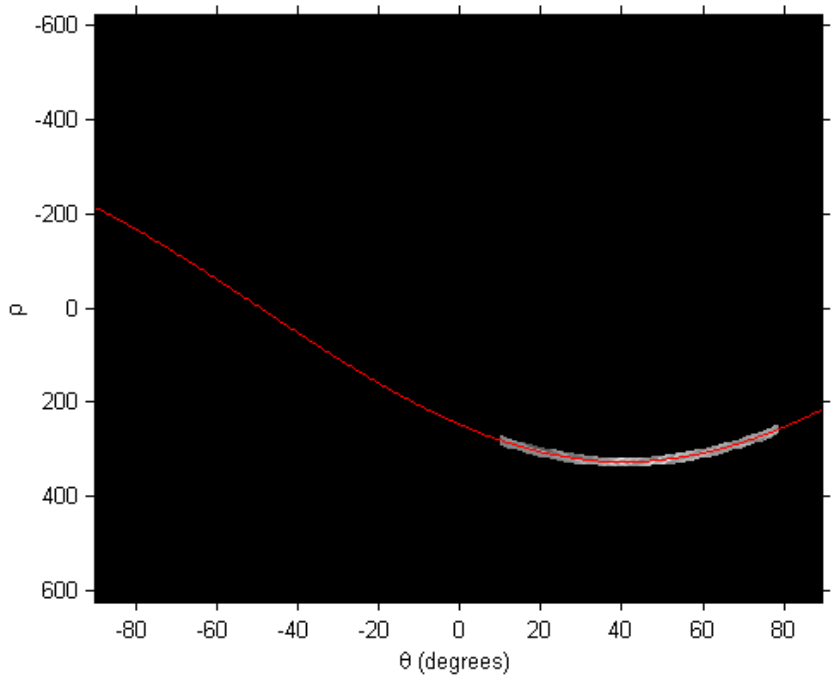


Figure 3.9. Hough transform band to search the missing line in it. VP curve is marked by red line. Missing right main line is computed as the strongest point within the determined band on Hough plane that is demonstrated in the figure.

among them is selected to be main lines based on ρ. One left and one right point which have minimum distance to the VP curve are selected from left side point group and right side point group in Hough space separately. The distance of each peak point to the VP curve, is calculated in ρ axis in HT space and the closest points are determined as follows:

$$\begin{aligned} k^* &= \arg\min_k |\rho_k - \rho_{vp}(\theta_k)| \\ &= \arg\min_k |\rho_k - (x_0 \cos\theta_k + y_0 \sin\theta_k)| \end{aligned} \tag{3.8}$$

where (ρ_k, θ_k) are polar parameters of k^{th} peak point and (x_0, y_0) is the image plane coordinates of VP. ρ_{vp} is the sin curve function of vanishing point as given in Equation 3.3. As a result $(\rho_{k^*}, \theta_{k^*})$ parameters gives the selected main lines.

On the other hand, if these ten peaks do not include both left and right sided lines, in other words, if all ten strongest lines are lied on one side, right or left, one of the main lines will be missing. In order to find a line instead of absent line, we create a band around VP curve in HT matrix and search for the appropriate point within it as shown in Figure 3.9. The width of this band is determined as the median of the distances of ten peaks to the VP curve. Then we take the most voted (strongest) point within that band instead of the missing line. In this way, this band constraint search area on the HT matrix prevents us from choosing a line far away from the vanishing point even if it has a high vote.

3.1.3. Construction of Background Depth Map

By using the two lines and vanishing point, we create a depth map for the background of the scene. The intersection region of left and right lines, the region of VP, is taken as reference for furthest region in image. Depth layers are built based on main lines geometry as shown in Figure 3.3. By taking VP region as beginning point and assigning its depth to 0, we assign a depth value in the range of $[0, 1]$ to each region with an increasing manner along main lines. In our experiments, we divide the image into 100 depth regions. We assign 0 to the deepest region and we add 0.01 for every neighbor depth region. Finally we get nearest region as 1 valued. As a result we generate a depth map as a gray image and the depth information is coded as luminance intensity level, lighter values for closer distances and darker values for further distances.

However, vanishing point is not always be inside the image. We need to determine the location of VP since depth map have to designed according to the VP location. There are nine possible areas where VP could be located as shown in Figure 3.10. First, we determine the area of VP for our depth map generation algorithm; and then we form the depth map according to this area. In the case of vanishing point being outside, the closest part of the image is considered as the deepest region. In Figure 3.12 and 3.11 examples of the cases that VP point is outside can be seen.

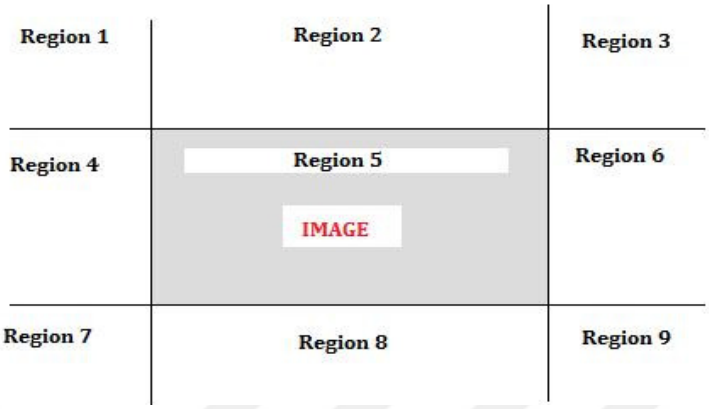


Figure 3.10. Possibilities for vanishing point location. Region 5 shows the border of image plane and other regions show the outside of the image that VP can be located.

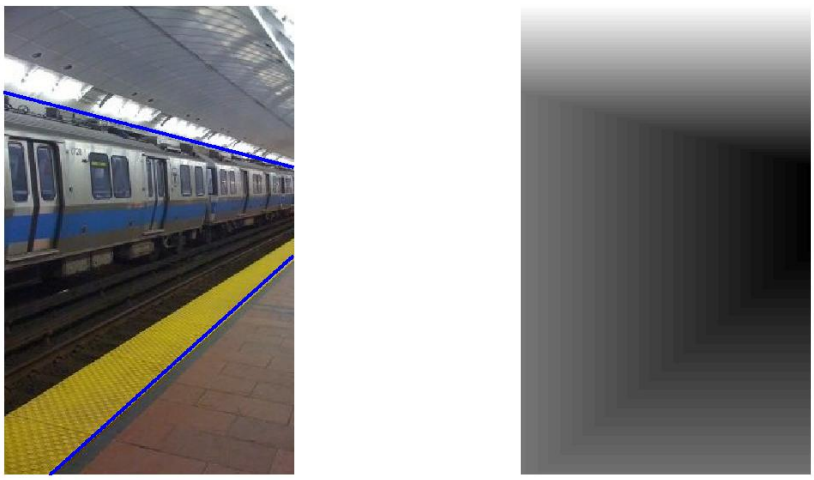


Figure 3.11. Main lines and depth map: VP is in Region 6

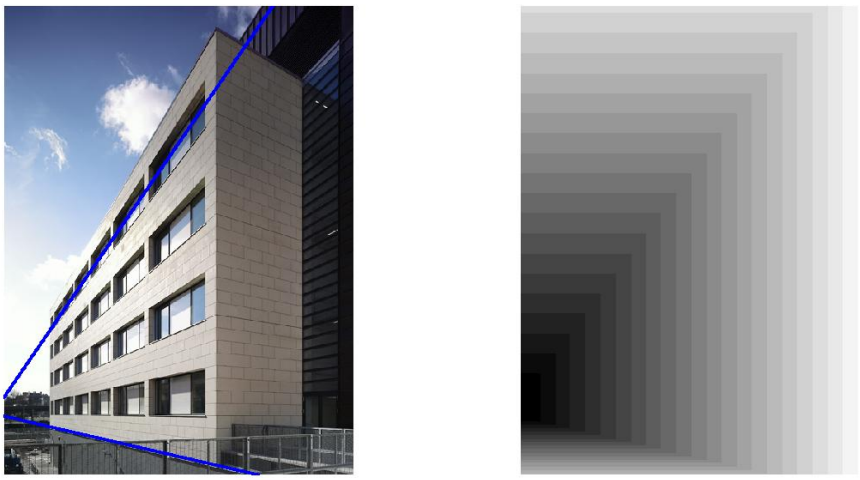


Figure 3.12. Main lines and depth map: VP is in Region 4

After background depth map estimation, we propose two methods for foreground depth/disparity estimation: (i) moving objects segmentation and depth estimation, (ii) disparity estimation of foreground objects by exploiting parallax in the moving camera videos.

