

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
Dott. Francesca Uccheddu

Academic Year 2014/2015

Abstract

Contents

| | |
|--|-----------|
| Introduction | 1 |
| 1 Stereoscopic Video | 3 |
| 1.1 Stereo vision | 5 |
| 1.1.1 Acquisition of stereoscopic images | 6 |
| 1.1.2 Disparity map computation | 8 |
| 1.2 3D capturing devices | 12 |
| 1.3 3D video displays | 15 |
| 2 Stereo video watermarking | 17 |
| 2.1 Watermaking | 17 |
| 2.1.1 Properties | 18 |
| 2.1.2 Embedding techniques | 18 |
| 2.1.3 Embedding domains | 19 |
| 2.2 Stereoscopic video watermarking | 20 |
| 2.2.1 State of the art | 20 |
| 3 Spatial disparity-coherent watermarking | 22 |
| 4 Conclusions | 24 |
| Bibliografia | 25 |

List of Figures

| | | |
|------|--|----|
| 1.1 | Stereoscopy in medical and industrial field | 4 |
| 1.2 | Stereoscopy application's fields | 4 |
| 1.3 | Stereoscopy in 3D video games | 5 |
| 1.4 | Binocular human vision vs. stereoscopic content acquisition. | 6 |
| 1.5 | Triangulation: with two cameras the depth of | 6 |
| 1.6 | Stereo camera model | 7 |
| 1.7 | Rectified stereo cameras | 8 |
| 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 | 8 |
| 1.9 | Geometry of standard form | 9 |
| 1.10 | Stereo pair and disparity map | 10 |
| 1.11 | Stereo matching general problems | 10 |
| 1.12 | Local stereo matching, window based | 11 |
| 1.13 | Results of the Kolmogorov and Zabih's graph cuts algorithm on the Tsukuba pair | 12 |
| 1.14 | Interaxial separation between lenses | 12 |
| 1.15 | Professional technologies for 3D TV | 13 |
| 1.16 | Digital personal stereo vision systems | 14 |
| 1.17 | Industrial and robotic stereo cameras | 14 |

| | | |
|------|---|----|
| 1.18 | Passive and active glasses for 3D viewer technologies | 16 |
| 2.1 | Watermarking workflow | 17 |
| 2.2 | Watermark properties trade-off | 18 |
| 2.3 | Spread spectrum technique | 19 |
| 2.4 | Spatial domain watermark insertion | 19 |
| 2.5 | Frequency domain watermark insertion | 20 |
| 2.6 | Hybrid technique | 20 |

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.

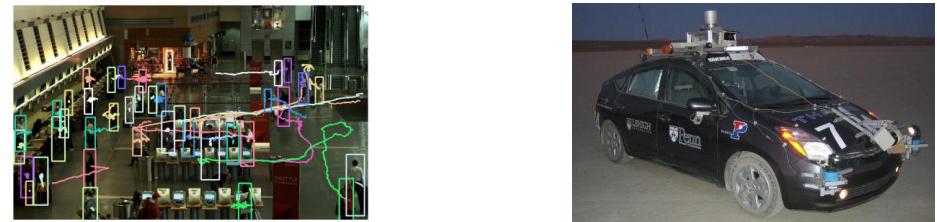
Depth in real world scenes can be explicitly measured by a number of range sensing devices such as by laser range sensors, by structured light or



(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 field.

Figure 1.1: Stereoscopic vision 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 provides a lot of quality data for low costs.

Figure 1.2: Stereoscopic vision 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.

