Bachelor's Thesis

2D-to-3D conversion system

Low-complexity 2D-to-3D video conversion based on multiple monocular depth cues

Wouter van Rooy

January 8, 2013

School of Information and Communication Technology

Student details

Name	Wouter van Rooy
Student ID	2143398
Course	Technische Informatica (Applied Computer Science)
Internship period	September 2012 - January 2013

Company details

Name	Axon Digital Design
Address	Lange Wagenstraat 55 5126 BB Gilze The Netherlands
Product owner	Dr. Ir. Rafael Peset Llopis
Intern supervisor	Ir. Luc Vosters

Institution details

Name	Fontys School of Information and Communication Technology
Address	Rachelsmolen 1 5612 MA Eindhoven The Netherlands
First assessor	Ir. Ben Schreur
Second assessor	Ing. Cees van Tilborg

Preface

This report is my bachelor's thesis for the conclusion of the course Technische Informatica (Applied Computer Science) at Fontys School of Information and Communication Technology, Eindhoven. Additionally, it forms the crown of my graduation project at the Research & Development department of Axon Digital Design.

From September 2012 to January 2013 I have been researching the possibilities of expanding and improving the Axon 2D-to-3D video conversion system.

Hereby I would like to thank all Axon employees for giving me a warm welcome and a great experience at their company. I would like to express my special thanks to my intern supervisor Luc Vosters, whom I could count on for both technical and process related support on a daily basis. Another word of thanks goes out to Rafael Peset Llopis, who always managed to find time despite his busy schedule.

Also my mentor and first assessor Ben Schreur has been very helpful throughout the project.

Wouter van Rooy

January 8, 2013

Contents

Summary

Axon Digital Design develops modular systems and equipment for audio and video signal processing and monitoring, mainly aimed at the broadcasting industry.

The recent successes of 3D film productions have pushed the consumer TV market into a new era: 3D television in the living room. One fundamental problem however is the lack of 3D-enabled content. Producing native stereoscopic 3D video is time consuming and costly. It requires broadcasters to invest in new expensive equipment like stereo cameras and stereo rigs and to hire specially trained stereographers. Realtime 2D-to-3D conversion is a cheaper option since it requires additional hardware only. Unfortunately this is an extremely difficult task for which no optimal generally applicable solution exists.

At Axon Digital Design, an efficient method for the conversion of shallow depth-of-field material has been developed. This method is based on the focal blur cue and it reconstructs the scene depth based on different blur values in the image. Unfortunately, this approach fails for material with a deep depth-of-field in which (nearly) all objects are in focus. Therefore the conversion system needed to be expanded, allowing it to leverage other depth cues as well.

An additional depth estimation algorithm based on linear perspective was developed. It leverages dominant lines in the image to find a vanishing point. The detected location is used to reconstruct the scene depth, by generating a depth map with increasing depth towards the vanishing point. Viewer perception tests have shown that this method yields a natural and more realistic depth experience compared to the gravity model. In the gravity model, depth is determined by the vertical position in the image only.

The fusion of these algorithms is based on an estimation of their confidence. For the focal blur method, the standard deviation of blur values is estimated. The confidence of the linear perspective method is determined by the spatial drift of vanishing point locations across time. Due to the nature of the current depth estimation algorithms, the depth cues they depend on are hardly ever concurrently available. Therefore the fusion module will simply select the best fitting algorithm rather than combining multiple methods simultaneously.

If the focal blur method is considered a confident option, it will always prevail over other methods. Else, if the detected vanishing points are temporally stable, the linear perspective method will be considered the best fitting algorithm. In the special case where neither method returns satisfactory results, the gravity model is applied as a fallback.

Nomenclature

Cortex	Configuration, Control and Monitoring software for the Synapse platform.
depth cue	Image features providing depth information. Monocular (or pictorial) cues can be seen with one eye, while binocular cues can only be perceived when viewing a scene with both eyes.
depth of field	The distance between the nearest and farthest sharp objects in an image.
disparity	Difference in horizontal position of an object in the images of two different viewpoints
DOF	Depth of field
focal blur	Monocular depth cue based on the difference in blur of objects in the image.
HVS	Human Visual System
linear perspective	Monocular depth cue based on real-world parallel lines that seem to converge in a single vanishing point.
motion parallax	Monocular depth cue based on the difference in motion vectors of objects in an image.
PHT	Probabilistic Hough Transform
stereoscopy	Stereoscopy is the illusion of depth that is perceived when viewing a slightly different image with each eye.
Synapse	Synapse Modular Interfacing and Conversion system. Modular audio and video format conversion system by Axon Digital Design.
temporal consistency	Characterization of algorithm behavior in the time domain.
TRACS	Transmission Recording and Compliance System. Recording and inventory system for broadcast content.
vanishing point	One of possibly multiple points in which parallel lines in a 3D scene seem to converge when represented as a 2D image.

1. Introduction

Producing native stereoscopic 3D video for live broadcast is still time consuming and costly. It requires broadcasters to invest in new expensive equipment like stereo cameras and stereo rigs and to hire specially trained stereographers. Real time 2D-to-3D conversion is a cheaper option since it requires additional hardware only. Unfortunately 2D-to-3D conversion is an extremely difficult task for which no optimal generally applicable solution exists.

Axon Digital Design has developed an efficient conversion method based on the focal blur cue. Blur values in low depth of field material are estimated and can be used to reconstruct the scene depth. Based on these depth estimations, a 3D image renderer is able to create a left and right eye view of the scene. This approach unfortunately fails for material in which the focal blur cue is not available (i.e. deep depth of field).

This project focuses on improving the conversion system by combining multiple depth cues. This requires the development of at least one additional depth estimation algorithm.

Chapter 2 will provide a short introduction to Axon Digital Design. In Chapter 3, a full overview of the project details will be given. Chapter 4 provides a short introduction to 3D video. Chapter 5 describes how to automatically extract depth from linear perspective. Chapter 6 will focus on combining multiple depth estimation algorithms. Finally, in Chapter 7 we propose a method for enforcing temporal stability in our 2D-to-3D conversion system.

2. Company information

Since its establishment in 1987, Axon Digital Design has been developing modular systems and equipment for audio and video signal processing and monitoring. Their products and services are used by companies in the broadcasting industry, like broadcasters, service providers and video professionals.

2.1. Products

There are two main product branches in the Axon portfolio. On the one hand there is the Synapse Modular Interfacing and Conversion system. The system consists of a 19-inch frame with passive connector panels and a vast spectrum of active hot swappable interface cards. Its flexibility and extensibility allow an easy conversion between a range of audio and video formats, for both digital and analog platforms.

The Synapse system is supported by the Cortex Configuration, Control and Monitoring software. This software is designed to easily create and manage multiple audio and video signal paths using a wide range of Synapse products.

Another major product branch is formed by the TRACS (Transmission Recording and Compliance System) products. Many countries have defined regulations for television programmes, like for example time restrictions regarding content less suitable for juvenile viewers. Often, broadcasters are considered guilty unless they prove otherwise. The TRACS product line facilitates live recording and inventory of broadcast material, allowing broadcasters to address such issues more efficiently.

Figure 2.1.: Axon product lines (left-to-right: Synapse, Cortex and TRACS)

2.2. Organization

The organizational structure of Axon Digital Design is shown in Fig. 2.2. There are five major departments:

- Operations (ICT, Purchasing & Logistics, Quality)
- Finance (Finance, Human Resources)
- Marketing & Sales
- Product Management
- Research & Development

This graduation project is situated at the Research & Development department of the Axon head office in Gilze, The Netherlands. Other company sites are located across the globe: there are sales offices in China, the United Arab Emirates, Russia, Singapore and the USA. Additionally, there is an office in the United Kingdom where the Cortex software is developed. All other operations are performed at the head office.

A typical R&D project structure for the development of a multidisciplinary system is shown in Fig. 2.3.

