# UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS FÍSICAS
Departamento de Arquitectura de Computadores y Automática


## TESIS DOCTORAL

**A non-feature based method for automatic image registration relying on depth-dependent planar projective transformations**

**Método para el registro automático de imágenes basado en transformaciones proyectivas planas dependientes de las distancias y orientado a imágenes sin características comunes**

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

**Carlota Salinas Maldonado**

Directores
Manuel Ángel Armada Rodríguez
Roemi Emilia Fernández Saavedra
Héctor Montes Franceschi

**Madrid, 2016**

**UNIVERSIDAD COMPLUTENSE DE MADRID**

FACULTAD DE CIENCIAS FÍSICAS

Departamento de Arquitectura de Computadores y Automática


**TESIS DOCTORAL**

# A Non-Feature Based Method for Automatic Image Registration Relying on Depth-dependent Planar Projective Transformations

Método para el registro automático de imágenes basado en trasformaciones proyectivas planas dependientes de las distancias y orientado a imágenes sin características comunes

Carlota Salinas Maldonado

**2015**

# A Non-Feature Based Method for Automatic Image Registration Relying on Depth-dependent Planar Projective Transformations

Método para el registro automático de imágenes basado en trasformaciones proyectivas planas dependientes de las distancia y orientado a imágenes sin características comunes


Tesis Doctoral presentada por / PhD Thesis presented by

**Carlota Salinas**

•

Dirigida por /Supervised by

**Dr. Manuel Ángel Armada Rodríguez**

**Dra. Roemi Emilia Fernández Saavedra**

**Dr. Héctor Montes Franceschi**

•

Grupo de Robótica de Exteriores y Servicios

**Centro de Automática y Robótica (CSIC-UPM)**

•

2015

Departamento de Arquitectura de Computadores y Automática
Facultad de Ciencias Físicas
Universidad Complutense de Madrid

# A Non-Feature Based Method for Automatic Image Registration Relying on Depth-dependent Planar Projective Transformations

Método para el registro automático de imágenes basado en trasformaciones proyectivas planas dependientes de las distancias y orientado a imágenes sin características comunes

*Memoria presentada para optar al grado de Doctor por*
*Dissertation submitted to obtain the PhD Degree by*

**Carlota Salinas**

*Dirigida por / Supervised by*

**Dr. Manuel Ángel Armada Rodríguez**
**Dra. Roemi Emilia Fernández Saavedra**
**Dr. Héctor Montes Franceschi**

Madrid, 2015

*A Lucas y Roberto,*
*a mis padres, Luly y Carlos,*
*y a mi hermano Sebastián*

# Agradecimientos

A lo largo de los años de trabajo de esta tesis doctoral, muchas son las personas que me han brindado su apoyo y su confianza.

En primer lugar me gustaría expresar mi profundo agradecimiento a mi Director de tesis el Dr. Manuel Armada Rodríguez, Director del Centro de Automática y Robótica CAR (CSIC-UPM), quién me ha acompañado en todos las etapas evolutivas de este trabajo y, gracias a su apoyo profesional y moral, su paciencia y comprensión a los largo de estos años, me ha brindado la oportunidad de crecer y desarrollarme como persona e investigadora, dándome la libertad de investigar distintos campos en la visión por computador y poder alcanzar mis objetivos.

Por otro lado, me gustaría agradecer de manera muy especial a mis otros dos Directores de tesis, la Dra. Roemi Fernández Saavedra y al Dr. Héctor Montes Franceschi. Ambos son extraordinarios amigos y compañeros de trabajo, que durante el desarrollo de la tesis me han guiado con amabilidad y firmeza, lo cual me ha permitido mantenerme concentrada en conseguir mis objetivos. A Roemi me gustaría agradecerle, sus sabios y acertados consejos además de su calidez como persona, a quién admiro por su profesionalidad y conocimientos. A Héctor le agradezco el toque de humor con el que aborda las cosas, además de sus consejos y comentarios, a quién admiro y respeto por su capacidad de trabajo y su serenidad.

También me gustaría agradecer a mi tutor de la tesis, al Dr. Jesús Manuel de la Cruz, Profesor de la Universidad Complutense de Madrid, quién con amabilidad y buena voluntad, siempre me brindado su apoyo para el desarrollo y culminación de la tesis. Por otro lado, me gustaría dar las gracias

al Dr. Pablo González de Santos, Jefe del Departamento de Robótica de Exteriores y Servicios del CAR, que con sus oportunos comentarios y pragmatismo, me ha dado impulso para continuar y terminar esta etapa. A la Dra. Elena García, me gustaría agradecerle por compartir sus experiencia y por sus consejos, los cuales siempre me han servido de apoyo para avanzar con el trabajo de la tesis.

A mi familia, gracias por su amor, paciencia, comprensión y apoyo incondicional en todos mis proyectos, tanto personales como profesionales. En especial a mi madre Luly, por servirme de inspiración por su valentía y empeño para afrontar los retos de la vida. A Roberto, mi marido y amigo, quién me ha acompañado en los momentos dulces y en los no tan dulces de esta etapa, y que gracias a su forma de ver la vida, me ha empujado a mirar siempre hacia adelante. A mi hermano Sebastián, gracias por sus consejos y constante cariño, por estar siempre a mi lado desde que empecé con en esta etapa profesional en España.

Finalmente, agradecer a mis compañeros del CAR-CSIC, con quienes he compartido vivencias muy enriquecedoras, que me han permitido cultivar grandes amistades. De manera particular, a Javier Sarria y a Jesús Reviejo, agradecerles primero por su cariño y amistad, luego por su colaboración y disposición para ayudarme durante el desarrollo de la tesis, gracias por compartir conmigo conocimientos y criterios. A Manuel Prieto y a Roberto Ponticelli, antes compañeros de trabajo y amigos siempre, quienes con su alegría y mirada crítica, me han servido de inspiración para trabajar con constancia y dedicación

# Abstract

Multisensory data fusion oriented to image-based application improves the accuracy, quality and availability of the data, and consequently, the performance of robotic systems, by means of combining the information of a scene acquired from multiple and different sources into a unified representation of the 3D world scene, which is more enlightening and enriching for the subsequent image processing, improving either the reliability by using the redundant information, or the capability by taking advantage of complementary information.

Image registration is one of the most relevant steps in image fusion techniques. This procedure aims the geometrical alignment of two or more images. Normally, this process relies on feature-matching techniques, which is a drawback for combining sensors that are not able to deliver common features. For instance, in the combination of ToF and RGB cameras, the robust feature-matching is not reliable. Typically, the fusion of these two sensors has been addressed from the computation of the cameras calibration parameters for coordinate transformation between them. As a result, a low resolution colour depth map is provided. For improving the resolution of these maps and reducing the loss of colour information, extrapolation techniques are adopted. A crucial issue for computing high quality and accurate dense maps is the presence of noise in the depth measurement from the ToF camera, which is normally reduced by means of sensor calibration and filtering techniques. However, the filtering methods, implemented for the data extrapolation and denoising, usually over-smooth the data, reducing consequently the accuracy of the registration procedure.

The study presented in this Thesis introduces a solution for dealing with the aforementioned problems. More specifically, an approach for sensor registration with non-common features is proposed, which is based on *Planar Projection Transformations* and the depth measurements from the ToF camera. The depth information is used as a virtual feature for estimating a depth-dependent homography lookup table (*Hlut*). The elements of the *Hlut* were computed by virtually discretizing the 3D-space world into {*n*-planes}, which are positioned in front and parallel to the sensory system. Then, suitable homographies from accurate transformation between views were selected. These homographies are able to transferring data belonging to several consecutive *i*-planes of the {*n*-planes}, which are constrained within a range of depth. In this way the working distance of each element on the *Hlut* is known. The procedure is capable of computing a low resolution colour depth map together with a labelled homography mask $mask_{LRGB}$ on the RGB image coordinates. The values of the $mask_{LRGB}$ correspond to the homographies $\{ H_k^{lut} \}$ used for transferring the data. Due to the difference in the cameras resolution, between each pair of adjacent points in the ToF image there are several unmapped points on the RGB image coordinates. Thus, the labelled mask $mask_{LRGB}$ is intended to be used for matching the unmapped points on the RGB image frame. This research presents an initial approach for this procedure, where a nearest neighbourhood algorithm was applied to create an entire mask of $\{ H_k^{lut} \}$ on the RGB pixel coordinates. Then, the high resolution colour depth map is straightforward computed by mapping points from the RGB to the ToF, by using the homographies $\{ H_k^{lut^{-1}} \}$.

The accuracy evaluation of the proposed method is twofold. In the first part, 104 image samples were registered and the discrepancy between control points and estimated points was calculated. The results indicate that the proposed method is capable of mapping points from the ToF to RGB frame with a mean error of 0.44 pixels and a standard deviation of 2.9 pixels. In the second part, the comparison between the depth-dependent *Hlut* approach and the standard calibration method for depth map registration was addressed. For this comparison three scenarios were considered: noise-free (ideal) depth data, raw depth data and filtered data. The numerical and the visual results show that the depth-dependent *Hlut* approach outperforms the standard calibration results.

The final contribution of this study includes the test and validation of the proposed method within the framework of two relevant robotic applications. First, an indoors application oriented to in-house surveillance is considered, where the capability of the proposed method for motion detection tasks was

assessed. For that, a new procedure for people's motion detection was proposed. The procedure is based on the depth-dependent *Hlut* approach, a robust affine structure from motion algorithm and a quadric surface approximation. After the analysis of a large number of person's poses, the obtained visual results demonstrate the satisfactory performance of the approach. In addition, the proposed procedure is capable of dealing with shadows and variations in the illumination conditions, while avoiding the false inliers detected as motion.

The second application is related to Precision Agriculture and is framed within the European Project entitled CROPS, which is enclosed in the topic of *Automation and robotics for sustainable crop and forestry management*. The experimentation aims to assess the feasibility of detecting and locating fruits (apples) and other plant elements in natural environments by utilising a multisensory system in combination with the proposed depth-dependent *Hlut* approach. This experimental stage was conducted in laboratory and in field conditions and the obtained visual results shown a satisfactory performance of the high resolution colour depth map formation. Despite the complexity of the scenes, (small, rounded and angled objects), the presence of misalignment problems is almost imperceptible, and shape and edges of objects are preserved. Additionally, a feature extraction procedure was proposed and implemented. The results illustrate the capability of the proposal for detecting and locating fruits.

# Resumen

La fusión multisensorial orientada a aplicaciones de procesamiento de imágenes, conocida como fusión de imágenes, es una técnica que permite mejorar la exactitud, la calidad y la disponibilidad de datos de un entorno tridimensional, que a su vez permite mejorar el rendimiento y la operatividad de sistemas robóticos. Dicha fusión, se consigue mediante la combinación de la información adquirida por múltiples y diversas fuentes de captura de datos, la cual se agrupa del tal forma que se obtiene una mejor representación del entorno 3D, que es mucho más ilustrativa y enriquecedora para la implementación de métodos de procesamiento de imágenes. Con ello se consigue una mejora en la fiabilidad y capacidad del sistema, empleando la información redundante que ha sido adquirida por múltiples sensores.

El registro de imágenes es uno de los procedimientos más importantes que componen la fusión de imágenes. El objetivo principal del registro de imágenes es la consecución de la alineación geométrica entre dos o más imágenes. Normalmente, este proceso depende de técnicas de búsqueda de patrones comunes entre imágenes, lo cual puede ser un inconveniente cuando se combinan sensores que no proporcionan datos con características similares. Un ejemplo de ello, es la fusión de cámaras de color de alta resolución (RGB) con cámaras de Tiempo de Vuelo de baja resolución (*Time-of-Flight* (ToF)), con las cuales no es posible conseguir una detección robusta de patrones comunes entre las imágenes capturadas por ambos sensores. Por lo general, la fusión entre estas cámaras se realiza mediante el cálculo de los parámetros de calibración de las mismas, que permiten realizar la trasformación homogénea entre ellas. Y como resultado de este

procedimiento, se obtienen mapas de profundad y de color de baja resolución. Con el objetivo de mejorar la resolución de estos mapas y de evitar la pérdida de información de color, se utilizan diversas técnicas de extrapolación de datos. Un factor crucial a tomar en cuenta para la obtención de mapas de alta calidad y alta exactitud, es la presencia de ruido en las medidas de profundidad obtenidas por las cámaras ToF. Este problema, normalmente se reduce mediante la calibración de estos sensores y con técnicas de filtrado de datos. Sin embargo, las técnicas de filtrado utilizadas, tanto para la interpolación de datos, como para la reducción del ruido, suelen producir el sobre-alisamiento de los datos originales, lo cual reduce la exactitud del registro de imágenes.

El estudio de investigación realizado en esta tesis, presenta una solución que permite solventar los problemas mencionados anteriormente. Esta propuesta presenta una nueva estrategia para el registro de imágenes que no cuentan con características similares, la cual está basada en *Transformaciones Proyectivas Planas* y en las medidas de profundidad adquiridas por una cámara ToF. La información de las profundidades son utilizadas como características virtuales para el cómputo de una tabla de búsqueda de homografías planas, las cuales dependen de las medidas de profundidad, llamada en inglés *depth-dependent homography lookup table (Hlut)*. Los elementos de *Hlut*, se calculan mediante una discretización virtual del espacio tridimensional en {*n*-planos}, los cuales se encuentran dispuestos frente al sistema y paralelos a éste. A continuación, de este proceso se obtienen las homografías más robustas mediante las transformaciones entre las vistas de las cámaras y los planos. Estas homografías tienen la capacidad de transferir puntos que pertenecen a varios *i*-planos virtuales consecutivos que pertenecen a {*n*-planos}, y que se encuentran delimitados por un rango de distancia determinado. De esta forma, las distancias de trabajo de los elementos en *Hlut* son conocidas. Mediante este procedimiento se obtienen mapas de color y de profundidad de baja resolución y, además, una máscara de etiquetas de homografías $mask_{LRGB}$ en el sistema de coordenadas de las imágenes RGB. Los valores de dicha máscara, se corresponden con las homografías $\{ H_k^{lut} \}$ utilizadas para la transformación de los datos. Debido a la gran diferencia entre las resoluciones de las cámaras, entre cada par de puntos adyacentes en la imagen ToF, existen varios puntos sin transformar en la imagen de color. Por ello, la máscara de etiquetas $mask_{LRGB}$, ha sido creada con el objetivo de ser utilizada como una herramienta en la resolución de este problema. En este trabajo de investigación se presenta una primera estrategia para el uso de esta

máscara, la cual está basada en el algoritmo de vecinos más próximos (*nearest neighbour algorithm*). Con ella se consigue la clasificación de los puntos vacíos de la máscara $mask_{LRGB}$, dando como resultado una máscara completa de etiquetas de homografías en el plano imagen de color. De esta manera, se obtiene un mapa de color y de profundidad de alta resolución, mediante la transformación inversa de las homografías $\left\{ H_k^{lut^{-1}} \right\}$, que permiten transferir puntos desde el plano imagen de color al plano imagen ToF.

La evaluación de la exactitud del método propuesto en esta tesis está divida en dos partes. Para la primera, se ha procedido al registro de 104 imágenes, y posteriormente al cálculo del error entre los puntos de control y los puntos estimados con el método. Los resultados obtenidos durante la evaluación resaltan la capacidad del método propuesto para transferir información desde la cámara ToF a la cámara de color, todo ello con un error medio de 0.44 píxeles y una desviación estándar de 2.9 píxeles. En la segunda parte, se presenta una comparación del método propuesto y el método de calibración estándar de las cámaras. Para esta comparación, tres escenarios han sido tomados en cuenta: con medidas de profundidad sin ruido (ideales), con medidas de profundidad sin procesar y con medidas de profundidad filtradas. Los resultados obtenidos, tanto visuales como numéricos, indican que el método propuesto *Hlut*, supera en rendimiento al método de calibración estándar de las cámaras.

La contribución final de este trabajo de investigación se centra en la experimentación y validación del método propuesto en un marco de trabajo relativo a dos aplicaciones en robótica. La primera, es una aplicación en interiores, enfocada a la seguridad y vigilancia en hogares, y en la cual, se ha evaluado las capacidades del método propuesto en tareas de detección del movimiento. Para ello, se ha propuesto un método basado en un algoritmo robusto de análisis afín del movimiento y en una aproximación cuadrática de superficies. Los resultados visuales obtenidos, a posteriori de un amplio análisis de posturas de personas, indican que el método propuesto proporciona resultados muy favorables para las tareas de detección del movimiento. Adicionalmente, el método propuesto, tiene la capacidad de lidiar satisfactoriamente con los problemas más comunes en entornos de interiores, que son las sombras y las variaciones de iluminación del entorno, evitando las detecciones de falsos positivos del movimiento.

La segunda aplicación en robótica se encuentra encamarada dentro de un proyecto europeo, denominado por el acrónimo CROPS, el cual se enfoca en

la temática relacionada a la automatización y robótica para la gestión sostenible de cultivos y de bosques. Dicha experimentación tiene como objetivo la validación de la implementación del método propuesto *Hlut* en combinación con un sistema multiespectral, en las tareas de detección y localización de frutas y otros elementos de las plantas, todo ello en entornos naturales. Esta etapa de experimentación ha sido realizada tanto en laboratorio como en entornos naturales de campos de cultivos. Los resultados obtenidos muestran la capacidad del sistema para la formación de mapas de color y de profundidad de alta resolución. A pesar de la complejidad de las escenas (objetos redondeados, pequeños e inclinados), los problemas de desalineación entre las imágenes, son casi imperceptibles. Además de esto, los bordes y formas de los objetos muestran muy pocas alteraciones. Por otro lado, este estudio incluye una propuesta para la extracción de características de las frutas (objetos de interés). En este caso, los resultados visuales también indican el potencial de método propuesto para detección y localización de frutas.

# Note to the Reader

This PhD Thesis is organized as follows:

The Part I comprises the main part of the Thesis and it is written in English language. This part is structured in six Chapters and it contains: (1) a summary of the addressed state of the problem in multisensor fusion, the motivations and the objectives; (2) a state-of-the-art of the image sensor fusion; (3) the description of the design, implementation and validation of the proposed image registration method; (4) the comparative evaluation between the proposed method and the standard calibration method; (5) the experimentation and validation of the method in two robotic applications; and, (6) the conclusions, contributions and future researches.

The Part II contains a summary of Part I, and it is written in Spanish language. This part is structured in five Sections, which describes: (1) a summary of the state of the problem in multisensor fusion; (2) the motivations and the scope; (3) the objectives of this research; (4) the organization of Part I; and, (5) the conclusions, contributions and future researches.

# Nota al Lector

Esta tesis doctoral se estructura de la siguiente manera:

La Parte I contiene la parte principal de la tesis y se encuentra escrita en inglés. Esta parte está formada por seis Capítulos que describen lo siguiente: (1) un breve resumen del estado del problema abordado en esta tesis, así como las motivaciones y objetivos de esta misma; (2) una revisión de la fusión sensorial de imágenes; (3) la descripción del diseño, implementación y evaluación del método de registro de imágenes propuesto; (4) una evaluación comparativa entre el método propuesto y el de calibración estándar de las cámaras; (5) la experimentación y validación del método propuesto en dos aplicaciones robóticas; y, (6) las conclusiones, aportaciones principales de esta tesis y trabajos futuros.

La Parte II presenta un resumen en español de la Parte I y está dividida en cinco secciones, las cuales contienen: (1) un breve resumen del estado del problema abordado en esta tesis; (2)-(3) las motivaciones, el alcance y los objetivos de la tesis; (4) la organización de la Parte I; y, (5) las conclusiones, aportaciones principales de esta tesis y trabajos futuros.

