University of Florence
School of Engineering

Master degree program in
Computer Engineering

Disparity coherent stereo video watermarking

Master Thesis of

Benedetta Barbetti, Michaela Servi

December 2015

Supervisor:

Prof. Alessandro Piva

Advisors:

Prof. Carlo Colombo
Dott. Pasquale Ferrara

Academic Year 2014/2015

Abstract

Contents

List of Figures

List of Tables

Introduction

In the last few years the stereoscopic technique has become a great part of image and video processing.
In medical diagnosis and endoscopic surgery as in fault detection in manufactory industry, army and arts, multiview imaging is considered as a key enabler for professional added value services.
Nowdays stereoscopic techniques are also used in people tracking and mobile robotics navigation for economic reasons and to improve performances.
Finally the worldwide success of 3D movie releases and 3D video games and the deployment of 3D televisions made the nonprofessional user aware about a new type of multimedia entertainment experience.
The increasing production and distribution of these contents leads to the concerns over copyright protection.
Digital watermarking can be considered as the most flexible property right protection technology, since it adds some information (a mark, i.e. copyright information) in the original content without altering its visual quality so that such a marked content can be further distributed/consumed by another user without any restriction; still, the legitimate/illegitimate usage can be determined at any moment by detecting the mark. In same case the watermarking protection mechanism, instead of restricting the media copy/distribution/consumption, provides means for tracking the source of

the content illegitimate usage.

The purpose of this thesis is to provide a new watermarking system for copyright protection of stereoscopic videos. The method operates in the frequency and in the spatial domain by embedding a pseudo-random sequence of real numbers in a selected set of DFT coefficients of the left image; then the reference watermark is distorted according to the depth information prior to insertion and spatially added to the right image.

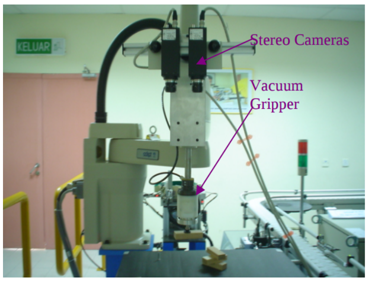
In Chapter **??**...

Chapter 1

Stereoscopic Video

In a wide variety of image processing applications, explicit depth information is required in addition to general image informations, such as intensities, color, densities.
Examples of such applications are found in 3D vision (robot vision, photogrammetry, remote sensing systems), in medical imaging (computer tomography, magnetic resonance imaging, microsurgery), in remote handling of objects (random bin picking), in space exploration (mobile robotics navigation) or 3D movies and videogames.
In each of these cases, depth information is essential for accurate image analysis or for enhancing the realism.
In remote sensing the terrain's elevation needs to be accurately determined for map production, in remote handling an operator needs to have precise knowledge of the threedimensional organization of the area to avoid collisions and misplacements.

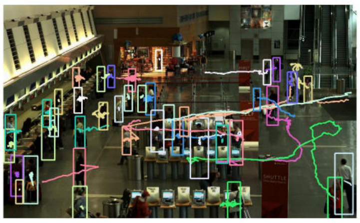
Depth in real world scenes can be explicitly measured by a number of range sensing devices such as by laser range sensors, by structured light or



(a) In bin picking applications stereo vision helps to reconstruct the 3D environment and detect the part of the object to be robotically picked

(b) Surgical robot *Da vinci* is provided with a stereoscopic camera that allows a tridimensional view of the operative filed.

Figure 1.1: Stereoscopy in medical and industrial field



(a) In people tracking application stereo vision improves segmentation thanks to depth information and it's less sensible to light changes.

(b) In mobile robotics navigation stereo vision has became the first choice technology because it provids a lot of quality data for low costs.

Figure 1.2: Stereoscopy application's fields

by ultrasound. However it's usually undesirable to have separate systems for acquiring the intensity and the depth information because of the relative low resolution of the range sensing devices and because it's not an easy task to fuse information from different type of sensors; for these reasons and for a non-negligible economic factor stereoscopic vision has becoming the technology of choice in these type of applications.

(a) Stereo video frames, left and right.

(b) Overlap of the two frames.