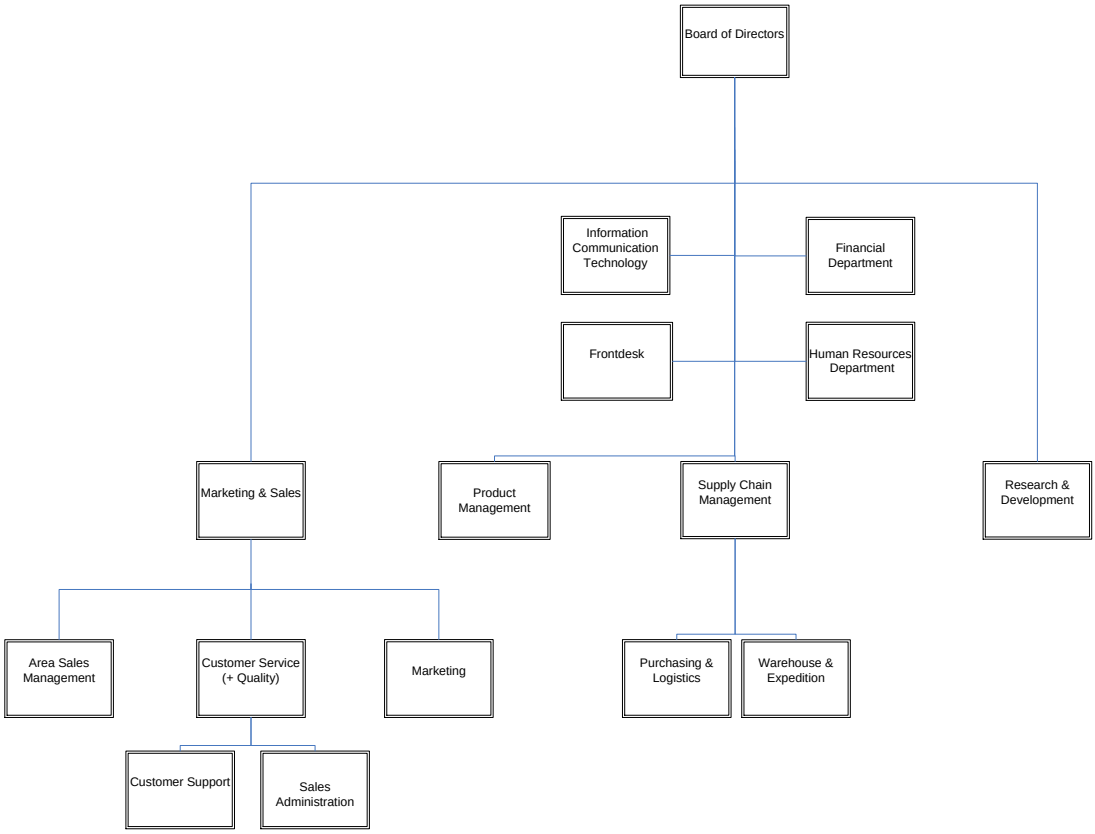


Figure 2.2.: Company organigram

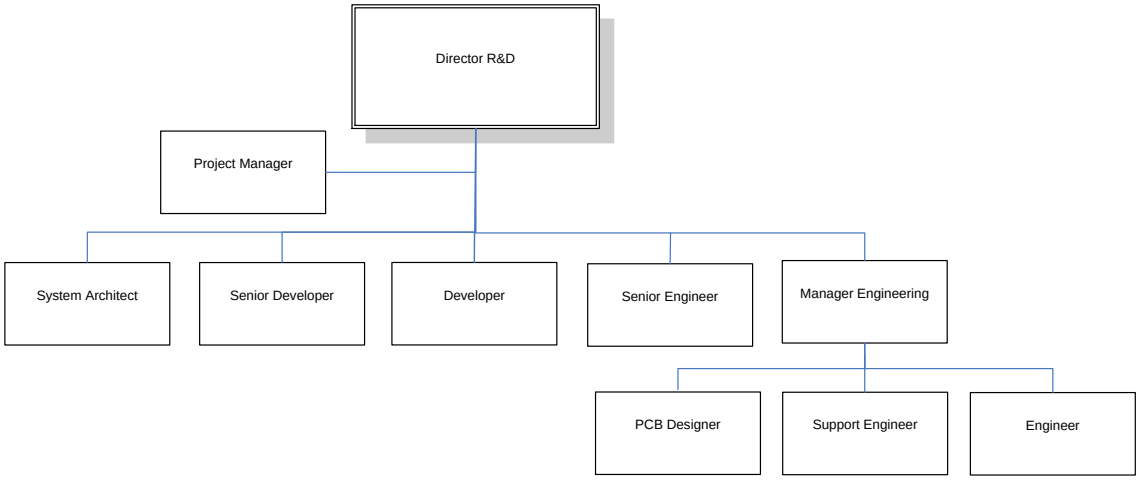


Figure 2.3.: R&D project organigram

3. Graduation project

3.1. Background

In state of the art automatic 2D-to-3D conversion both motion based and pictorial cue based methods can be distinguished. In motion based algorithms camera and object motion are used to estimate depth from motion parallax. Pictorial cue based methods use depth cues like focal-blur, perspective, texture-density, occlusion, relative height, etc. Many techniques have been proposed to estimate depth from individual cues, and an excellent overview can be found in [1].

However, the human visual system (HVS) integrates multiple depth cues rather than perceiving depth from a single cue.[2] Even in monocular video depth is perceived which is determined from a composition of cues, where the individual cue contribution can vary from shot to shot. Therefore, a key challenge in obtaining realistic depth for 2D-to-3D conversion lies in integrating various depth cues into the system.

3.2. Initial situation

Axon Digital Design has developed a 2D-to-3D conversion system based on the focal blur cue.[3][4] It exploits different blur levels in a shallow depth of field (DOF) video to reconstruct the scene depth. Based on these depth estimations, a 3D image renderer is able to create a stereoscopic image pair. Then a scaler encapsulates these images into several formats for display on a 3D television set.

Unfortunately this approach fails for video with a deep DOF in which (nearly) all objects are in focus. Therefore, an additional depth estimator based on another reliable depth cue is required, and the conversion system has to be extended to support it. Hence we need a content based fusion module that automatically decides - based on the monocular depth cues in an image - which depth estimator to apply.

3.3. Goals

The main goal of this project is to extend the Axon 2D-to-3D conversion system and thereby improve the viewer's 3D experience. To get a more reliable conversion at least one additional depth estimation algorithm needs to be included. Linear

perspective is chosen because it is an important cue to the HVS[5] and it is readily available in still images and video sequences.

Another aspect of the project will cover the integration and combination of individual depth cues. The fusion of depth maps is a relatively underexposed topic in literature. The topic however is invaluable for a successful conversion system.

Due to the complexity of the proposed system, extensive research is needed. The central question that needs to be answered during this research is:

> How can an automatic 2D-to-3D conversion system be improved by combining multiple depth cues to let users experience a significantly better depth perception?

3.4. Approach

The complexity of the project called for a flexible approach, in which research received the main focus. Implementation is involved in the project, but merely as an instrument to indicate the feasibility and accuracy of the researched algorithm. A simplified version of the W-model is chosen as main structure. (Fig. 3.1)

The W-model is similar to the V-model, except for the development phase of the project.[6] Detailed research, prototype implementation and testing are all embedded in a cyclic and iterative process. This allows step-by-step applied research, directly coupled with the development of a prototype.

3.4.1. Project definition

The project definition phase consisted of writing the attached Project Initiation Document (Appendix A). This document defines the project in detail and served as a basis for project management and assessment.

3.4.2. Initial research

During the initial research phase, the main objective was getting familiar with existing approaches to 2D-to-3D conversion and image processing in general. A library of technical papers (including [1], [7], [8] and [9]) provided by Axon served as a starting point from which further investigation started.

3.4.3. Development

The development phase of the project consists of three sub-phases, each concerning one of the desired modules. During the planning phase of the project, these modules

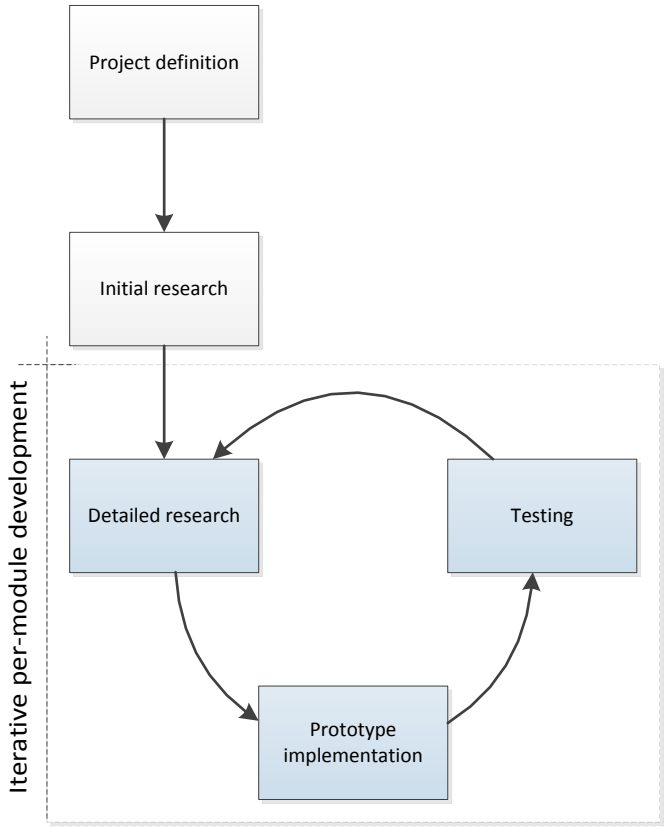


Figure 3.1.: Project approach

seemed to be loosely dependent: the depth map fusion module required an additional depth estimator, but for example the scene change detection algorithm could be developed without other modules in place.

Therefore, the order in which the sub-phases would be developed allowed a relatively large degree of freedom. Because of earlier described dependencies the depth from perspective module was scheduled at the start. Depth map fusion would be the next topic of research, finally followed by the scene change detection module.

3.4.3.1. Depth from perspective

Phase 1 comprises of the research and development of an additional depth estimation algorithm based on linear perspective cues in a still image.

In its simplest form perspective can be simulated by applying a general gravity depth map in which the depth increases from the bottom to the top of the image. De-

termining whether the proposed linear perspective map improves the 3D experience compared to the simple gravity map will be an important step in this phase.

If the viewer perception tests point out that there is no significant improvement from the gravity map to the linear perspective map, a depth estimation algorithm based on an alternative depth cue will be developed. If there is sufficient improvement the linear perspective algorithm will be further optimized for use with video sequences.