# Contents

Agradecimientos ..... v
Abstract ..... vii
Resumen ..... xi
Note to the Reader ..... xv
Nota al Lector ..... xvi
Contents ..... xvii
List of Figures ..... xxi
List of Tables ..... xxix
Part I: A Non-Feature Based Method for Automatic Image Registration Relying on Depth-dependent Planar Projective Transformations ..... 1
Chapter 1 ..... 1
Introduction ..... 1
1.1 Multisensor Image Fusion ..... 1
1.2 Motivation and Scope ..... 4
1.3 Research Objectives ..... 5
1.4 Thesis Outline ..... 6
Chapter 2 ..... 9

Image Sensor Fusion – State of the Art ..... 9
2.1 Introduction ..... 9
2.2 Combination of RGB and ToF Cameras ..... 17
Chapter 3 ..... 23
Depth-dependent Homography Lookup Table for Dense Map Registration ..... 23
3.1 Introduction ..... 23
3.2 Method Description ..... 26
3.3 Validation of the Depth-dependent Homography Lookup Table Approach ..... 33
3.4 High Resolution Colour Depth Map Estimation ..... 42
3.5 Conclusions ..... 44
Chapter 4 ..... 47
Comparison of Methods for Depth Map Registration ..... 47
4.1 Introduction ..... 47
4.2 Standard Camera Calibration Computation and Evaluation ..... 49
4.3 Noise-free Data (ideal) Evaluation ..... 63
4.4 Raw Depth Measurements Evaluation ..... 67
4.5 Filtered Depth Measurements Evaluation ..... 70
4.6 Conclusions ..... 77
Chapter 5 ..... 79
Experimental Results and Proposed Method Validation ..... 79
5.1 Introduction ..... 79
5.2 Man-made Indoor Environments ..... 81
5.3 Precision Agriculture: Detection and Localization of Fruits for Automatic Harvesting ..... 114
Chapter 6 ..... 149
Conclusions, Contributions and Future Research Directions ..... 149

6.1 Design, Implementation and Validation of the Proposed Image Registration Method ..... 150

6.2 Experimental Testing and Validation of the Proposed Method in Indoors and Outdoors Robotic Applications ..... 151

6.3 Main Contributions ..... 154

6.4 Future Research Directions ..... 155

Part II: Método para el registro automático de imágenes basado en trasformaciones proyectivas planas dependientes de las distancia y orientado a imágenes sin características comunes ..... 157

Resumen ..... 159

1. Fusión de imágenes multisensorial ..... 159

2. Motivación y alcance ..... 163

3. Objetivos de la investigación ..... 164

4. Organización de la tesis ..... 165

5. Conclusiones, aportaciones principales y trabajos futuros ..... 166

5.1 Diseño, implementación y validación del método propuesto para el registro de imágenes ..... 167

5.2. Análisis experimental y validación del método propuesto en aplicaciones de sistemas robóticos en entornos de interiores y de exteriores ..... 169

5.3 Aportaciones Principales ..... 172

5.4 Trabajos futuros ..... 173

References ..... 175

# List of Figures


| | | |
|---|---|---|
| 1.1 | Image fusion procedure structure ……..... | 2 |
| 2.1 | Active techniques for 3D image sensing. (a) Laser-scanner imaging. (b) Structured light imaging …… | 10 |
| 2.2 | Passive techniques for 3D image sensing. (a) Structure from motion. (b) Stereovision ……..... | 12 |
| 2.3 | Omnidirectional systems. (a) Catadioptric image formation. (b) Rectified catadioptric stereovision configuration …….. | 14 |
| 2.4 | ToF camera working principle ……... | 15 |
| 2.5 | Some of the most frequently used ToF cameras and commercially available. (a) MESA SwissRanger SR4000; (b) MESA SwissRanger SR 4500; (c) PMD CamCube; (d) PMD CamBoard nano (currently Sold Out) …… | 16 |
| 3.1 | Sensory system configuration. The sensory rig consists of a ToF camera and a RGB camera, and it is mounted on a robotic platform with four degrees of freedom …….. | 25 |
| 3.2 | Plane induced parallax ……... | 26 |
| 3.3 | Plane projective transformation induced by two planes $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$ on a scene……... | 27 |
| 3.4 | Formation of the depth-dependent homography lookup table….. | 28 |

3.5 Samples of images pair of the pattern grid board. (a) RGB and ToF amplitude images. (b) The depth map representation in the Cartesian system of the inner 2 × 3 grid ……..... 30
3.6 Geometric error evaluation. (a) Geometric error. (b) Distance error. (c) Error distribution in *u*-axis. (d) Error distribution in *v*-axis ……. 36
3.7 Normalized RMSE on RGB image coordinates vs the angle of the board plane w.r.t. the image plane …….. 37
3.8 Depth measurements evaluation (a) Distances from the pattern board to the cameras. (b) Differential value of raw and filtered data. (c) Samples mean depth vs overall error. (d) Samples maximum variation vs overall error …….. 38
3.9 Image sample 49 - pattern board positioned at 527 mm. (a) Top: selected ROI on the ToF image. Bottom: zoom of the selected points on the ToF image. (b) Top: mapped points on the RGB image. Bottom: zoom of the estimated points on the RGB image. (c) Top: ROI of the ToF image. Bottom: Composition ROI from the mapped points on the RGB image. (d) Colour depth map ……. 40
3.10 Image sample 25-pattern board positioned at 891 mm. (a) Top: selected ROI on the ToF image. Bottom: zoom of the selected points on the ToF image. (b) Top: mapped points on the RGB image. Bottom: zoom of the estimated points on the RGB image. (c) Top: ROI of the ToF image. Bottom: Composition ROI from the mapped points on the RGB image. (d) Colour depth map …….. 41
3.11 Image sample 49. (a) Top: mapped points on the ToF image. Bottom: points of the ROI on the RGB image. (b) High resolution colour depth map …….. 43
3.12 Image sample 25. (a) Top: mapped points on the ToF image. Bottom: points of the ROI on the RGB image. (b) High resolution colour depth map …….. 44

4.1 Distribution of the geometric error on the pixels coordinates. (a) *u*-axis error on RGB images. (b) *v*-axis error on RGB images. (c) *u*-axis error on ToF amplitude images. (d) *v*-axis error on ToF amplitude images ……... 52

4.2 Errors on the pixels coordinates. (a) RGB camera distorted error. (b) ToF camera distorted error ……... 53
4.3 Errors on the pixels coordinates. (a) RGB camera normalized RMSE. (b) ToF camera normalized RMSE ……..... 54
4.4 Normalized Calibration Error on the ToF camera coordinates ……... 55
4.5 Estimation of the angle of the board plane w.r.t. the image plane – image sample 44. 55
4.6 Errors vs the angle of the board plane w.r.t. the image plane. (a) RGB camera $E_d$ error. (b) ToF camera $E_d$ error ……..... 56
4.7 Errors vs the angle of the board plane w.r.t. the image plane. (a) RGB camera $E_d$ error. (b) ToF camera $E_d$ error ……..... 57
4.8 Normalized Calibration Error on the ToF camera coordinates vs the angle of the board plane w.r.t. the image plane ……... 58
4.9 Geometric error on the pixels coordinates. (a) Error on RGB images. (b) Error on ToF amplitude images ……. 59
4.10 RGB camera potential outliers. (a) Image sample 44. (b) Image sample 46 ……... 61
4.11 ToF camera potential outliers. (a) Image sample 9. (b) Image sample 18 ……... 62
4.12 Normalized RMSE on RGB camera coordinates vs the angle of the board plane w.r.t. the image plane. (a) Standard calibration method. (b) Depth-dependent Hlut approach ……. 65
4.13 Rendering of the depth measurements and amplitude data acquired by the ToF camera. (a) Image sample 1. (b) Image sample 59 ……... 67
4.14 Normalized RMSE on RGB camera coordinates vs the angle of the board plane w.r.t. the image plane. (a) Standard calibration method. (b) Depth-dependent Hlut approach ……. 68
4.15 Depth map registration of sample 14. (a) RGB image. (b) ToF depth measurement. (c) Standard calibration result. (d) Depth-dependent Hlut result ……. 70
4.16 Denoising filtering of sample 41. (a) Original data. (b) Bilateral filtering. (c) Non-local means filter ……. 72
4.17 Normalized RMSE on RGB camera coordinates vs the angle of the board plane w.r.t. the image plane. (a) Standard calibration method. (b) Depth-dependent Hlut approach ……. 74
4.18 Normalized RMSE on RGB camera coordinates vs the angle of

the board plane w.r.t. the image plane. (a) Standard calibration method. (b) Depth-dependent Hlut approach …… 75

5.1 Image sample 19. (a) Selected points on the ToF. (b) Estimated points on the RGB image. (c) Top: selected ToF ROI. Bottom: estimated RGB ROI. (d) Colour depth map of the ROI……. 84
5.2 Image sample 25. (a) Selected points on the ToF. (b) Estimated points on the RGB image. (c) Top: selected ToF ROI. Bottom: estimated RGB ROI. (d) Colour depth map of the ROI…….. 85
5.3 Image sample 19. (a) RGB image. (b) ToF depth measurements. (c) Homography labelled $mask_{LRGB}$ on RGB image coordinates. (d) Homography labelled $mask_{LRGB}$ on ToF image coordinates. (e) Registered RGB image. (f) ToF amplitude image. (g) Colour depth map ……. 87
5.4 Image sample 8. (a) RGB image. (b) ToF depth measurements. (c) Homography labelled $mask_{LRGB}$ on RGB image coordinates. (d) Homography labelled $mask_{LRGB}$ on ToF image coordinates. (e) Registered RGB image. (f) ToF amplitude image. (g) Colour depth map ……... 89
5.5 Error evaluation of the experimental tests corresponding to group #1. (a) Geometric error. (b) Distance error. (c) Error distribution in *u*-axis. (d) Error distribution in *v*-axis ……. 90
5.6 Error evaluation of the experimental tests corresponding to group #2. (a) Geometric error. (b) Distance error. (c) Error distribution in *u*-axis. (d) Error distribution in *v*-axis ……. 91
5.7 Estimation of the angle ($\beta$) of the objects approximation w.r.t. the image plane. (a) RGB image and control points (sample 31). (b) Plane model of the pattern board (sample 31). (c) RGB image and control points (sample 12). (d) Quadric model of an object (sample 12) …….. 94
5.8 Normalized RMSE on RGB pixel coordinates vs the angle ($\beta$) of the object's approximation w.r.t. the image plane. (a) First group of experiments. (b) Second group of experiments ……... 95
5.9 Depth measurements acquired by the ToF camera (in meters). (a) Image sample 12. (b) Image sample 3 ……. 96
5.10 Depth map registration of sample 12 with depth-dependent Hlut approach. (a) Homography labelled mask, where each

colour represents a homography of the Hlut. (b) Correspondence control points (red) and estimated points (green). (c) Low resolution colour depth map – view 1. (d) Low resolution colour depth map – view 2 ……..... 97
5.11 Depth map registration of sample 31 with standard calibration method. (a) Registered points on RGB pixel coordinates. (b) Correspondence control points (red) and estimated points (green). (c) Low resolution colour depth map – view 1. (d) Low resolution colour depth map – view 2 ……..... 98
5.12 Depth map registration of sample 31 with depth-dependent Hlut approach. (a) Homography labelled mask, where each colour represents a homography of the Hlut; (b) Correspondence control points (red) and estimated points (green); (c) Low resolution colour depth map – view 1; (d) Low resolution colour depth map – view 2 ……..... 99
5.13 Depth map registration of sample 31 with standard calibration method. (a) Registered points on RGB pixel coordinates. (b) Correspondence control points (red) and estimated points (green). (c) Low resolution colour depth map – view 1. (d) Low resolution colour depth map – view 2 ……..... 100
5.14 High resolution colour depth map reconstruction. (a) Two volumetric objects placed at different distances from the sensory system (sample 19). (b) An object with a large relief with respect to the image extent (sample 8). (c) A curved object (sample 12). (d) A continuous surface which is slanted with respect to the cameras axis (sample 31) …… 102
5.15 Colour depth map registration results of image samples 11, 12 and 17. (a) Standard calibration method results. (b) Depth-dependent Hlut approach results ……..... 103
5.16 Procedure for objects motion detection …….. 105
5.17 Image pair of a person's falling sequence. (a) Image sample $sample_t$. (b) Image sample $sample_{t+1}$ ……. 105
5.18 Results of the robust structure from motion algorithm implementation on RGB registered images. (a) Inliers of motion detection in $sample_t$. (b) Inliers of motion detection in $sample_{t+1}$. (c) Depth measurements of the inlier motion region of $sample_t$. (d) Depth measurements of the inlier motion region of $sample_{t+1}$ . ……... 106

5.19 Results of the robust structure from motion algorithm implementation on ToF amplitude images. (a) Inliers of motion detection in $sample_t$. (b) Inliers of motion detection in $sample_{t+1}$. (c) Depth measurements of the inlier motion region of $sample_t$. (d) Depth measurements of the inlier motion region of $sample_{t+1}$ ..... 107

5.20 Motion detection results of image pair ($sample_{t=11}$, $sample_{t=12}$). (a) Image pair of a sequence of a person's postures. (b) Inliers of motion detection in amplitude images. (c) Depth measurements of the inlier motion region ..... 109

5.21 Image registration results of the motion inliers regions of image pair ($sample_{t=11}$, $sample_{t=12}$). (a) High resolution colour depth map. (b) High resolution colour information of the inlier motion region ..... 110

5.22 Motion detection results of an image pair ($sample_{t=17}$, $sample_{t=18}$). (a) Image pair of a sequence of a person's postures. (b) Inliers of motion detection in amplitude images. (c) Depth measurements of the inlier motion region ..... 111

5.23 Image registration results of the motion inliers regions of image pair ($sample_{t=17}$, $sample_{t=18}$ ). (a) High resolution colour depth map. (b) High resolution colour information of the inlier motion region ..... 112

5.24 Close-up views of the multisensory system for fruit harvesting. (a) Multisensory rig and filter wheel. (b) Complete view of the system ..... 118

5.25 Multisensory system structure ..... 120

5.26 Structure of the multisensory system pre-processing procedure. 122

5.27 Scenes of apple orchard on the field. (a) RGB image and ToF range data. (b) RGB image and ToF range data. 123

5.28 Images of artificial apples acquired in laboratory conditions. (a) Occlusions free scene (sample 33). (b) Scene with occlusions (sample 50). 124

5.29 Results of image registration of image sample 33. (a) Depth measurements. (b) Homography labelled mask, each colour represents a homography of the Hlut. (c) Low resolution colour depth map. (d) Close-up of the high resolution colour depth map ..... 125

5.30 Results of image registration of image sample 50. (a) Depth

measurements. (b) Homography labelled mask, each colour represents a homography of the Hlut. (c) Low resolution colour depth map. (d) Close-up of the high resolution colour depth map ……. 126
5.31 Results of image registration of image sample 52. (a) Standard calibration method. (b) Depth-dependent Hlut approach …….. 127
5.32 Features extraction procedure ……. 128
5.33 Low resolution results on the feature extraction procedure of image sample 33 (see Figures 5.28(a) and 5.29) ……..... 129
5.34 Low resolution results of the feature extraction procedure on image sample 50 (see Figures 5.28(b) and 5.30) ……..... 130
5.35 Results of the feature extraction procedure of image sample 33, implemented on high resolution colour depth information …..... 131
5.36 Details of the registration data improvement in the high resolution context. (a) Guided local data grid fitting ($gridfitt^{i}_{fruit}$). (b) Guided quadric surface approximation ($quad\_approx^{i}_{fruit}$) ……... 133
5.37 Spectral images of the apple orchard. (a) 635 nm image. (b) 880 nm image …… 135
5.38 Apple orchard images. (a) RGB image. (b) Classification map. 135
5.39 ToF data of the apple orchard. (a) Amplitude image. (b) Depth measurements ……. 136
5.40 Ground truth data acquisition of an apple orchard ……. 137
5.41 Scenes of apple orchard on the field. (a) RGB image and ToF range data of scene 5 (Day 1). (b) RGB image and ToF range data of scene 12 (Day 3) ……. 138
5.42 Results of image registration of scene5 (Day 1). (a) Depth measurements. (b) Homography labelled mask, where each colour represents a homography of the Hlut. (c) Low resolution colour depth map. (d) Close-up of the high resolution colour depth map ……... 139
5.43 Results of image registration of scene12 (Day 2). (a) Depth measurements. (b) Homography labelled mask, where each colour represents a homography of the Hlut. (c) Low resolution colour depth map. (d) Close-up of the high resolution colour depth map ……... 140
5.44 Results of the feature extraction procedure of a fruit in scene 5 (Day 1). (a) Fruit position in the RGB image coordinates. (b) Guided local data grid fitting. (c) Guided quadric surface

approximation. (c) Depth and colour model of the fruit. (e) Raw depth and colour model of the fruit. ……... 141
5.45 Results of the feature extraction procedure of a fruit in scene 12 (Day 3). (a) Fruit position in the RGB image coordinates. (b) Guided local data grid fitting. (c) Guided quadric surface approximation. (c) Depth and colour model of the fruit. (e) Raw depth and colour model of the fruit ……... 142
5.46 Results of the feature extraction procedure of a fruit in scene 12 (Day 3). (a) Fruit position in the RGB image coordinates. (b) Guided local data grid fitting. (c) Guided quadric surface approximation. (c) Depth and colour model of the fruit. (e) Raw depth and colour model of the fruit ……... 144
5.47 Results of the feature extraction procedure of a fruit in scene 12 (Day 3). (a) Fruit position in the RGB image coordinates. (b) Guided local data grid fitting. (c) Guided quadric surface approximation. (c) Depth and colour model of the fruit. (e) Raw depth and colour model of the fruit ……... 145

1.1 Diagrama de flujo para el procedimiento para la fusión de imágenes…….. 160

# List of Tables


| | | |
|---|---|---|
| 3.1 | Properties of the robotic platform joints ……... | 25 |
| 3.2 | Results of the error distribution …… | 35 |
| 4.1 | Distribution of the absolute pixel error ……. | 52 |
| 4.2 | Accuracy of the standard calibration parameters …….. | 60 |
| 4.3 | Results of the error distribution: noise-free (GTD) depth values . | 66 |
| 4.4 | Results of the error distribution: raw depth measurements …….. | 66 |
| 4.5 | Results of the error distribution: filtered depth values - bilateral filtering …….. | 76 |
| 4.6 | Results of the error distribution: filtered depth values - non-local means filter ……... | 76 |
| 5.1 | Results of the error distribution …… | 93 |
| 5.2 | Results of the feature extraction error in terms mean error …….. | 132 |

# Part I: A Non-Feature Based Method for Automatic Image Registration Relying on Depth-dependent Planar Projective Transformations

# Chapter 1

# Introduction

## 1.1 Multisensor Image Fusion

Image fusion is one of the most relevant image processing operations that aims the combination of several images of a scene acquired from different and multiple sensors, and taken at different times, to provide a better understanding and representation of 3D world scenes. Image fusion has been widely applied in most of the fields where images are ought to be analysed. These fields include medical imaging (James and Dasarathy 2014, Wyawahare, Patil and Abhyankar 2009), remote sensing (Inglada and Giros 2004, Fonseca and Manjunath 1996), computer vision (Salvi et al. 2007), robotics (Hines et al. 2003, Luo, Chih-Chen and Kuo Lan 2002). Given the variety of applications (problems) and the increasing number and diversity of sensors for collecting the data, it is unlikely that a single methodology of multisensor image fusion will suit satisfactorily for all the aforementioned applications. Therefore, the selection for a fusion solution may be dependent on the specific application and on what is considered relevant information (Goshtasby and Nikolov 2007).

In a general overview, image fusion techniques can be classified in three groups of algorithms: pixel, feature and symbolic levels. Pixel-level algorithms have been extensively investigated in comparison with related works on feature-level and symbolic-level algorithms. For instance, an

extensive critical review of these algorithms is presented in (Sahu and Parsai 2012). Pixel-level methods rely on the intensity values variations and they can work either in the spatial domain as local imaging operations, or in the transform domain, becoming then into global fusion operations.

Typically, the structure for computing image fusion is composed of four steps: pre-processing (noise removal), image registration (image alignment), image fusion (pixel-level) and post-processing (classification, segmentation and features extraction). In Figure 1.1 the main steps of the image fusion procedure concept are presented.