3.2. Foreground Disparity Assignment Based On Background Model

The depth of foreground objects differs from the background and is critical for a realistic 3D experience. In the proposed method that is based on motion cues, the segmentation of moving regions is applied to provide object-wise depth ordering. First, moving foreground is extracted by modeling background with mixture of Gaussians. Second, a depth value is assigned to foreground object according to its location in the background scene where depth map was estimated.

3.2.1. Foreground Extraction with Gaussian Mixture Models

Adaptive Gaussian Mixture Model (GMM), proposed by Stauffer and Grimson, is a common method to model complex and dynamic backgrounds [41]. The values of

a particular pixel are modeled as a mixture of Gaussian distributions. Determination of which Gaussians may correspond to background is done according to the persistence and the variance. Pixels that fit the background distributions are considered background, while pixel values which are not included by none of these background Gaussians are considered as foreground.

The values of a particular pixel, e.g. scalars for grayvalues or vectors for color images, over time are defined as pixel process. At time t, the history of a particular pixel, $\{x_0, y_0\}$, consist of time series of its values:

$$\{X_1, ..., X_t\} = \{I(x_0, y_0, i) : 1 \leq i \leq t\} \tag{3.9}$$

where I is the image sequence and $X_t = (x_t^r, x_t^g, x_t^b)$ for color images. This algorithm models the history of each pixel independently as a mixture of weighted K Gaussian distributions in the color space. The probability of observing the current pixel value, X_t, is

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \tag{3.10}$$

where K is the number of distributions which is used between 3-5 currently. For the i^th Gaussian in the mixture at time t; $\omega_{i,t}$ is the weight, $\mu_{i,t}$ is the mean value, $\Sigma_{i,t}$ is the covariance matrix, and η is the Gaussian probability density function which are given in Equation 3.11.

$$\mu_{i,t} = (\mu_{i,t}^r, \mu_{i,t}^g, \mu_{i,t}^b)$$

$$\Sigma_{i,t} = \begin{pmatrix} \sigma_r^2 & 0 & 0 \\ 0 & \sigma_g^2 & 0 \\ 0 & 0 & \sigma_b^2 \end{pmatrix}$$

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2n)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_t - \mu_t)^T \Sigma^{-1} (X_t - \mu_t)} \tag{3.11}$$

The model is updated with the new pixel value each time instant. The current K distributions are checked for every new pixel value until a match is found. If a pixel value within 2.5 standard deviations of a Gaussian distribution, it is defined as a match [41].

If the pixel can not match any of the existing distributions, then a new distribution will be created instead of the least probable distribution. This new distribution is created with a mean value that is defined as the current pixel value and with a high prior variance and initially low weight.

The weights of the K distributions at time t is determined as follows:

$$\omega_{k,t} = (1 - \alpha)\omega_{k,t-1} + \alpha(M_{k,t}) \tag{3.12}$$

where α is the learning rate and M is 1 for matched distribution and 0 for others. After this update process, the weights are renormalized to 1.

In case of the new pixel value is matched, the model parameters are updated as follows:

$$\mu_t = (1 - \rho)\mu_{t-1} + \rho X_t \tag{3.13}$$

$$\sigma_t^2 = (1 - \rho)\sigma_{t-1}^2 + \rho(X_t - \mu_t)^T(X_t - \mu_t) \tag{3.14}$$

where $\rho = \alpha\eta(X_t|\mu_k, \sigma_k)$.

All model parameters remain the same in case of unmatch.

Background Gaussians are assigned by the determination of the distributions with larger variances and higher weights. In order to find distributions which belongs to background, firstly, distributions ordered according to ω/σ values.

Then first B distributions, from the sorting, are chosen as the background model according to following criterion:

$$B = argmin_b \left(\sum_{k=1}^{b} \omega_k > T \right) \tag{3.15}$$

where T is a threshold ($0.5 \leq T \leq 1$) indicates the minimum data which should be formed by the background.

Eventually, if the pixel X_t is matched one of those background distributions, it is marked as belonging to the background, otherwise it is marked as foreground pixel.

After GMM method, resulted foreground region is enhanced and noise is removed by morphological filtering.

3.2.2. Disparity Assignment

In order to specify depth values of segmented foreground, we utilize the background relative depth information. We assign the relative depth value of the foreground object according to its location in the background depth map. Background depth value on the coordination of the bottom point of the foreground is assigned to whole foreground region. Consequently, final depth map of the frame is generated by merging background depth map and foreground depth. An example of result is shown in Figure 3.13.

3.3. Foreground Disparity Assignment with Background Registration

In videos which are taken with moving cameras, displacement between two consecutive frames is more on the objects closer to camera and less on the objects further away from camera. Owing to that parallax effect, motion vector of each pixel gives information about its depth. In the proposed method, we registered consecutive frames based on the displacements on background pixels that is caused by camera motion,

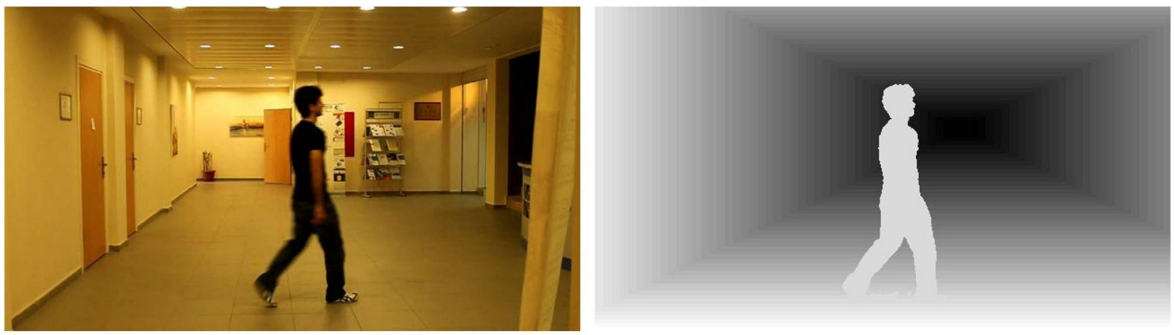


Figure 3.13. Original image and its synthesized depth map that includes both foreground and background. Foreground is extracted by using gaussian mixture model.

while keeping foreground region unregistered. By doing this, we obtain marginal disparity on foreground regions. After that, we provide an artificial disparity to background pixels according to the background depth map. Finally, we create stereo pair using this combined disparity information. General flowchart of the method is given in Figure 3.14.

In Figure 3.14, the algorithm flowchart of generating the stereo frame by background registration method is given. I_{t-n} and I_t are the images of the same scene which are taken by a horizontally moving camera at times $t-n$ and t respectively. Here, I_t will be taken as left image and right image will be calculated using I_{t-n}. Using the geometrical structures of backgrounds which are obtained by the method in Section 3.1, background motion is extracted by block matching. Depending on background motion, the background of I_{t-n} is registered to align with the background of I_t. Resulted image is denoted by $\tilde{I}_{t-n}$ which has the same background pixels with I_t but with different disparity on foreground regions. After obtaining disparity on foreground, to create stereo image with the frame I_t, and enhance the 3D effect also on background pixels, disparity is added for background pixels on $\tilde{I}_{t-n}$. Therefore, finally obtained image $\hat{I}_{t-n}$ and original image I_t has disparities on both foreground and background pixels, so they are used as stereo images to create 3D effect [42].