This phase requires the following questions to be answered:

- How can a computer vision algorithm automatically generate a depth map from linear perspective cues in a still image?
- What is the difference in depth perception between a simple gravity depth map and an enhanced map based on linear perspective?
- How can the depth-from-perspective algorithm be optimized to detect temporally consistent vanishing points in video sequences?

3.4.3.2. Depth map fusion

To decide for which images we have to apply depth from focal blur and for which depth from perspective, a depth map fusion module has to be designed.

This phase requires the following questions to be answered:

- Is hard switching between depth cues sufficient or would a mixture/weighted average of depth cues be more appropriate?
- Would spatially variant mixture weights improve the overall depth perception?

3.4.3.3. Scene change detection

Applying depth estimation algorithms to video sequences poses several new challenges: the estimated depth of an object could fluctuate across time as a result of estimation inaccuracies. Previous research at Axon has shown that this disturbs the viewer's depth perception significantly. In order to maintain temporal consistency for generated depth maps a temporal depth filtering algorithm has been developed for the depth from focal blur estimator.

With the added functionality of the depth map fusion module comes an additional challenge: it is not desirable to constantly change the (mixture of) depth estimator(s) within the same scene. This would distort the temporal depth consistency and has a negative impact on depth perception and video quality.

A better method would be to only alter the fusion weights at scene changes. A scene change detection algorithm would help in solving this problem. Flagging scene changes allows the conversion system to reset its temporal depth filters and switch to a (mixture of) depth cue(s) that is more suitable for the new scene's content.

Furthermore, individual depth estimation algorithms may benefit from these flags for their internal filtering.

This phase requires the following questions to be answered:

- How can scene changes in a video sequence be identified automatically?
- What can be done to allow individual depth estimation algorithms to benefit from scene change detection?

3.5. Revisions

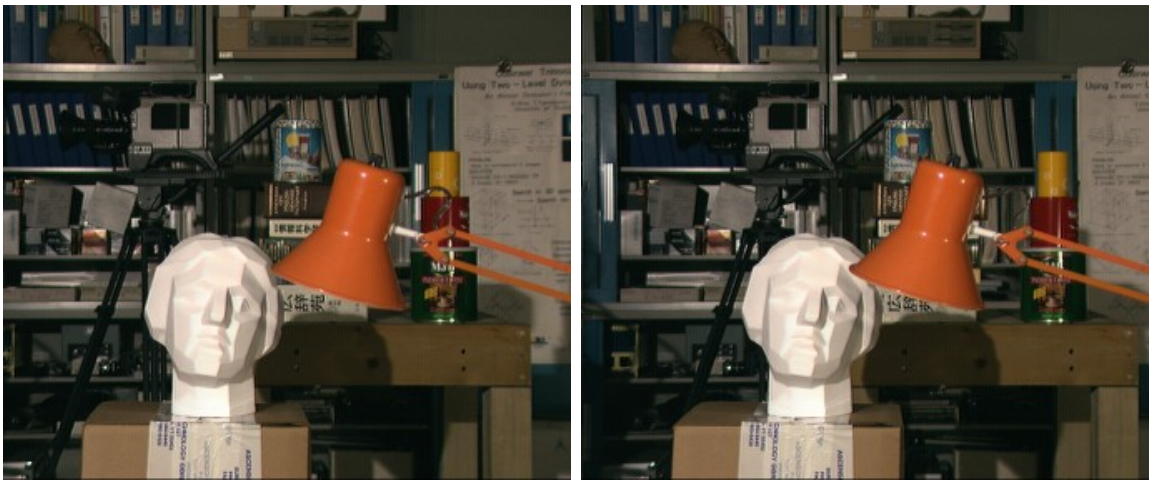
The order of module development in the research & development phase (Sec. 3.4.3) was slightly changed during the project execution. The depth map fusion module requires a metric of confidence for each available depth estimator. The depth from perspective module however, would only be able to return such a metric based on the temporal stability of the detected vanishing point.

Including support for scene change detections was a relatively simple step and greatly simplified implementation and testing of the fusion module. Therefore, these two modules were developed in reverse order making depth map fusion the final topic of research rather than scene change detection.

4. Introduction to 3D video

4.1. Stereopsis

Our two eyes are positioned in different locations on the head, causing their views of a 3D scene to slightly differ. Our brain uses this information to form a 3D representation of the world. Looking carefully at Fig. 4.1a reveals that certain objects have shifted horizontally compared to Fig. 4.1b.



(a) Left image **(b)** Right image

Figure 4.1.: Stereopsis in a real-life scene[10]

This horizontal shift is called disparity and its amount is directly related to the object's depth. Defining x_{left} and x_{right} as the horizontal position of a point in the left and the corresponding point in the right image respectively, the horizontal disparity d for that point is defined as:

$$d = x_{right} - x_{left} \tag{4.1}$$

The observed effect can be explained using Fig. 4.2, which is a top-down view of a stereo camera system. The cameras in this setup are analog to the both eyes and the

dotted line is formed by the points on which the disparity is equal to zero. The circle is located before this line, while the rectangle is positioned behind it. According to Eq. 4.1 the circle has a negative disparity, while the rectangle's disparity is positive.

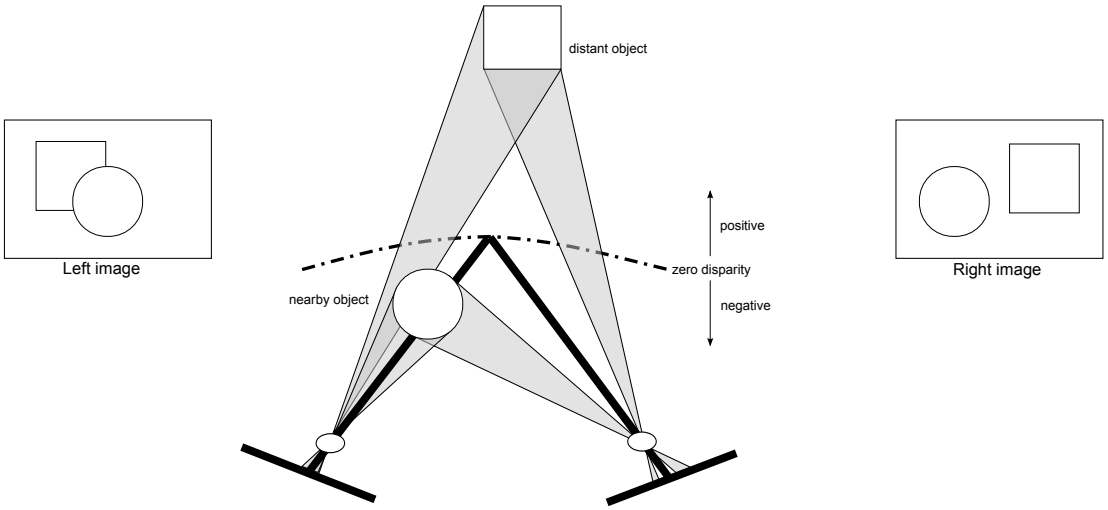


Figure 4.2.: Geometry of stereo vision setup [11]

The HVS would account for these different disparities by tensioning or relaxing the extraocular muscles, effectively aiming the eyes at the object of interest. This information combined with the disparity is finally used by the brain to reconstruct a 3D representation of the scene.

In 3D video systems the points at which the disparity is equal to zero are located 'on-screen'. Research has shown that negative disparity values (i.e. objects appearing in front of the screen) can cause discomfort and visual fatigue.[12] Therefore, we will limit ourselves to positive disparity values (i.e. objects appearing behind the screen) in this project.

4.2. Depth maps

The 3D rendering system makes use of a disparity image, containing a disparity value for each pixel in the input image. The preprocessors in the 2D-to-3D conversion system however use a slightly different approach to describe the location of 3D objects: depth maps.

Depth maps can be considered grayscale images that define a depth value for each image pixel, ranging from white (closeby) to black (far away). Algorithms aiming at reconstruction of depth based on one or more input images are called depth estimators, of which one is described in Chapter 5.

The generated depth maps are then converted into disparity maps, so that the 3D rendering algorithm can generate a set of stereoscopic images. A linear conversion from depth to disparity would be a simple approach to do so. Internal testing revealed that this method introduces some geometrical distortions. Instead, the approach in [13] is used to convert depth Z to disparity d. This function (Eq. 4.2) describes a non-linear inverse relation, which eventually clips at a maximum disparity d_{max}, where d_{max} is a user defined parameter. In this report we define d_{max} as 3% of the input image width.

$$d = \min \begin{cases} \frac{d_{max}}{Z} - d_{max} \\ d_{max} \end{cases} \tag{4.2}$$

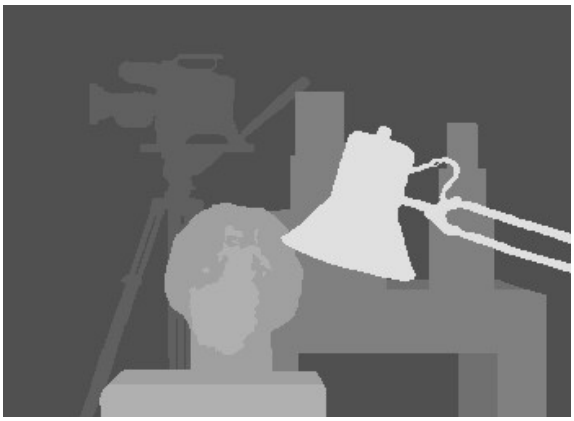


Figure 4.3.: Depth map for the scene in Fig. 4.1[10]

4.3. Display techniques

A multitude of stereoscopic display techniques exists, all having one feature in common: they leverage stereopsis for depth perception by exposing a different image to the left and right eye. A short overview of the most common techniques that are currently available are described in Sec. 4.3.1 through Sec. 4.3.3.