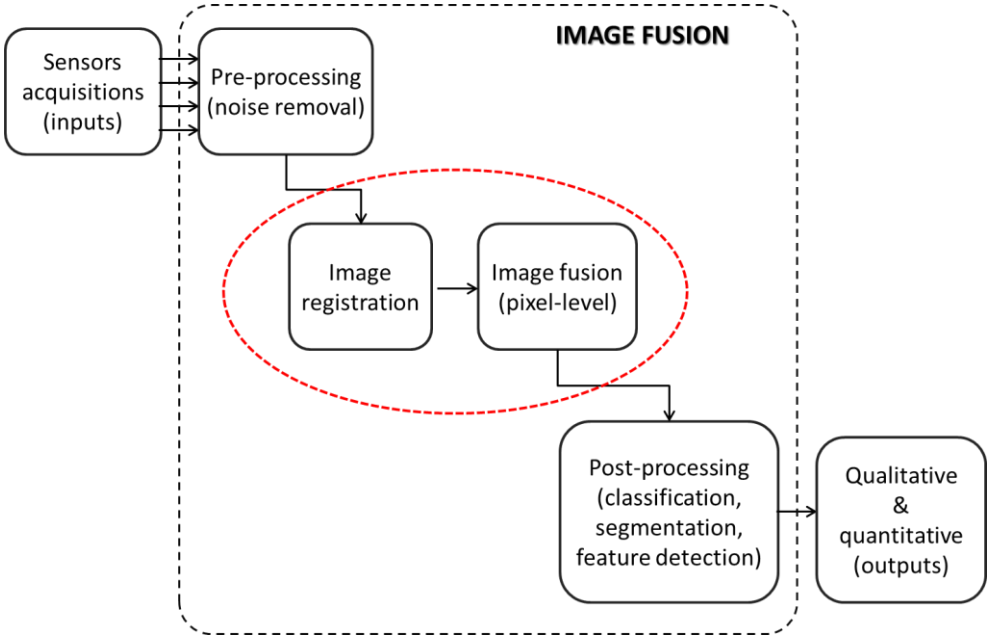



**Figure 1.1** Image fusion procedure structure.

In most cases, algorithms for image fusion procedure assume the data is perfectly aligned. Nevertheless, in practice these situations are difficult to achieve. Situations where the intrinsic and extrinsic camera parameters are not modified might provide spatially registered images (Hall and Llinas 1997). Otherwise image registration algorithms need to be previously applied. These algorithms spatially align two images by means of area-based and feature-based methods, depending on the nature of the sources and applications. Normally, area-based or feature-based image registration methods, consist of four steps: feature detection, feature matching, transform

model estimation and image resampling, and transformation (Zitová and Flusser 2003). Whether the registration is achieved by the standard calibration parameters (Deshmukh and Bhosle 2011) as if it is obtained from the (area or feature)-based approaches, the image registration is a key stage in image fusion, because the obtained errors in this process can be dragged into any fusion algorithm (pixel-level, feature-level or symbolic-level). This intricate relation between image registration and image fusion, could be addressed as a solution that considers two main types of differences the data when fusing images: spatial differences and non-spatial differences (Zhang and Blum 2001). The first is related to images with spatial misalignment caused by the geometric transformation between the views (translation, rotation, scale, etc.). The second is attributed to the environments parameters such as the variation in the illumination conditions, dynamic scenes, the use of different sources for data acquisition, and the use of different configuration parameters when using similar sensors. While the spatial differences are addressed with image registration methods, the non-spatial differences rely on image fusion approaches.

In nature, image registration is a feature-based method that depends on the robustness of the common characteristic between the image sources. In cases where these features (pixel intensities, regions similarities) are not available or difficult to compute, specific solutions should be adopted. This is the case for the combination of low resolution Time-of-Flight (ToF) and high resolution RGB cameras, where the ToF cameras are not able to detect contextual information, as is the case for the RGB cameras. ToF cameras are able to deliver low resolution amplitude-response grey-scale images and depth estimations of a scene by emitting IR light and measuring the time for the light to travel back from objects to the sensor. Typically, depth measurements are noisy, mainly because of the device hardware configuration and the external conditions. This require filtering and sensor calibration for acquiring depth measurements with confidence (Reynolds et al. 2011). Nevertheless, this image registration problem could be addressed from the standard calibration method perspective, since the internal and external cameras parameters can be used for the transformation of the depth measurements from the ToF to the RGB camera coordinates. One of the first investigation that has introduced this fusion concept was Reulke (2006). This 3D sensing approach delivers both, high resolution contextual information and the 3D structure of a 3D world scene, at rather high frame rate. This is an advantage in several robotic applications, especially the ones that need to

fulfil real-time conditions. On the other hand, this approach does not rely on correspondence features matching, as most of the 3D sensing passive techniques (stereo vision, structure from focus, motion or shape), which is an advantage in dynamically changing environments. In comparison with other active techniques, the combination of ToF and RGB cameras does no requires moving parts as laser scanner, or controlled lighting environments as in the structured light methods. A comprehensive comparison of 3D sensing methods is presented in (Sansoni, Trebeschi and Docchio 2009).

The purpose of the research presented in this Thesis is to investigate techniques that allow the combination of images acquired from ToF and RGB cameras by means of an accurate and flexible image fusion solution, and which results are also suitable for real-time applications. This solution should be able to deal with sensors which are incapable of delivering data with robust common features. Thus, this Thesis is focused on the design and implementation of image registration and image fusion methods, and its testing and validation in indoor service robot applications such as the in-house surveillance for monitoring people movements, and on the field precision agriculture applications, such as the detection and localisation of fruits for harvesting tasks. In both cases, robotic applications are constrained within near real-time conditions.

## 1.2 Motivation and Scope

In 3D imaging for robotic applications, image fusion is one of the most valuable techniques for recovering contextual and structural information of a 3D world scene. As mentioned above, two steps of the image fusion process are key issue to achieve accurate and high quality representation of 3D environments. In Figure 1.1 the procedure structure of the image fusion methodology is illustrated, as well as the aforesaid operations: the image registration and the image fusion (pixel-level) processes, which are enclosed with a red marker. The good performance of the data fusion process is conditioned by the success or failure in the image registration process Therefore, lot of effort is required to achieve accurate image spatial alignment in the registration process, which is an important part of the research of this Thesis.

For that purpose, the approach of ToF and RGB cameras combination is explored. The compromise between the high resolution colour images provided by the RGB camera and the depth measurements delivered by the ToF cameras at video frame rate, turn this solution into an approach capable

of producing information of well-defined objects (context and structure), accurate shape and edges of the objects description, which is suitable for near real-time.

There are several problems associated with the registration of these two cameras. First, because of the low resolution of the ToF cameras the computation of accurate camera calibration parameters is not always an easy task to fulfil. Even if the calibration parameters are obtained accurate enough, the noise in the depth measurements introduce errors in the data transformation process. In order to reduce the effect of noise, filtering techniques are applied, which in this case, due to the low data resolution, make the data more susceptible to over-smoothing surfaces and edges.

On the other hand, because of the large difference between the cameras resolution, only hundreds of points (144×176) in the RGB image can be initially mapped, the ones corresponded to the ToF image size. In order to take full advantage of the size of the RGB images, the study of solutions for providing high resolution colour depth maps is a relevant issue in the research of this Thesis.

Obtaining the fused results does not come to the end on the multisensor fusion procedure. Another challenge is the assessment of the fused results. For that purpose, two representative robotic applications, one indoors and the other outdoors, are selected for the image fusion methodology validation. In which the capabilities of the proposed solution for achieving specific tasks should be tested.

## 1.3 Research Objectives

The first and main objective of this research is to design, to implement and to validate an image registration method for combining a ToF and a RGB camera. The proposed solution should be capable of dealing with images with non-common features, with the noise in the depth measurements and with the large difference between the cameras resolution, and to be suitable for near real-time applications. As part of the validation stage, visual and numerical results should be evaluated. Additionally, since there are other techniques for registering ToF and RGB cameras, and in this case the most relevant is the standard calibration method. An in-depth comparison of the obtained visual and numerical results between the proposed approach and the standard calibration method should be conducted.

The second objective is to develop an experimental assessment for two selected robotic applications: in-house robotic surveillance for monitoring

and detecting people falling, and the detection and localisation of fruits on the field for harvesting robots, utilized in Precision Agriculture. In which, the capability of the proposed image registration procedure in combination with image fusion (pixel-level) algorithms should be evaluated. The goal of these evaluations is to demonstrate that the methodology proposed in this work, is able to provide accurate and quality information of the data classification, data segmentation and features extraction of the targets in successfully way. All of these, oriented to the designated tasks in the robot service application and in the Precision Agriculture application.

## 1.4 Thesis Outline

In order to address the objectives presented in this Chapter, this dissertation is organized as follows:

Chapter 2 presents the state-of-the-art of image sensor fusion techniques, which are the basis of several 3D sensing methods, and in this case, all of them oriented to the robotics filed. Among these methods, in this study a comprehensive research focused on techniques for the combination of ToF cameras and RGB cameras is addressed.

Chapter 3 is devoted to the design, implementation and validation of the proposed method for registration of ToF and RGB images. The fundamental concepts for designing the registration approach are introduced, as well as the detailed methodology for computing the depth-dependent *Hlut* approach. A preliminary accuracy evaluation of the method results is also discussed. Finally, a proposal for computing high resolution colour depth maps by means of the depth-dependent *Hlut* approach is presented.

Chapter 4 presents in-depth comparisons of the standard calibration method and the depth-dependent *Hlut* approach for image registration. For these comparisons three scenarios are considered: noise-free depth measurements (ideal), raw depth measurements and filtered depth measurements. The first input data is collected from the cameras calibration procedure. The second input data corresponds to the depth estimation acquired with the ToF camera. Lastly, for the third scenario, two filtering techniques are applied to the raw data, the bilateral filtering and the non-local means filter.

Chapter 5 presents the experimental stage, where two relevant robotic applications (indoors and outdoors) are considered. The first group of experiments is oriented to in-house surveillance and monitoring of people movements applications. In this case, the designed experiments have two

objectives: firstly, the validation of the accuracy of the method and the evaluation of the method capability for properly registering large and angled surfaces; secondly, the validation of the method in objects motions detections tasks. The second group of experiments is focused on the detection and localisation of fruits (apples) for harvesting robots is addressed. For that purpose, several experiments are conducted in laboratory and on the field conditions, and additionally, a procedure for features extraction of objects of interest is introduced. The process combines the depth-dependent *Hlut* registration approach and pixel-level techniques for image fusion.

Finally, Chapter 6 summarizes the major obtained results, the main contributions of this Thesis and the outline for future researches.

# Chapter 2

## Image Sensor Fusion – State of the Art

### 2.1 Introduction

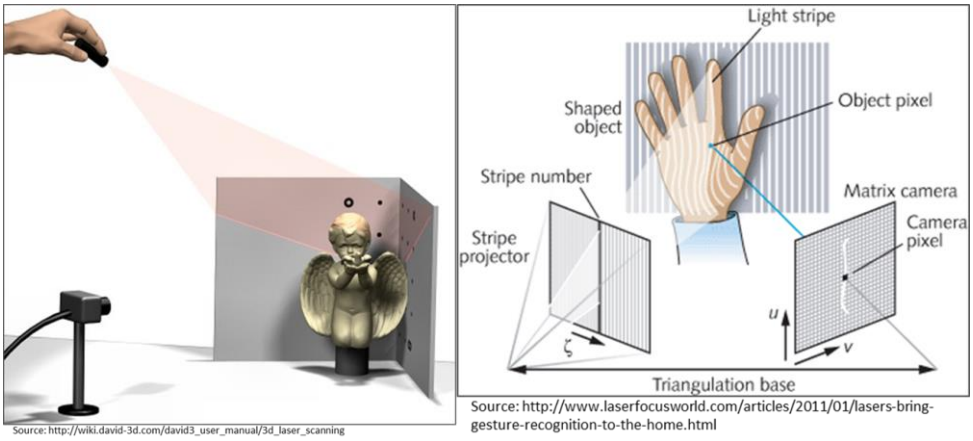
As has been previously mentioned, multisensory image fusion is one of the most relevant techniques for recovering 3D imaging of a scene, by means of combining the information of a scene acquired from multiple and different sources into a unified representation of the 3D world scene, which is more enlightening and enriching for the subsequent image processing, improving either the reliability by using the redundant information, or the capability by taking advantage of complementary information.

The acquisition of information of 3D world scenes is a fundamental stage for a worthwhile variety of applications in robotic fields such as: robot navigation (Gonzalez de Santos et al. 2007), precision agriculture (Sarig 1990, Fernández et al. 2013b), forestry (Fernández, Montes and Salinas 2015), human assistance (Salinas et al. 2011), demining activities (Ponticelli et al. 2008, Fernández et al. 2012) and many other. Typically, the acquired data is composed of information from the visible spectrum captured by CCD/CMOS cameras (Janesick et al. 1987, Litwiller 2001) and the structure of the scene, which is normally derived from the stereoscopic cameras or some kind of range sensor (Blais 2004). Nevertheless, in last decade several advances in technology and affordable prices allowed the emergence of new cameras capable of capturing not only the visible spectrum, but also the long-wave infrared (LWIR), the mid-ware infrared (MWIR), the short-wave

infrared (SWIR), and a large variety of multi-band cameras and hyperspectral systems (Govender, Chetty and Bulcock 2007, Shaw 2003).

Regarding the 3D sensing, there are several active and passive techniques for recovering the information of the 3D world, composed by colour and structure information of the scene (Bernardini and Rushmeier 2002). Some of them are based on time-of-flight (ToF) cameras, laser scanning, stereovision system and pattern projection (structured light). All of these techniques have different uses in robotics applications, as well as their advantages and weaknesses, but all of them provide more or less accurate information for reconstructing surfaces (Hebert 2000, Besl 1988, Sansoni et al. 2009).

The research of this Thesis is focused on techniques for sensor registration that provide 3D information (structure and colour) for dynamically changing environments, which means that either the robot is in motion or objects in the scene are in motion. Therefore, fast algorithms that qualify for near real-time conditions are desired. In such a scenario, laser-based methods are non-suitable solutions for real-time applications and dynamic environments (Beraldin and Gaiani 2005, Forest Collado 2004), because normally, they require moving parts to scan the scene row by row. The same is true for structured light methods, where besides the scanning of light projection (DePiero and Trivedi 1996), 3D sensing also needs to be carried out under very controlled light conditions. In Figure 2.1 some examples of these techniques are shown.




**(a)** **(b)**

**Figure 2.1** Active techniques for 3D image sensing. (**a**) Laser-scanner imaging. (**b**) Structured light imaging.

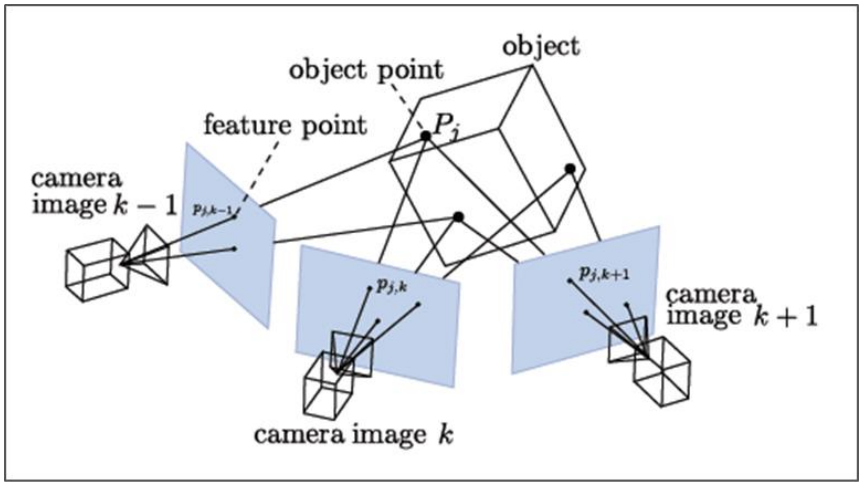
Passive camera-based methods including depth from motion, shape and focus, and stereo triangulation normally require solving the correspondence problem, Figure 2.2 illustrates both aforementioned methods. The first group of methods involves acquiring multiple images, which produces ambiguities and singularities, and also introduces additional computation load and temporal cost (Triggs 1996).

On the other hand, the stereo triangulation technique is the most common and well-known technique for acquiring 3D information (Tippetts et al. 2013). Its working principle is to determine what pair of points on two images are the corresponding projections of a same 3D point. Normally, finding correspondence features is carried out over segments instead of points, because (Ayache and Sander 1991):

- The number of features to be matched is reduced.
- The neighbourhood similarity, explicitly takes into account the continuity of contours.
- The geometric attributes of the contour provide stronger matching constraints, because discriminant properties can be derived from this valuable information.
- Since segments provide geometric attributes, the measurement of their position and orientation are usually more precise in comparison with the position of an isolated point.

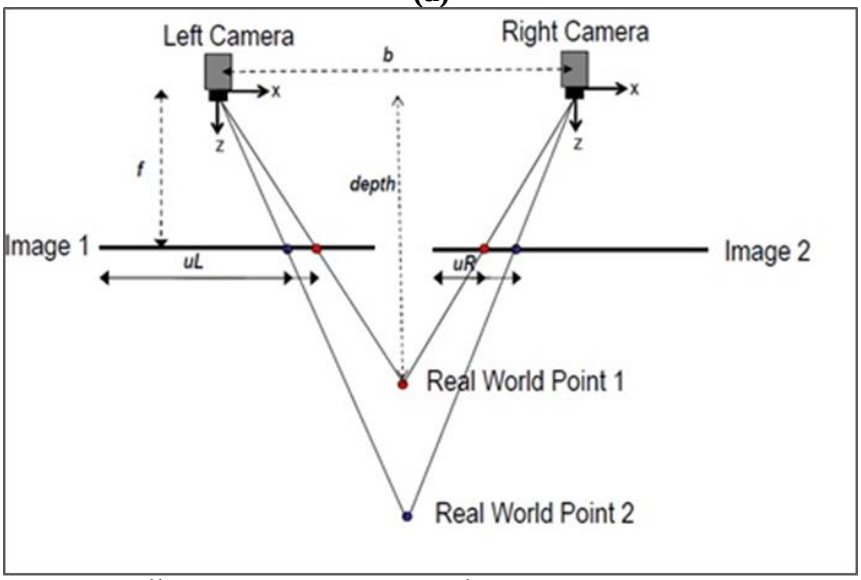
A very interesting review of passive camera-based techniques is presented by Scharstein and Szeliski (2002). Over the last three decades, significant improvements have been made for solving the correspondence problem. Nevertheless, the problems of occlusion mismatching and the incapability of matching textureless regions remain unsolved.

On the other hand, besides the use of conventional perspective stereo cameras, the 3D imaging could also be achieved with catadioptric stereo and panoramic stereo systems. In contrast with the perspective cameras, these systems provide the capability of tracking and detecting objects over large 3D environments. In the particular case of omnidirectional vision systems, since the Ress' first proposal in U.S. Patent No. 3,505,465 in 1970 (Gilvydis 1985), and later on, in the earliest 90's when these system started to been developed again by (Yagi, Nishizawa and Yachida 1995, Yamazawa, Yagi and Yachida 1993, Hong et al. 1992) , several configuration and theories of catadioptric panoramic system has been presented in order to obtain images of the entire scene (Geyer and Daniilidis 2001, Baker and Nayar 1999, Svoboda and Pajdla 2002).




Source: http://openmvg.readthedocs.org/en/latest/openMVG/sfm/sfm/

**(a)**




Source: http://www.depthbiomechanics.co.uk/?p=102

**(b)**

**Figure 2.2** Passive techniques for 3D image sensing. (**a**) Structure from motion. (**b**) Stereovision.

In comparison with conventional cameras, the greatest advantage of the panoramic systems relies on their capability for acquiring wide range view

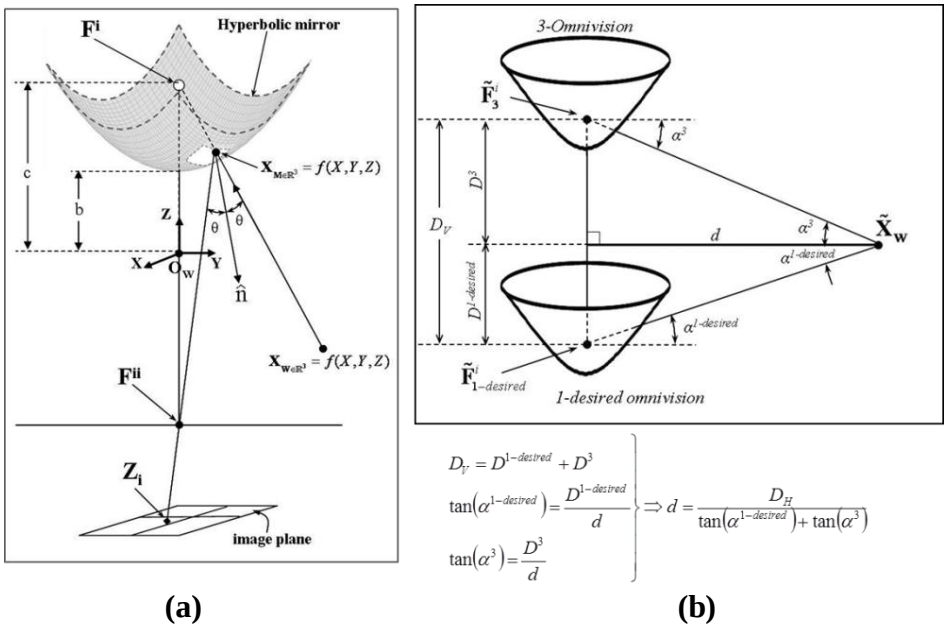
images, which allows the robotic systems to have a better perception of the environments for tasks such as navigation, tracking of objects and ego-motion detection, since the objects disappears later on the images.

