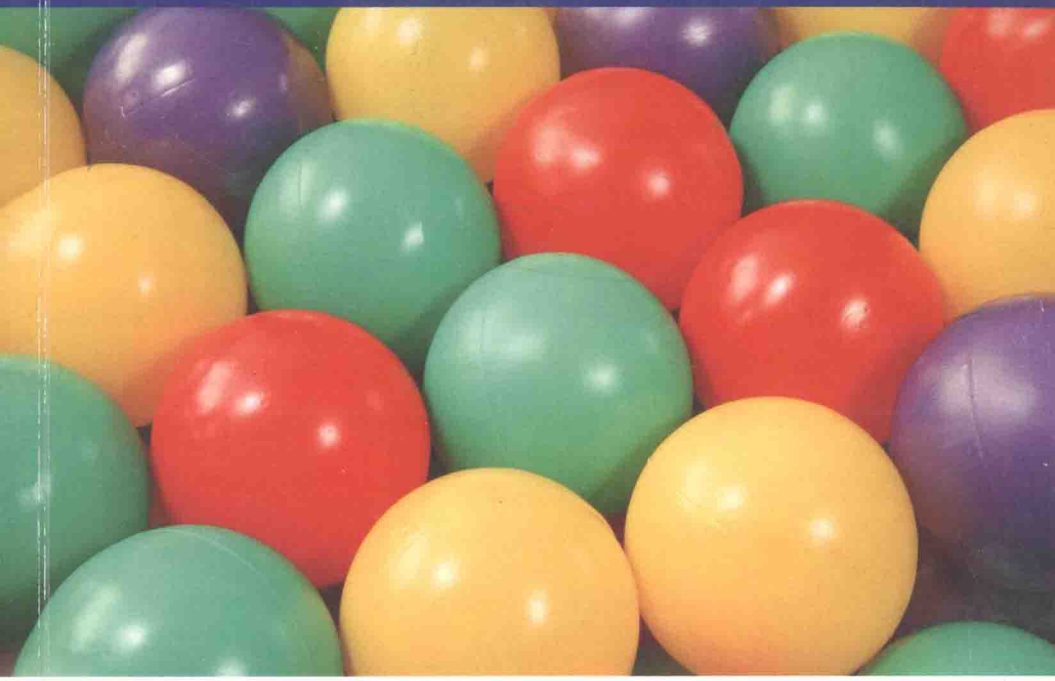

Kai Hugo Hustoft Endresen

Tracking objects in 3D using Stereo Vision

a real-time approach based on color segmentation for use on a mobile robot

LAP LAMBERT Academic Publishing

Kai Hugo Hustoft Endresen

Tracking objects in 3D using Stereo Vision

a real-time approach based on color segmentation for use on a mobile robot

LAP LAMBERT Academic Publishing

Impressum/Imprint (nur für Deutschland/ only for Germany)
Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Coverbild: www.ingimage.com

Verlag: LAP LAMBERT Academic Publishing AG & Co. KG
Dudweiler Landstr. 99, 66123 Saarbrücken, Deutschland
Telefon +49 681 3720-310, Telefax +49 681 3720-3109
Email: info@lap-publishing.com

Herstellung in Deutschland:
Schaltungsdienst Lange o.H.G., Berlin
Books on Demand GmbH, Norderstedt
Reha GmbH, Saarbrücken
Amazon Distribution GmbH, Leipzig
ISBN: 978-3-8433-5332-8

Imprint (only for USA, GB)
Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Cover image: www.ingimage.com

Publisher: LAP LAMBERT Academic Publishing AG & Co. KG
Dudweiler Landstr. 99, 66123 Saarbrücken, Germany
Phone +49 681 3720-310, Fax +49 681 3720-3109
Email: info@lap-publishing.com

Printed in the U.S.A.
Printed in the U.K. by (see last page)
ISBN: 978-3-8433-5332-8

Abstract

This report describes a stereo vision system to be used on a mobile robot. The system is able to triangulate the positions of cylindrical and spherical objects in a 3D environment. Triangulation is done in real-time by matching regions in two images, and calculating the disparities between them.

Preface

This master thesis was during the spring semester of 2010 at the Department of Computer and Information Science, Norwegian University of Science and Technology, and it is a further development of work done during my pre-study project of the fall of 2009.

The purpose of this work was to gain an understanding of stereo vision, as well as create a basis for a real-time working system for tracking objects in 3D. This work was designed to be part of a mobile robot in collaboration with people of different disciplines for competing in Eurobot 2010.

The project work was supervised by associate professor Ketil Bø, and the assignment description was as follows:

> *The project involves the design and implementation of a mobile system to track moving and stationary objects in 3D using stereo vision. The processed data should be able to give reasonably accurate coordinates and distances to objects for further processing by a planning system.*

I would like to thank my supervisor, associate professor Ketil Bø, for allowing me to pursue this assignment and reminding me to focus on my writing.I would also like to thank Kongsberg Gruppen ASA for providing the funding for the hardware to make this assignment possible, as well as the Department of Cybernetics at NTNU for the use of their equiptment and staff for mechanical work.