4.3.1. Anaglyph

The anaglyph display technique has been around since its invention in 1852 by Wilhelm Rollmann.[14] It is based on different (often chromatically opposite) color filtering for the left and right eye. Modern anaglyph material typically uses a red filter for the left eye and a cyan filter for the right eye.

A major drawback to this technique is the loss of color information. It is therefore best used with grayscale material. Additionally, the display must be calibrated to match the filter's colors to minimize crosstalk.

4.3.2. Polarization

Polarization is the orientation of oscillations of a wave in the perpendicular field. For the purpose of 3D displays, this is typically a linear polarization system (using horizontal and vertical waves) or a circular polarization (using clockwise and counter-clockwise waves).

An advantage of this technique is formed by the low cost of the glasses. Circularly polarized glasses add the benefit of the audience being able to tilt their heads without seeing the effects of crosstalk.

Polarized displays and projectors are generally more expensive than other systems. Combined with the low cost glasses this technique is especially suitable for application in cinemas.

4.3.3. Active shutter

The active shutter system uses active battery operated glasses which physically blocks the sight of one of the two eyes by using shutters. These shutters each contain a liquid crystal screen which darkens when a certain voltage is applied to it. Synchronizing these shutters to alternating left and right frames on a (typically high frame rate) display effectively splits the views for both eyes.

The fact that the glasses are relatively complicated devices significantly increases their cost. Additionally, they require batteries to operate which might not be as convenient compared to other techniques.

Displays are however generally less expensive compared to other systems, which makes this technique an excellent option for the consumer living room.

(a) Anaglyph **(b)** Polarization **(c)** Active shutter

Figure 4.4.: 3D display glasses

5. Depth from linear perspective

Linear perspective is an important cue to the HVS for depth estimation and perception. Its effect is clearly visible when looking at railroad tracks: the tracks seem to converge as distance increases, eventually converging in a distant point on the horizon. This is due to objects appearing smaller as they are farther removed from the viewer. The point in which the lines converge is called a vanishing point.

The depth from perspective module's primary task will be to extract depth information based on the convergence effect from an input image. However generating depth from perspective is not a trivial task.

Figure 5.1.: Convergence effect illustrated using railroad tracks with the vanishing point marked in red

5.1. Depth map generation

> How can a computer vision algorithm automatically generate a depth map from linear perspective cues in a still image?

Several approaches to this problem exist in literature. [15] describes an approach using a cascaded Hough transform, exploiting the mathematical relations of points in different image spaces. [16] and [17] present a method based on machine learning algorithms such as Expectation Maximization. [18] rather uses a geometric approach using circle intersections. All these papers present promising theories on automatic

vanishing point detection for the 2D-to-3D conversion system. Implementing these algorithms would however prove difficult; either because of their computational complexity or simply due to a lack of details in the papers.

A more viable approach is presented in [19], where candidate perspective lines are identified using the Hough transform. A possible extension of this approach is shown by [20]. This method leverages a combination of the Hough transform, Sobel gradient direction and a modified Hough transform to improve detection accuracy. Due to the limited time available for this project, our scope is limited to the approach in [19].

5.1.1. Overview

A common starting point for all vanishing point estimation techniques is edge detection. The next step comprises of the identification of vanishing lines in the edge image. The vanishing point location is derived from the (most) common intersection point of these lines. After generating a depth map with increasing depth toward the vanishing point, another algorithm renders the stereoscopic image pair. These steps are described in more detail in Sec. 5.1.2 through Sec. 5.1.5.

Fig. 5.2 gives an overview of the steps in the proposed algorithm. Additionally, an anaglyph output frame is rendered to give an impression of the achieved depth effect.

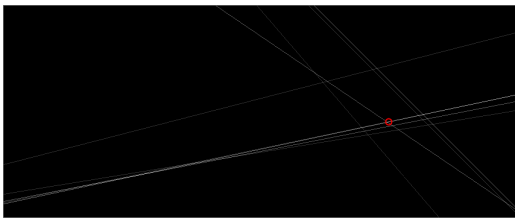
5.1.2. Edge detection

In order to successfully detect dominant lines in an image, some preprocessing is required. The first step in this stage is to retrieve the image's derivative. The first order derivative is, simply put, the difference in intensity between subsequent pixels in an image. Assigning higher values for larger changes yields a gradient image in which the edges of the original image are highlighted. Hence, this procedure is called edge detection.

A multitude of algorithms exists for this purpose with varying performance and accuracy figures.[21] For the linear perspective module it is sufficient to detect dominant lines only. Thus, an accurate edge detector like Canny is not likely to improve vanishing point detection. Instead, the Sobel edge detector is used with a somewhat higher threshold value to minimize noise in the gradient image.

(a) Input frame

(b) Binarized edge detection



(c) Lines detected by Hough transform with VP marked in red

(d) Generated depth map

(e) Anaglyph output frame

Figure 5.2.: Results of the depth-from-perspective algorithm on a frame of the Inception[22] movie trailer

The Sobel edge detector calculates the gradient of the image intensity of each particular position. It uses two convolution kernels; one for the horizontal gradient and one for the vertical gradient. We define A as the source image and G_x and G_y as its gradient images in respectively horizontal and vertical direction. In order to obtain G_x and G_y a simple convolution kernel is applied (Eq. 5.1 and, in its transposed form, Eq. 5.2). As the directions of G_x and G_y are orthogonal to each other, a combined gradient image G can be obtained by taking the Pythagoras sum of the two components (Eq. 5.3). The direction θ of the gradient is calculated as shown in Eq. 5.4 and Eq. 5.5. This information may be used in a later stage to improve the detection of dominant lines in the image.

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * A \tag{5.1}$$

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * A \tag{5.2}$$

$$G = \sqrt{G_x^2 + G_y^2} \tag{5.3}$$

$$\theta = \mathrm{atan2(G_y, G_x)} \tag{5.4}$$

$$= \begin{cases} 2 \arctan \frac{G_y}{\sqrt{G_x^2+G_y^2}+G_x} & G_x \neq 0 \vee G_y \neq 0 \\ 0 & G_x = 0 \wedge G_y = 0 \end{cases} \tag{5.5}$$

After the edge detection algorithm has finished, the gradient magnitude will be thresholded to obtain a binary image mask that is 0 at non-edge positions and 255 at edge positions. Fig. 5.2b shows an example of the binarized gradient image.

5.1.3. Vanishing point estimation

We determined earlier that real-world parallel lines seem to intersect in a vanishing point when the scene is projected on a 2D image. The Hough transform is used to determine the parameters of these lines:

Every edge point in the gradient image has a set of lines crossing through it. Each of these lines can be represented in the form of a slope-intercept equation (Eq. 5.6). Rewriting the equation so that b is a function of a with x and y being constants, yields Eq. 5.7.

$$y = ax + b \tag{5.6}$$

$$b = -xa + y \tag{5.7}$$

The problem with this formulation is that for a vertical line the slope would become infinite, and a near-vertical line would give a very large value for a. In computation systems this will lead to issues like data type saturation or, even worse, overflows. This problem can be solved by using a polar representation as defined in Eq. 5.9.

$$y = \left(-\frac{\cos\theta}{\sin\theta}\right) x + \left(\frac{\rho}{\sin\theta}\right) \tag{5.8}$$

$$\rho = x\cos\theta + y\sin\theta \tag{5.9}$$

In Hough space, the parameters ρ and θ are represented as Cartesian coordinates. Each point in Hough space defines a line in the image space. A point in image space corresponds to a sine wave in Hough space, defining the parameters for all lines passing through the point. If we take two random points in image space, a set of two sine waves is observed in Hough space. The intersection point of these sine waves in Hough space defines the line passing through both of the points in image space.

In the Axon 2D-to-3D conversion system, we will use a slightly modified version of the algorithm above: the Probabilistic Hough Transform (PHT). [23] The difference is that rather than using all edge pixels in the image, only a random subset is used. This allows for a performance increase by minimizing the input dataset.

If we apply this mechanism to a gradient image obtained from Sec. 5.1.2, a large set of sine waves is returned. The dominant vanishing lines can be identified by a multitude of sine waves intersecting in the same point. These intersection points are represented as peaks in Hough space.

From the parameters returned by the PHT, a set of dominant lines can be plotted in image space (Fig. 5.2c). The point(s) in which a large set of these lines intersect are probable vanishing points in the image.

5.1.4. Depth map generation

Based on the acquired vanishing point a rough depth map can be generated. The initial setup will be to fit a radial depth gradient with the vanishing point as center and the farthest image corner as radius (Fig. 5.2d).

5.1.5. Depth image based rendering

The rendering of stereoscopic image pairs is outside the scope of this project. Thus, an existing algorithm (written by Luc Vosters) will be used. It is capable of generating anaglyph, top-down and side-by-side material from a source image and its depth map. An example of a rendered anaglyph output image is included in Fig. 5.2e.

5.2. Quality comparison

> What is the difference in depth perception between a simple gravity depth map and an enhanced map based on linear perspective?

