Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Jacopo Losi

Structured light assisted real time stereo photogrammetry for robotics and automation

Novel implementation of stereo matching

Master's Thesis
Espoo, 2em

DRAFT! — April 2, 2020 — DRAFT!

Supervisors:	Professor Juho Kannala, Aalto University
	Professor Nicola Conci, University of Trento
Advisor:	Sami Ruuskanen M.Sc. (Tech.)

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author:	Jacopo Losi		
Title: Structured light assisted real time stereo photogrammetry for robotics and automation Novel implementation of stereo matching			
Date:		**Pages:**	v + 11
Major:	Autonomous Systems	**Code:**	
Supervisors:	Professor Juho Kannala Professor Nicola Conci		
Advisor:	Sami Ruuskanen M.Sc. (Tech.)		

The abstract provides goal, motivation, background, and conclusions of the work. It has to fit to one page together with the bibliographical information.

If the thesis is in English and the language of school education is Finnish or Swedish, the abstract is written in English and in Finnish or in Swedish. If the language of school education is other than Finnish or Swedish, the abstract is written in English only.

The thesis example file (`thesis-example.tex`), all the chapter content files (`1introduction.tex` and so on), and the Aalto style file (`aalto-thesis.sty`) are commented with explanations on how the Aalto thesis works. The files also contain some examples on how to customize various details of the thesis layout, and of course the example text works as an example in itself. Please read the comments and the example text; that should get you well on your way!

In the thesis template, you can find the text of the abstract in the abstract in the `thesis-example.tex` file, together with the bibliographical information of the abstract tables. !FIXME **This is an example how to use fixme: add your abstract here.** FIXME! Fixme is a command that helps you identify parts of your thesis that still require some work. When compiled in the custom `mydraft` mode, text parts tagged with fixmes are shown in bold and with fixme tags around them. When compiled in normal mode, the fixme-tagged text is shown normally (without special formatting). The draft mode also causes the "Draft" text to appear on the front page, alongside with the document compilation date. The custom `mydraft` mode is selected by the `mydraft` option given for the package `aalto-thesis`, near the top of the `thesis-example.tex` file.

The instructions on how to compile LaTeX *.tex files to *.pdf files like this are giving in the `thesis-example.tex` file as comments and also in this pdf in a Section ??.

Keywords:	stereo vision; matching cost; census transform; hamming distance; binary pattern; semi-global matching
Language:	English

Acknowledgements

I wish to thank all students who use LaTeX for formatting their theses, because theses formatted with LaTeX are just so nice.

Thank you, and keep up the good work!

Espoo, 5mm Jacopo Losi

Abbreviations and Acronyms

2k/4k/8k mode	COFDM operation modes
3GPP	3rd Generation Partnership Project
ESP	Encapsulating Security Payload; An IPsec security protocol
FLUTE	The File Delivery over Unidirectional Transport protocol
e.g.	for example (do not list here this kind of common acronymbs or abbreviations, but only those that are essential for understanding the content of your thesis.
note	Note also, that this list is not compulsory, and should be omitted if you have only few abbreviations

Contents

Chapter 1

Introduction

1.1 Problem statement

Dense and accurate disparity maps are the key factor for obtaining correct depth estimations for many computer vision applications such as autonomous driving, 3D reconstruction, object detection and robotics. Therefore, stereo matching and disparity estimation can be identified as fundamental problems in the current developments of computer vision [1].

Multiple methods for disparity estimation has been developed for many years [1]. Specifically, older strategies are focused on local-based or global-based methods. On the contrary, deep learning based strategies applied to local or global methods has been recently proposed. The latter approach aims to a precise local correspondence exploiting deep learning and applying Semi-global matching (SGM) as the regularization step of the pipeline. Therefore, deep learning techniques such as FlowNet and DispNet [1] are used as the end-to-end part of the pipeline. According to the current benchmark database ranks for stereo matching algorithms, e.g. the one published in the KITTI website, the state of the art implementations are based on deep learning methods. However, these strategies lack in accuracy compared to the standard pipelines. This is probably due to the difference between real environment and the training database as underlined in [1] [2].