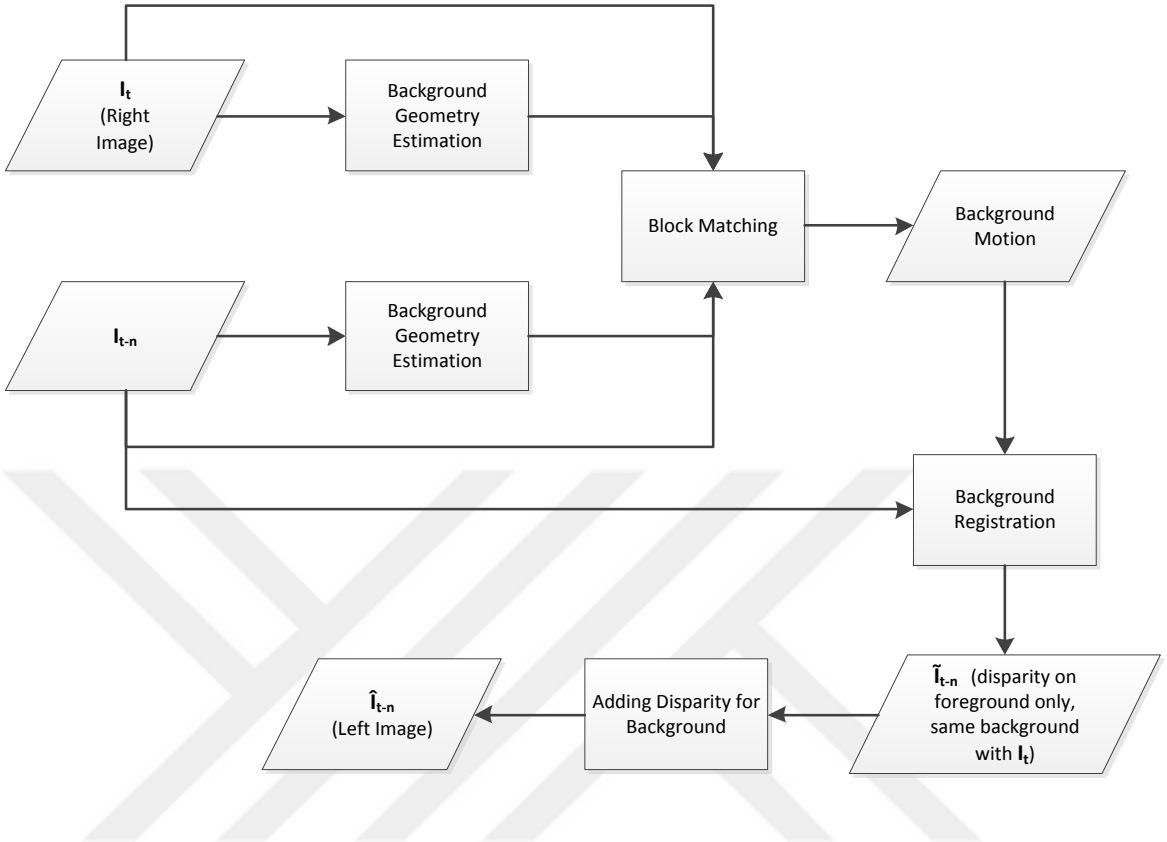


Figure 3.14. Background registration method for stereo frame generation from monocular image frames.

Below, background registration and disparity adding steps are explained in detail.

3.3.1. Background Registration

In order to estimate horizontal motion of background between frames, we used specified interest points which characterize the background geometry of the scene. Background geometry is obtained by the method of Section 3.1, and by exploiting it, interest points are selected on the main lines which construct background geometry. Since foreground objects mostly stand closer to the center of the image and outer regions tends to be background, four points are selected to be close to the borders of the image and on the main lines as shown in Figure 3.15(a).

(a) Selected points on I_{t-n}

(b) The matched points of 3.15(a) are shown on I_t

Figure 3.15. Block matching results. (a) Interest points are determined to apply block matching method. (b) Resulted corresponding points of (a). Background motion between two frames is estimated based on these five background points.

For the motion estimation between two frames block matching method is applied to these interest points. By taking I_{t-n} as reference frame and I_t as current frame, blocks defined around each point on the reference frame and their corresponding blocks are searched on the current frame. For $I_{t-n}(i,j)$ point in the reference image, the block is defined as $I_{t-n}(i-b/2:i+b/2, j-b/2:j+b/2)$. After that, this block is searched in the the current image for the points of $I_t(i, j-k:j+k)$. b states the size of the block and k defines the search area in I_t. Since it is horizontal motion, block searching is performed on horizontal axis. For finding best match of the interest point, a cost function is calculated at each possible location in the search area. The cost function is defined as the sum of absolute differences of two block pixels. For the blocks with Δl displacement, cost is calculated as follows:

$$C(\Delta l) = \sum_{n=-b/2}^{b/2} \sum_{m=-b/2}^{b/2} |I_{t-n}(i+n, j+n) - I_t(i+n, j+n+\Delta l)| \tag{3.16}$$

The block which gives the minimum cost in the current frame defined as the best match of reference block. Thus, Δl value of best match taken as the displacement value of the point $I_{t-n}(i,j)$. This process is applied to the four interest points and displacement amounts are estimated for each of them. The points showed in Figure 3.15(a) are matched with the points that are showed in Figure 3.15(b).

As fifth interest point, vanishing point is estimated by the method of Section 3.1.1 for each frame and displacement of VP is calculated separately. After all, these displacement values of five background points used for the estimation of the displacement map of whole background. Since the disparity between stereo frames is also a displacement which is occurred by a horizontal camera motion; we can correlate the displacement and disparity (background disparity map is known by the method in Section 3.1) linearly. For each pixel i with the total pixel amount N, we can define displacement (or horizontal motion) $(\Delta l_i, i \in \{1, ..., N\})$ based on disparity $(d_i \in [0,1], i \in \{1, ..., N\})$ as follows:

$$\Delta l_i = \alpha d_i + \beta \tag{3.17}$$

where α and β parameters are constants. In this study, geometrical structure of a scene background is explained by its relative disparity/depth map. If the background geometry and at least two pixels' displacement values are known, then we can determine (α, β) parameters using 3.17 and estimate whole displacement map of the background. In order to make the displacement estimation strong against to errors/noise in the proposed method, we use five different pixels' displacement values. The background disparity map of the frame I_t is estimated in Section 3.1. Therefore, the displacement between I_t and I_{t-n} is extended from five points displacement to the whole image's background displacement map via least square estimation (LSE) as follows:

$$\underbrace{\begin{bmatrix} \Delta l_1 \\ \vdots \\ \Delta l_5 \end{bmatrix}}_{\boldsymbol{b}} = \underbrace{\begin{bmatrix} d_1 & 1 \\ \vdots & \vdots \\ d_5 & 1 \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} \alpha \\ \beta \end{bmatrix}}_{\boldsymbol{w}}$$

$$\boldsymbol{w}_{LS} = (\boldsymbol{A}^T \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{b} \tag{3.18}$$

Using estimated (α, β) parameters, we spatially transform I_{t-n} to align with the background of I_t. For that registration process, the method of pixel shifting which is explained in Section 3.4 is used. As a result of background registration, we obtained a new image, $\tilde{I}_{t-n}$, whose background coordinates are same with I_t but whose foreground pixels have a marginal disparity.

3.3.2. Adding Disparity to the Background

After background registration, we have the frames I_t and $\tilde{I}_{t-n}$ which have disparity just on the foreground regions. If these frames were used as stereo pair, the whole background would seem like a plain curtain behind the foreground region. In order to remove this "curtain effect" on the background region, we add user controlled disparity to $\tilde{I}_{t-n}$, based on the background geometry. Using the method defined in Section 3.4, we shift pixels of $\tilde{I}_{t-n}$ based on I_t's background relative depth map and

create the background disparity. Consequently, we obtain $\hat{I}_{t+k}$ next to I_t as the final stereo pair which creates a 3D effect on both background and foreground.

3.4. Stereo Image Generation

In order to create 3D effect from 2D images, the depth map is converted to right and left eye images. The original image is taken as the right eye image; and the left eye image is created based on its depth map. In our proposed method, we create depth layers within the nearest and farthest region of a scene, i.e., within [Z_{near} and Z_{far}]. The extent of horizontal shift (disparity) of the pixels belonging to i^{th} depth layer is defined as:

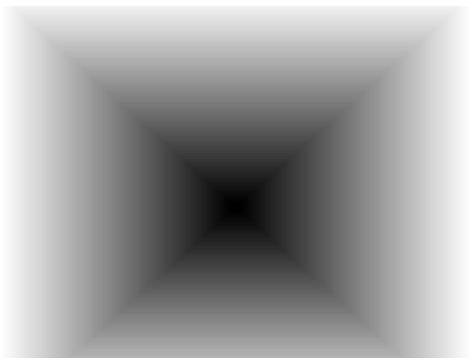
$$d(i) = i\ \left(\frac{d_{max} - d_{min}}{N}\right) \qquad 1 \leq i \leq N \tag{3.19}$$

Where d is the disparity value and i ranges from layer 1, which corresponds to Z_{near} to layer N which corresponds to Z_{far}. Maximum disparity, d_{max}, is the shifting value of the closest region in the image; and minimum disparity, d_{min}, is the shifting value of the deepest region in the image. Both of them can be decided by user in order to adjust the depth effect of 3D viewing.

After calculating the disparity values for each pixel on the single 2D image, we shifted its pixels according to that disparity map. As a result of that shifting process, there will be holes in the shifted version of original image. These holes are filled by interpolation methods.