The qualitative performance of the depth from perspective module is assessed using five test images (Fig. 5.3) featuring strong linear perspective cues. For each image two pairs of stereoscopic images were rendered: one using the gravity depth map

and one using the proposed linear perspective depth map. Eleven participants were asked to rank the perceived depth and global image quality in a paired comparison [24] between both approaches.

(a) inception1 **(b)** inception2

(c) skyfall1 **(d)** skyfall2 **(e)** skyfall3

Figure 5.3.: Test images from the Inception[22] and Skyfall[25] movie trailers

In each comparison the participants answered the following question:

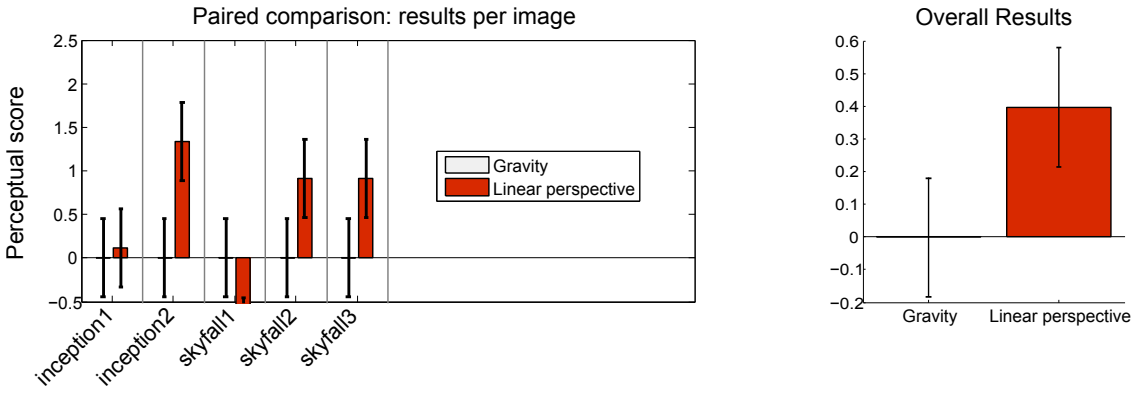
Which image provides a more comfortable and natural 3D experience?

The stereoscopic image pairs were displayed sequentially on a 50-inch Panasonic VT20 3D plasma television with active shutter glasses. Test participants could switch between the two algorithms at the press of a key.

Fig. 5.4a and Fig. 5.4b show the results of respectively the paired comparison per image and for all images combined. We are not interested in the absolute perceptual scores of the two methods, merely the difference between them indicates which is better. Therefore, the perceptual scores of the gravity-based renders are set to 0. The 95% confidence intervals are calculated using the formula in [24].

As defined by [24], overlapping error bars indicate that the difference between two methods is not statistically significantly better. Non-overlapping intervals prove the opposite.

From these figures we can conclude that the linear perspective algorithm yields a statistically significantly better depth perception in images inception2, skyfall2 and skyfall3. However, for skyfall1 the gravity approach achieves a statistically significantly better result. This is probably due to a coincidental match between the gravity map and the 3D scene structure. For inception1 there is no difference in performance. Overall, the linear perspective proves to give a significantly better result as can be seen in Fig. 5.4b.



(a) Perceptual score per image

(b) Overall perceptual score

Figure 5.4.: Paired comparison results for depth perception of gravity and linear perspective approach

5.3. Temporal consistency

> How can the depth-from-perspective algorithm be optimized to detect temporally consistent vanishing points in video sequences?

Vanishing points typically lie at the horizon. Due to the great distance between the viewer and these points, changes in camera position (panning or tilting excluded) are not likely to alter their locations.

The information retrieved from video sequences can also be used to rule out false positives in the estimation process. Vanishing points do not appear or disappear from one frame to another, which means that inconsequently detected vanishing points are not likely to be valid. Therefore large fluctuations in vanishing point location indicate that the generated depth maps are probably unreliable.

5.3.1. Filtering

The detected vanishing point can be stabilized by applying a weighted median filter to its location parameters. A weighted median filter uses different filter weights for each element.[26] In this case, the elements are historic vanishing point locations. A Gaussian pulse response is used for the filter weights, as defined by Eq. 5.10.

$$w(x) \quad = \quad e^{-\frac{2x^2}{N^2}} \tag{5.10}$$

This effectively neutralizes occasional outliers, but it does not account for small drifts in the vanishing point location caused by inaccurate line detection. To correct

this behavior the filtered vanishing point locations of the current and previous frames are averaged.

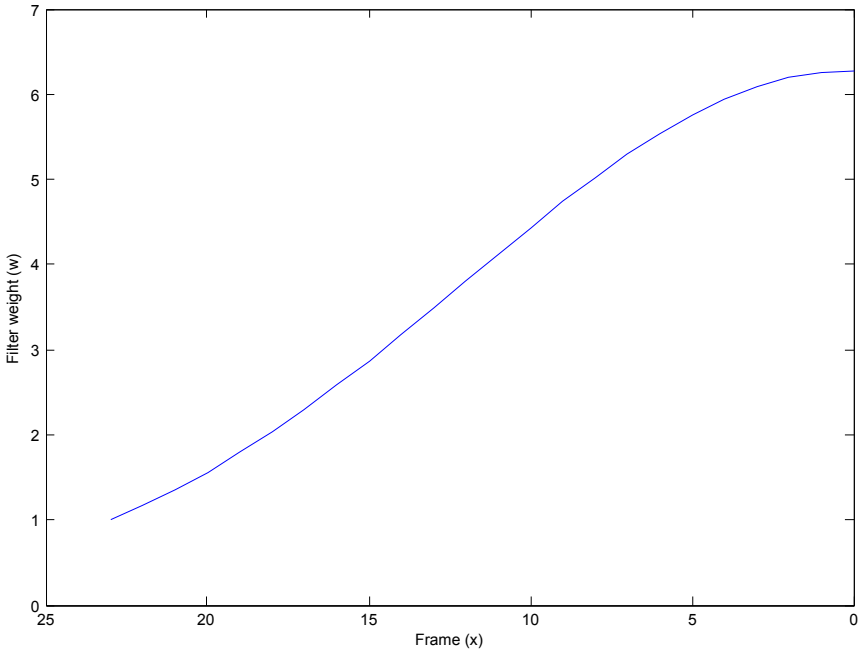


Figure 5.5.: Normalized Gaussian filter weights for a 24-frame history

5.3.2. Scene changes

Of course the above does not hold while going through scene changes in a video sequence. A change of scene (and thus typically a change of subject or environment) introduces a whole new set of dominant vanishing lines and points. Therefore, the depth-from-perspective module is configured to reset its vanishing point location history when a scene change is detected. The conversion system has been extended with a scene change detector as described in Chapter 7.

6. Depth map fusion

A variety of monocular depth cues has been explored for efficient 2D-to-3D video conversion. The fusion of multiple depth maps has however received little attention in the image processing literature.

6.1. Fusion strategy

> Is hard switching between depth cues sufficient or would a mixture/weighted average of depth cues be more appropriate?

The depth map fusion strategy highly depends on the available depth cues in the input image. Both [27] and [28] describe a fusion strategy in which motion-based depth estimation plays an important role. However, [2] introduces a fairly new approach inspired by our knowledge of processes in the human brain. A fusion system is proposed which promotes or demotes individual depth cues based on their presumed reliability.

Our limited set of available depth estimation algorithms poses another problem though. An analysis of the testing material revealed that hardly any images contained both the focal blur cue and the linear perspective cue. With the currently available algorithms it is therefore more meaningful to select one depth estimator per shot. This selection is based on the image content in the shot.

At the start of each scene, all depth estimation algorithms are executed in parallel. Each module returns a per-frame boolean whether or not it is confident about the generated depth map, as described in Sec. 6.1.1 and Sec. 6.1.2. The selection of the best fitting depth estimation algorithm is finally based on a tally of these confidence figures for each algorithm.

Notice that this step needs a few frames to base its algorithm election on, and therefore it introduces a certain latency to the system.

6.1.1. Detecting shallow depth of field material

The depth from focus module has the advantage of segmenting objects based on their estimated blur value. Internal testing revealed that the different depths assigned to

different objects in the image let the viewer perceive a highly realistic depth effect. Therefore, this module is the preferred method if its cue is available.

Detection of focal blur in images is done by estimating the standard deviation of blur values. If the deviation is greater than a certain threshold the image is marked as shallow DOF material.

6.1.2. Detecting linear perspective

The depth from perspective module derives its confidence from its temporal behavior. As stated in Sec. 5.3, large fluctuations in vanishing point location indicate an inaccurate depth estimation. The detected raw vanishing point location is compared to the temporally filtered location. If the change in location is below a certain threshold, the detection is marked as confident.

6.1.3. Fallback: gravity

If both the focal blur and linear perspective cues are not available, the gravity depth map is used as fallback method. Because the module is not dependent on any image features it will always consider its depth map accurate. It will however never be preferred over other methods provided that their confidence is large enough.

6.2. Results

Testing the depth map fusion module was done by comparing the automatic selection results to a set of manually identified scenes. During the manual identification, the same order of preference is applied as mentioned in Sec. 6.1.

Tab. 6.1 shows a tally of identified scene types in the Inception[22] movie trailer. In Tab. 6.2, the frequency distribution of the elected depth estimator for each scene type is shown. From the quoted results we can learn that the proposed system features an average correct detection rate of 82% (i.e. 28 out of 34 scenes).