Trondheim,
Kai Hugo Hustoft Endresen

Contents

Chapter 1

Introduction

In this project I hope to create a real-time stereo vision system to track various objects. The reason for using stereo vision is primarily because I believe that it would be the best way to pinpoint the location of an object in a 3D environment. Since the objects are known to me beforehand (shape, size and color), a single camera could have been used, but the accuracy for pinpointing location would be more accurate with stereo vision. The results should be accurate enough to be used for a planner system, the task of which is to decide the optimal way to pick up the maximum amount of balls and cylinders. The finished system is to be used as part of a robot that will compete in Eurobot 2010.

Eurobot 2010 is an annual international robotics competition, in which teams from many different countries compete. The general premise is that two robots should compete to get the most points within a 90 second interval. This competition is well suited for computer vision approaches, stereo vision in particular. The tasks vary quite a bit from year to year, and even though some reuse of earlier solutions is possible, a lot of equiptment and software has to be made anew. So even though some teams have been competing for a decade, their advantage over more recent teams isn't that pronounced.

The motivation for this project is mainly to create a working stereo vision system. Especially measuring distances to various objects, and also learning more about approaches to stereo vision, and depth imaging in general.

Previous work in the area of stereo vision has been mostly restricted to either stationary use, or very slow movement. With the introduction of increasingly faster computing devices, more and more of the things that required considerable processing time before, can now be done in real-time.

An example of a rather slow, but functional system for navigation with stereo vision, is the Stanford cart[21]. The Stanford cart came to be originally from the need of controlling a robotic unit on the moon, which is too far away for being remote controlled from earth without any degree of autonomous behavior, and was one of the first stereo vision systems. The stereo system was a camera that moved on a rail to get several pictures of the same scene and then calculate the distance to obstacles. It was very slow and needed 10 to 15 minutes of image processing time between each move.

Stereo Vision as a phenomenon was first described by Charles Wheatstone in 1838. He did various psychological and optical experiments to try to understand how humans percieve depth. He also mentioned that Leonardo da Vinci, several hundred years before, had noticed that it was not possible to draw something on a canvas and get the same realism as in the real world.

Stereo vision systems essentially try to emulate how the human vision system works, by modelling a scene using two cameras.

The most common approach is to have two cameras parallel to each other, with an horizontal offset. Other methods include using vertically aligned cameras, or cameras tilted towards one another. All stereo vision then comes down to is finding matching points in both views, and measuring the disparity between them. This can then be used for tracking objects[14], finding their position in 3D[19] as well as robot navigation[6]. It has even been used by NASA in their STEREO project for solar observations[22]. In Norway, FINN AS, which provides 3D maps of Trondheim, Oslo and other cities in Norway, uses stereo vision technology from C3 Technologies to do stereo reconstruction of landscape and houses from aerial pictures[29].

During the writing of this thesis, a surge in use of 3D imaging has been noticed for motion pictures, TVs and gaming devices such as the Nintendo 3DS. One movie of note is *Avatar*, in which the entire movie was shot by special 3D cameras. The same system is currently being designed in a more mobile format for the next Mars Rover. The system is delivered by Malin Space Science Systems, and optically it is very impressive. For instance it is possible to adjust the focal length while still maintaining stereo calibration. Thus, one is able to increase accuracy for objects far away, while still having the possibility of a large viewing area.[28]

The remainder of this report is divided into the following chapters:

(2) Theory describes most of the theoretical background used for the experiments and solution, and some theory concerning lenses and depth imaging.

(3) Environment presents the environment in which the system is meant to operate.

(4) System Design describes the system requirements, a brief overview of the steps involved in the programming of the system, as well as an overview of the software and hardware used.

(5) Experiments describes the results of testing various approaches to different parts of the stereo vision system.

(6) Solution shows how the final implementation works under various conditions.

(7) Discussion provides an evaluation of the working solution, as well as some notes about the performance of the system.

(8) Conclusion summarizes the report, and makes some statements with regards to future work.

Chapter 2

Theory

This chapter provides all the relevant background material to the stereo vision system presented later in this report. Subjects to be looked at include depth images, color conversion, edge detection, camera calibration, mathematical morphology, color segmentation, epipolar geometry and bayer filtering.

2.1 Stereo Vision

Stereo vision is denoted as *processes directed on understanding or analysing three-dimensional visible object surfaces based on image data.* The visual systems of humans and animals prove that stereo vision works in complex enviroments. Stereo vision is a very active field of research in computer vision.

An overview of steps can be seen to the right, in figure 2.1.

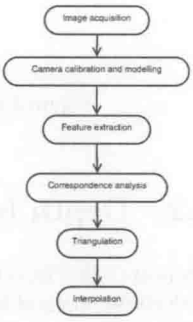

Figure 2.1: Pipeline

2.1.1 View area in stereo vision

The view area in a stereo vision system is essentially limited by the view angle of the individual cameras, and the relative orientation, as well as the distance between the cameras. See figure 2.2 for illustrations.