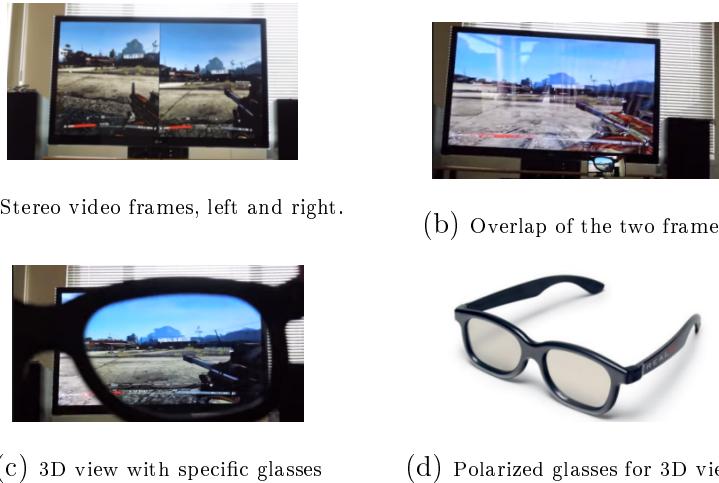


Figure 1.3: Stereoscopy in 3D video games

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.

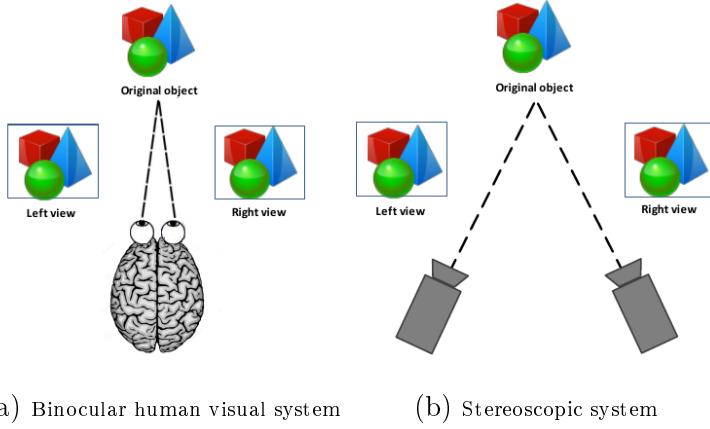


Figure 1.4: Binocular human vision vs. stereoscopic content acquisition.

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

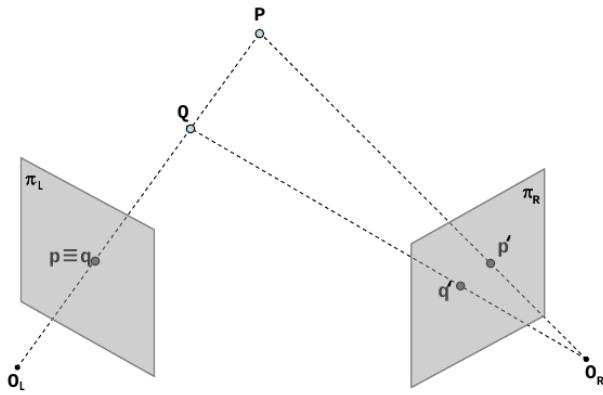


Figure 1.5: Triangulation: with two cameras the depth of

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

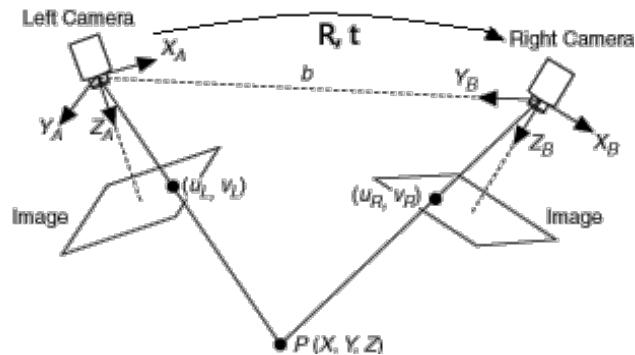


Figure 1.6: Stereo camera model

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.