(c) 3D view with specific glasses

(d) Polarized glasses for 3D view

Figure 1.3: Stereoscopy in 3D video games

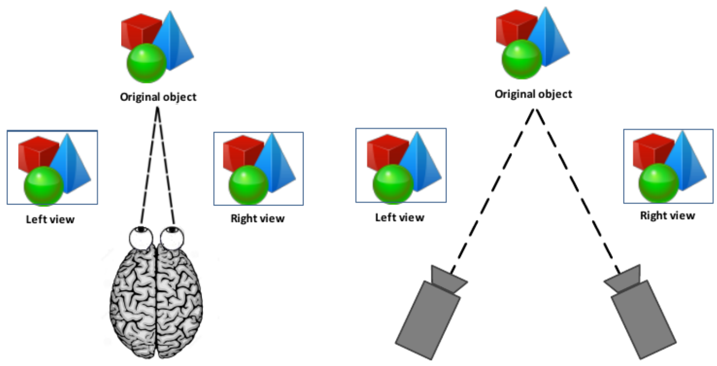
1.1 Stereo vision

In image processing stereo vision is the process of extracting 3D information from multiple 2D views of a scene.

The 3D information can be obtained from a pair of images, also known as a stereo pair, by estimating the relative depth of points in the scene.

From the anatomic point of view, the human brain calculates the depth in a visual scene mainly by processing the information brought by the images seen by the left and the right eyes. These left and right images are slightly different because the eyes have biologically different emplacements.

Consequently, the straightforward way of achieving stereoscopic digital imaging is to emulate the Human Visual System (HSV) by setting-up (under controlled geometric positions), two traditional 2D cameras.



(a) Binocular human visual system (b) Stereoscopic system

Figure 1.4: Binocular human vision vs. stereoscopic content acquisition.

1.1.1 Acquisition of stereoscopic images

In order to be able to perceive depth using recorded images, a stereoscopic camera is required, which consists of two cameras that capture two different, horizontally shifted perspective viewpoints; with two (or more) cameras we can infer depth, by means of triangulation, if we are able to find corresponding points in the two images (Figure).

The camera setup should be geometrically calibrated such that the two

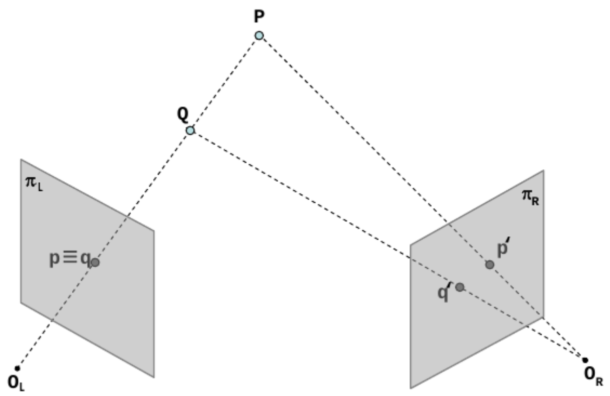


Figure 1.5: Triangulation: with two cameras the depth of

cameras capture the same part of the real world scene.

Calibration of a stereo camera system involves the estimation of the intrinsic and extrinsic parameters of the model: intrinsic parameters embody the characteristics of the optical system and its geometric relationship with the image sensor, extrinsic parameters relate the location and orientation of the second camera with respect to the first one in the 3D space (Figure).

These parameters can be used to rectify a stereo pair of images to make

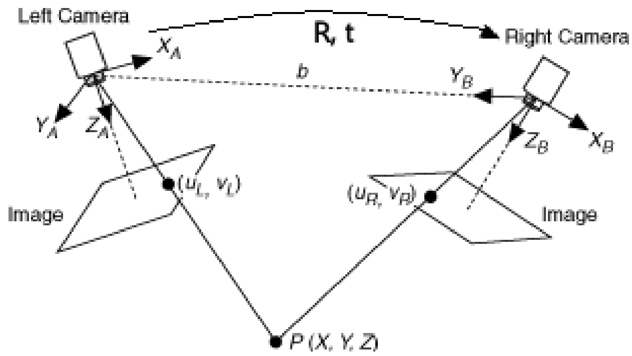


Figure 1.6: Stereo camera model

them appear as the two image planes are parallel (Figure); once the images are rectified, epipolar geometry it's used to find corresponding points and compute the disparity map.