The further apart the cameras are, the further away the minimum distance for successful stereo vision becomes. The advantage of increasing the distance

between cameras is that the disparity increases linearly with the distance between cameras. This means that by doubling the distance between cameras, the depth resolution also doubles at a given distance. The disadvantage is that it is then impossible to measure the distance to objects very close, and the imaging system would need other methods to remedy this. This issue is also present in the human vision system, where we cannot distinguish how far an object very close to our eyes is without other visual cues.

As can be seen in figure 2.2, this issue can be reduced by mounting the cameras in such a way that the epipoles[1] are no longer parallel to each other. Instead of at infinity they should meet at a fixed point in front of the cameras, but then the geometric calculations and disparities would be much harder to calculate.

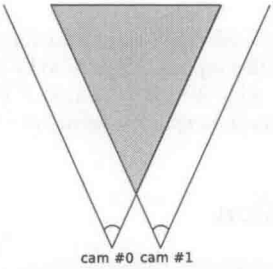

Figure 2.2: The green triangle shows the view area

2.2 Depth images

There are many ways to achieve depth images. The image formation process projects an image of the 3D world onto a 2D surface. Reversing the process is impossible as some of the information is invariably lost in the projection. Experiments with our own eyes show that the human visual system is capable of recovering a lot of this information from stereo vision; contours, texture, shadows and so on.

Using a single image, depth information can be extrapolated from shading, textures, contours, focus/defocus and various visual cues.

Shape from shading exploits the fact that if you are able to control the

[1]Epipoles are the center of the view area of the cameras.

lighting enviroment you are in, you can look at the reflections from surfaces, and therefore extrapolate the distance.

The way textures are transformed when they are projected from 3D to 2D can be used to infer depth. For instance, a wallpaper with horizontal stripes have an equal distance between each stripe when viewed at a perpendicular angle. When viewed from the side, the distance between stripes far away are greater than stripes up close.

Providing you know the physical size of an object, contour shapes can be used for calculating the distance to an object in an image. For example a perfectly spherical object can be extracted from the image by using the Hough transform. The radius of the matched circle can be measured, and you can then compare this to the physical radius of the object and find the distance.

Shape from focus/defocus exploits the fact that by adjusting the focus of a lens in consequtive steps, so that various depths of the image come into focus, and you can then calculate the distance. This might be particulary useful if the environment is static, and there are no time constraints. It is commonly used in cameras to give an approximation of the distance to an object in focus.

Other approaches tend to require more than one image, or have special hardware requirements.

Stereo vision uses two different viewpoints provided by two different cameras. The offset between them, in the horizontal or vertical direction, can be used to extract depth information. This is done by finding the pixels in both images that correspond to the same point in the 3D world, and then checking the disparity between them. This can then be used to calculate the distance.

Stereo Vision can be aided by using structured light to aid stereo vision in its task to find correspondences. This is done by for example creating a grid with lasers, and using the way in which the grid is deformed when it hits an object to model a 3D environment.

2.3 Image acquisition

For static stereo vision with moving objects, as well as dynamic stereo vision with static objects, it is essential that both cameras capture images at the same time. Otherwise, the epipolar lines might not line up, and calculations of disparities will be wrong due to movement. This means that there needs

to be provisions for getting the image acquisition done in a synchronized manner. Using regular consumer USB cameras, this can be very difficult. The cameras typically only have provisions for starting a capture, and stopping a capture. There is typically not possible to capture single frames, feeding it a clock signal or enabling any internal synchronization routines.

Various firewire cameras, following the IIDC standard allow for feeding the cameras a clock signal, which can be used to capture frames in a synchronized fashion. Some professional cameras, such as used for the system presented in this report, support auto synchronization.

The best way to synchronize common consumer USB cameras, is to make sure that the cameras are on their own USB buses. This way requests and frames can be recieved in a concurrent fashion. The synchronization will still often drift slightly with varying CPU load, and since there is typically very little relation between when a frame is captured and when it reaches the system buffer, trying to make both cameras capture frames at the exact same time might actually make things worse.

2.3.1 Color conversion

Most digital cameras used these days use a Bayer filter infront of the CCD sensor. The filter is a mosaic infront of the CCD sensor, and only lets through the relevant wavelengths for each pixel. The filter pattern is 50% green, 25% red and 25% blue. This means that for each pixel, only one of the relevant colors will be in the same pixel position, the rest is taken from neighbouring pixels. The side effects of this is that for a given pixel in the image, the colors are often interpolated from neighbouring pixels.

There are numerous different demosaicing algorithms available, with different algorithms being suitable for different tasks. The algorithms have been split into two distinct groups. Fast algorithms, and slow algorithms.

Slow algorithms are typically very good at reproducing the colors accurately, and various implementations such as VNG, AHD and PG are available. They are most used to import raw images from SLR cameras, in which visual quality has much higher priority than speed.

The fast algorithms, which are the most relevant for dynamic stereo vision, tend to either focus on visual quality, or precision. The former kind such as nearest-neighbour or bilinear demosaicing causes the image to become somewhat blurred, while linear algorithms that focus on precision causes the image to become grainy.