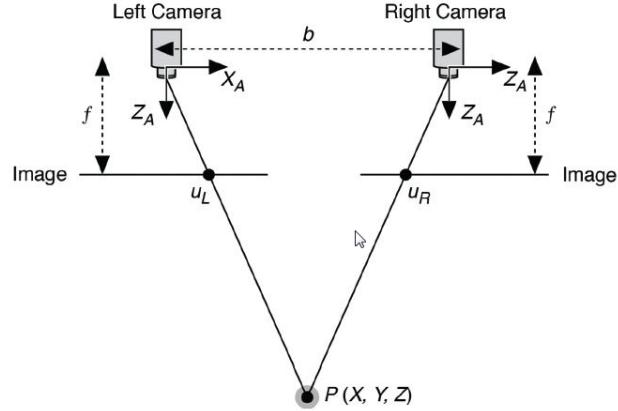
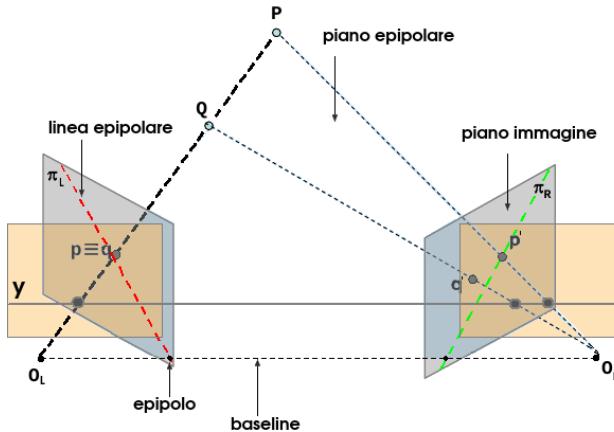


Figure 1.7: Rectified stereo cameras

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 ($PO_L O_R$ and $Pp p'$):