In Figure 3.16, results of process is shown. The original image in Figure 3.16(a) is shifted to its left stereo pair based on the depth map shown in Figure 3.16(b). Direct result of that pixel shifting has holes as seen in Figure 3.16(c). Thus, we filled the holes and obtain left image as shown in Figure 3.16(d). Finally, we obtain two images; one of which is the original image as right eye image, and the other is the shifted version of it as the left eye image. Any artifact in the newly created image would be subject to be masked by the higher picture quality of the original image which is presented

(a) Original image (Right image)



(b) Depth layers

(c) Shifted image with holes

(d) Shifted image with filled holes (Left image)

Figure 3.16. Stereo image generation. Original image in (a) is shifted according to gradient depth layer map of (b). Emanated holes in (c) are filled as shown in (d).

to the right eye [43]. Consequently, stereoscopic depth is created for human brain to interpret these two shifted images view as one 3D image.

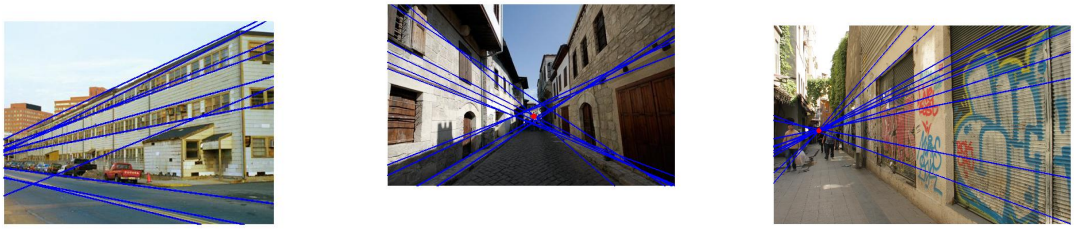
4. EXPERIMENTS AND RESULTS

The proposed methods are implemented in MATLAB environment on an Intel Core 2 Duo PC with 2GB memory by using a data comprised of real images and videos. We collected some of these images from internet, and we captured some other images and videos by a conventional camera.

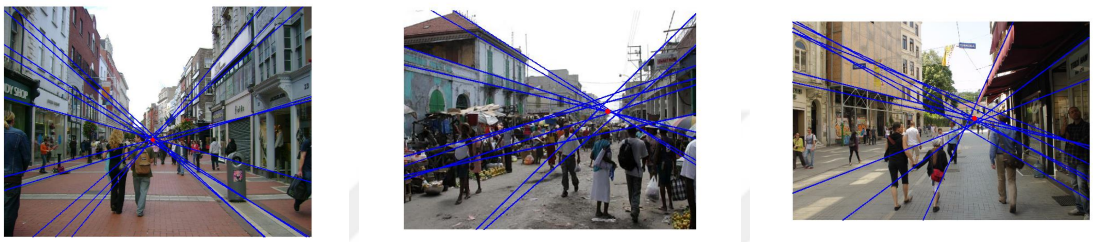
4.1. Vanishing Point Estimation Experiments and Results

We tested vanishing point detection method which is given in Section 3.1.1 on real images with different scenes. We categorized the image data into five classes in terms of indoor with geometrical cues, outdoor with geometrical cues, indoor with geometrical cues including people, outdoor with geometrical cues including people, and nature. In Figure 4.1, we show some examples from each class and we pointed estimated vanishing point with a red dot on the images. In addition, we draw ten lines which are used to estimate vanishing point on the images as blue lines.

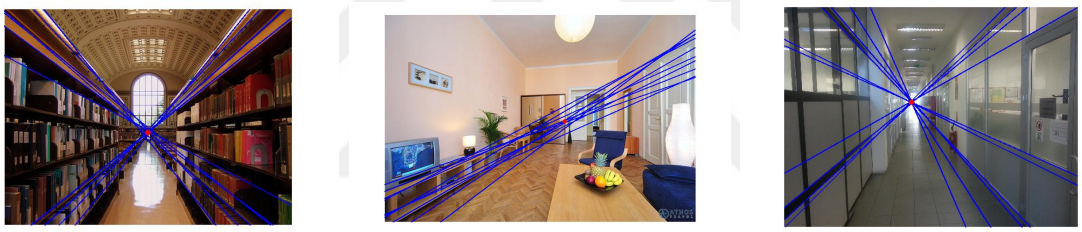
The location of VP is estimated by the method in Section 3.1.1. Besides, true vanishing point is marked on the image manually as the furthest point in the image. A circle around true VP with a radius of 5% of the longer dimension of image is defined as the convenient VP region. If the estimated coordinates of VP is in the VP region determined as above, the detection is accepted successful. Number of images in each category and their success rate is given in Table 4.1. According to the results given in the table, it can be seen that VP detection is more successful in the scenes with geometric elements. In the scenes that includes people, two situations are to be considered. If the scene is crowded and people are close to camera, VP is very difficult to be estimated correctly, because people occludes the main lines converging to VP. If there is a few people or people are away from camera enough, then the results are not negatively affected so VP detection is achieved successfully. For outdoor images with geometric elements, in spite of complexity of the scenes and even though some of them had more then one VP, results are satisfactory in general view. Finally, VP detection



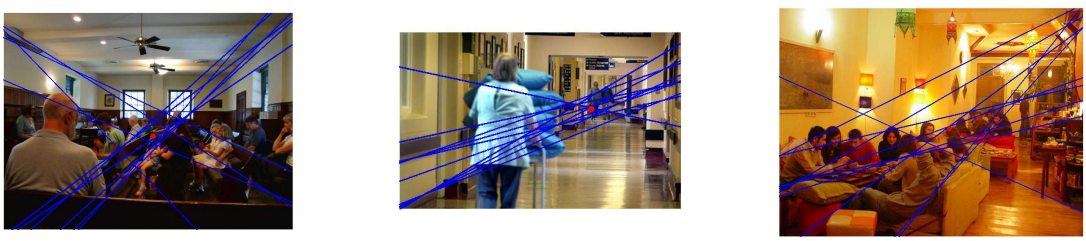
Outdoor images with geometric elements



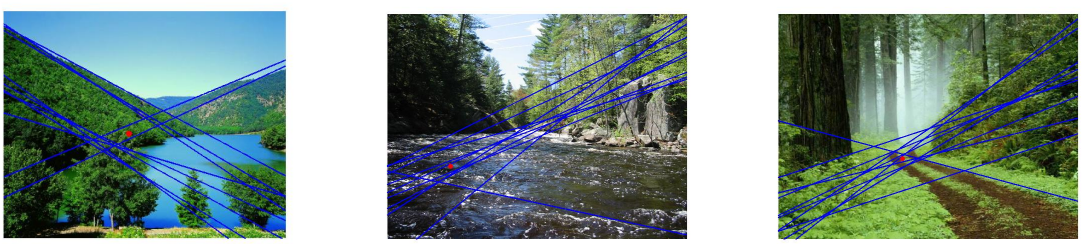
Outdoor images with geometric elements including people



Indoor images



Indoor images including people



Nature images

Figure 4.1. Sample images from test data which is used for vanishing point estimation method

Table 4.1. Vanishing Point Estimation Results. Success rate of VP detection is given as estimated VP points fall inside the manually labeled true VP region.

	# of samples	**the percentage of success**
indoor	22	96
indoor with people	6	67
geometric outdoor	63	86
geometric outdoor with people	23	82
nature	13	39
total	127	81

performance for nature images is low as expected, since they do not have geometric structures and a certain VP.

In Figure 4.2, results of proposed depth map estimation method which depends on scene perspective cues such as vanishing point and main lines are shown.

4.2. Background Registration Experiments and Results

To show the results of the background registration method which is proposed in section 3.2.2, we recorded a video with horizontal camera motion of a scene including foreground objects. Video frames are size of 640x360 and are recorded by the frame rate of 25 frames per second. Our proposed method, background registration, is applied to two frames which are shown in Figure 4.3, in order to evaluate the results in a sample of two frames with seven frame interval.