As aforementioned, the state of the art methods to recover dense disparity maps from stereo pairs are focused on deep convolutional neural networks trained end-to-end [3]. Most of these techniques, which will be subsequently described, exploit as regularization phase the Semi-global matching (SGM) method. Actually, among local and global approaches, the Hirschmuller's algorithm [4] appears to be the best performing in terms of computational cost and accuracy. For this reason, it is the preferred trade-off for most real time applications.

Considering the multiple algorithm for stereo correspondence, they can be conventionally classified [5] into two general categories, local and global approaches. Specifically, the local-based methods tend to estimate the disparity image trough a comparison of the matching cost from left and right views of the scene. In order to recover from low accuracy proper of the previous strategy, global-based methods try to calculate the disparity values by minimizing an energy function. In this context, Semi-Global Matching (SGM) combines strong factors of global and local approaches allowing to obtain a good trade-off between computational cost and accuracy.

Technically speaking, SGM applies a pixelwise, Mutual Information (MI) based matching cost for analysing pixel intensity value differences of input images [4]. Moreover, pixelwise matching is enhanced with a smoothness constraint, which leads to a global cost function. Then, post-processing techniques are applied to remove outliers and filter the image.

Referring to the analysis performed by Scharstein and Szeliski [5], SGM carries out four main steps, as well as most of the stereo matching algorithms. These are defined as: matching cost computation, cost aggregation, disparity computation and disparity refinement.

Considering the former, it is usually based on absolute, squared or sampling insensitive difference between pixel intensities [4]. Although those methods allow to reach a reliable accuracy, they are sensitive to radiometric difference. Thus, cost based on image gradients or window-based methods, such as rank and census transform [6], became an optimal choice. Furthermore, Mutual Information results as a good trade-off for dealing with complex radiometric relationships between images.

In the second phase, cost aggregation collects the matching costs considering multiple directions and the disparity levels. Following, disparity evaluation is defined for each pixel, as the one with the lowest cost. This is the approach typically used for local methods. Global algorithms, rather, used to get rid of the aggregation step and define a global energy function. Over that function, pixel similarity and disparity smoothness are enforced with different strategies. In this latter case, the best disparity is identified finding the minimum of the cost function. This is achieved with multiple techniques such as: Dynamic Programming (DP) [7], Belief Propagation [8] or Graph Cuts [9].

Disparity refinement tends to differ more among the different methods. Usually, post-processing techniques such as filtering, outlier removal and consistency check are in general applied.

As anticipated above, among the top-ranked stereo matching algorithms, SGM results to be the best performing in terms of computational time and

accuracy. Its benefits stand in the hierarchical computation of the matching cost, which exploit Mutual Information. Cost aggregation is achieved taking into account a global energy function and a pathwise pixel optimization. The final disparity is chosen with a winner takes all strategy. Disparity refinement is completed by consistency check between left and right disparity images.
Besides the challenge of building up the optimal algorithm for recovering a disparity image from a stereo image pair, it is necessary to develop an analysis of the basis of stereo correspondence and its importance for multiple applications such as: autonomous driving, robotics, object detection and 3D reconstruction.
First of all, stereo matching is defined as the process of estimating a 3D model of a scene, starting from two or more images. Therefore, the matching pixel between the images are found and their 2D positions are converted into 3D depths. Thus, how this operation of building a dense depth map, assigning relatives depth to the input image pixels, is achieved. This is based on the disparity, defined as the amount of horizontal motion between two properly configured images of a stereo pair. This one is then inversely proportional to the distance from the observer, i.e. the camera. Although this concepts are relatively simple to understand, the challenging task within this process stands in establishing dense and accurate inter-image correspondences[10].
As already underlined, stereo matching is one of the most widely studied topic in computer vision from years and it continues to be one of the most active research in that field. In fact, modelling of human visual systems, robotic navigation and manipulation and autonomous driving [2] and 3D model building are some of the possible applications.
The explanations of the fundamental principles of stereo matching, such as epipolar geometry, rectification and disparity map, follows.