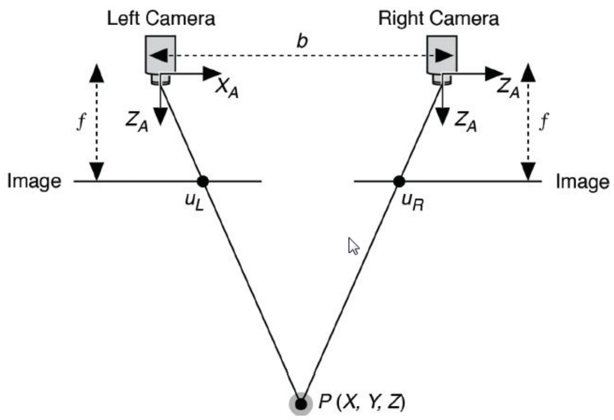


Figure 1.7: Rectified stereo cameras

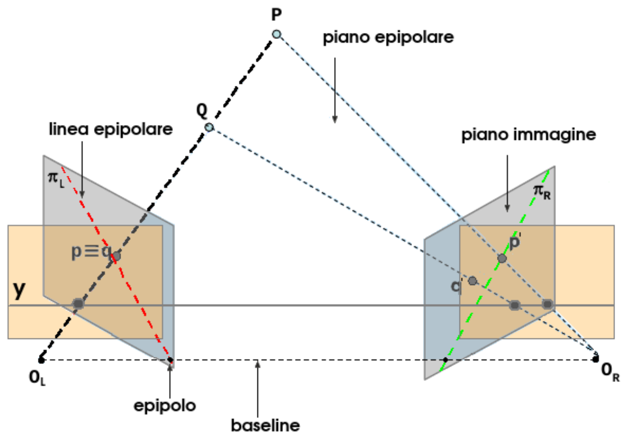


Figure 1.8: Rectified images: corresponding points (p, p'), projection of the same 3D point (P) are constrained on the same image horizontal line, the epipolar line

1.1.2 Disparity map computation

With the stereo rig in standard form and by considering similar triangles in Figure XX (PO_LO_R and Ppp'):

$$\frac{b}{Z} = \frac{(b + x_L) - x_R}{Z - f}$$

so

$$Z = \frac{b \cdot f}{x_L - x_R} = \frac{b \cdot f}{d}$$

where $d = x_L - x_R$ it's called *disparity*.

Disparity is, therefore, the difference between the x coordinates of two corre-

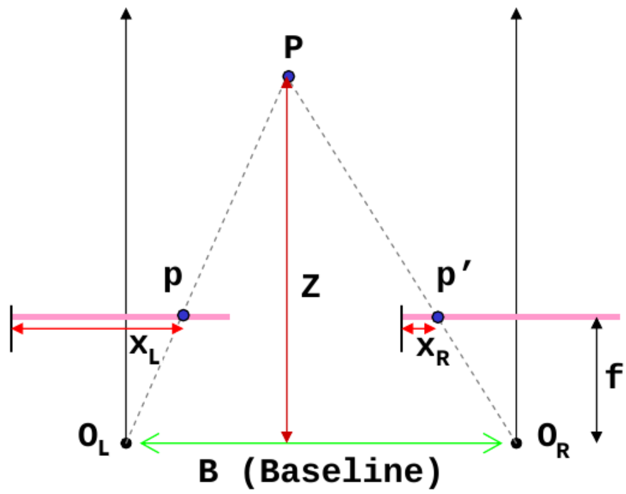


Figure 1.9: Geometry of standard form

sponding points and it is usually encoded with greyscale image (Figure XX), where points closer to the cameras are brighter and correspond to a higher disparity.

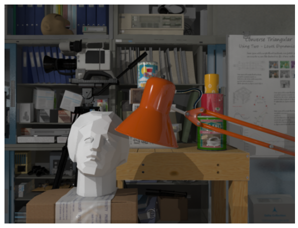
In order to compute the disparity map is necessary to find corresponding points; stereo correspondance is though a challenging task that has to manage with perspective distortions, uniform and ambiguous regions, repetitive patterns, occlusions and discontinuities(Figure XX).

In general, stereo matching algorithms can be categorized into two major classes:

- local methods
- global methods.

Local stereo algorithms estimate the correspondence using a local support region or a window. Local algorithms generally rely on an approximation

(a) Left image



(b) Right image

(c) Disparity map

Figure 1.10: Stereo pair and disparity map

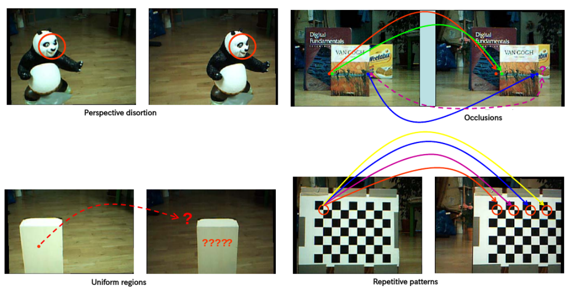


Figure 1.11: Stereo matching general problems

of the smoothness constraint assuming that all pixels within the matching region have the same disparity. However, this assumption is not valid for highly curved surfaces or around disparity discontinuities.