We first extracted the background geometry based on geometrical cues as presented in Section 3.1. After that we found the displacement between two frames based on their background geometry and shifted pixels of second frame to register the background with first frame (Section 3.3.1). As a result we obtained a disparity just on

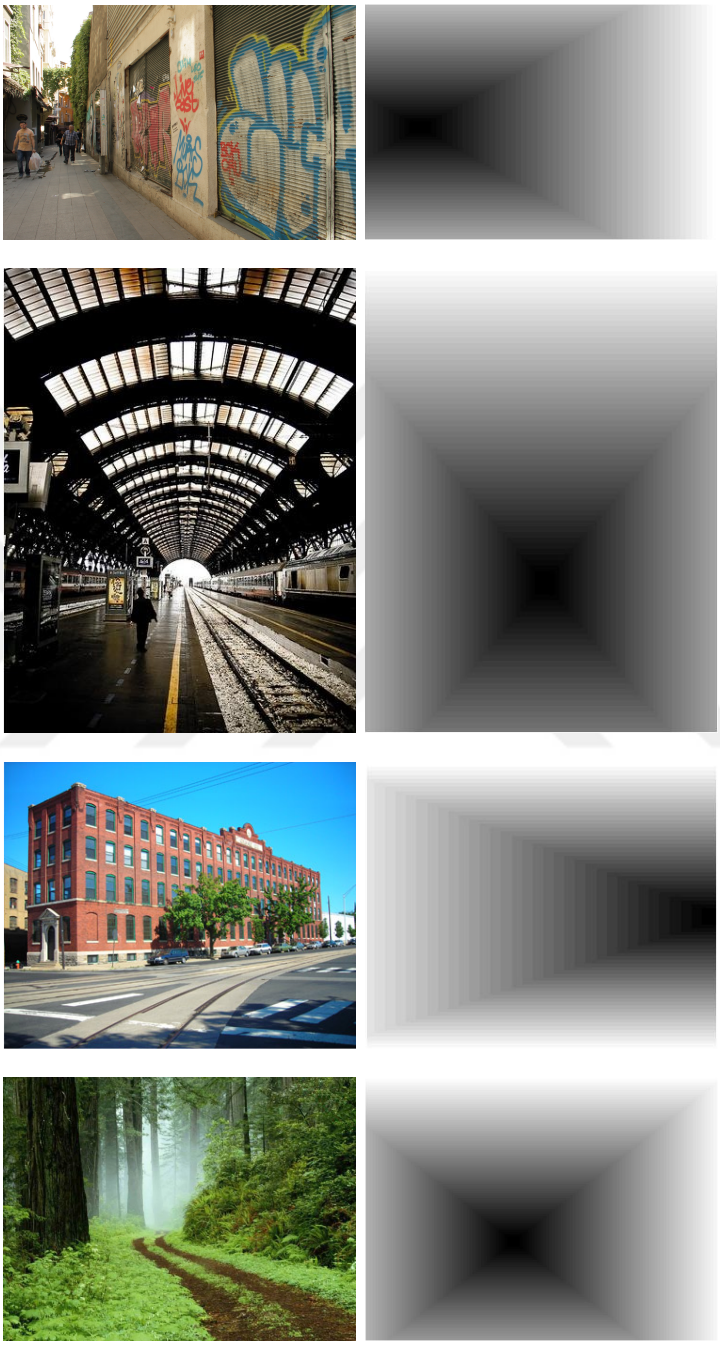


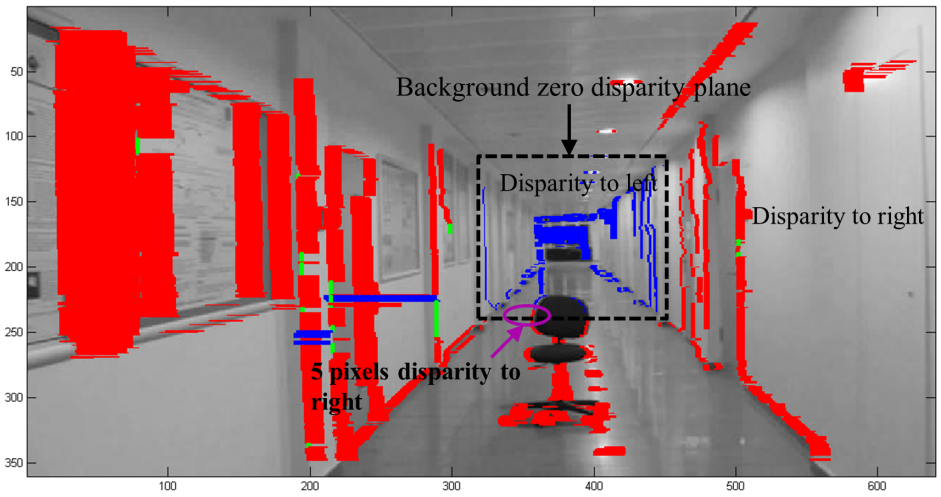
Figure 4.2. Depth maps of different images that is obtained by the proposed depth estimation method based on VP detection. Left column shows original images and right column shows their depth maps.

Figure 4.3. Two consecutive frames taken by a horizontally moving camera.

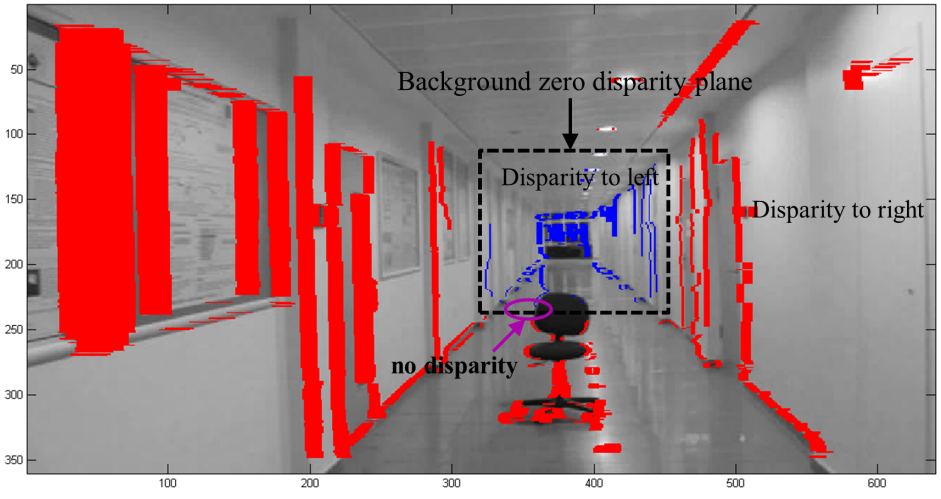
foreground object. After all, we added a disparity to the frame according to depth geometry of background (Section 3.3.2). In this experiment we chose the convergence layer (zero disparity layer) in the middle of depth lines and we gave disparities between -10 to 10.

For the evaluation of obtained stereo frame, we extracted disparity map from stereo pair using the method given in [44]. Taking the original image as right image, and the resulted image of background registration method as left image; we calculated the disparity map between left and right image. Figure 4.4(a) shows this disparity values on edges of scene. As seen in the figure, we have determined disparities on background region. In this way, the disparity of the chair at foreground region is different from background disparity which makes it distinguished from background by means of 3D effect.

In order to see the difference of disparity values on foreground region with background registration method and without background registration method, we applied disparity to the same initial frame based on only background geometry - ignoring foreground region. In the Figure 4.4(b), the disparity between this result and original frame is also shown. As it can be seen from the figure, there is no difference between the disparities of chair at the foreground and background of the scene. If one watch this stereo frame as 3D view, he/she will feel the chair belongs to background without any 3D effect.



(a) Background registration method is applied. Foreground disparity differs from background. Disparity results of generated stereo image is demonstrated on edges of the scene by red lines which show the disparity amount to the right and blue lines which show disparity amount to the left. In the figure, it is shown that chair on the foreground has 5 pixels disparity while its horizontal neighbor pixels are at zero disparity plane of the background.



(b) Background registration is not applied. Foreground disparity is the same with background disparity. It is shown that chair on the foreground has 0 pixel disparity that is the same with its horizontal neighbor pixels which are at zero disparity plane of the background.

Figure 4.4. Disparity results. (a) with background registration, (b) without background registration

4.3. 3D Effect Verification of Generated Stereo Frames

Based on the final depth and disparity maps of the single image, we implemented an image warping process to render virtual view for stereo generation. In this context, we shifted pixels of original image according to depth order information and filled the holes (Section 3.4). Firstly, we tested background depth-order maps, so we created stereo images according to just background depth order information which is explained in Section 3.1. Secondly, we created stereo images based on depth maps including foreground depth-order which is obtained by method in Section 3.2.