	Identified
Focal blur	17
Linear perspective	11
Neither	6

Table 6.1.: Manually identified depth cues

		Automatic		
		Focal blur	Linear perspective	Gravity
Manual	Focal blur	14 (82%)	1 (6%)	2 (12%)
	Linear perspective	0 (0%)	9 (82%)	2 (18%)
	Gravity	0 (0%)	1 (17%)	5 (83%)

Table 6.2.: Frequency distribution of manual identifications vs. automatic election results

7. Scene change detection

Temporal filtering of depth maps and vanishing points greatly improves depth stability. The filter history however is only valid during a single coherent scene and should be reset on every scene change. Therefore, scene change detection is a viable factor in temporally stabilizing depth maps.

7.1. Detection algorithm

How can scene changes in a video sequence be identified automatically?

There are numerous ways of detecting scene changes. [29] describes an advanced algorithm making use of object segmentation and tracking, while [30] takes a more simplistic approach by measuring inter-frame differences. The latter requires less computation cost and is sufficient for the 2D-to-3D conversion system.

The paper covers two different scene change types: abrupt and gradual scene changes such as fade-in/out. For our purpose, only the abrupt scene transitions are important. After all there would be no fully correct approach to handling a gradual transition from one scene to another with respect to depth estimation. This allows for a further simplification of the described algorithm.

The first step in the proposed algorithm splits the input frame in a grid of equally sized bins. Each bin is averaged per-channel, and compared to the corresponding value in the previous frame. The sum of absolute differences in all bins yields the total inter-frame difference. A plot of this value is shown in Fig. 7.1b.

While it would be possible to flag a scene change at each and every frame difference value greater than a certain threshold, this would not be a very robust approach. For example a translating camera motion could cause a relatively great change between frames.

In order to tell motion and scene changes apart, a different approach is used. The total frame difference metrics are stored in a historic buffer for the last N frames. A scene change is flagged only if the two highest values in the history have a certain minimum discrepancy factor. A plot of this top value difference factor is shown in Fig. 7.1b. When a scene change has been detected, the frame difference history is cleared.

7.2. Enhancing individual depth cues

> What can be done to allow individual depth estimation algorithms to benefit from scene change detection?

The depth from focus algorithm uses a temporal filtering strategy similar to the depth from perspective module. Future depth estimation algorithms will probably also incorporate temporal filtering of detected features or even entire depth maps. The history used for these filtering methods is only valid during a single scene and should be reset upon scene change.

The developed prototype implements this functionality by exposing an event to both depth estimation modules.

7.3. Results

Our implementation of the scene change detector described in [30] has been tested by manually annotating the locations of scene changes in a video sequence. The sequence used for the test is the Inception[22] movie trailer. Tab. 7.1 shows a tally of manually identified abrupt and gradual[1] scene changes. The correct, missed and false detections by the automatic detector are shown in Tab. 7.2. The false positive detection is caused by an unusually large amount of camera motion in one particular shot.

The performance of the scene change detector is measured in terms of recall and precision. Recall (Eq. 7.1) is defined as the percentage of correct detections out of the total number of opportunities. The precision (Eq. 7.2) describes the fraction of detected scene changes that are valid.

$$Recall = \frac{N_{correct}}{N_{correct} + N_{miss}} * 100\% \tag{7.1}$$

$$Precision = \frac{N_{correct}}{N_{correct} + N_{false}} * 100\% \tag{7.2}$$

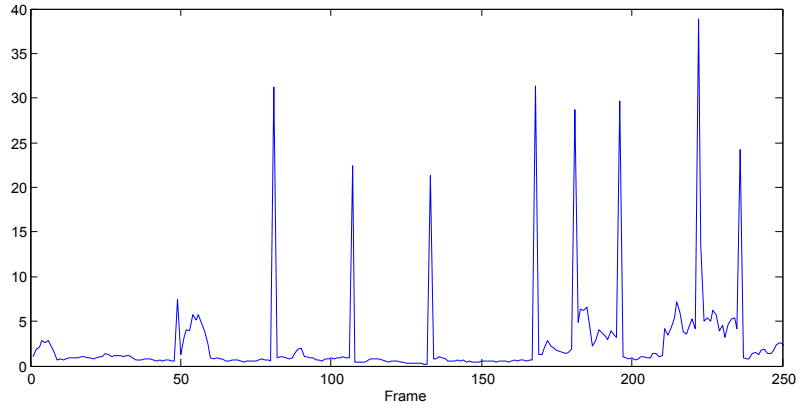
[1] The scene change detector is not optimized for detection of gradual scene changes. These metrics are merely included for completeness.

	Identified
Abrupt	30
Gradual	4

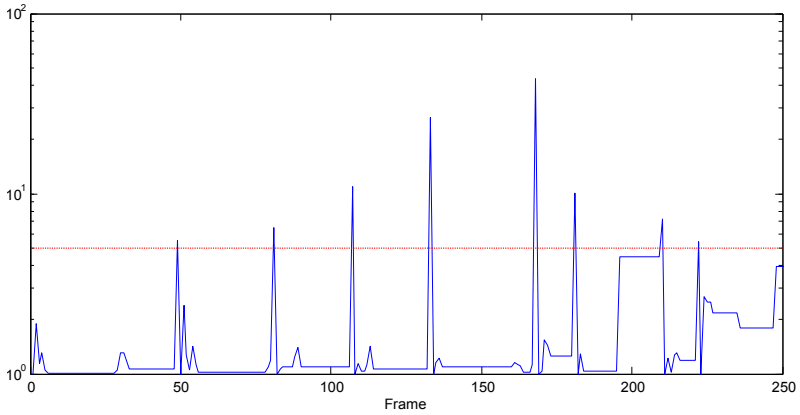
Table 7.1.: Manually identified scene changes

	Correctly identified	Missed	False positive	Recall	Precision
Abrupt	26	4	1	87%	96%
Gradual	1	3	0	25%	100%

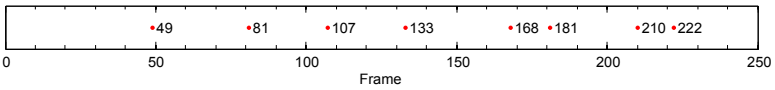
Table 7.2.: Automatically identified scene changes



(a) Inter-frame difference



(b) Top value difference factor



(c) Flagged scene changes

Figure 7.1.: Frame difference graphs for the first 250 frames of the Inception[22] movie trailer

8. Conclusion

The main goal of this project was answering our central question:

> How can an automatic 2D-to-3D conversion system be improved by combining multiple depth cues to let users experience a significantly better depth perception?

Answering this question required extensive research and the implementation of a prototype conversion system. A successful attempt has been made to develop an algorithm that automatically interprets linear perspective cues in a scene. This information is used to detect the location of the vanishing point and finally reconstruct the scene depth. Viewer perception tests have shown that this method yields a natural and more realistic depth perception compared to the gravity model.

The fusion of depth maps based on multiple cues proved to be a difficult task. With the current set of algorithms - focal blur, linear perspective and gravity - we choose a classification system which selects the most accurate depth cue based on its confidence metric. In this setup there is a defined order of preference for all algorithms, as the cues they depend on are hardly ever concurrently available. If however in the future more diverse algorithms are developed this might not be the best approach. In that case, a re-assessment of fusion strategies is recommended.

The temporal stability of the entire system has been improved by developing a scene change detection module. When a scene change has been detected, a new best-fitting depth estimation algorithm is selected. This event is also available for other modules in the system, allowing them to clear their temporal filters.

Evaluation

My first day at Axon Digital Design was in late September 2012. After a tour around the office and an introduction with some of the employees, work could actually start.

The first step during this graduation project was formed by simply diving into literature, searching for clues on how an automatic 2D-to-3D video conversion system could possibly work. Already after a few weeks, the first linear perspective prototype started to take shape. This combination of theoretical research and practical prototyping suited me well.

I experienced a great advantage of working in the image processing field: the visibility of one's work. Whether it was a small optimization or the introduction of a completely new module, I could always see a rewarding change in the conversion results.

My strongest points in this project were acting on my own initiative and working independently. Because I had the luxury of sharing an office with my intern supervisor, we often had a quick exchange of thoughts after which we could both continue our work.

Something I struggled with was selecting the best options in literature for a specific problem. Often there are numerous approaches to solving the same problem and it is simply not feasible to test them all out in a prototype. I have spent a considerable amount of time trying to comprehend each approach before taking a decision. Looking back, I could have made these same decisions more quickly.

Overall I think I can be glad with the achieved results.

Bibliography

[1] L. Zhang, C. Vazquez, and S. Knorr, "3DTV content creation: Automatic 2D-to-3D video conversion," *IEEE Transactions on Broadcasting*, vol. 57, pp. 372–383, June 2011.

[2] C.-T. Li, Y.-C. Lai, C. Wu, S.-F. Tsai, T.-C. Chen, S.-Y. Chien, and L.-G. Chen, "Brain-inspired framework for fusion of multiple depth cues," *IEEE Transactions on Circuits and Systems for Video Technology*, no. 99, pp. 1–13, 2012.

[3] L. Vosters and G. de Haan, "Efficient dense blur map estimation for automatic 2D-to-3D conversion," 2012.