Common configurations for panoramic systems include rotating cameras, multiple cameras, or catadioptric systems for obtaining images of 360 º of a scene. However the first approach brings in mechanical problems as the movement of heavy parts, the manufacture costs and, the rotation mechanisms are not suitable for real-time applications, and also to achieve accurate positioning extra efforts are required. Multiple cameras present a high computing cost to form a single panoramic image.

On the contrary, catadioptric systems, resulting from the combination of refracting (dioptric) and reflecting (catoptrics) surfaces, are considered as very interesting solutions. These systems are easily built employing a conventional high-resolution camera as the refracting part and a curved mirror as the reflecting one. In order to acquire a single image containing the information of the whole scene, the camera and the mirror must be arranged in a configuration such that the entire system has a single effective viewpoint (Baker and Nayar 1999), named as central catadioptric cameras (Svoboda and Pajdla 2002). In order to generate omnidirectional images, only perfect quadrics surfaces are considered it is considered perfect quadrics surfaces only as candidates for mirror shapes. In this way, every incident ray of light that strikes a surface toward the mirror focus is reflected to the second focus. Since the geometry of the system is known it is possible to compute the ray direction for each pixel and its irradiance value.

In the literature, there are several configurations for achieving omnidirectional stereovision systems. The general theory of epipolar geometry for central catadioptric stereo cameras was depicted in (Svoboda and Pajdla 2002). A rectified systems in was presented in (Gluckman, Nayar and Thoresz 1998) where two omnidirectional systems were placed vertically aligned, one on top of the other. A special double lobbed mirror was introduced in (Cabral, de Souza and Hunold 2004, Nene and Nayar 1998),and in (Nene and Nayar 1998), the use of two curved mirror with a single camera was proposed. The lack of high-resolution of the last two configurations makes them less interesting. Among the possible configurations of stereo systems, the rectified configuration is a more desirable solution, because their epipolar lines corresponds to the radial axis of the omnidirectional image, hence the computation of disparity is simplified. In Figure 2.3 the data acquisition structure and the rectified configuration are illustrated. Although,

the stereo rectified omnidirectional vision systems are a promising and suitable solution for a large number of robotics applications and, the occlusion problem might be reduced because of its wide range of view, these systems still rely on conventional stereo passive methods for recovering the 3D information. Thus, omnidirectional stereovision systems are also incapable of dealing with textureless regions.



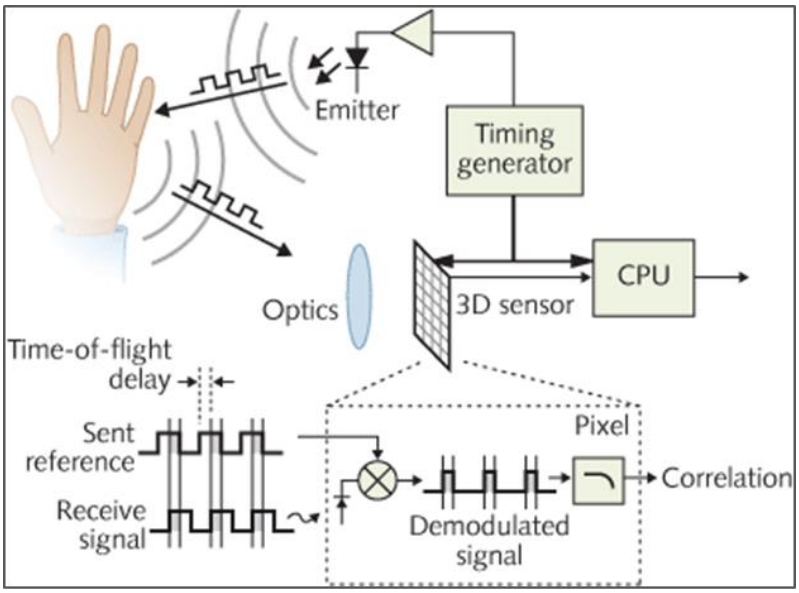

**(a)** **(b)**

**Figure 2.3** Omnidirectional systems. (**a**) Catadioptric image formation. (**b**) Rectified catadioptric stereovision configuration.

Alternatively, ToF cameras are becoming more and more popular, less expensive and more powerful. As mentioned before, these cameras estimate the depth by emitting a modulated light and observing the reflected light. Then, the phase shift between the emitted and reflected light is measured and translated to distance (Ringbeck 2007). The emitted light could be a pulse or continuous wave (CW). Most of the available cameras use the sinusoidal or square CW modulation, and use the demodulation lock-in pixels. For the demodulation, the "four-bucket" technique is usually adopted, where each pixel samples the amount of light reflected by the scene on every measurement, and four samples per measurement are taken, each sample phase is stepped by $\pi/2$ (Foix, Alenya and Torras 2011). Typically, the light

source is generated by a solid-state laser or a light-emitting diodes (LED) operating in the near-infrared spectrum (NIR).

In Figure 2.4 the ToF cameras working principle is shown and in Figure 2.5, some of the most frequently used ToF cameras are displayed. Some interesting works have evaluated these cameras, and they have shown their advantages in certain fields. In conclusion, the most relevant attribute of these systems is their capability of delivering simultaneously depth maps and intensity images at a video frame rate. However, their spatial resolution is very low, not more than thousands of pixels are provided, and they tend to be noisy and poorly calibrated.

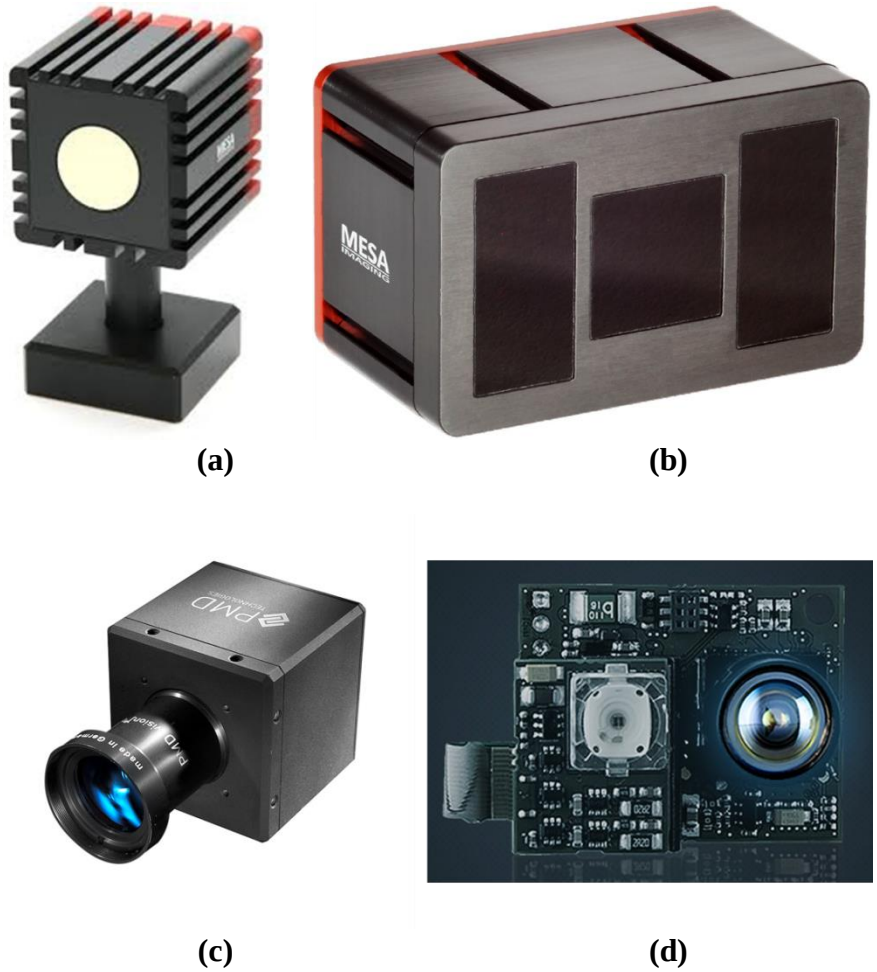



Source: http://www.laserfocusworld.com/articles/2011/01/lasers-bring-gesture-recognition-to-the-home.html

**Figure 2.4** ToF camera working principle.

In (Chiabrando et al. 2009, MESA Imaging 2011) the authors presented a methodology to reduce the errors in the depth measurements of the SR4000 camera (MESA Imaging 2011). They modelled the measurement errors with a sinusoidal approximation and calibrated the intrinsic camera parameters. A most extensive evaluation of the ToF cameras was presented by (Foix et al. 2011). This work shows the potential of these systems, but due to their limited resolution they conclude that previous technologies are still leading

the 3D sensing field. However, the combination of ToF and colour cameras has shown great improvements to compensate for this lack of resolution. A comparative study between the stereo vision systems and ToF cameras is beyond the scope of this research, and interested readers may refer to (Beder, Bartczak and Koch 2007).




**(a)** **(b)**

**(c)** **(d)**

**Figure 2.5** Some of the most frequently used ToF cameras and commercially available. (**a**) MESA SwissRanger SR4000; (**b**) MESA SwissRanger SR 4500; (**c**) PMD CamCube; (**d**) PMD CamBoard nano (currently Sold Out).

## 2.2 Combination of RGB and ToF Cameras

The approaches for combining ToF and colour cameras are commonly presented in two configurations: the monocular setting which combines a single colour camera and a ToF camera, and the coupling of a stereo vision system and a ToF camera. In this work a monocular setup is adopted. In general, the fusion of these two systems is addressed by computing the extrinsic parameters for the homogenous transformation between them, which means that the method efficiency relies on the cameras calibration and the accuracy of the depth measurements of the ToF camera. In the stereo configuration, the 3D-3D correspondences are used to estimate the transformation between the two systems. Some approaches utilize the depth from the ToF as a constraint to compute the stereo matching. In (Guðmundsson, Aanæs and Larsen 2008, Hahne and Alexa 2008), the authors use fast algorithms and inaccurate extrinsic parameters to improve the disparity computation, and the results show that sensor fusion is possible. However, when upscaling the dense maps to the colour image size, some problems at the objects' edges are reported, and only visual results are reported. On the other hand, very interesting results are described in (Zhu et al. 2008), where the authors calibrate the system within a range of 400 mm and use the depth values as an additional observed variable in the global approximation function. In this case, the method was tested in real scenes and numerical results report mean errors within 2–3 pixels. Nevertheless, this approach assumes a global regularization method for stereo matching, which normally is not fast enough for real time applications (Scharstein and Szeliski 2002). Regardless, in this configuration the most important drawbacks of the stereo system, which are the occlusions and textureless regions, remain unsolved.

In the monocular case, the depth information is used during the calibration process to back-project the 3D points into the 2D points of the RGB image. Normally, in related works, orthogonal generation is applied for the cameras frames co-alignment (Reulke 2006). However, other researches adopted projective texturing. In this case, the RGB camera is projected onto the ToF camera projective geometry. Unfortunately, in both cases, only few works present numerical results of their methods implementation (Linarth et al. 2007, Park et al. 2011). As it was mentioned in (Foix et al. 2011), the challenging issue is how to handle the difference between the cameras' resolutions, because between each pair of nearby points of the ToF image there are several points of

the colour image. Therefore, the complexity of this work lies in the upsampling techniques for computing high-resolution depth maps without losing the colour information and achieving near real-time processing conditions. Most of the related works upscale depth maps of up to 1.5 Megapixels by means of bilinear or bicubic interpolation (Van den Bergh and Van Gool 2011). However, the proposal of this Thesis deals with high resolution colour dense maps of 5 Megapixels. Other approaches are concerned with improving the quality of the high resolution depth maps. For instance, in (Lindner, Lambers and Kolb 2008), the authors present an interpolation algorithm for edge enhancement that uses the gradient and Laplacian to adjust the sampling location, but only visual results of a single object scene are presented. Remarkable efforts have been made in (Park et al. 2011) to create high-accuracy depth maps, where outlier detection was addressed as a minimization function of the Mark Random field. Then through a robust optimization function that combines several factors, namely the data, the neighbourhood smoothness and the non-local mean regularization, depth fusion was achieved. Their results stand out from other algorithms, but the complexity of the method makes it unreliable for real-time applications. The authors report a computation time for real-world scenes of 19.00 s. A similar case is presented in (Huhle et al. 2010), where a Graphic processing unit (GPU) implementation for parallel computation is adopted, with the aim of implementing a denoising and enhancement filter based on non-local means formulation. In this case it takes nearly 2 s to complete the processing.

As it was mentioned above, the ToF and RGB sensor fusion relies on the extrinsic calibration and the depth estimations from the ToF camera. The depth information is noisy and because of the ToF camera's low resolution, the extrinsic parameters are inaccurate. In some cases it is possible to achieve good results without accurate extrinsic calibration as it is shown in (Hahne and Alexa 2008). Other works report some simplification when sensors are mounted in particular configurations (Van den Bergh and Van Gool 2011, Guðmundsson et al. 2010, Song 2011, Hahne 2009). The typical noise of the depth measurements can be modelled as a Poisson distribution around the true value. However, the artifacts derived from the object's albedo are not easy to model. Most of the related works addressed the problem by applying filtering techniques to the depth measurements. Nevertheless, the filtering can often over-smooth the interpolated data, significantly affecting the depth discontinuities of the boundaries. As it is shown in (Chan et al. 2008), the noise aware filtering for depth map registration improved the quality of the results, however the misalignments problems derived from an noise-sensitive

registration technique, could introduce error in the results, which are difficult and computationally expensive to remove.

In the search of a method for registering sensors that deliver data with non-common features, and that additionally be capable of addressing biased depth measurements, the proposal in this work undertakes the idea of working with uncalibrated methods for automatic data registration, which has not been studied yet.

The distinction of the uncalibrated methods is that they do not need to know, at first, the internal and external parameters of the cameras. This may normally lead to a system capable of achieving up to projective reconstruction. Nevertheless, the theory introduced by Hartley and Zisserman (Hartley and Zisserman 2003), regarding multiple view geometry, demonstrates the possibility of achieving both affine and Euclidean reconstruction with no previous knowledge of the camera calibration matrix.

Regarding projective geometry, there are two relations between two views (cameras) and a scene plane. The first relation is the epipolar geometry, which represents the intersection of a pencil with two image planes, where the axis of the pencil is a line joining the cameras' centres, denoted as the baseline. The intersection of the baseline with the image planes are the epipoles $(e, e')$. Given that information, it is possible to back-project an image point $x$ on image 1 to a ray in the 3-space. The ray passes through the camera centre, the point $x$ and the 3-space point $X$, which is on one plane of the pencil. This ray is projected onto image 2 as a line that intersects its epipole $(e')$. Then, the problem of finding a correspondence for $x$, which is the projection of $X$ on image 2, is reduced to a search on a line. The second relation is given through the plane projective transformation, which is the relation of image points on a plane in a view to the corresponding image points in a second view by a planar homography, $x_2 = Hx_1$.

Consequently, when considering the search for corresponding points in a 3D-space scene, epipolar geometry is the straightforward solution to reckon with. Nonetheless, the problems of feature matching based on images are very well known. Most of the problems arise from the occlusions and the changes in the illumination conditions, and all of them contribute to non-matched or wrongly matched features. Some works have presented solutions for these problems, such as the method introduced by Sagüés (Sagüés 2006), where the author proposes matching lines between images instead of matching points to compute the fundamental matrix *F*. However, the problem of finding control

points between data acquired by different sources with non-common or robust enough features is still an unconsidered field.

In some special cases, a scene is considered as a planar scene. Such a case may possibly be produced when the images baseline is null or the depth relief of the scene is small compared with the extent of the image. In both cases, epipolar geometry is not defined because the epipoles are not accessible and the plane projective transformation is the exact solution to transfer points from one view to another. However, this solution should not be taken as a general rule, because most of the scenes in the man-made environments usually comprise several planes.

On the other hand, it has been demonstrated that the homography induced by a plane $\pi = (v^T, d)^T$ is determined uniquely by the plane and vice versa, only if the plane does not contain any of the cameras' centres; otherwise, the homography is degenerated. Suppose the system is a sensory rig set-up; then, the homography matrix is (Hartley and Zisserman 2003):

$$H = K'(R - tv^T/d)K^{-1} \tag{2.1}$$

The homography matrix is defined by the camera internal ($K$) and external parameters ($[R, t]$) and the plane $\pi = (v^T, d)^T$. Since the camera parameters are constant, the result in Equation (1) also shows that a family of homographies is parametrized by $v/d$, where $d/\|v\|$ is the distance of the plane from the origin.

Let us assume that a 3D scene reinterpretation is possible by discretizing the scene into *n*-planes. Then, it is also possible to compute *n*-homographies, and transfer the image points from the first view to image points of the second view. Taking advantage of the depth information provided by the ToF camera it would also be possible to compute the approximation of the object planes of the scene. However, such approximation should not be done lightly, because some planes may generate a virtual parallax.

Now, let us suppose that a scene contains two objects; one is represented by a plane angled to the cameras' views, and the other by a plane in front and parallel to the cameras. Then, the homography induced by the second object (in the front plane) maps incorrectly the points off this plane, in this case the first object. Nevertheless, if the intersection of these two planes is in the cameras' views, the points of the intersected line could be properly mapped. Now, instead of using the homography of the first plane (angled object) to transfer it, let us suppose that this angled object is virtually intersected by *m*-planes, all positioned at different distances in front and parallel to the

cameras view. Then, there are *m*-lines as a result of these intersections. These *m*-lines describe a discrete shape of the object. Hence, each homography induced by these virtual *m*-planes is able to map its corresponding intersection (*m*-line) on the angled object. This assumption implies that objects into a scene could be explained with a family of virtual *m*-planes, and their induced *m*-homographies are able to map the discrete object's shape. This homography family only depends on the planes parameters and the distance of the planes to the cameras, similar to Equation (2.1). However, in this case, the planes do not directly represent the planes on the scenes; they are virtual planes, positioned in front and parallel to the sensory system.

# Chapter 3

## Depth-dependent Homography Lookup Table for Dense Map Registration

### 3.1 Introduction

Commonly, registration methods aim the geometrical alignment of two (or more) images of the same scene by means of a feature-based method. These images might be acquired by different sensors, from different views or taken at different times. An extensive review is presented by Zitová and Flusser (2003).

The primary goal of the research of this Thesis is the generation of high resolution colour depth maps under real time conditions by using the data acquired by ToF and RGB cameras. In the sensory rig configuration composed of ToF and RGB cameras, finding robust features between depth maps and colour information is not feasible, and only artificial landmarks might be matched properly. However, since it is desirable that the method should works under natural conditions, landmarks are not the proper solution.

Normally, depth map registration is done by computing the extrinsic parameters of the coordinate transformation between the two cameras. The 3D points from the available depth measurements are back-projected to the colour image and a low resolution depth colour map is obtained (Foix et al. 2011). As it was mentioned in Chapter 2, some works have been dedicated to increase the resolution of the dense map by upsampling techniques (Park et

al. 2011, Zhu et al. 2008, Huhle et al. 2010). However, most of them are not suitable for near real-time applications. Other methods have adopted interpolation algorithms for the depth upsampling before transferring the data (Lindner et al. 2008), though the maximum dense map size reported is 1.5 Megapixels. Some works have also reported difficulties on the objects edges, being this problem mainly produced by noisy depth measurements or by the over-smoothing of depth values, caused by the data interpolation. For instance, in (Chan et al. 2008), a noise-aware filter for a colour depth map upsampling was proposed, and improved quality maps were obtained. However, after the proposed enhancement processing, some blurred regions and artifacts remained on the data, mainly because of their erroneous alignment procedure.