In order to create disparity between stereo frames, we shifted the pixels of original image with an amount that determined based on depth order map of pixels. Convergence layer (zero disparity region) was determined by user and different disparity ranges were applied to original images. A group of experts from industry watched the generated 3D stereo videos in commercially available 3D TV sets and marked the problematic regions manually. Generally, it was experimented for background scenes which include geometric perspective. It is found out that, the stereo images whose convergence layer is at closest region are best in terms of depth impression and watching comfort. According to these experiments, we determined the maximum disparity of 20 pixels give better results for viewer. Therefore, we prepared a stereo data for user test to evaluate the success of the proposed method specially for background depth perception. We adjusted the convergence layer as closest region of the background-depth-order map, and shifted whole depth regions linearly in the range of 0 and 20 pixels.

Two test setups are configured for the user test. In the first user test setup, proposed method is compared with other commercial 2D-3D converters. In the second test setup, the results of the proposed method are graded with respect to original stereo videos taken by stereo cameras.

4.3.1. Comparison with Commercial 2D-3D Converters

For the user test, a group of users are invited to watch and grade the quality of the generated stereo images. We showed two version of 3D images which is obtained from the same 2D image data - one is our results and the other is obtained from two different commercial 3DTV's own 2D-3D converters. By such a comparison, we could make a better understanding of what is achieved by our method. Beside our results, one group of users watched the results of Commercial 3D-TV Set-1 while others watched the results of Commercial 3D-TV Set-2 for the same 2D data. This 2D data comprised of eight real 2D single images with different scene contexts for evaluation of background depth effect and two 2D videos that include moving objects for the evaluation of foreground depth effect. Images include indoor with geometric elements, outdoor with geometric elements, and nature without geometric elements and they are required to test background depth estimation method. These original images used as right image and resulted image is used as left image for stereo viewing. Generated stereo image results which are viewed by test subjects are shown in Figure 4.5 and Figure 4.6. In order to test foreground object's 3D performance which is generated by the method of Section 3.2.1, the subjects are required to watch two videos with moving objects. A sample stereoscopic frame from each video is given in Figure 4.7.

We asked to users to give a satisfaction score of the depth feeling for each 3D image and video. The score is in the range of 0 to 10 where 0 stands for no depth feeling and 10 for the perfect depth impression. Forty people in total participated in our tests; twenty six of them viewed test data on Commercial 3D-TV Set-2, and fourteen of them viewed them on Commercial 3D-TV Set-1. For each 2D test data, the subjects were shown 3D images generated by our proposed methods, and 3D images converted by other commercial TVs. The subjects were not informed about source of 3D images. Finally, average score for each test data is obtained and used as measure of the performance of the generated stereoscopic images. Results are shown in the Figure 4.8 separately.

(a) img1

(b) img2

(c) img3

(d) img4

Figure 4.5. Generated stereo images for the user test for evaluation of background 3D effect

(a) img5

(b) img6

(c) img7

(d) img8

Figure 4.6. Generated stereo images for the user test for evaluation of background 3D effect

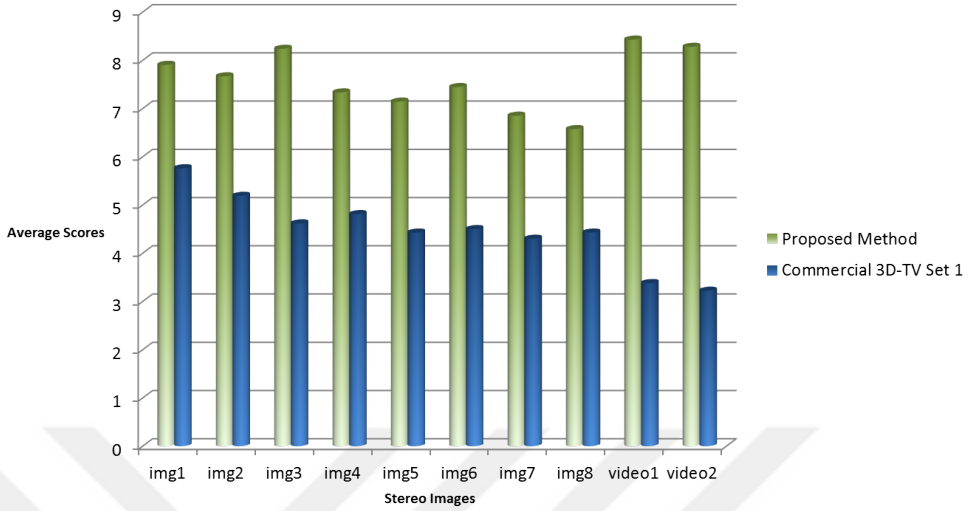
(a) video1

(b) video2

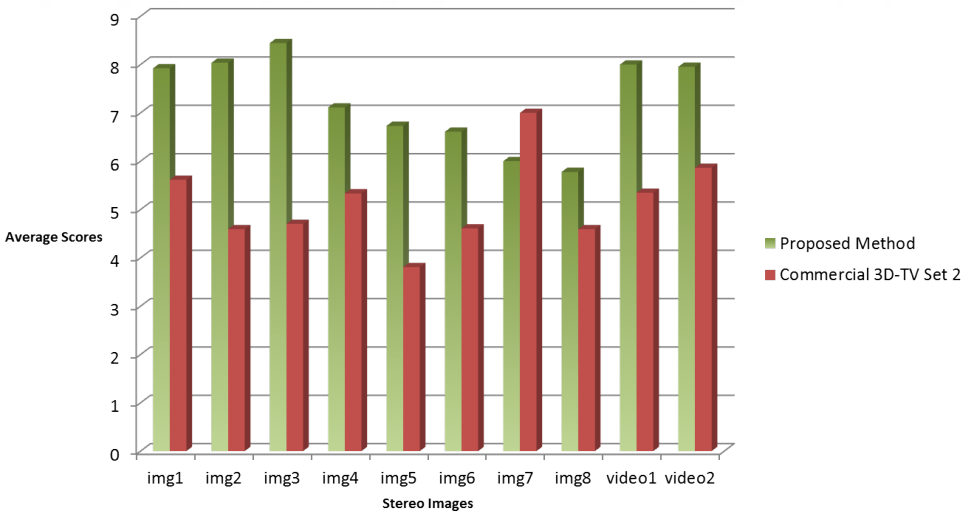
Figure 4.7. Generated stereo images for the user test for evaluation of both foreground and background 3D effect

According to this test results, our proposed methodology produces better performance on 3D effect than other two converters of commercial 3D-TV sets. Background depth effect of proposed method is effective especially on the scenes with strong geometric structures, as it obtained circa eight points in average for the first three images and above the averages of the other TV converters. The reason why first three images have such higher success over other images is that they have strong geometrical cues and larger real scene depth value. Therefore, scene is converging to vanishing point more apparently in line with the proposed methodology. Thus, 3D effect on this type scenes is more enhanced by our background depth estimation method.

For the nature images of img7 and img8, our proposed method has not satisfactory success. Since they do not have any geometrical cues and a certain vanishing point; estimated background geometry for this type of images do not meet the real depth structure.



(a) Proposed method vs 2D-3D converter of Commercial 3D-TV Set 1. Average user test scores over maximum of ten points for eight images and two videos for comparing 3D experience of our proposed method with that commercial 2D-3D converter. Proposed method is graded higher scores than the other converter in all cases.

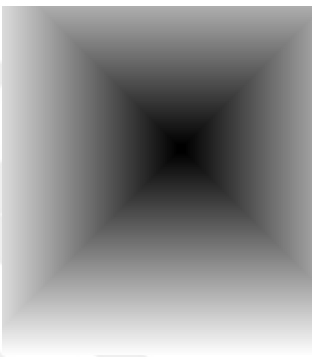


(b) Proposed method vs 2D-3D converter of Commercial 3D-TV Set 2. Average user test scores over maximum of ten points for eight images and two videos for comparing 3D experience of our proposed method with that commercial 2D-3D converter. Proposed method is graded higher scores than the other converter in most cases.

Figure 4.8. Charts of the 3D user test results

(a) Original stereo image: Street

(b) Resulted stereo image of the proposed method



(c) Background depth map obtained by the proposed method

Figure 4.9. Stereo test image and stereo result by background depth estimation

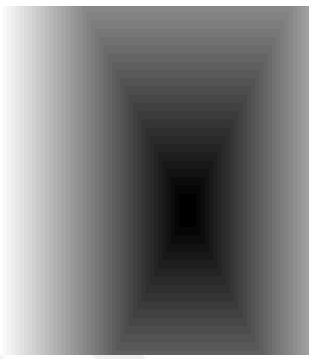
The best 3D effect performance of our proposed method is obtained from the videos including foreground disparity in addition to background disparity. We determine this foreground regions by the method explained in the Section 3.2 and combine foreground depth with background depth. Since adding a foreground depth effect next to the background depth effect makes a 3D perception more impressive, video1 and video2 get higher scores.