[4] L. Vosters and G. de Haan, "Efficient and stable sparse-to-dense conversion for automatic 2D-to-3D conversion." 2012.

[5] J. A. Saunders and B. T. Backus, "The accuracy and reliability of perceived depth from linear perspective as a function of image size," *Journal of Vision*, vol. 6, pp. 933–954, 2006.

[6] E. Bouman, *SmarTEST*. Academic Service, 2 ed., 2008.

[7] X. Cao, A. C. Bovik, Y. Wang, and Q. Dai, "Converting 2D video to 3D: An efficient path to a 3D experience," *IEEE MultiMedia*, vol. 18, pp. 12–17, Oct. 2011.

[8] S. Battiato, A. Capra, S. Curti, and M. L. Cascia, "3D stereoscopic image pairs by depth-map generation," in *Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium*, 3DPVT '04, (Washington, DC, USA), pp. 124–131, IEEE Computer Society, 2004.

[9] L. J. Angot, W.-J. Huang, and K.-C. Liu, "A 2D to 3D video and image conversion technique based on a bilateral filter," *Proceedings of SPIE*, vol. 7526, 2010.

[10] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, pp. 7–42, 2002.

[11] E. Olson and M. Hao, "An introduction to stereo vision and disparity computation." Lecture, 2001.

[12] D. Hoffman, A. Girshick, K. Akeley, and M. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, no. 3, p. 33, 2008.

[13] L. Zhang and W. J. Tam, "Stereoscopic image generation based on depth images for 3D TV," *Broadcasting, IEEE Transactions on*, vol. 51, no. 2, pp. 191–199, 2005.

[14] W. Rollmann, "Zwei neue stereoskopische methoden," *Annalen der Physik*, vol. 166, pp. 186–187, 1853.

[15] T. Tuytelaars, M. Proesmans, and J. V. Gool, Luc, "The cascaded hough transform as support for grouping and finding vanishing points and lines," in *AFPAC* (G. Sommer and J. J. Koenderink, eds.), vol. 1315 of *Lecture Notes in Computer Science*, pp. 278–289, Springer, 1997.

[16] W. Zhang and J. Kosecka, "Efficient computation of vanishing points," in *In Proc. of IEEE ICRA02*, pp. 3321–3327, 2002.

[17] D. C. Lee, M. Hebert, and T. Kanade, "Geometric reasoning for single image structure recovery," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[18] M. Kalantari and F. Jung, "Precise, automatic and fast method for vanishing point detection," *The Photogrammetric Record*, vol. 24, June 2009.

[19] V. Cantoni, V. Cantoni, L. Lombardi, M. Porta, and N. Sicard, "Vanishing point detection: Representation analysis and new approaches," in *Proceedings of the 11 th International Conference on Image Analysis & Processing*, pp. 26–28, 2001.

[20] E. J. Chappero, R. A. Guerrero, and F. J. Seron, "Lineal perspective estimation on monocular images," *CACIC*, pp. 444–454, 2010.

[21] M. Juneja and P. S. Sandhu, "Performance evaluation of edge detection techniques for images in spatial domain," *International Journal of Computer Theory and Engineering*, vol. 1, pp. 614–621, Dec 2009.

[22] C. Nolan, "Inception." Movie, 2010.

[23] N. Kiryati, Y. Eldar, and A. Bruckstein, "A probabilistic hough transform," *Pattern Recognition*, vol. 24, no. 4, pp. 303–316, 1991.

[24] E. D. Montag, "Empirical formula for creating error bars for the method of paired comparison," *J. Electronic Imaging*, vol. 15, no. 1, p. 10502, 2006.

[25] S. Mendes, "Skyfall." Movie, 2012.

[26] L. Yin, R. Yang, M. Gabbouj, S. Member, and Y. Neuvo, "Circuits and systems exposition weighted median filters: A tutorial," 1996.

[27] Z. Zhang, Y. Wang, T. Jiang, and W. Gao, "Visual pertinent 2D-to-3D video conversion by multi-cue fusion.," in *ICIP* (B. Macq and P. Schelkens, eds.), pp. 909–912, IEEE, 2011.

[28] Y.-L. Chang, W.-Y. Chen, J.-Y. Chang, Y.-M. Tsai, C.-L. Lee, and L.-G. Chen, "Priority depth fusion for the 2D to 3D conversion system," pp. 680513–680513–8, 2008.

[29] S. ching Chen, M. ling Shyu, C.-C. Zhang, and R. L. Kashyap, "Video scene change detection method using unsupervised segmentation and object tracking," in *IEEE International Conference on Multimedia and Expo (ICME*, pp. 57–60, 2001.

[30] C.-L. Huang and B.-Y. Liao, "A robust scene-change detection method for video segmentation," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 11, pp. 1281–1288, Dec. 2001.

A. Project Initiation Document

2D to 3D Conversion System

Project Initiation Document

Wouter van Rooy[1,2]
Student no: 2143398

15 October 2012

Axon Digital Design BV[1], Gilze

Fontys School of Information and Communication Technology[2], Eindhoven

Internship supervisor
Luc Vosters[1]

First assessor
Ben Schreur[2]

Version
1.0 - Final

Document history

Revisions

Version	Status	Date	Changes
0.1	Draft	2 October 2012	Initial version
0.2	Draft	8 October 2012	Revised project planning, phases and exclusions after internal review
0.3	Draft	10 October 2012	Revised project planning, definition, consultation matrix and other minor issues after external review
1.0	Final	15 October 2012	Released document, no change in content

Distribution list

Version	Date	Recipient	Role
0.1	2 October 2012	Luc Vosters	Intern supervisor
0.1	2 October 2012	Rafael Peset Llopis	Product owner
0.2	8 October 2012	Luc Vosters	Intern supervisor
0.2	8 October 2012	Rafael Peset Llopis	Product owner
0.2	8 October 2012	Ben Schreur	First assessor
0.3	10 October 2012	Luc Vosters	Intern supervisor
0.3	10 October 2012	Rafael Peset Llopis	Product owner
0.3	10 October 2012	Ben Schreur	First assessor
1.0	15 October 2012	Luc Vosters	Intern supervisor
1.0	15 October 2012	Rafael Peset Llopis	Product owner
1.0	15 October 2012	Ben Schreur	First assessor

Contents

1. Introduction

1.1. Goal of this document

This document has been prepared to provide all relevant information and basic principles for the proposed project. Its goal is to define the project, serving as a basis for project management and assessment.

This Project Initiation Document covers the following fundamental aspects of the project:

- What are we trying to achieve?
- Why is it important to achieve these goals?
- Who are involved in managing the project and what are their roles and responsibilities?
- How and when will the measures described in this document be realized?

The document will be used to:

- Ensure that the project has a sound basis before the stakeholders are asked to commit to the project.
- Serve as a basis for the stakeholders and project manager to assess progress, changes and validity of the project implementation.

1.2. Structure of this document

The document has been split in two sections: a static and a dynamic part. The static part will not receive any updates after approval of the document. The dynamic part however tends to contain planning related information and will be updated if necessary.

The static part comprises of:

- Project definition (chapter 2)
- Project organisation (chapter 3)
- Project control (chapter 4)

The dynamic part contains:

- Project planning (Appendix A)

2. Project definition

2.1. Background

Producing native stereoscopic 3D video for live broadcast is still time consuming and costly. It requires broadcasters to invest in new expensive equipment like stereo camera's and stereo rigs and to hire specially trained stereographers. Real time 2D-to-3D conversion is a cheaper option since it requires additional hardware only. Unfortunately 2D-to-3D conversion is an extremely difficult task for which no optimal generally applicable solution exists.

In state-of-the-art automatic 2D-to-3D conversion both motion based and pictorial cue based methods can be distinguished. In motion based algorithms camera and object motion are used to estimate depth from motion parallax. Pictorial cue based methods use depth cues like focal-blur, perspective, texture-density, occlusion, saliency, relative height, etc. Many techniques have been proposed to estimate depth from individual cues, and an excellent overview can be found in [1].

However, the human visual system (HVS) integrates multiple depth cues rather than perceiving depth from a single cue. Even in monocular video depth is perceived which is determined from a composition of cues, where the individual cue contribution can vary from shot to shot. Therefore, a key challenge in obtaining realistic depth for 2D-to-3D conversion lies in integrating various depth cues into a single depth map.

2.2. Goals

The main goal of this project is to enhance the Axon 2D-to-3D conversion system and thereby improve the viewer's 3D experience. Currently, an efficient method for 2D-to-3D conversion of low-depth-of-field video has been developed based on the focal blur cue. Unfortunately, this approach fails for other types of video.

To get a more reliable 2D-to-3D conversion at least an additional depth estimation algorithm needs to be included. Linear perspective seems a good choice, as it is a strong cue to the HVS[2] and it is available in many still images and video sequences.

Due to the complexity of the proposed system, extensive research is needed. The central question that needs to be answered during this research is:

How can an automatic 2D-to-3D conversion system be improved by combining multiple depth cues to let users experience a significantly better depth perception?

2.3. Approach

2.3.1. Phase 1: Depth from perspective

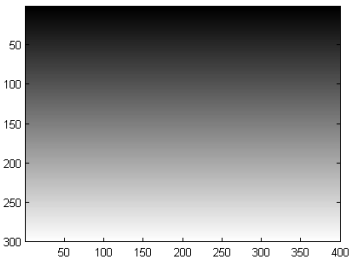
Phase 1 will comprise of the research and development of an additional depth estimation algorithm based on linear perspective cues in a still image. An example of such a depth map can be found in Fig. 2.1b.