The approach proposed in this work relies on uncalibrated techniques for transferring points from one view to another. Normally, uncalibrated techniques are based on the epipolar geometry, which is a feature-based solution for computing correspondences of 3D-space points between two views. Nevertheless, in similarity with registration methods, matching robust features between depth and colour information is not achievable. On the contrary, planar projective transformation does not require the search of features once the homography is computed. In some cases, the scenes might be considered as planar scenes, but most of natural scenes consist of several planes and the objects into the scene are considered non-planar objects. In consequence, multiple homographies describe the correspondence between views, which is the foundation of the proposed non-common features registration method based on planar projection transformation.

The proposed sensory system for the data acquisition consists of a high resolution colour camera and a 3D ToF camera. The ToF camera of the system is the SR4 Mesa SwissRanger (MESA Imaging 2011) with a resolution of 176x144 pixels and a frame rate up to 30 fps. The ToF camera provides three images: the amplitude response, the confidence map and the depth map. The depth map could also be converted to XYZ Cartesian coordinate data, with the origin of the coordinated system in the centre front of the camera, with *Z* coordinate increasing along the optical axis away from the camera, *Y* coordinate increasing vertically upwards and *X* coordinate increasing horizontally to the left (see Figure 3.1). For the RGB camera, the AVT Prosilica GC 2450 (Allied Vision 2011) was used. The camera resolution is 2448×2050 pixels and its frame rate is up to 15 fps.

The cameras are vertically aligned and placed as close as possible to each other. The sensory system is mounted on a four degrees of freedom robotic platform. This platform consists of two prismatic joints and a pan-tilt unit. The prismatic joints provide the vertical and horizontal movements in the *XZ* Cartesian plane. The rotational joints on the pan-tilt unit provide the pitch and yaw movements of the system (Montes et al. 2012). The joints properties of the platform are described in Table 3.1. The system configuration is depicted in Figure 3.1.

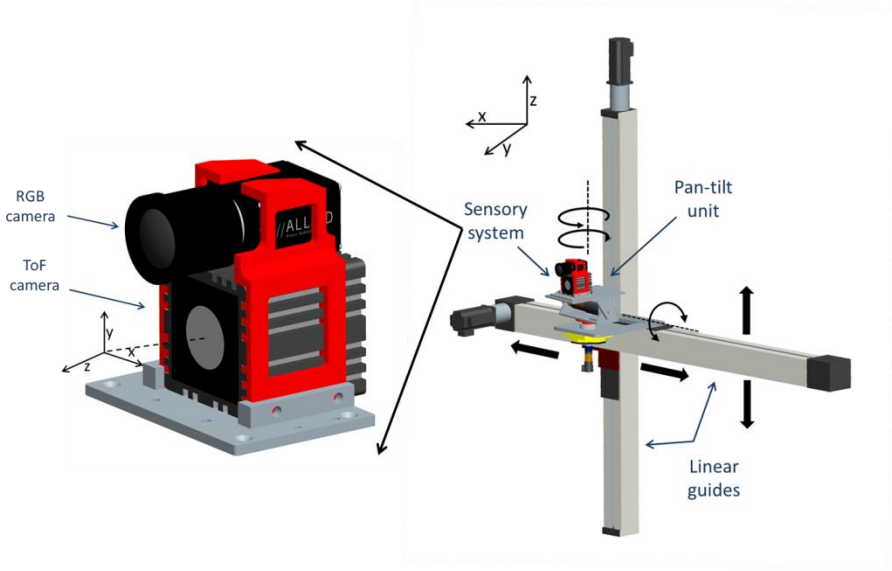



**Figure 3.1** Sensory system configuration. The sensory rig consists of a ToF camera and a RGB camera, and it is mounted on a robotic platform with four degrees of freedom.

**Table 3.1** Properties of the robotic platform joints.

| Joints/axis | Max. velocity | Operating ranges | accuracy |
|---|---|---|---|
| Vertical | 500 mm/s | ± 700 mm | ± 0.1 mm |
| Horizontal | 500 mm/s | ± 700 mm | ± 0.1 mm |
| Pitch | 40 rpm | ± 30° | ± 0.0012 º |
| Yaw | 81 rpm | ± 360° | ± 0.00243 º |

## 3.2 Method Description

The proposal approach is inspired on the uncalibrated techniques for transferring data between two views and the search of non- feature-based methods for matching correspondence points. For instance, the epipolar geometry is the most extended uncalibrated technique, but it is the feature-based solution. On the contrary, planar projective transformation does not require features matching after the homography is computed. The planar projective transformation assumes the transformation has been done within two views and a plane into the scene. In some cases, the scenes might be considered as planar scenes, but most of natural scenes consist of several planes and the objects into the scene are considered non-planar objects.

Let us assume that $H_\pi$ is the homography induced by the plane π. Then, suppose that when mapping 3D-space points between the two views, some of these points are off the plane π. In such a case, the homography generates a virtual parallax; a schematic illustration of this assumption is displayed in Figure 3.2. The 3D point $X$ is off the plane π, thus the ray through $X$ intersects π at some point $X_\pi$. These two 3D points are coincident in the first view at point $x$, but in the second view, the images of $X$ and $X_\pi$ are not coincident. The vector between $\hat{x}'$ and $x'$ is the parallax relative to $H_\pi$.

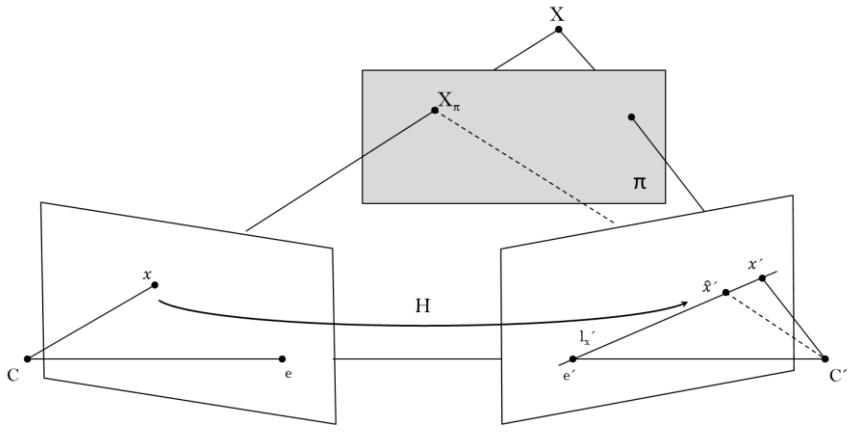



**Figure 3.2** Plane induced parallax.

Assume from the scene above that the two points $X_1 = X_\pi$ and $X_2 = X$ are on plane $\pi_1$ and $\pi_2$, respectively, and $H_{\pi_1}$ and $H_{\pi_2}$ are homographies induced

by the corresponding planes. If the ray through each 3D-space point is not coincident neither in the first view nor in the second view, then the images of the points are $x'_1 = H_{\pi_1} x_1$ and $x'_2 = H_{\pi_2} x_2$ (see Figure 3.3).

Along this idea, suppose that scenes composed by *n*-objects could be approximated to *n*-planes and consequently *n*-homographies could be computed. This assumption should be prudently considered, because objects with large relief or positioned closed to the sensory system, certainly are explained with more than one plane. Under these circumstances, a unique homography approximation of an object also generates a virtual parallax. In this case, let assume the object is virtually intersected by *m*-planes, all positioned in front and parallel to the cameras. Then, each of these intersections generates *m*-silhouettes of the object shape. Hence, each homography induced by these virtual *m*-planes is able to map its corresponding intersection, the *m*-silhouettes of the object.

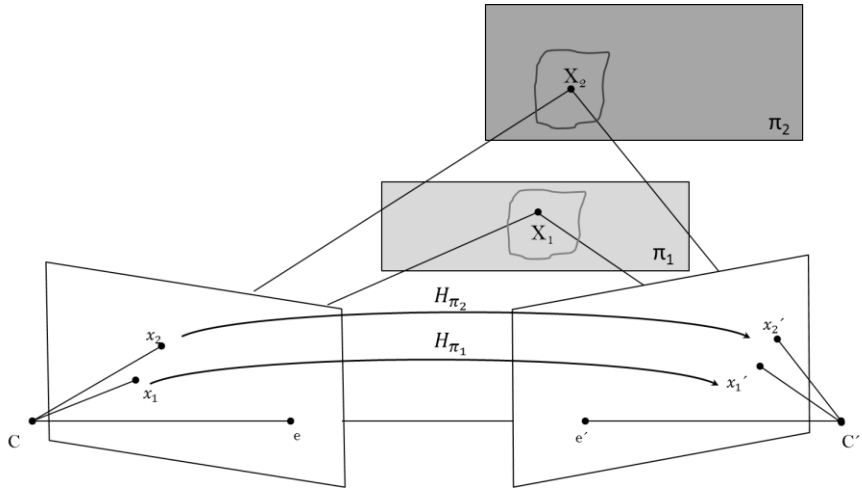



**Figure 3.3** Plane projective transformation induced by two planes $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$ on a scene.

The approach of this work proposes an alternative 3D world parametrization by virtually discretizing it into *n*-planes and thus, computing a depth-dependent homography lookup table. These *n*-planes are parallel to the sensory system and sequentially positioned in front of it. Taking advantage of the depth information available from the ToF camera, the distance of each *n*-plane from the camera is known. A 3D-space plane ($\tilde{\pi}_i$) in

the discretizing process is represented as a volumetric unit. This unit is composed by 3D points hold within a depth differential section, denoted as differential of depth of a plane ($\Delta dop$). The dimension of $\Delta dop_i$ is directly proportional to the distance from the plane $\tilde{\pi}_i$ to the sensory system. For instance, the closer the object is to the vision system, the larger the object relief is in comparison with the extent of the image, the higher the number of *n*-planes is for explaining the object and the smaller the $\Delta dop_i$ of each *i*-plane is. Henceforth, the matching feature for the image registration method is the distance from 3D-space points to the ToF camera ($d_i$). A plane $\tilde{\pi}_i$ is approximated from a cluster of 3D points if and only if, its induced homography maps their images points from one view to another within errors less than 3 pixels on the RGB frame. The distance between planes ($\Delta dbp$) should be approximately equal to zero. Figure 3.4 shows an illustration of the discretizing process and the depth-dependent homography lookup table formation.

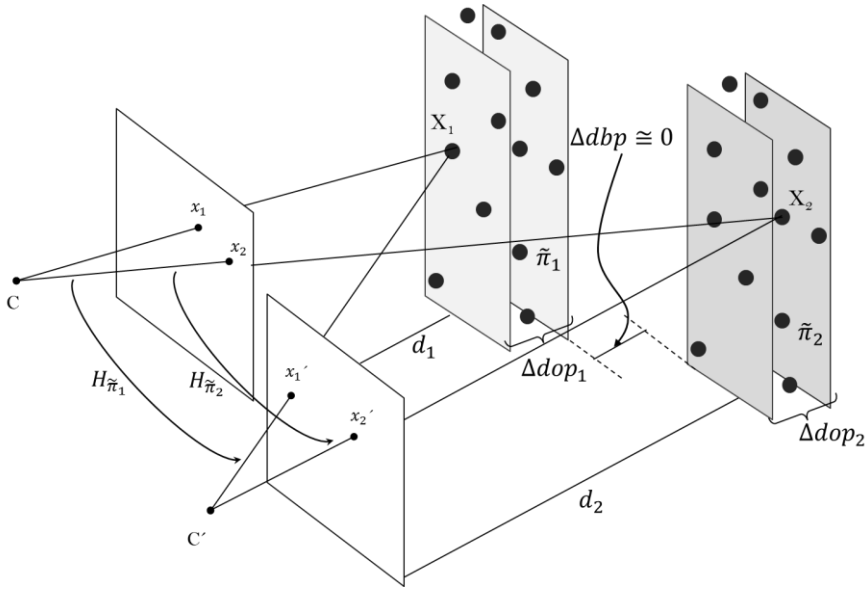



**Figure 3.4** Formation of the depth-dependent homography lookup table.

In terms of mathematic formulation, in planar projective transformation, every possible virtual plane *i*-plane of the ToF depth measurements $\{d_i\}$ of a scene should induce one homography $\{H_i\}$. Nevertheless, it has been proven that under certain configurations of the scene or the sensory rig, a single

homography is an exact transformation of the 3D scene (Hartley and Zisserman 2003). This is the case of planar scenes or for small (null) baseline images pairs. When considering planar scenes, it does not denote a scene composed with planar objects; it refers a scene where the objects have small relief compared with the extent of the image. On the other hand, as it is pointed out by Sagüés (2006), this condition is not strictly necessary when considering image pair of nearly null or small baseline, where a 3D-space scene might be explained by means of a single homography. These two concepts are the constraints for the proposed method of this work when computing the entries of the homographies lookup table. Thus, the dimension of the range of depth clusters for each homography $\{H_i\}$ on the *Hlut* is in essence, the representation of these two constraints.

For computing the depth-dependent *Hlut*, 104 images of a known pattern grid were captured. In order to avoid unreliable depth measures because of dark objects, the pattern grid is a 3 × 5 white-red chessboard with squares of 50 mm of side. The effective pattern is the inner 2x3 grid, thus, the 12 control points on the board $\{X_i\}$ are 12 image control points on each view $\mathrm{C}\{xg_j^{ToF}\} \leftrightarrow \{xg_j^{RGB}\}; j = 1 \dots M, M = 12.$ Then $\{xcp_i^{ToF}\}$ and $\{xcp_i^{RGB}\}; i = 1 \dots N,\ N = 104$ are $N$ samples of the 12 grid points, where $xcp_i^{ToF} \ni \{xg_j^{ToF}\}\ and\ \ xcp_i^{RGB} \ni \{xg_j^{RGB}\}$. These points are extracted from RGB images and grayscale amplitude images, these last ones provided by the ToF camera. From this point forward, when referring to ground control points, it is assumed that it is referred to $xcp_i^{ToF}\ and\ \ xcp_i^{RGB}$.

The board was positioned at several distances in front of the sensory system and approximately parallel to it. The 104 image samples are different poses of the pattern board, where the pattern was sequentially positioned at distances from 400 mm to 2300 mm from the board to the sensory system. The distance from the pattern to the sensory system is calculated by using the depth information enclosed in the inner 2 × 3 grid. This region is extracted for computing the mean depth and subsequently, the distances $d_i$ from the board to the system. An example of image pairs from the RGB image and the ToF amplitude, and their ground control correspondence points are shown in Figure 3.5(a). The 3D view of the region enclosed in the inner grid of the board is displayed in Figure 3.5(b). Notice that both images and the depth information have been previously undistorted before extracting the ground control points.

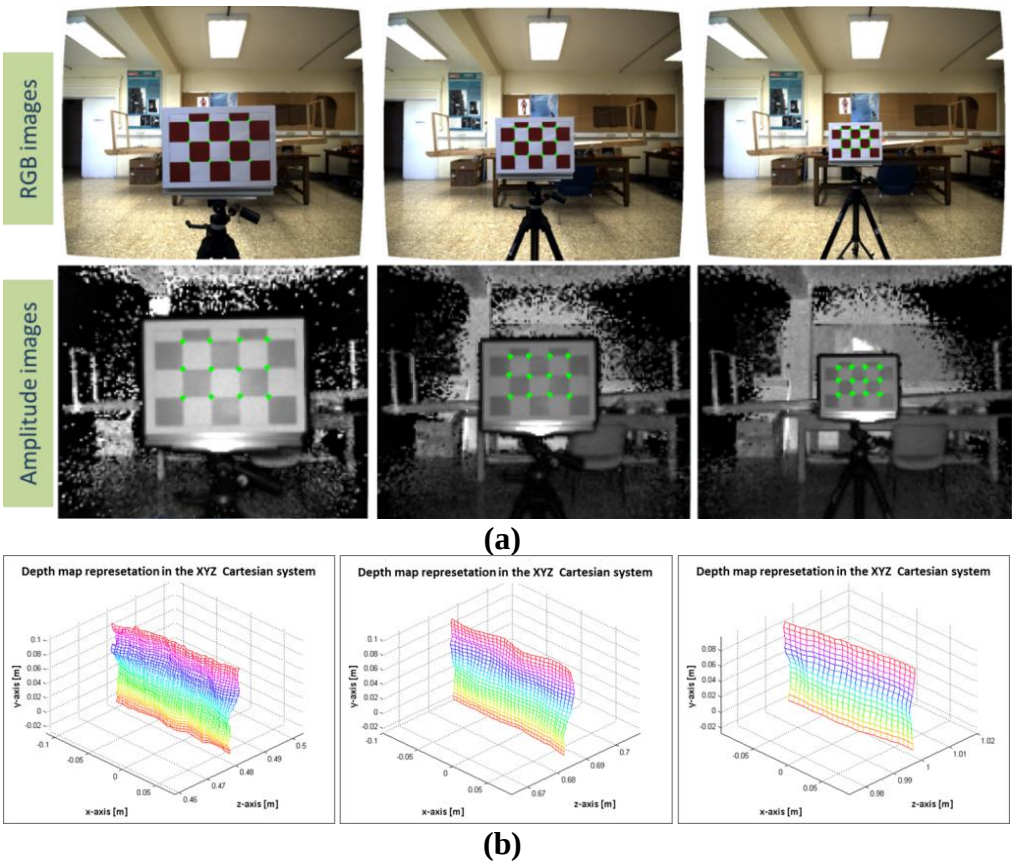



**Figure 3.5** Samples of images pair of the pattern grid board. (**a**) RGB and ToF amplitude images. (**b**) The depth map representation in the Cartesian system of the inner 2 × 3 grid.

The homography computation was carried out by following the normalized direct linear transform (DLT) algorithm and the RANSAC method for robust model approximation in presence of outliers (Hartley and Zisserman 2003, Torr and Zisserman 2000). The initial step was the calculation of sets of homographies by gathering combinations of ground control points of the $N$ image sample $\{xcp_i^{ToF}\}$ $and$ $\{xcp_i^{RGB}\}; i = 1 \dots N, N = 104$. Only homographies capable of mapping points within absolute geometric error per point less than 2 pixels on any image axis ($u$, $v$) are selected. However, in order to avoid error misleading because of the outliers on the selection of the grid control points, the overall absolute error is used. The overall error is computed with the sum of the error points, and since

the number of grid points is 12 and the points has two image axis (*u*, *v*), the maximum overall error is 48. This error is measured when mapping points from the ToF to the RGB frame and it is the absolute difference between the estimated points and the ground control points. The sets of points mapped are denoted as $\{xmap_i^{ToF}\}\ and\ \{xmap_i^{RGB}\}$ and the grid points as $\{xgm_i^{ToF}\}\ and\ \{xgm_i^{RGB}\}$, respectively. Thus, the absolute error of the estimated sample is:

$$\epsilon = \left|xcp_i^{RGB} - xmap_i^{RGB}\right|$$
$$\xRightarrow{overall} \sum_{j=1}^{12} \left|xg_j^{RGB} - xgm_j^{RGB}\right| < 48 \tag{3.1}$$

The next step is to remove the duplicated homographies, which involves the utilization of the transformation matrices computed by the same combination of control points $\{xcp_i^{ToF/RGB}\}$. At this stage, a list of potential homographies is achieved. Each of them is related to a list of its properly mapped samples $\left\{xcp_i^{ToF/RGB} \underset{H}{\rightarrow} xmap_i^{ToF/RGB}\right\}$.

Since the distance $d_i$ from the board to the sensory system at any sample is known, a list of minimum and maximum working distances related to each homography is created such that $(dmin_i, dmax_i)$. The final step of the procedure is the selection of the optimal entries for the homography lookup table such that $Hlut \ni H_i^{lut}, dmin_i^{lut}, dmax_i^{lut}\ \ i = 1 \cdots numH$. For that purpose, some conditions should be satisfied. The *Hlut* should cover the entire depth of field [300–2300 mm] of the parametrized 3D world. The distance between homographies ($\Delta dbp \cong 0$) should be approximately equal to zero, and the number of entries on the LUT should be as minimal as possible. Algorithm 1 shows the pseudocode for computing the depth-dependent *Hlut* by using the ToF and the RGB cameras.