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

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-

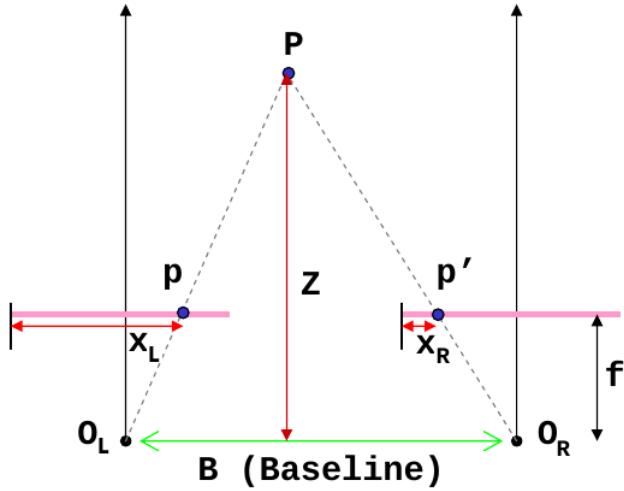


Figure 1.9: Geometry of standard form

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

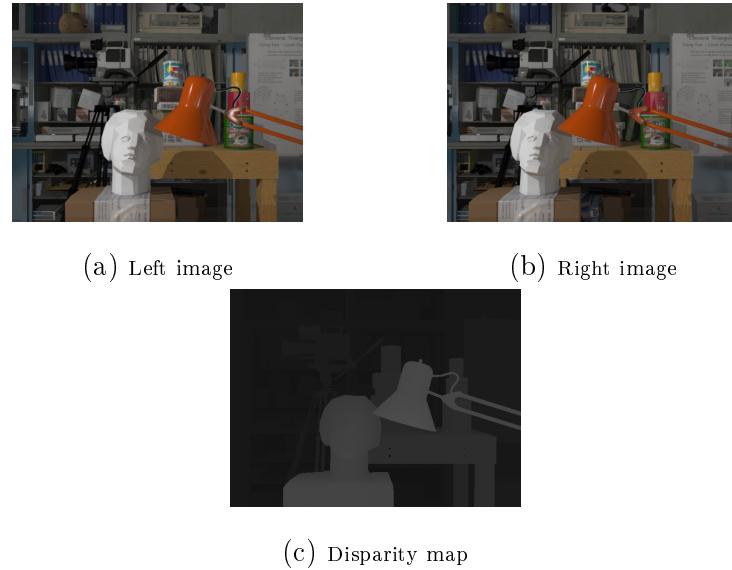


Figure 1.10: Stereo pair and disparity map

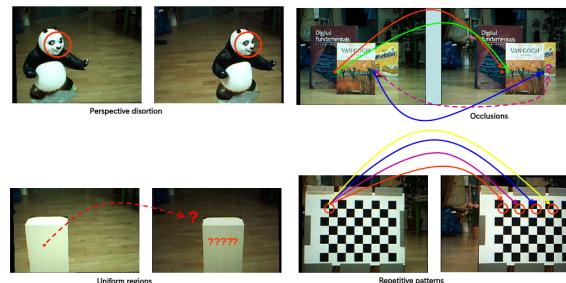


Figure 1.11: Stereo matching general problems

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

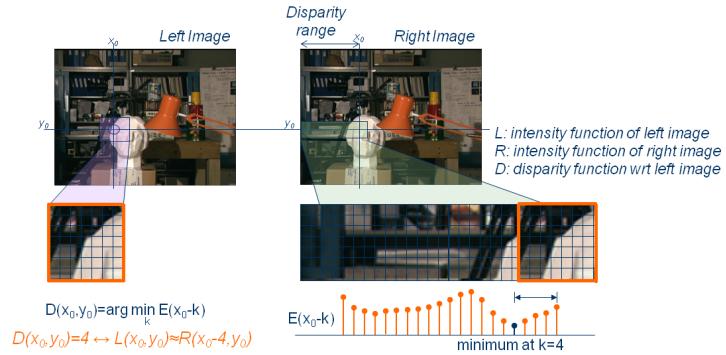


Figure 1.12: Local stereo matching, window based

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 truth.

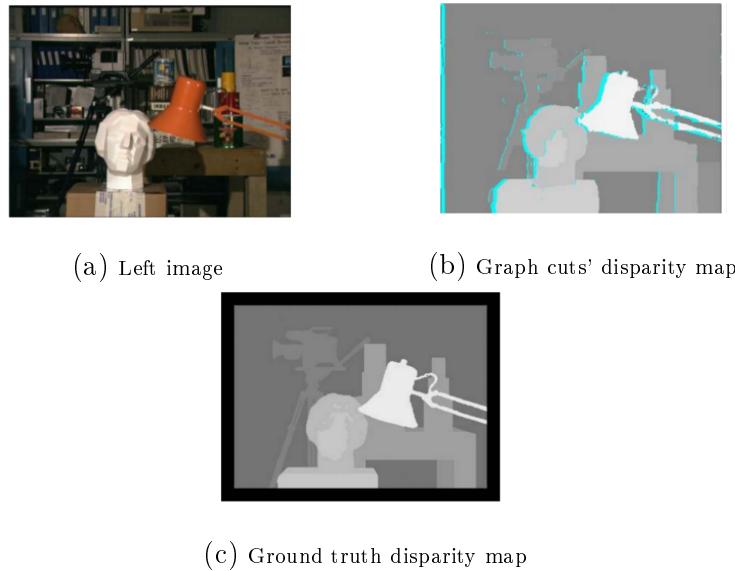


Figure 1.13: Results of the Kolmogorov and Zabih's graph cuts algorithm on the Tsukuba pair

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.



Figure 1.14: Interaxial separation between lenses

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 because of the physical size of the cameras;



(a) Side-by-side rig



(b) Beamsplitter rig



(c) Monoblock camera

Figure 1.15: Professional technologies for 3D TV

- 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.

An other important category of 3D image capture devices it's the one em-



Figure 1.16: Digital personal stereo vision systems

ployed in the robotics and automation field. They are usually impressively precise, cost-efficient and fast.

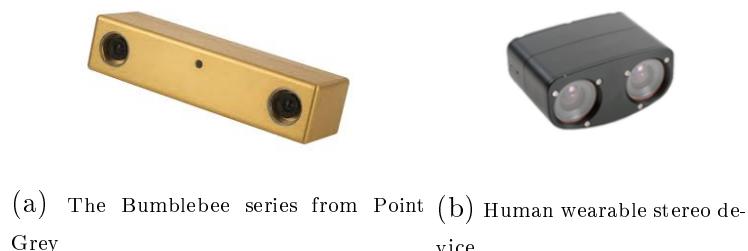


Figure 1.17: Industrial and robotic stereo cameras

1.3 3D video displays

The basic technique of stereo displays is to present offset images that are displayed separately to the left and right eye. Both of these 2D offset images are then combined in the brain to give the perception of 3D depth.