Currently, perspective is simulated by applying a general gravity depth map (Fig. 2.1a), ranging from the bottom of the image (closeby) to the top of the image (farther away). Determining whether the proposed linear perspective map improves the 3D experience compared to the simple gravity map will be an important step in this phase.

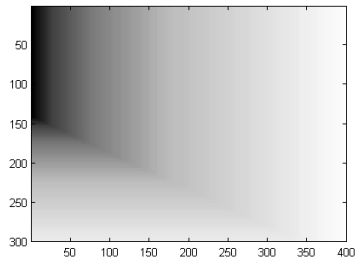
If the viewer perception tests point out that there is no significant improvement from the gravity map to the linear perspective map, a depth estimation algorithm based on an alternative depth cue will be developed. If there is sufficient improvement the linear perspective algorithm will be further optimized for use with video sequences.

This phase requires the following questions to be answered:

- How can a computer vision algorithm automatically generate a depth map from linear perspective cues in a still image?
- What is the difference in depth perception between a simple gravity depth map and an enhanced map based on linear perspective?
- How can the depth-from-perspective algorithm be optimized to detect temporally consistent vanishing points in video sequences?



(a) Gravity



(b) Linear perspective with a single vanishing point to the left of the image[3]

Figure 2.1.: Depth map examples

2.3.2. Phase 2: Depth map fusion

After implementing an additional depth estimation algorithm work will be started on the topic of depth map fusion. At first testing will only take place using still images. During phase 3 support for video sequences will be added (sec. 2.3.3).

This phase requires the following questions to be answered:

- Is hard switching between depth cues sufficient or would a mixture/weighted average of depth cues be more appropriate?
- Would spatially variant mixture weights improve the overall depth perception?

2.3.3. Phase 3: Scene change detection

Applying depth estimation algorithms to video sequences poses several new challenges: the estimated depth of an object would fluctuate across time as a result of algorithm inaccuracies. Previous research at Axon has shown that this disturbs the viewer's depth perception significantly. In order to mainain temporal consistency for generated depth maps a temporal depth filtering algorithm has been developed.

With the added functionality of the depth map fusion module comes an additional challenge: it is not desired to constantly change the (mixture of) depth cue(s) within the same scene. A scene change detection algorithm would help in solving this problem. Flagging scene changes allows the conversion system to reset its temporal depth filters and switch to a (mixture of) depth cue(s) that is more suitable for the new scene's content. Furthermore, individual depth estimation algorithms may benefit from these flags for their internal filtering.

This phase requires the following questions to be answered:

- How can scene changes in a video sequence be identified automatically?
- What can be done to allow individual depth estimation algorithms to benefit from scene change detection?

2.4. Deliverables

The following products and documents must be delivered:

- Project Initiation Document (PID)
- Depth from perspective module[1]
- Depth map fusion module
- Scene change detection module

[1] A different depth estimation algorithm may be used as described in sec. 2.3.1

- Gradutation thesis
- Graduation presentation

The three modules will each comprise of the following sub-products:

- Research & design document
- C++/OpenCV implementation
- Viewer perception tests & results

2.5. Exclusions

The proposed algorithms will be implemented as a prototype on a general purpose PC workstation using C++ and the OpenCV library. Implementation on other hardware, such as an FPGA, is outside the scope of this project. The algorithms will, wherever possible, be implemented with real-time performance in mind.

2.6. Limitations

Available working time	Available resources	Delivery date	Turnaround
736 hours	Workstation	18 January 2013	19 weeks
	Technical support		
	3DTV testing environment		

Table 2.1.: Project limitations

2.7. Preconditions

- A 2D-plus-depth (2D+D) to stereo 3D renderer suitable for the 3DTV test environment is available
- An implementation of a depth estimation algorithm based on focal blur is available

2.8. Risks

Risk	Probability (1-3)	Impact (1-3)	Score	Mitigation
Lack of knowledge/expertise	2	2	4	Sufficient reading on the subject
Graduation thesis deadline not met	2	3	6	Start early, provide early draft
Data loss / computer crash	1	3	3	Online version control system
Difficulty finding sufficient test participants	1	1	1	Early invitations (RSVP)
Conflicting educational/commercial interests	1	2	2	Commit to agreements in PID

Table 2.2.: Qualitative risk register

3. Project organisation

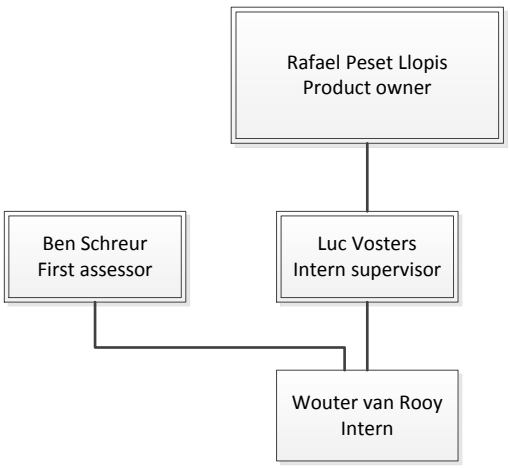


Figure 3.1.: Organisation chart

3.1. Product owner

Rafael Peset Llopis is the product owner and has commisioned to execute the project. He represents the requesting party and is responsible for transferring wishes and requirements of the deliverables.

3.2. Intern supervisor

Luc Vosters serves as intern supervisor in this project. He is the primary contact for the intern and may be requested to advise on technical subjects. For progress updates and non-critical issues he also takes the role of delegate product owner.

3.3. First assessor

The role of first assessor is fulfilled by Ben Schreur. He monitors the project from an educational point of view and may be requested to advise on process related subjects.

3.4. Intern

Wouter van Rooy has been contracted as an intern and will execute the project.

4. Project control

4.1. Reporting

(a) Reports

Report	Product owner	Intern supervisor	First assessor	Intern
PID	Ap	Ap	Ap Ad	C Ap D
Research & design documents	N	Ap Ad	N	C Ap D
Viewer perception tests & results	N	Ap Ad	N	C Ap D
Implementations	N	Ap Ad	N	C Ap D
Graduation thesis & presentation	N	N Ad	Ap Ad	C Ap D

(b) Legend

C	Ad	N	Ap	D
Create	Advise	Be notified	Approve	Distribute

Table 4.1.: Reporting matrix

4.2. Progress monitoring

(a) Consultations

Consultation	Attendees	Frequency	Time	Remarks
Progress update	A I	Fortnightly	Fridays around 16:00	By e-mail
Progress update	PO IS I	Monthly	To be determined	-
Company visit	PO IS A I	-	To be determined	-
Project meeting	IS I	Weekly	Mondays around 10:00	-

(b) Legend

PO	IS	A	I
Product owner	Intern supervisor	First assessor	Intern

Table 4.2.: Consultation matrix

4.3. Issue management

If problems arise during the execution of the project, an extra meeting will be arranged with the intern supervisor and, if necessary, the product owner. Together an adequate countermeasure will be planned to solve the problem.

When the stakeholders remain in conflict after executing the procedure above, the first assessor may be consulted as an additional resort.

4.4. Deviation and escalation procedure

In case of product deviations an extra meeting will be arranged with the intern supervisor and product owner. In this meeting the deviation will either be approved or rejected. Rejection might cause the projects planning or the products functionality to be altered if no suitable solution can be found.

A. Project planning

Delivery date	Product
3 October 2012	Project Initiation Document (draft)
10 October 2012	Project Inititation Document (final)
19 October 2012	Depth from perspective: Research & design document
2 November 2012	Depth from perspective: C++/OpenCV implementation
2 November 2012	Depth from perspective: Viewer perception tests & results
16 November 2012	Depth map fusion: Research & design document
30 November 2012	Depth map fusion: C++/OpenCV implementation
30 November 2012	Depth map fusion: Viewer perception tests & results
14 December 2012	Scene change detection: Research & design document
21 December 2012	Graduation thesis (draft)
4 January 2013	Scene change detection: C++/OpenCV implementation
4 January 2013	Scene change detection: Viewer perception tests & results
10 January 2013	Graduation thesis (final)
15 January 2013	Graduation presentation (draft)
18 January 2013	Graduation presentation (final)

Table A.1.: Global product planning

NB: The contract between the intern and Axon Digital Design BV expires in mid-February 2013. The last few weeks are reserved for project transfer.

Bibliography

[1] L. Zhang, C. Vazquez, and S. Knorr, "3dtv content creation: Automatic 2d-to-3d video conversion," *IEEE Transactions on Broadcasting*, vol. 57, pp. 372–383, June 2011.

[2] J. A. Saunders and B. T. Backus, "The accuracy and reliability of perceived depth from linear perspective as a function of image size," *Journal of Vision*, vol. 6, pp. 933–954, 2006.

[3] S. Battiato, A. Capra, S. Curti, and M. L. Cascia, "3d stereoscopic image pairs by depth-map generation," in *Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium*, 3DPVT '04, (Washington, DC, USA), pp. 124–131, IEEE Computer Society, 2004.

Nomenclature

2D+D	2D plus depth
FPGA	Field Programmable Gate Array
HVS	Human Visual System
PID	Project Initiation Document