A naive approach consists of comparing each pixel or window in the left image with every pixel or window on the same epipolar line in right image and picking position with minimum match cost (e.g., SSD, SAD, normalized correlation).

Global stereo methods consider stereo matching as a labeling problem where

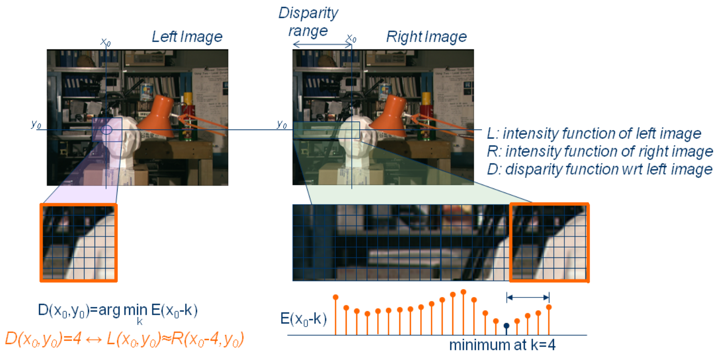


Figure 1.12: Local stereo matching, window based

the pixels of the reference image are nodes and the estimated disparities are labels. An energy functional embeds the matching assumptions by its data, smoothness, and occlusion terms and propagates them along the scan line or through the whole image. The labeling problem is solved by energy functional minimization, using dynamic programming, graph cuts, or belief propagation.
Even if this class of algorithms is significantly slow, the results, especially when textures and discontinuities are present, are much accurate.

In this thesis the Kolmogorov and Zabih's Graph Cuts Stereo Matching Algorithm has been used, because there were no time constraints requirements and the quality of the computed disparities has been considered satisfying regard to the ground trouth.

(a) Left image

(b) Graph cuts' disparity map

(c) Ground truth disparity map

Figure 1.13: Results of the Kolmogorov and Zabih's graph cuts algorithm on the Tsukuba pair

1.2 3D capturing devices

For stereoscopic shooting, two synchronized cameras must be used. The distance between the center of the lenses of the two cameras is called the interaxial, and the cameras' convergence, is called the angulation. These two parameters can be modified according to the expected content peculiarities. The two cameras must be correctly aligned, identically calibrated (i.e. brightness, color, etc...) and perfectly synchronized (frame-rate and scan-wise).

To hold and align the cameras, a stereo-rig is used; the rigs can be of two main types:

- the side-by-side rig, where the cameras are placed side by side (Figure XX). This kind of 3D-rig is mostly useful for large landscape shots since it allows large interaxials; however, it doesn't allow small interaxials

Figure 1.14: Interaxial separation between lenses

because of the physical size of the cameras;

- the beamsplitter rig (Figure XX), where one camera films through a semi-transparent mirror, and the other films the reflection in the mirror. These rigs allow small and medium interaxials, useful for most shots, but not the very large interaxials (because the equipment would be too large and heavy).

  Monoblock cameras have been designed as well, where the two cameras are presented in a fixed block and are perfectly aligned, which avoids cameras desynchronization (Figure XX).

A second category of 3D shooting devices is presented in Figure XX. These electronic devices are less expensive and are targeting the user-created stereoscopic picture/movie distribution.

(a) Syde-by-syde rig

(b) Beamsplitter rig

(c) Monoblock camera

Figure 1.15: Professional technologies for 3D TV

(a) Mobile phones

(b) Digital cameras

(c) Camcorders

Figure 1.16: Digital personal stereo vision systems

1.3 3D video displays

Chapter 2

Stereo video watermarking

2.1 Watermaking

blablablablablablablablablablablalblalbalb

2.2 State of the art

Chapter 3

Conclusions

Bibliography

[1] Richard Hartley and Andrew Zisserman. Multiple View Geometry in Computer Vision, Cambridge University Press, 2000.

[2] Lens Blur in the new Google Camera app
`http://googleresearch.blogspot.it/2014/04/lens-blur-in-new-google-camera-app.html`