For stereoscopic 3D displays the viewer needs to wear special glasses which separate the views of the stereoscopic image for the left and the right eye. These 3D glasses can be active or passive.

On the one hand, active glasses are controlled by a timing signal that allows to alternatively darken one eye glass, and then the other, in synchronization with the refresh rate of the screen. Hence presenting the image intended for the left eye while blocking the right eye's view, then presenting the right-eye image while blocking the left eye, and repeating the process at a high speed which gives the perception of a single 3D image. This technology generally uses liquid crystal shutter glasses(Figure XX).

On the other hand, passive glasses are polarization-based systems and contain a pair of opposite polarizing filters; each of them passes light with similar polarization and blocks the opposite polarized light (Figure XX). Passive 3D TV screens sport a filter with alternating horizontal and vertical stripes, separated by a black, picture-blanking bars. When used with glasses which have corresponding polarising lenses, alternate frames are presented to each eye to create a 3D image.

The color anaglyph-based systems are a particular case of the passive glasses and use a color filter for each eye, typically red and cyan, Figure XX. The anaglyph 3D image contains two images encoded using the same color filter, thus ensuring that each image reaches only one eye.

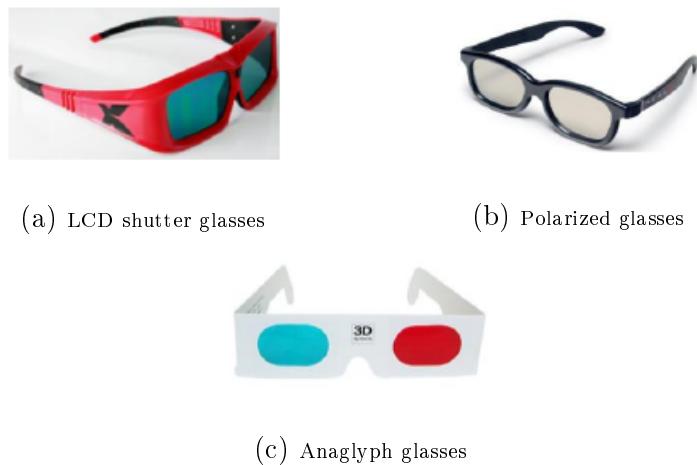


Figure 1.18: Passive and active glasses for 3D viewer technologies

Chapter 2

Stereo video watermarking

2.1 Watermarking

Digital watermarking consists in imperceptibly and persistently associating some extra information with some original content.

The basic watermarking workflow is presented in Figure XX.

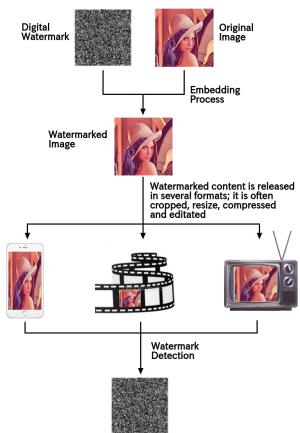


Figure 2.1: Watermarking workflow

2.1.1 Properties

Three parameters are required to evaluate watermarking technique performances:

- transparency, that is the measure of how much the watermark affects the quality of the host data;
- robustness, i.e., the capability of the hidden data to survive host signal manipulation including compression, signal processing, geometric manipulations;
- data payload, that is the amount of data of information bits that it is able to convey.

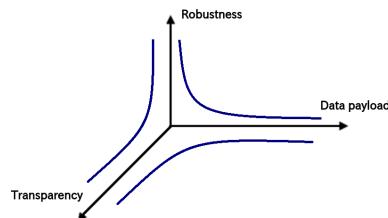


Figure 2.2: Watermark properties trade-off

In the stereoscopic video watermarking context it

2.1.2 Embedding techniques

The most straightforward ways to add a watermark in a given content have been proved to be Spread Spectrum (SS) approach and Side Information (SI).