4.3.2. Comparison with Original Stereo Videos

For the 3D effect verification, some 3D film frames which are captured by stereo camera sets are downloaded from internet and some stereo videos are taken by a stereo camera to compare with proposed results. Each right eye frames of original stereo

(a) Original stereo image: Cliff

(b) Resulted stereo image of the proposed method



(c) Background depth map obtained by the proposed method

Figure 4.10. Stereo test image and stereo result by background depth estimation

images are taken as a single image to be processed by the proposed method. Resulted stereo images and original stereo images of the same single frames are presented to subjects by using a 3D-TV display with a polarized 3D system. Eleven subjects participated in that user test. For each frame, subjects first watched original stereo frame and they are asked to accept its grade as 10 points. After that, they watched the resulted stereo image produced by the proposed method of the same scene. They are asked to grade proposed result by taking the original stereo image as reference of 10 points.

In Figure 4.9(a) and Figure 4.10(a), two original stereoscopic images taken from different 3D movies are shown. In Figures 4.9(b) and 4.10(b), results of proposed method are shown. Right eye images of the original stereo images are taken as single

Table 4.2. User test results for the single frames of 3D movies. The original stereo frames are taken as references for 10 points. Only background geometry is used for depth map estimation.

	Scores	**Applied method**
street	7.9 / 10	Section 3.1
cliff	5.2 / 10	Section 3.1

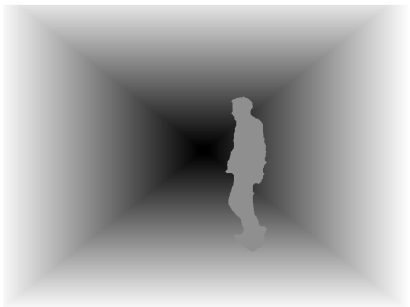
images and they are used as input for the proposed 2D to 3D conversion system. Estimated background depth maps that are used to create stereo images are also shown in the figures.

The user test results for these two images are given in Table 4.2. These two input images are taken as single images and motion analysis are not used for depth determination. Since just background depth map is used for stereo image generation, *street* image which is not include a close foreground object got good scores. In 3D movies, foreground objects which are very close to camera are generally more enhanced by means of depth illusion. In *cliff* image, there is a zebra which stands very close to camera. In original stereo image, this zebra has a very impressive 3D depth effect. In the proposed method we just used background depth map, in other words we gave a depth illusion for the cliff only, and could not determine the depth of zebra. Therefore, subjects could not see a depth on zebra while they were impressed by depth illusion of original stereo image. However they were satisfied by the depth impression of cliff on the 3D image created by proposed method. Although there is no straight lines or geometric elements to converge to vanishing point in the scene, the proposed method is prone to find VP near to the center of the image in such cases. Since, cliff has a deep depth near to the center of the image also, subjects found the background depth effect very good.

For the evaluation of the proposed motion analysis for depth effect creation we captured stereo videos by using a stereo vision camera with 648x488 resolution, 2.5 mm focal length and color chroma. Two different videos are taken which include: (i)

(a) A frame from the stereo video of moving object with stationary camera

(b) Resulted stereo frame by the proposed method



(c) Hybrid depth map obtained by the proposed method

Figure 4.11. Stereo test image and stereo result by background and foreground disparity assignment.

moving object with stable camera (Figure 4.11(a)), and (ii) stable object with moving camera (Figure 4.12(a)).

Using these original stereo frames' right eye images, we also create stereo videos by our proposed method. In the video of stable camera and moving object, Gaussian mixture models are used to determine moving background object as explained in Section 3.2. Figure 4.11(b) shows the final result which is obtained by using both background and foreground depth estimation methods. After that, we presented our results and original stereo frames to the subjects for evaluation. While scoring, first original stereo frames are showed to users and they are asked for taking it as a reference

(a) A frame from the stereo video of stable object with moving camera

(b) Resulted stereo frame by the proposed method

Figure 4.12. Stereo test image and stereo result by background and foreground disparity assignment.

for maximum points that is 10 points.

In the same way, the stereo video of stable object with moving camera is processed using our proposed method and results are showed to subjects. For this video frames, background registration method (Section 3.3) is applied for disparity assignment beside background depth estimation. Obtained comparison scores of user test is given in Table 4.3.

According to the subjective evaluation results which is given in Table 4.3, we can say that the case of stationary camera and moving foreground objects are successful. Its overall scene depth impression score is 8.3/10 where original stereo frame is taken as maximum points. In addition, when it is evaluated separately for background depth

Table 4.3. User test results for stereo videos taken by a stereo camera. The original stereo frames are taken as references for 10 points. Background geometry is used for background depth and motion analysis are used for foreground depth estimation.

	Background scores	**Foreground scores**	**Overall scene scores**	**Applied method**
Video with moving object and stationary camera	8.4	8.0	8.3	Section 3.1 Section 3.2
Video with stable object and moving camera	8.4	2.5	5.4	Section 3.1 Section 3.3

and foreground depth, it has good scores for both of them.

For the second video of stable foreground with moving camera is not such successful for overall depth impression. Both of two videos can be said that have same background geometries, so their background depth impression has same scores which can be accepted as successful. However, subjects stated that the foreground region in the second video has not enough depth impression for a satisfactory depth experience. From the numerical disparity analysis that is given in Section 4.2, we can see that the marginal disparity amounts on foreground objects are not much enough. We can create a different disparity on foreground region by background registration method, however this disparity is inadequate for an impressive depth illusion, especially when compared by original stereo version of frames. Therefore, for further studies, enhancement of that obtained foreground disparity can be give better results.

5. CONCLUSION

In this thesis, we proposed a method for stereoscopic image generation from monocular image or monocular image sequences. In order to convert 2D images for 3D stereoscopic visualization, main and most difficult part is depth map estimation to create the stereo images from single frames. As shown in the literature review, there are different methods that use pictorial cues of the given image such as perspective geometry, motion parallax, and focus/defocus for depth estimation. In this work, we presented a hybrid method for depth estimation based on geometrical cues and motion analysis over a 2D image.

In the proposed method, geometrical cues are utilized to determine depth gradient layers of scene background. Within a given image, a rough background depth map is built benefiting from perspective distortion in real images. Vanishing point resulting from perspective distortion of parallel lines in the real world serves as anchor point. The vanishing point is estimated through the use of Hough Transformation. After determining vanishing point, its location in the 2D image is assigned as farthest location of scene in terms of the distance from camera. In addition we found main lines which constitute general background geometry to determine depth order layers. This enabled us to create entire background geometry of the image.

In order to reinforce depth effect, we also included two different approaches to assign depth and disparity to foreground regions based on motion analysis. In the first approach, we extract moving foreground regions from the 2D frame sequences by using Gaussian mixture models. After finding foreground object, background depth information generated from geometrical cues are combined with this region based on its location. In the second approach, background registration is applied for consecutive frames that are captured by a moving camera. By this method, disparity in foreground regions is distinguished from background disparity that leads to a distinctive 3D effect on foreground regions.

In the proposed hybrid methodology, all these background and foreground depth and disparity information were fused to create final disparity map to obtain stereo image. By taking single input image as right eye image, we created left eye image by warping input image based on the eventual disparity map.

Depth maps which are extracted from single images are only approximations of true world for adequate 3D visualisation. It is shown that proposed methodology has successful results for stereoscopic viewing. As user evaluation test indicates, our results have satisfactory 3d impression and better than some commercially available 2D to 3D converters. However, since our method depends on geometrical features of scenes, its performance was highly satisfactory in scenes mostly composed of manmade structures while it showed limited performance in natural scenes.

Our purpose was to present a real-time implementation of 2D to 3D video conversion for 3D-TVs. Therefore, we avoided to use the methods with more computational complexity. In this perspective, more extensive implementation based on our method may yield better results on more complex computational environment. Improvements can be made for both effectiveness and robustness. Our estimated depth map can be used as a first estimation for accurate iterative depth map estimation methods.

REFERENCES

1. Forsyth, D. A. and J. Ponce, *Computer Vision A Modern Approach*, Pearson Education, Inc., Upper Saddle River, NJ 07458, 2003.

2. Feng, Y., J. Ren and J. Jiang, "Object-Based 2D-to-3D Video Conversion for Effective Stereoscopic Content Generation in 3D-TV Applications", *Broadcasting, IEEE Transactions on*, Vol. 57, No. 2, pp. 500–509, 2011.