1.1.1 Stereo geometry

Main goal of epipolar geometry is the computation of pixels correspondences among the input images. Neighbouring pixels information, cameras positions and their calibration data are fundamental to achieve that. Figure 1.1 demonstrate a pixel in one image $\mathbf{p}_1$ projected to its correspondent epipolar line segment in the other image, which is lower bounded by the projection of the first camera center into the second camera plane, i.e. the epipole $\mathbf{e}_2$. Projecting the epipolar line in the second image back to the first, another line would be obtained, bounded by the correspondent epipole $\mathbf{e}_1$. The extensions to infinity of these two segment are identified as the epipolar lines, which are defined by the intersection of the two image planes with the epipo-

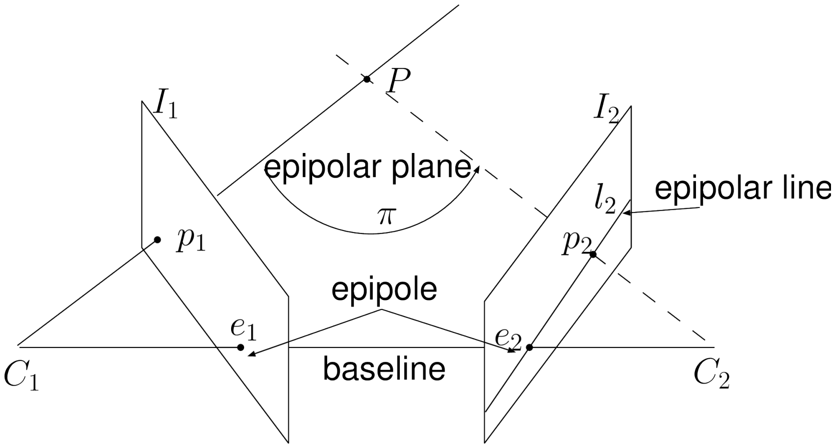


Figure 1.1: Epipolar geometry

lar plane. A fundamental property is that the epipolar plane passes through both camera centers $\mathbf{C}_1$ and $\mathbf{C}_2$, as well as point $\mathbf{P}$. Therefore, they lie in the same plane.

1.1.2 Rectification

Epipolar geometry for a pair of cameras is relative to pose and calibration of the camera and can be computed using the fundamental matrix, which can be obtained applying the eight point algorithm [11]. Computing this geometry allows, then, to find the correspondent pixels between the two images using the constraint of the epipolar lines. This is possible, because, as explained in 1.1.1, considered a pixel in one image, the correspondent one lies on the relative epipolar line.
Beside that, pixels correlations can be more efficiently performed by rectifying the input images [11]. In Figure 1.2 is clearly visible the outcome of this process and its advantages. As shown, corresponding horizontal scanlines are epipolar lines. The essential importance of this standard rectified geometry is clearly explained by the following equation,

$$d = f\frac{B}{Z} \tag{1.1}$$

that leads to a linear relationship between 3D depth Z and disparity d, where f is the focal length (in pixel) and B the baseline. Moreover, the relationship between the corresponding pixels in the left and right images