**Algorithm 1** Automatic estimation of the depth-dependent homography lookup table

Objective: given $N$ samples of sets of 2D to 2D correspondence points $\{x_i\} \leftrightarrow \{x'_i\}$, compute a depth-dependent homography lookup table $Hlut = \{H_k^{lut}\}$ such that $x'_i = H_k^{lut} x_i$. These sets are the projected image points of the 3D-space points $X_i$, which are distributed at several distances from the system, and parallel to it.

1: Acquire ToF and RGB images pairs of $N$ different poses of a known pattern, where $numSamples = \{1 \cdots N\}$. The pattern is a white-red chessboard sequentially positioned in front and approximately parallel to the sensory system.

2: Extract the $M$ grid correspondence points of each image sample from the previous step ($i$) to compose the $N$ sets $x_i \leftrightarrow x'_i$.

3: Apply the DLT algorithm to compute homographies by combining sets of the 2D to 2D correspondence points such that $xx'_j = H_k xx_j$, where $x_a \cup \cdots \cup x_n = \{xx_j : xx_j \in x_g \; where \; a \leq g \leq n \; and \; a, n \in numSamples\}$ and $x'_a \cup \cdots \cup x'_n = \{xx'_j : xx'_j \in x'_g \; where \; a \leq g \leq n \; and \; a, n \in numSamples\}$.

4: Compute the absolute geometric error between the mapped points $\hat{x}_i$, $\hat{x}'_i$ and the measured points $x_i, x'_i$ such that $\epsilon = |x_i - \hat{x}|_i$ and $\epsilon' = |x'_i - \hat{x}'|_i$.

5: Create a list of homographies that map points within error less than 3 pixels on the highest resolution image frame.

6: Remove duplicated homographies and define a list of potential homographies $H_k$.

7: For each element of the list in (*vi*), compute the maximum and minimum working distance from the depth information of the set of 2D-2D correspondence points of (*iii*), such that $dmax = \max_{a \leq g \leq n} dg, dg = \{d_a \cup \cdots \cup d_n\}; a, n \in numSamples$ and $dmin = \min_{a \leq g \leq n} dg, d_g = \{d_a \cup \cdots \cup d_n\}; \; a, n \in numSamples$.

8: Select the optimal transformation matrices to create the depth-dependent homography lookup table where $Hlut \ni H_i^{lut}$, $dmin_i^{lut}$, $dmax_i^{lut}$ $i = 1 \cdots numH$. For that:

   a: Limit the depth of field by $dof^{hlut} = \left[dmin_1^{lut}, dmax_{numH}^{lut}\right]$.

   b: Approximate the distance between homographies to zero $\Delta dbp_i^{lut} \cong 0$, where $\Delta dbp_i^{lut} = dmin_{i+1}^{lut} - dmax_i^{lut}$.

   c: Minimize the elements of the lookup table $\min_{a \le numH \le n} numH$.

## 3.3 Validation of the Depth-dependent Homography Lookup Table Approach

The transformation from the ToF to the RGB frame was considered for the method evaluation. Since the method is depth-feature-based, the procedure input is the depth information provided by the ToF camera. The uncertainty because of the difference between the cameras resolution is a crucial issue for evaluating the proposed registration method. For any ToF point there are several potential correspondence points on the RGB frame. Consequently, the discrepancy between the control points on the RGB image coordinates $\left\{ xcp_i^{RGB} \right\}$ and the estimated points $\left\{ xmap_i^{RGB} \right\}$ for the 104 image samples was analysed. These mapped points are the registered points from the control points on the ToF image coordinates $\{ xcp_i^{ToF} \}$. The pseudocode for mapping points from the ToF to the RGB frame by using the depth-dependent *Hlut* method is presented in Algorithm 2.

For quantitative assessments of the discrepancy between the control points on the RGB image coordinates $\left\{ xcp_i^{RGB} \right\}$ and the estimated points $\left\{ xmap_i^{RGB} \right\}$, the *Accuracy of the Undistorted Image Coordinates* ($E_u$) (Salvi, Armangué and Batlle 2002), detailed in Equation (3.2), and the geometric error distribution were evaluated. Figure 3.6(a) and Figure 3.6(b) show the geometric error on *(uv)*-axis and the distance error of the estimated points, while the distribution of the error on the *u*-axis and *v*-axis are illustrated in Figures 3.6(c) and 3.6(d), respectively. Table 3.2 summarizes the results of the error distribution.

**Algorithm 2** Procedure for mapping points between two views (ToF → RGB) based on the depth-dependent *Hlut*.

1: Extract the mean depth of ROI of the control points $d_i^{pt}$.

2: Find the corresponding entry $k$ on the *Hlut* that suits $d_i^{pt}$ such that $dmin_k^{lut} \leq d_i^{pt} \leq dmin_k^{lut}$.

3: Compute the transformation of the points by applying the homography $H_k^{lut}$ such that $xmap_i^{RGB} = H_k^{lut}\, xcp_i^{ToF}$.

$$E_u = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(xmap_{xi}^{RGB} - xcp_{xi}^{RGB}\right)^2 + \left(xmap_{yi}^{RGB} - xcp_{yi}^{RGB}\right)^2} \tag{3.2}$$

The results in Figure 3.6 along with the data in the Table 3.2 indicate that the error deviation in the *v*-axis is higher than the error in *u*-axis. Since the cameras are vertically aligned, such behaviour was expected. Regarding the errors distribution, the standard deviation in *v*-axis is $\sigma_v = 3.19$, and at least the 66% of the estimated data has an absolute error $\leq 3\ pixels$ (see Figure 3.6(d) and Table 3.2). Only the 8.5 % of the data has an absolute error higher than 6 pixels. The maximum absolute error is 20 pixels. For the error distribution in the *u*-axis, the maximum error is ±8 pixels and the standard deviation is $\sigma_u = 2.78$. Most of the absolute error is $\leq 3\ pixels$, exactly the 83% of the data, and only the 0.9% of the absolute error is higher than 6 pixels. In practice, errors which are three or more times the standard deviation away from the mean, could be considered as outliers and should be removed (Osborne and Overbay 2004). In this case $outliters_{sample} = [2,41,42,43]$ are the detected samples with outliers. In order to evaluate the influence of these outliers, the image samples $outliters_{sample}$ were removed, and the errors were calculated once again. These results are also included in Table 3.2.

**Table 3.2 Results** of the Error Distribution

| Error Distribution (pixels) | | Error Percentage [%] | | |
|---|---|---|---|---|
| | | ***u-axis*** | ***v-axis*** | ***Geometric Distance*** |
| **104 image samples** | $error \leq \|3\|$ | 83.0 | 66.5 | 49.2 |
| | $\|3\| < error \leq \|6\|$ | 16.1 | 25.0 | 39.3 |
| | $\|6\| < error \leq \|9\|$ | 0.9 | 4.7 | **7.3** |
| | $error > \|9\|$ | 0 | 3.8 | **4.2** |
| **Outliers removed** | $error \leq \|3\|$ | 83.4 | 69.1 | **51.0** |
| | $\|3\| < error \leq \|6\|$ | 15.7 | 25.9 | **40.8** |
| | $\|6\| < error \leq \|9\|$ | 0.9 | 4.6 | **7.4** |
| | $error > \|9\|$ | 0 | 0.4 | **0.8** |

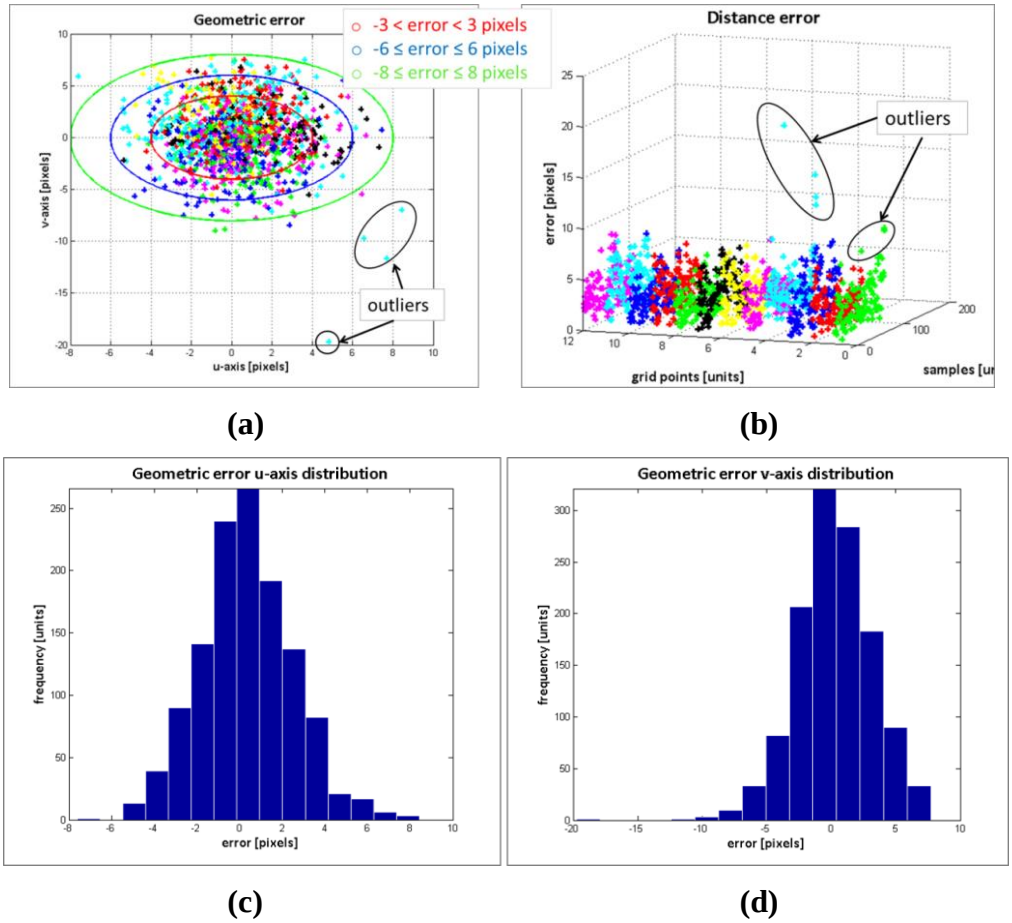



**Figure 3.6** Geometric error evaluation. (**a**) Geometric error. (**b**) Distance error. (**c**) Error distribution in *u*-axis. (**d**) Error distribution in *v*-axis.

The results of the error distribution for the outliers removal shows an improvement in the accuracy of the depth-dependent *Hlut* approach with respect to the previous analysis, with a mean value $Mean_{(u,v)-axis} = [0.33, 0.44]$ and a standard deviation $\sigma_{(u,v)-axis} = [2.1, 2.9]$ on pixel coordinates, in contrast with the obtained values when using the entire set of 104 image samples, with a mean value $Mean_{(u,v)-axis} = [0.33, 1.11]$ and a standard deviation $\sigma_{(u,v)-axis} = [2.1, 4.6]$ on pixel coordinates. In addition to the geometric error evaluation, the normalized RMSE of the discrepancy between the control points on the RGB image coordinates $\{xcp_i^{RGB}\}$ and the

estimated points $\{xmap_i^{RGB}\}$ was evaluated as well. In the Figure 3.7, the results of the normalized RSME on each pattern board is show and the overall error is $NRSME = 0.079$ and $NRSME = 0.1146$ for the image samples without considering image samples $outliters_{sample}$. In this case the influence of outliers is more visible, because the accuracy of the method is improved in 31.1 %.

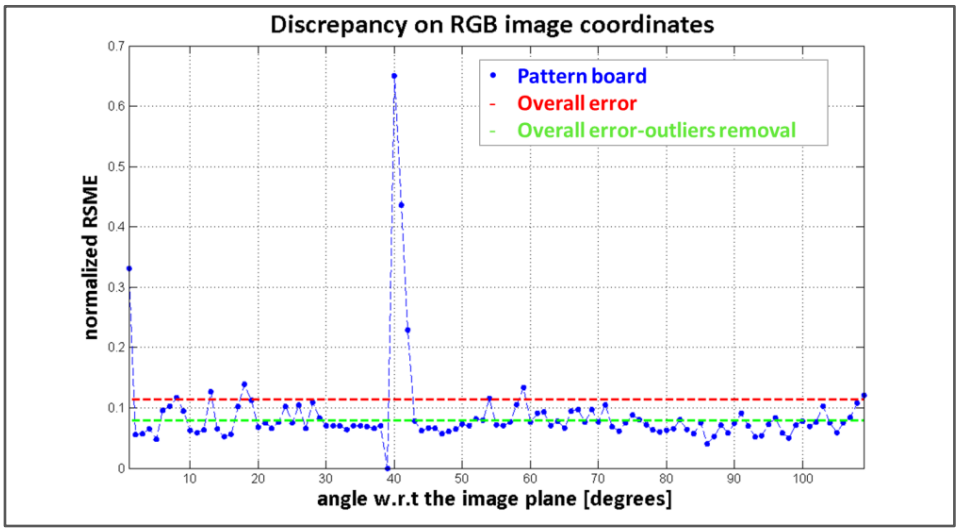



**Figure 3.7** Normalized RMSE on RGB image coordinates vs the angle of the board plane w.r.t. the image plane.

In general, when analysing the error within the RGB frame (2448×2050 pixels), the relative errors are significantly low. An error of 3 pixels represents a relative error of 0.15% over the RGB image, and for the 20 pixels deviation, a relative error of 0.9% is reached. Though these values are evidently small, yet it is something to be concern of. Several conditions might introduce error to the method, for instance:

1. The outliers in the selection process of the correspondence control points.
2. The presence of noise in the depth measurements.
3. The implicit error of the transformation matrices.

In order to evaluate the influence of the depth variations on the proposed method, the depth measurements of the effective grid pattern were analysed. Two groups of data were compared: the raw depth and the filtered depth. For smoothing the depth data, the denoising algorithm proposed in (Buades, Coll

and Morel 2005) was adopted, and Figure 3.8 shows the results of the analysis. In Figure 3.8(a) the mean depth and the depth boundaries of the pattern board acquired in the 104 image samples are shown. The raw depth has higher data variance, though the mean of the raw and filtered data are nearly the same, as it is illustrated in Figure 3.8(b). The impact in the overall error because of the object distance and the object depth variations are illustrated in Figure 3.8(c) and 3.8(d). According to these results, neither the mean distance nor the maximum depth variations have direct correlation to the error's scope. Therefore, it is possible to conclude that the error is not reliant on the depth variations within 25 mm, corresponding to the mean maximum depth variations of the raw data.

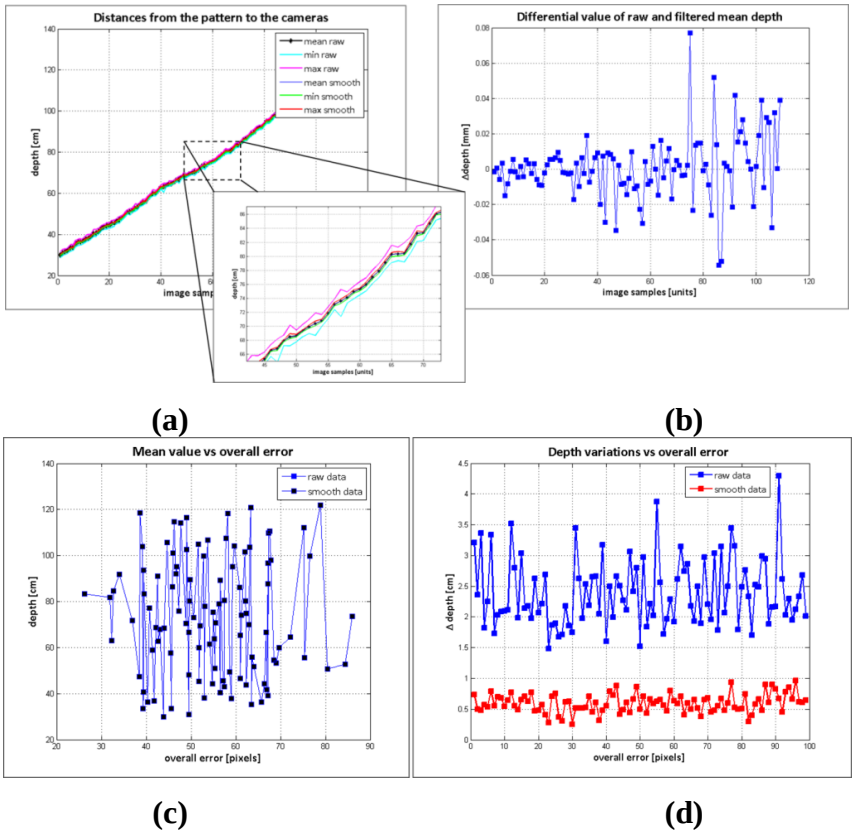



**Figure 3.8** Depth measurements evaluation (**a**) Distances from the pattern board to the cameras. (**b**) Differential value of raw and filtered data. (**c**) Samples mean depth vs overall error. (**d**) Samples maximum variation vs overall error.

The flawed points selected as correspondence control points could be the most frequent problem for introducing error on the estimated data. Consequently, instead of a point to point evaluation, entire regions of the images with higher errors were analysed. Figures 3.9 and 3.10 show the evaluation results of two images of the pattern board; in one image the board is positioned at 527 mm and in the other at 891 mm. First, a region of interest (ROI) in the ToF image is selected. Then, the depth measures of the ROI are sorted in ascending order, and clusters of 12 mm of standard deviation are created $c^j = \{x_i^{ToF}\}$. Finally, the mean depth $dm_j^c$ of each of these clusters $(c^j)$ is matched with a suitable distance entry $k$ on the *Hlut*, such that $dmin_k^{lut} \leq dm_j^c \leq dmax_k^{lut}$. Thus a homography $H_k^{lut}$is designated to each $c^j$, and the selected ROI is mapped as $\forall\, c^j: \{xmap_i^{RGB}\} = H_k^{lut}\{x_i^{ToF}\}$. In the images displayed in Figures 3.9(a-b) and 3.10(a-b), the mapped points are marked with dots. The colour of the dots indicates the entry $k$ of the homographies $H_k^{lut}$ used to transfer the data. The composition of the ROI from the estimated points on the RGB image and the ROI of the selected points on the ToF are illustrated in Figures 3.9(c-d) and 3.10(c-d).

Since the RGB and ToF images are acquired from distinct sources and there is a large difference between their image resolutions, the properties of the sensed objects tend to be different. The most relevant effects are perceived in the borders of the textured objects and in the objects dimensions. Let us utilise the image capturing of the pattern board by way of illustration (see Figures 9 and 10). The RGB high resolution camera acquisition sharpens the squares borders, while the capturing by the low resolution ToF camera unsharpens the squares borders of the pattern. Thus, the squares on the composed ROI from the mapped points on the RGB image are slightly blurred.

Additionally, a great number of elements can affect the response of the ToF camera, for instance the rays emitted from the sensor that lie on the object's edge tend to be less accurate because they are affected by the multi-path interferences. Consequently, the objects dimensions on the ToF image are not always alike compared with the ones on the RGB image. In Figure 3.10, some of these issues are illustrated. As an example, the pattern board is smoothly rotated with respect to the optical axis of the cameras. This rotation is only perceived by the RGB camera.