As in spread spectrum communications, the former approach considers the original content as a signal and the watermark as a noise that is spread over very many frequency bins so that the energy in any one bin is very small and

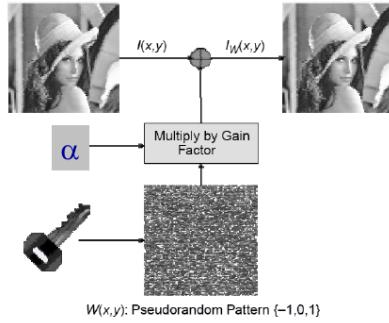


Figure 2.3: Spread spectrum technique

certainly undetectable.

The latter takes advantage of the fact that the original content is known at the embedder side (but unknown at the detector): this way the watermark can be modulated according to the original and the quantity of inserted data can be maximized.

2.1.3 Embedding domains

Host features modified during embedding can belong to

- spatial domain: the watermark is embedded by directly modifying the pixel values;



Figure 2.4: Spatial domain watermark insertion

- frequency domain: the image is transformed through a mathematical transformation, some coefficients are modified and finally the inverse transform is carried out;

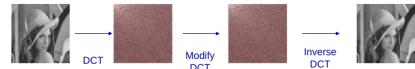


Figure 2.5: Frequency domain watermark insertion

- hybrid techniques: a block wise transform is applied, the image is divided into blocks and for each block a mathematical transformation is computed, some coefficients are modified and the inverse transform is done.

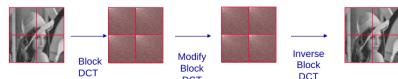


Figure 2.6: Hybrid technique

2.2 Stereoscopic video watermarking

2.2.1 State of the art

In this thesis due algoritmi di marchiatura presentati; il primo spaziale ss con rumore gaussiano etc.. il secondo ss nella frequenza additivo moltiplicativo...

data payload è ... la trasparenza è stata valutata con ... la robustezza è stata provata per view synthesis e compressione

Chapter 3

Spatial disparity-coherent watermarking

As said in the previous chapter a number of articles focused on how to incorporate depth information into the perceptual shaping process of the embedded watermark. This process allows to achieve disparity-coherence and makes sure that a physical point of the captured scene carries the same watermark sample regardless of where it appears in the left and right view.

This process brings two advantages: it produces stereoscopic views more in line with reality therefore yields less visual discomfort; and is expected to have superior robustness against view synthesis.

3.1 Prior work

A prior work, that's based on the disparity coherent technique is the one carried on by Doerr et al in "Blind Detection for Disparity-Coherent Stereo Video Watermarking" []. The watermark strategy assumes that the key-seeded reference watermark pattern $w \in \{0, 1\}$ is embedded spatially in the left view and subsequently transferred to the right one. The watermark embedding and detection operations for the left view are therefore given

by the conventional spread-spectrum equations: FORMULA LEFT For the right view, the watermarking equation is the same, except that the watermark pattern w_K is warped according to the depth information prior to insertion. FORMULA RIGHT The watermark detection on the right view relies on the computation of a horizontal cross-correlation array. FORMULA The correlation array is then mapped into a scalar value in order to compare it with a threshold and decide whether the tested content contains the watermark, three possible mapping functions are proposed: FORMULA

3.2 Gaussian-noise disparity-coherent watermarking

Based on the described technique a spatial watermarking technique is proposed. For the spatial watermark it been taken under consideration the insertion of a Gaussia-noise reference watermark in an additive way. As in Doerr et al, the left view is processed in the conventional way, with spred-spectrum equations (riferimento all'equazione); the watermark is then warped according to the disparity value and inserted in the right view (rif all'eq).

In the detection process, we apply a conventional correlation-based detector to the left view (ref to eq). On the other hand to detect the watermark in the right view two differet correlation-based strategies are proposed: in the first strategy its computed the correlation value between the non-distorted watermark and the right view warped according to the right-to-left disparity: this way the prviously warped watermark is restore, even if there will be discontinuities due the occluded zones. FORMULA The second strategy is again a simple correlation-based detector, but the correlation value is computed between the right view and the warped watermark instead of the original one. FORMULA

Chapter 4

Frequency disparity-coherent watermarking

Chapter 5

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](http://googleresearch.blogspot.it/2014/04/lens-blur-in-new-google-camera-app.html)