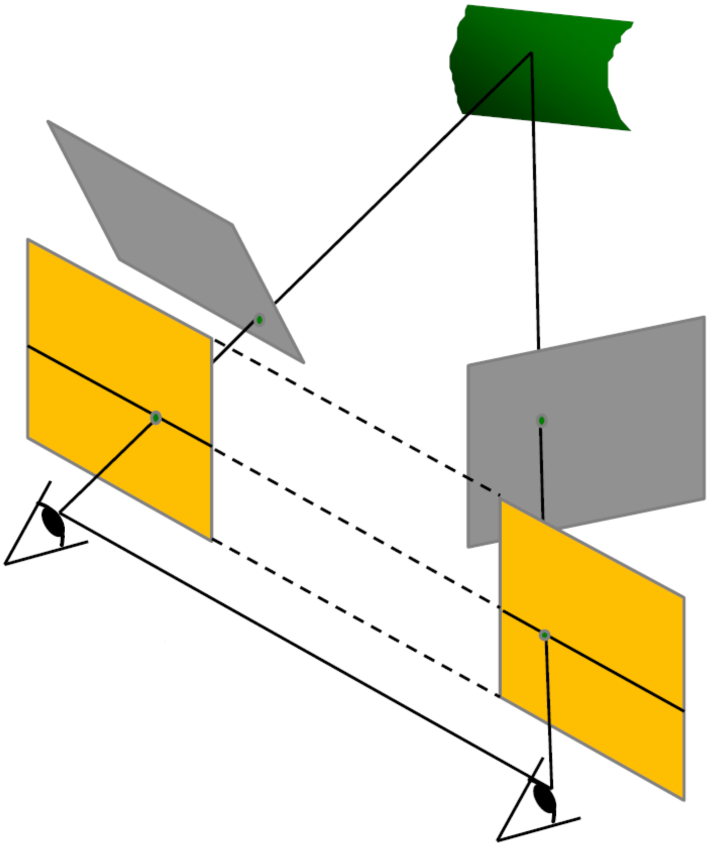


Figure 1.2: Image rectification – Source: L. Lebeznik

can be defined as follows:

$$x' = x + d(x,y),\ y' = y \tag{1.2}$$

Thus, the main step for recovering a depth image of a scene is the estimation of the disparity map $d(x,y)$.
As introduced at the beginning, the best disparity map is estimated after the rectification process. This is performed by comparing the similarity of corresponding pixels, as defined in equation 1.2, and storing them in a disparity space image $C(x,y,d)$, which is then processed with multiple algorithms.

1.1.3 Stereo methods and dense correspondence

In this section a brief delineation of the general pipeline implemented in most of the stereo matching method is presented. Moreover, as a theoretical completion of what introduced above, some generic algorithm are further explained.

Stereo algorithm follows in general a subspace of the following methods: matching cost computation, cost aggregation, disparity computation and optimization and disparity refinement [5].

A preliminary distinction, based on those phases, separates stereo methods between local or window-base global methods.

In local methods, the disparity computation in a certain region depends on the pixel intensities within a limited window.

On the contrary, global algorithms, are based on an energy function. In this one smoothness assumptions are defined and then a global optimization problem is solved. These algorithm are mainly distinguished considering the minimization strategy, such that, simulated annealing, graph cuts or belief propagation.

Between these two classes there are iterative and hierarchical algorithms. The latter aim to constraint gradually the disparity estimation from the coarser to the finer levels [4].

Considering the first general step of stereo matching algorithms, the matching cost, there are multiple measures to define it. Among the most prevalent pixel-based algorithm can be included square intensity differences, absolute intensity differences, mean-squared error and mean absolute difference.

Other common matching cost comprehend normalized cross-correlation, which is similar to sum-of-squared-difference (SSD), and binary methods. However, the latter tend to not be used any longer.

On the other hand, lately, more robust algorithms are used for their insensitivity to non-stationary exposure and illumination changes. Entropy measures and non-parametric functions such as, rank and census transform [12], sampling insensitive difference[7] and hierarchical mutual information [4], are some examples. In particular, they allow to obtain accurate performance when considerable exposure or appearance variations are present.

Drawing up some conclusion regarding the local methods, the core steps are the matching cost calculation and the aggregation phase. Disparity estimation, then, becomes trivial. Each pixel takes the disparity levels whose cost value is the minimum. This approach is said to be a local *winner-take-all* optimization. A drawback of this approach is that the matches are imposed for the reference image. While points is the support image might have multiple correct matches. For this reason, cross-checking and post-processing become more relevant here.