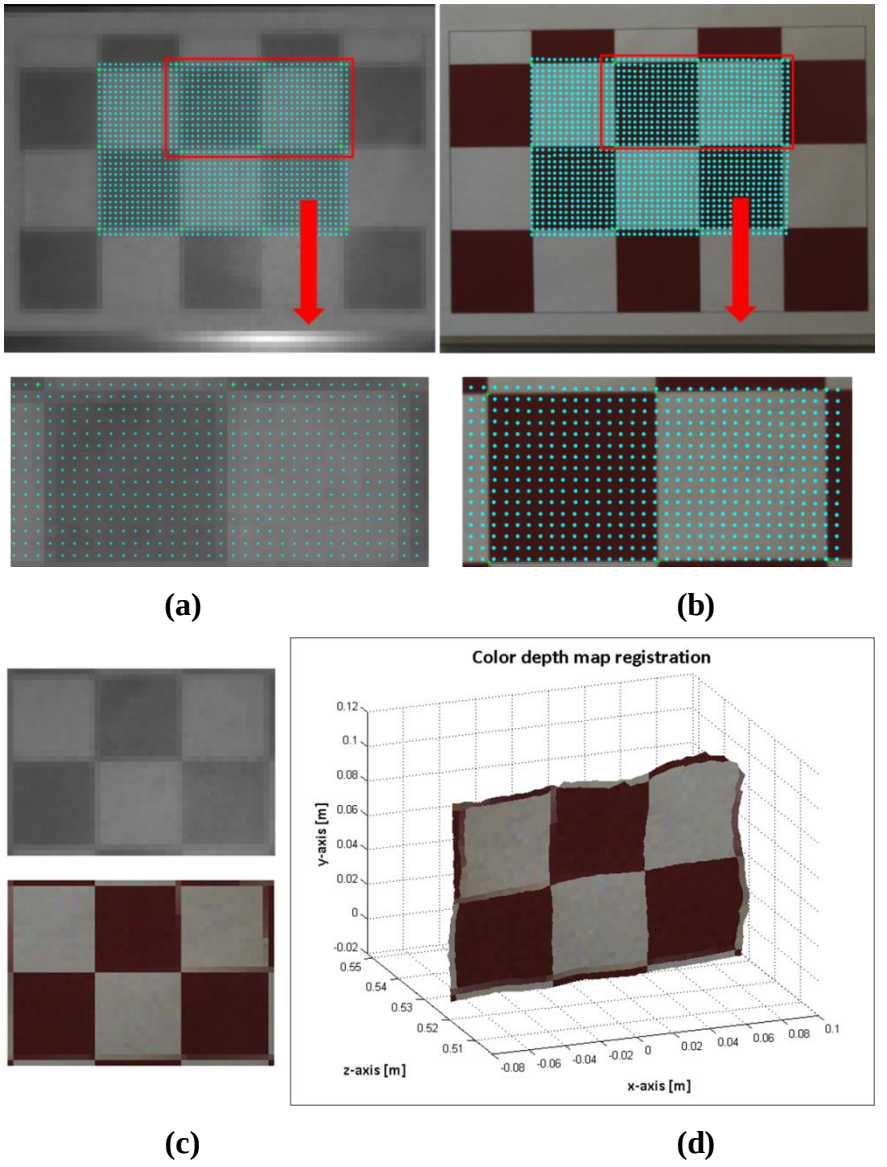



**Figure 3.9** Image sample 49 - pattern board positioned at 527 mm. (**a**) Top: selected ROI on the ToF image. Bottom: zoom of the selected points on the ToF image. (**b**) Top: mapped points on the RGB image. Bottom: zoom of the estimated points on the RGB image. (**c**) Top: ROI of the ToF image. Bottom: Composition ROI from the mapped points on the RGB image. (**d**) Colour depth map.

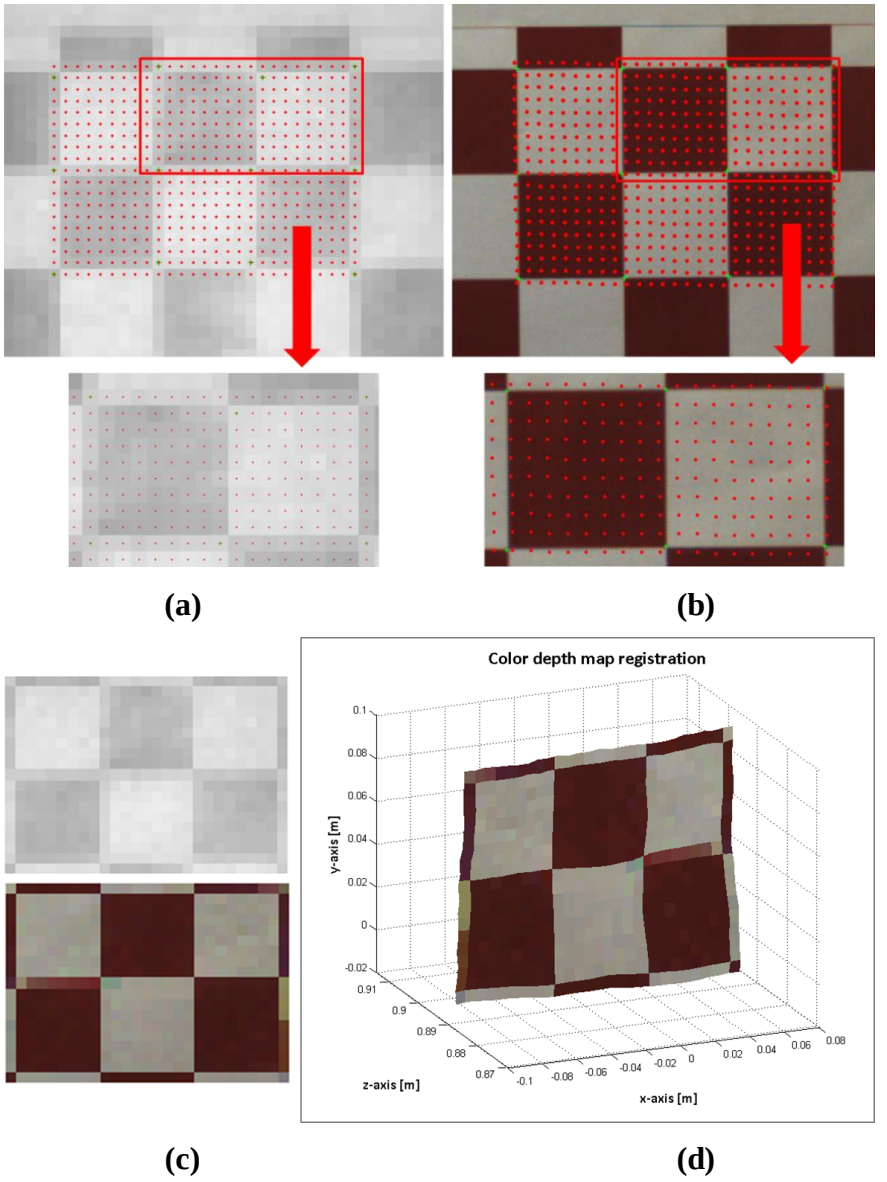



**Figure 3.10** Image sample 25-pattern board positioned at 891 mm. (**a**) Top: selected ROI on the ToF image. Bottom: zoom of the selected points on the ToF image. **(b**) Top: mapped points on the RGB image. Bottom: zoom of the estimated points on the RGB image. (**c**) Top: ROI of the ToF image. Bottom: Composition ROI from the mapped points on the RGB image. (**d**) Colour depth map.

Regarding the large difference in the cameras resolution, there are several unmatched points (~10–11 pixels) between adjacent estimated points on the RGB image. Nonetheless, the results of the proposed method show that the region composed of the estimated points on the RGB image and the selected region on the ToF image are very close to each other, and proportionally registered.

## 3.4 High Resolution Colour Depth Map Estimation

This section is devoted to the evaluation of the capability of the proposed image registration method, for computing high resolution colour depth maps. For that purpose, an initial procedure was introduced and a visual assessment was obtained. . The procedure is based on the results of the image registration obtained by means of the depth-dependent *Hlut* method. This proposal combines the homography labelled mask $mask_{LRGB}$and a nearest neighbour algorithm for the RGB unmapped pixels classification. The pseudocode is detailed in Algorithm 3.

---

**Algorithm 3** Procedure for mapping points between two views (ToF ↔ RGB) based on the depth-dependent *Hlut* approach.

---

1: Select a ROI in the ToF image for data registration or select the entire image.

2: Sort in ascending order the depth measures of the selected region and create $Q$ clusters with 12 mm of standard deviation such that $c^j = \{x_i^{ToF}\}, j = 1 \cdots Q$ and compute the mean depth of each cluster $dm_j^c$.

3: Find the corresponding distance entry $k$ on the *Hlut* that suits the mean depth of each cluster, such that $\forall\, c^j : \{xmap_i^{RGB}\} = H_k^{lut}\{x_i^{ToF}\} \mid dmin_k^{lut} \leq dm_j^c \leq dmax_k^{lut}$, where $j = 1 \cdots Q$ and$1 \leq k \leq numH$.

4: Compute the points transformation from the ToF to the RGB images by using the homography lookup table $\{H_k^{lut}\}$ designated in the previous step.

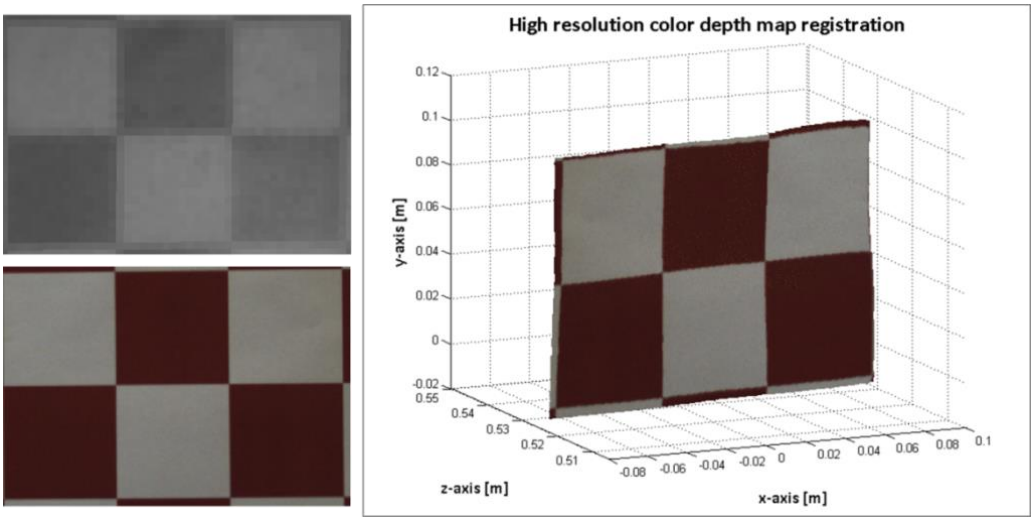
---

5: Create a labelled mask ($mask_{LRGB}$)corresponding to the RGB frame, where the values for the mapped points are the $k$ entries of $\{H_k^{lut}\}, 1 \leq k \leq numH$, and the values for the unmatched points are zero.

6: Approximate the unmatched pixels of the labelled mask $mask_{LRGB}$ by applying the nearest neighbour classification algorithm.

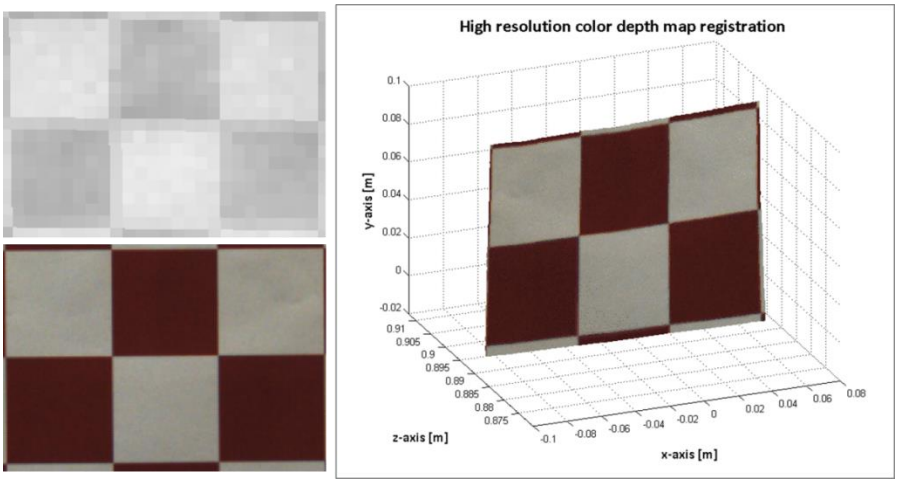
7: Compute points transformation from RGB to ToF images such that $xmap_i^{ToF} = {H_k^{lut}}^{-1} x_i^{RGB}$.

In Figures 3.11 and 3.12, the results of computing high resolution colour dense maps by using the procedure listed in Algorithm 3 are shown. The procedure was applied on image samples 25 and 49, which are illustrated in Figures 3.9 and 3.10 in section 3.3.




**(a)** **(b)**

**Figure 3.11** Image sample 49. (**a**) Top: mapped points on the ToF image. Bottom: points of the ROI on the RGB image. (**b**) High resolution colour depth map.




**(a)** **(b)**

**Figure 3.12** Image sample 25. (**a**) Top: mapped points on the ToF image. Bottom: points of the ROI on the RGB image. (**b**) High resolution colour depth map.

## 3.5 Conclusions

In this chapter, a new approach for colour depth map registration was presented. In contrast with the standard calibration method for transforming the ToF 3D-space coordinates to the RGB camera frame, the presented approach relies on planar projective transformations and uncalibrated techniques. In this way, the 3D world is parametrized and discretized in *n*-planes. Hence a discretized region corresponding to an *i*-plane, where $i = 1 \cdots n$, is explained with a unique homography of a lookup table $\{H_i^{lut}\}$, which is dependent on the depth measured from the *i*-plane to the ToF camera.

The depth map registration obtained by means of the standard calibration method provides a low resolution colour depth map. In order to achieve high resolution depth maps, the depth values are usually extrapolated. This normally leads to the over smoothing of objects' edges. Since this method depends on the value of depth measurements for the data registration, it is more sensitive to noise and depth filtering. In consequence, the data filtering could decrease the accuracy. On the contrary, in the depth-dependent *Hlut* method a range of depth values is considered for transferring the clusters of

data from the ToF to the RGB image frame. Therefore, small variations on depth measurements are not a critical issue for the proposal approach.

Regarding the high resolution colour depth maps, the proposal method provides a labelled mask $mask_{LRGB}$ on the RGB image coordinate, and the values of the mask correspond to the homography $\{H_i^{lut}\}$ of the mapped points $\{xmap_i^{RGB}\}$. Hence, the unmapped points on the RGB image coordinates could be estimated by means of the inverse homography and the labelled homography mask $mask_{LRGB}$, with the implementation of sophisticated, smart and guided algorithms to interpolate the depth information. For instance, in this chapter a neighbourhood classification algorithm was presented to map the entire RGB image and to compute a high resolution colour depth map. The proposed approach leads to a non-loss of the original colour information (2448×2050 pixels) and to the computation of high resolution colour dense maps under near real-time conditions.

A methodology and the pseudocode for computing the depth-dependent *Hlut* are introduced, as well as the pseudocode for implementing the method. On the other hand, the accuracy of the method has been evaluated and the method implementation has been validated within a large set of image samples. The contributions of this section and the method proposed in this Thesis are presented in the journal publication (Salinas et al. 2015).

# Chapter 4

## Comparison of Methods for Depth Map Registration

### 4.1 Introduction

The most extended technique for computing the colour depth map registration is the standard calibration method, as it was studied in the state-of-the-art presented in Chapter 2. This method relies on the computation of the external and internal cameras parameters, and the homogenous transformation between RGB and ToF cameras. Then, the 3D depth measurements from the ToF camera are used to back-project points on the ToF pixels coordinates to the RGB pixel coordinates. For the depth-dependent Hlut image registration approach, the depth measurements are used to select an homography in the lookup table $\{ H_i^{lut} \}$, where each entry of the lookup table is related to a range of depth values. Evidently, in both cases, the accurate acquisition of depth estimations is a key issue for computing a satisfactory registration procedure. Several research have shown that the noise in the depth measurements acquired by the ToF cameras is a persistent problem (Chiabrando et al. 2009, Foix et al. 2011), and a crucial issue when evaluating the depth map registration results. Therefore, the methods evaluation should be focused on analysing the capability of registration methods for properly overcoming the noise, while avoiding the generation of misalignment problems. For the ToF cameras depth estimations, a modulated

infrared light is emitted from an internal lighting source. The light is reflected by objects in the scene and travels back to the sensor. Then the time of flight between the camera and the object is measured for each of the sensor's pixel, by calculating the phase delay between the emitted and the received wavelength. Systemic and non-systemic errors are presented in depth estimations of these cameras. The quality of the measurements relies on the sensor hardware, the sensor configuration, the objects albedo, the objects shape and edges, the temperature, and others. Some interesting works have investigated the source of the errors and have presented solutions to minimize the problem (Chiabrando et al. 2009, Guomundsson, Aanaes and Larsen 2007). The systematic errors can be reduced by a calibration process and the non-systematic errors with a filtering technique, as it is summarized in (Foix et al. 2011). Therefore, the methods comparison proposed in this Chapter is committed to the evaluation of the standard calibration method and the depth-dependent *Hlut* approach from the perspective of their response to the noise in the depth estimations.

The methods comparison is addressed through two case studies. In the first case, the problem of depth map registration of noisy depth estimation was evaluated. For that matter, a white-black chessboard was chosen as a target for reproducing noise in the depth measurements. It is known that dark objects produce errors in the depth measurements. This is because dark objects absorb the IR radiation. Consequently, the intensity of reflected light from the objects is lower than the emitted, what produces that some of them never reach back to the sensor, and that others are poorly detected. In the second case, a procedure for minimizing the noise in the depth estimations was adopted, and the evaluation of the methods was performed on the filtered data. While a detailed description of existing techniques is beyond the scope of this work, two representative methods for noise minimization and surface smoothing were implemented: the Bilateral Filtering (Tomasi and Manduchi 1998) and the Non-local Means Filter (NL-means) (Buades et al. 2005). The bilateral filtering, which is one of the most extended techniques for noise removal, is a neighbourhood smoothing filter, characterized as an edge preserving method. On the other hand, the NL-means is based on a non-local averaging of all pixels of the data.

This section involves the computation and the accuracy evaluation of the standard calibration parameters and the comparison of the depth map registration results provided the by the two methods: the standard calibration

parameters and the depth dependent *Hlut*. The comparison process is addressed by analysing three possible scenarios:

- Ideal data: noise free depth information of the scenes.
- Noisy depth information of the scenes.
- Filtered noisy depth information of the scenes.

In order to compute a quantitative assessment of the methods results, the normalized Root Mean Square Error (NRMSE) was adopted. The use of the RMSE over the Mean Absolute Error has been studied in several works (Chai and Draxler 2014). Although both methods are sensitive to outliers, the RMSE has several advantages. For instance, it satisfies the distance function metric requirement of the triangle inequality, and it does not use absolute values which is an advantage to calculate the gradient or sensitivity of the sample with respect to certain model parameters. It is also appropriated for error that follows normal distribution.

The data fusion by means of the standard calibration technique was carried out by implementing the method described in (Park et al. 2011), where the depth measurements are used to back-project the 3D world points to the 2D points on the RGB image. In the case of the depth dependent *Hlut* approach, the depth map registration was done by following the procedure described in Table 3.3 (see section 3.3).

## 4.2 Standard Camera Calibration Computation and Evaluation

The standard calibration of the sensory rig described in section 3.3.1, was carried out with the help of the Matlab Camera Calibration Toolbox (Bouguet 2008), which provides the intrinsic and extrinsic camera parameters. A thoroughgoing calibration methods evaluation is beyond the scope of this section. However, during last decades, several researchers have dedicated their efforts on that matter, such is the case of (Salvi et al. 2002), where the authors presented an extensive evaluation of several calibration methods. The results of this comparison show that of the Tsai's algorithm (Tsai 1987), surpasses the achieved accuracy by other methods. However this method requires an accurate 3D measurement, which normally involves a large amount of elaborated training data and time consuming setups. In a subsequent work (Wei and Cooperstock 2005), the accuracy of the Tsai's method was compare with the planar calibration approach introduced by (Zhang 2000). The evaluation of these two methods pointed out the

advantages of the Zhang's method, which does not require complex measuring procedures or specialized equipment. In order to avoid the noise in pixel coordinate, the authors propose the use of a large number of training points, easily achievable. The Matlab Calibration Toolbox is a very well-known tool in the Computer Vision Community. This toolbox is inspired in Zhang's method, which has been validated in several researches, and has been proven to be a flexible and suitable method for calibrating dynamic scenes.