3. Iinuma, T., H. Murata, S. Yamashita and K. Oyamada, "54.2: Natural Stereo Depth Creation Methodology for a Real-time 2D-to-3D Image Conversion", *SID Symposium Digest of Technical Papers*, Vol. 31, No. 1, pp. 1212–1215, 2000.

4. Kim, D., D. Min and K. Sohn, "Stereoscopic Video Generation Method Using Motion Analysis", *3DTV Conf.*, pp. 1–4, 2007.

5. Kim, D., D. Min and K. Sohn, "A Stereoscopic Video Generation Method Using Stereoscopic Display Characterization and Motion Analysis", *IEEE Transactions on Broadcasting*, Vol. 54, No. 2, pp. 188–197, June 2008.

6. Jung, C., L. Wang, X. Zhu and L. Jiao, "2D to 3D conversion with motion-type adaptive depth estimation", *Multimedia Systems*, Vol. 21, No. 5, pp. 451–464, 2015.

7. Liu, W., Y. Wu, F. Guo and Z. Hu, "An efficient approach for 2D to 3D video conversion based on structure from motion", *The Visual Computer*, Vol. 31, No. 1, pp. 55–68, 2015.

8. Pourazad, M. T., P. Nasiopoulos and R. K. Ward, "An H.264-based Scheme for 2D to 3D Video conversion", *IEEE Trans. Consumer Electronics*, Vol. 55, No. 2, pp. 742–748, 2009.

9. Ideses, I., L. Yaroslavsky and B. Fishbain, "Real-time 2D to 3D video conversion",

Journal of Real-Time Image Processing, Vol. 2, pp. 3–9, October 2007.

10. Huang, X., L. Wang, J. Huang, D. Li and M. Zhang, "A Depth Extraction Method Based on Motion and Geometry for 2D to 3D Conversion", *Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on*, Vol. 3, pp. 294–298, 2009.

11. Trucco, E. and A. Verri, *Introductory Techniques for 3-D Computer Vision*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998.

12. Stoykova, E., A. A. Alatan, P. Benzie, N. Grammalidis, S. Malassiotis, J. Ostermann, S. Piekh, V. Sainov, C. Theobalt, T. Thevar and X. Zabulis, "3-D Time-Varying Scene Capture Technologies, A Survey", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 17, No. 11, pp. 1568–1586, November 2007.

13. Wei, Q. Q., "Converting 2D to 3D: A Survey", December 2005, Delft University of Technology, the Netherlands.

14. Lai, S., C. Fu and S. Chang, "A Generalized Depth Estimation Algorithm with a Single Image", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 4, pp. 405–411, 1992.

15. Subbarao, M. and G. Surya, "Depth from Defocus: A Spatial Domain Approach", *International Journal of Computer Vision*, Vol. 13, No. 3, pp. 271–294, 1994.

16. Nayar, S. K. and Y. Nakagawa, "Shape from focus", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 8, pp. 824–831, August 1994.

17. Ens, J. and P. Lawrence, "An Investigation of Methods for Determining Depth from Focus", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 2, 1993.

18. Schechner, Y. Y. and N. Kiryati, "Depth from Defocus vs. Stereo: How Different

Really Are They?", *International Journal of Computer Vision*, Vol. 39, No. 2, pp. 141–162.

19. Ens, J. E., *An investigation of methods for determining depth from focus*, Ph.D. Thesis, The University of British Columbia, 1990.

20. Tam, W. J., A. S. Yee, J. Ferreira, S. Tariq and F. Speranza, "Stereoscopic image rendering based on depth maps created from blur and edge information", , 2005.

21. Aguirre Valencia, S. and R. M. Rodriguez-Dagnino, "Synthesizing stereo 3D views from focus cues in monoscopic 2D images", , 2003.

22. Cantoni, V., L. Lombardi, M. Porta and N. Sicard, "Vanishing point detection: representation analysis and new approaches", *Image Analysis and Processing, 2001. Proceedings. 11th International Conference on*, pp. 90–94, September 2001.

23. Matessi, A. and L. Lombardi, "Vanishing Point Detection in the Hough Transform Space", *Lecture Notes in Computer Science - Euro-Par'99 Parallel Processing*, Vol. 1685, pp. 987–994, 1999.

24. Battiato, S., S. Curti, M. L. Cascia, E. Scordato and M. Tortora, "Depth Map Generation by Image Classification", *In Proceedings of SPIE Electronic Imaging – Three-Dimensional Image Capture and Applications VI*, Vol. 5302-13, 2004.

25. Battiato, S., A. Capra, S. Curti and M. La Cascia, "3D stereoscopic image pairs by depth-map generation", *3D Data Processing, Visualization and Transmission, 3DPVT 2004. Proceedings. 2nd International Symposium on*, pp. 124–131, 2004.

26. Cheng, C.-C., C.-T. Li, P.-S. Huang, T.-K. Lin, Y.-M. Tsai and L.-G. Chen, "A block-based 2D-to-3D conversion system with bilateral filter", *Consumer Electronics, 2009. ICCE '09. Digest of Technical Papers International Conference on*, pp. 1–2, 2009.

27. Nam, S. W., H. S. Kim, Y. J. Ban and S. I. Chien, "Real-time 2D to 3D conversion for 3DTV using time coherent depth map generation method", *2013 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 187–188, January 2013.

28. Saxena, A., S. H. Chung and A. Y. Ng, "Learning Depth From Single Monocular Images", *Neural Information Processing System (NIPS)*, Vol. 18, 2005.

29. Saxena, A., S. Chung and A. Ng, "3-D Depth Reconstruction from a Single Still Image", *International Journal of Computer Vision*, Vol. 76, No. 1, pp. 53–69, 2008.

30. Saxena, A., M. Sun and A. Ng, "Make3D: Learning 3D Scene Structure from a Single Still Image", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 31, No. 5, pp. 824–840, May 2009.

31. Zhang, J., Y. Yang and Q. Dai, "A novel 2D-to-3D scheme by visual attention and occlusion analysis", *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*, pp. 1–4, May 2011.

32. Zhang, Z., S. Yin, L. Liu and S. Wei, "Real-time time-consistent 2D-to-3D video conversion based on color histogram", *2015 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 188–189, January 2015.

33. Jung, C. and J. Cai, "Superpixel matching-based depth propagation for 2D-to-3D conversion with joint bilateral filtering", *Image Processing (ICIP), 2015 IEEE International Conference on*, pp. 3515–3519, September 2015.

34. Phan, R. and D. Androutsos, "Robust Semi-Automatic Depth Map Generation in Unconstrained Images and Video Sequences for 2D to Stereoscopic 3D Conversion", *IEEE Transactions on Multimedia*, Vol. 16, No. 1, pp. 122–136, January 2014.

35. Criminisi, A., I. Reid and A. Zisserman, "Single View Metrology", *Int. J. Comput. Vision*, Vol. 40, No. 2, pp. 123–148, 2000.

36. Phan, R., R. Rzeszutek and D. Androutsos, "Semi-automatic 2D to 3D image conversion using a hybrid Random Walks and graph cuts based approach", *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 897–900, May 2011.

37. Yan, X., Y. Yang, G. Er and Q. Dai, "Depth map generation for 2D-to-3D conversion by limited user inputs and depth propagation", *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*, pp. 1–4, May 2011.

38. *MATLAB and Image Processing Toolbox Release 2012b*, Natick, Massachusetts, United States.

39. Canny, J., "A Computational Approach to Edge Detection", *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. PAMI-8, No. 6, pp. 679–698, November 1986.

40. Hough, P. V., *Method and Means for Recognizing Complex Patterns*, U.S. Patent 3069654, 1962.

41. Stauffer, C. and W. Grimson, "Adaptive background mixture models for real-time tracking", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, 1999.

42. Coban, A., B. Aksu, B. Acar and F. Basaran, "Background registration for foreground separation in 2D to 3D video conversion", *Signal Processing and Communications Applications Conference (SIU), 2012 20th*, pp. 1–4, April 2012.

43. Stelmach, L., W. J. Tam, D. Meegan and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution", *Circuits and Systems for Video Technology, IEEE Transactions on*, Vol. 10, No. 2, pp. 188–193, March 2000.

44. Gurol, O., S. Ozturk, B. Acar, B. Sankur and M. Guney, "Sparse disparity map

estimation on stereo images", *Signal Processing and Communications Applications Conference (SIU), 2012 20th*, pp. 1–4, April 2012.

45. Gonzalez, R. C. and R. E. Woods, *Digital Image Processing (3rd Edition)*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.