Summarizing the general pipeline of global stereo matching methods, they often get rid of the aggregation step. They usually perform sort of optimization steps after disparity estimation, exploiting the smoothness constraints

as aggregation part.
Goal of this approach is to find the solution to a global energy function, i.e. the disparity d that minimizes the energy,

$$E(d) = E_d(d) + \lambda E_s(d) \tag{1.3}$$

where $E_d(d)$ is the data term and $E_s(d)$ the smoothness term. Adopting the aforementioned disparity space image (DSI) matching cost, the data energy is calculated as:

$$E_d(d) = \sum_{(x,y)} C(x, y, d(x, y)) \tag{1.4}$$

where C is the DSI. Then, the smoothness term is usually defined as:

$$E_s(d) = \sum_{(x,y)} \rho(d(x, y) - d(x + 1, y)) + \rho(d(x, y) - d(x, y + 1)) \tag{1.5}$$

where ρ is some monotonically increasing function of disparity difference.
After the energy function has been clearly identified, different categories of algorithms can be exploited to recover a (local) minimum. Graph cut, belief propagation and Markov random field (MRF) based methods have been proved to give the most accurate results.

1.2 Structure of the Thesis

You should use transition in your text, meaning that you should help the reader follow the thesis outline. Here, you tell what will be in each chapter of your thesis. Often the thesis does not have as many chapters as is in this template. For example, environment and implementation can be combined as well as chapters of evaluation and discussion. The rest of this thesis is organized as follows. Chapter **??** gives the background, etc.

Bibliography

[1] A. Seki, M. Pollefeys, T. Corporation, E. T. Zürich, and Microsoft, "SGM-Nets: Semi-global matching with neural networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-Janua, no. 1, pp. 6640–6649, 2017.

[2] M. Poggi, D. Pallotti, F. Tosi, and S. Mattoccia, "Guided Stereo Matching," 2019.

[3] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. D. Stefano, "Real-Time Self-Adaptive Deep Stereo," pp. 195–204, 2020.

[4] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[5] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Proceedings - IEEE Workshop on Stereo and Multi-Baseline Vision, SMBV 2001*, no. December, pp. 131–140, 2001.

[6] J. Ko and Y. S. Ho, "Stereo matching using census transform of adaptive window sizes with gradient images," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*, pp. 2–5, 2017.

[7] S. Birchfield and C. Tomasi, "Depth discontinuities by pixel-to-pixel stereo," *International Journal of Computer Vision*, vol. 35, no. 3, pp. 269–293, 1999.

[8] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," *Proceedings - International Conference on Pattern Recognition*, vol. 3, pp. 15–18, 2006.

[9] V. Kolmogorov and R. Zabih, “Computing visual correspondence with occlusions using graph cuts,” *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, pp. 508–515, 2001.

[10] R. Szeliski, “Computer vision: algorithms and applications,” *Choice Reviews Online*, vol. 48, no. 09, pp. 48–5140–48–5140, 2011.

[11] R. Hartley and A. Zisserman, “Multiple view geometry in computer vision. cambridge university press, isbn: 0521540518,” 2004.

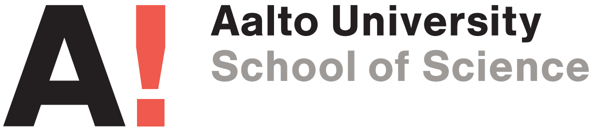
[12] R. Zabih and J. Woodfill, “Non-parametric local transforms for computing visual correspondence,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 801 LNCS, pp. 151–158, 1994.

Appendix A

First appendix

This is the first appendix. You could put some test images or verbose data in an appendix, if there is too much data to fit in the actual text nicely.

For now, the Aalto logo variants are shown in Figure A.1.



(a) In English

Figure A.1: Aalto logo variants