For the cameras calibration procedure, 62 image samples of a black-white chessboard were acquired. These image samples are composed by an RGB image and a ToF amplitude image. For the image samples acquisition, the pattern board was located at different poses and orientations, and at different distances from the sensory system. Then, 30 correspondence ground truth points were selected from each image pairs. Consequently, 1860 training points for each camera were provided, a sufficient number of points for the Zhang's algorithm to achieve satisfactory accuracy (Wei and Cooperstock 2005). For evaluating the accuracy of the calibration parameters results, two of the most commonly used criteria were followed: *the Accuracy of the Distorted and Undistorted Image Coordinates* (Salvi et al. 2002) and the *Normalized Calibration Error* (NCE) (Weng, Cohen and Herniou 1992) which evaluates the accuracy in the world coordinates.

The first criterion, the *Accuracy of the Distorted and Undistorted Image coordinates*, is a 2D measurement technique that computes the discrepancy on the image coordinates. For the matter of measuring these deviations two methods were considered. The deviations are referred to the difference between the ground truth control points on pixel coordinates $(\mathrm{x_{pxi}}, \mathrm{y_{pxi}})$ and the projection of the 3D control points in to the image plane $(\hat{\mathrm{x}}_{\mathrm{pxi}}, \hat{\mathrm{y}}_{\mathrm{pxi}})$. One method to measure the error is the Mean value of the geometric distance error ($E_d$,) described in Equation (4.1) and detailed in (Wei and Cooperstock 2005, Salvi et al. 2002).

$$E_d = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(\hat{x}_{pxi} - x_{pxi}\right)^2 + \left(\hat{y}_{pxi} - y_{pxi}\right)^2} \tag{4.1}$$

The second method for measuring the 2D geometric distance error is the normalized Root Mean Square Error ($RMSE_d$). The use of the RMSE over the Mean error for normal distribution is an appropriated solution, as it is shown

in (Chai and Draxler 2014). Regarding the pixel error distribution of the standard calibration results, the distribution information is represented in Figure 4.1 and in Table 4.1. For computing the RMSE, the residuals of the distance error of the projected points $(\hat{x}_{pxi}, \hat{y}_{pxi})$, and the predicted distance error ($dist_{pred} = 0$) are estimated. Then, the $RMSE_d$ is computed as:

$$dist\_err_i = \sqrt{\left(\hat{x}_{pxi} - x_{pxi}\right)^2 + \left(\hat{y}_{pxi} - y_{pxi}\right)^2}$$

$$RMSE_d = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[dist\_err_i - dist_{pred}\right]^2};\ dist_{pred} = 0 \tag{4.2}$$

The second criterion calculates the accuracy with respect to the 3D camera coordinates. This technique overcomes the sensitivity to: the image resolution, the object-to-camera distance and the field of view of the camera, by normalizing the discrepancy between the estimated and the observed 3D points with respect to the area that each back-projected pixel covers at a given distance from the camera. Let $(\hat{X}_i, \hat{Y}_i, \hat{Z}_i,)$ be the 3D point on the camera coordinates which is estimated by the back-projection from the 2D pixels of the control points on the image coordinates, and $(X_i, Y_i, Z_i)$ the ground truth observations of the 3D world points on the camera coordinates. Since the observed 3D points were calculated with respect the ToF camera coordinate system, the NCE was computed only on this coordinates system. Then the NCE is defined as:

$$NCE_{ToF} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\left(\hat{X}_{i_{ToF}} - X_i\right)^2 + \left(\hat{Y}_{i_{ToF}} - Y_i\right)^2}{\frac{\hat{Z}_{ToF_i}^2\left(f_{u_{ToF}}^{-2} + f_{v_{ToF}}^{-2}\right)}{12}}\right]^{\frac{1}{2}} \tag{4.3}$$

The results of both criteria are detailed in Table 4.2 Figures 4.2 and 4.3 show the results of the first criterion, while in Figure 4.4, the error computed with the second criterion are presented.

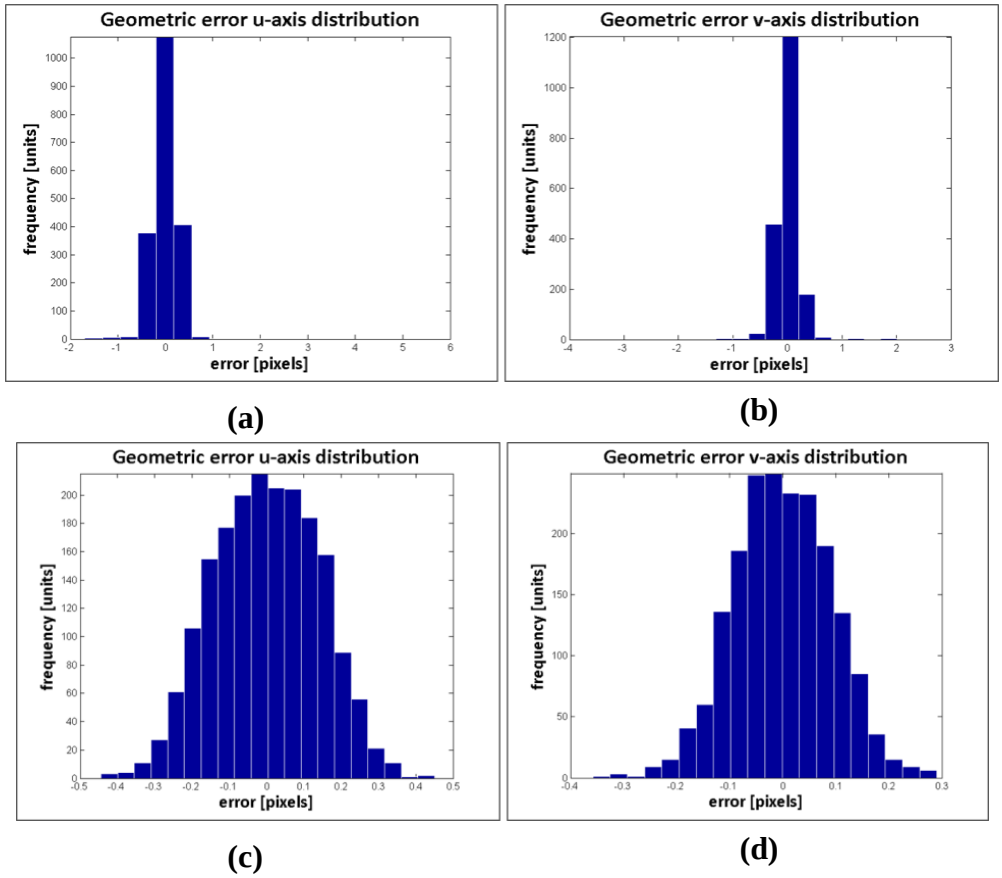
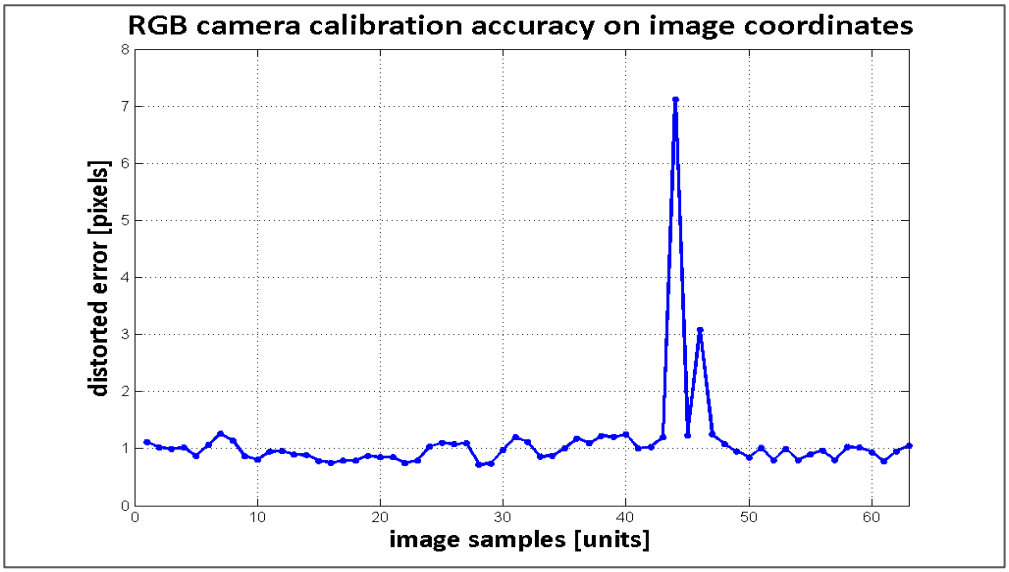


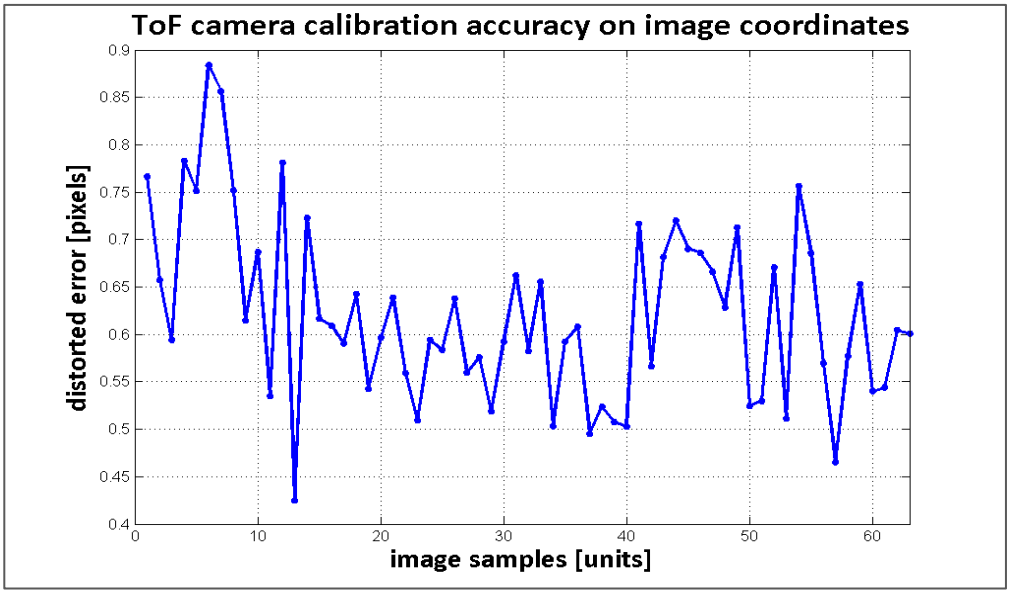

**Figure 4.1** Distribution of the geometric error on the pixels coordinates. (**a**) *u*-axis error on RGB images. (**b**) *v*-axis error on RGB images. (**c**) *u*-axis error on ToF amplitude images. (**d**) *v*-axis error on ToF amplitude images.

**Table 4.1** Distribution of the Absolute Pixel Error.

| **Absolute Error Distribution [pixels]** | **Error Percentage [%]** | | | |
|---|---|---|---|---|
| | ***RGB camera*** | | ***ToF Camera*** | |
| | ***u-axis*** | ***v-axis*** | ***u-axis*** | ***v-axis*** |
| $\boldsymbol{error \leq \lvert 0.3 \rvert}$ | 85.2 | 92.1 | 98.1 | 99.9 |
| $\boldsymbol{\lvert 0.3 \rvert < error \leq \lvert 0.6 \rvert}$ | 13.4 | 6.5 | 1.9 | 0.1 |
| $\boldsymbol{\lvert 0.6 \rvert < error \leq \lvert 1.0 \rvert}$ | 0.7 | 0.6 | 0 | 0 |
| $\boldsymbol{\lvert 1.0 \rvert < error \leq \lvert 1.2 \rvert}$ | 0.3 | 0.4 | 0 | 0 |
| $\boldsymbol{error > \lvert 1.2 \rvert}$ | 0.4 | 0.4 | 0 | 0 |

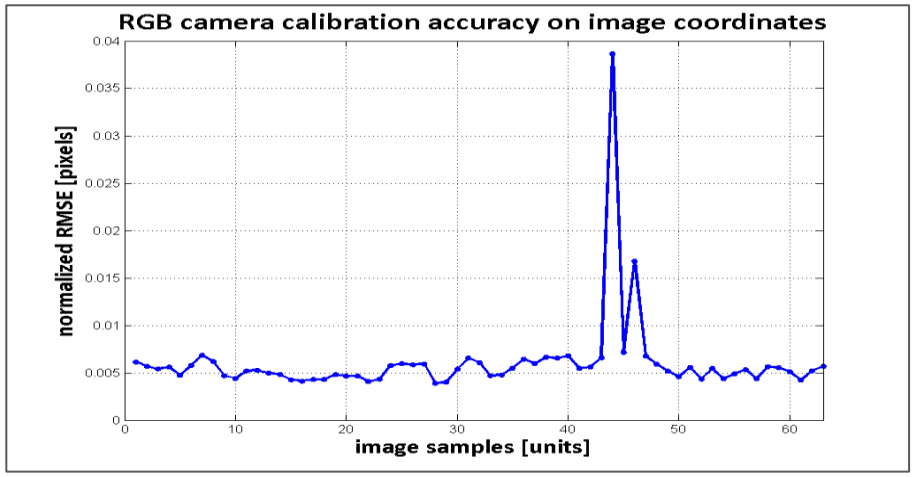



**(a)**

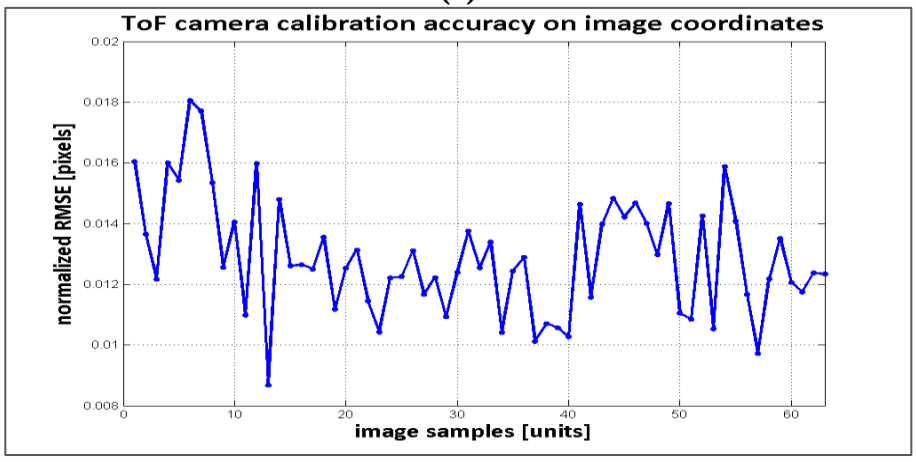



**(b)**

**Figure 4.2** Errors on the pixels coordinates. (**a**) RGB camera distorted error. (**b**) ToF camera distorted error.




**(a)**




(**b**)

**Figure 4.3** Errors on the pixels coordinates. (**a**) RGB camera normalized RMSE. (**b**) ToF camera normalized RMSE.

Additionally, the influence of the pose of the pattern boards to the performance of the calibration parameters was investigated in this section. For that purpose, the criterion used in (Zhang 2000) was followed. When the angle between objects and image plane increases, it is known that foreshortening makes the digitalization of objects less precise, and consequently an important issue to reckon with. For that purpose, the angle of the plane model of the pattern board with respect to the image plane was estimated. In Figure 4.5, the angles computation is exemplified, and in

Figures 4.6-4.8, the results of the angles estimation versus the accuracy errors are shown. For convenience, the angle is represented such as $\beta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

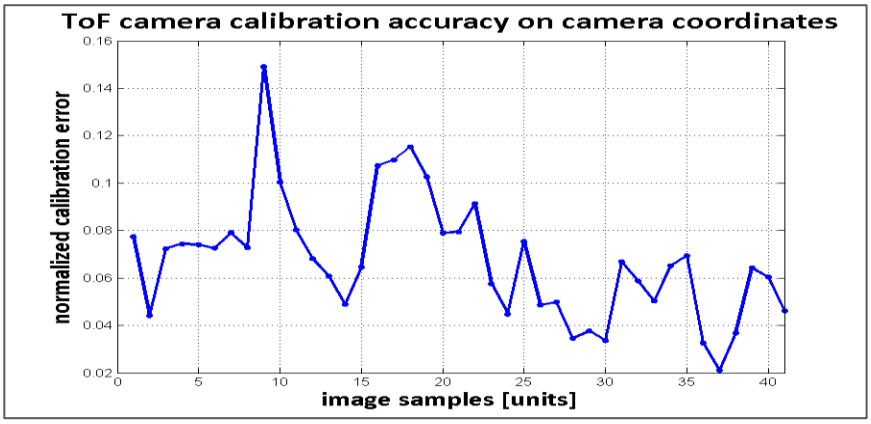



**Figure 4.4** Normalized Calibration Error on the ToF camera coordinates.

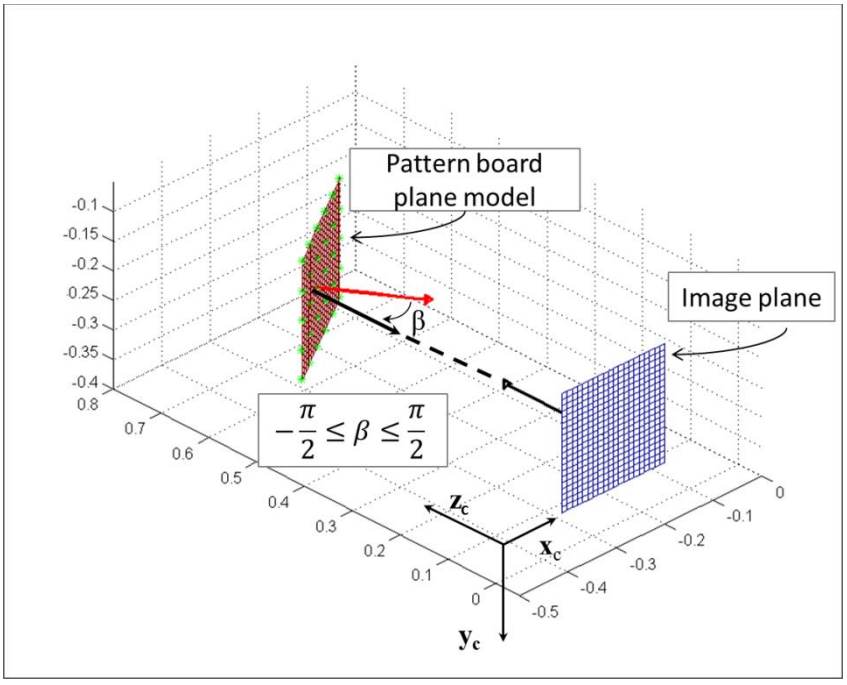
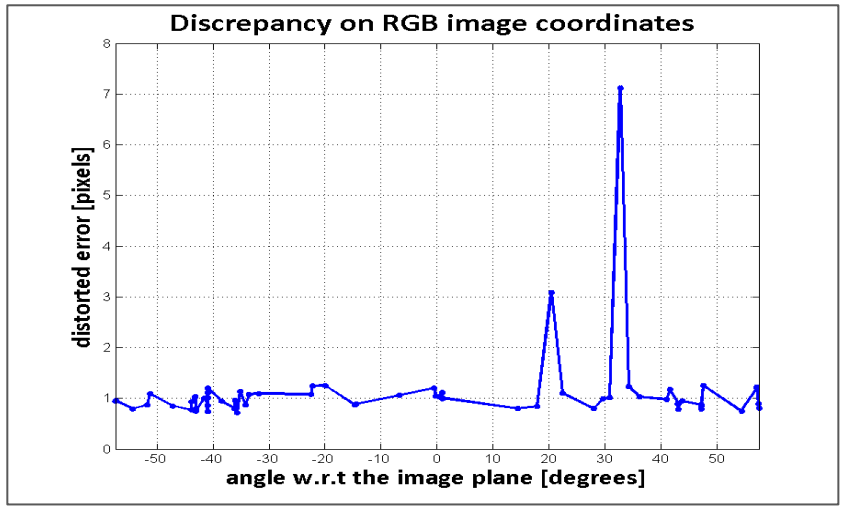
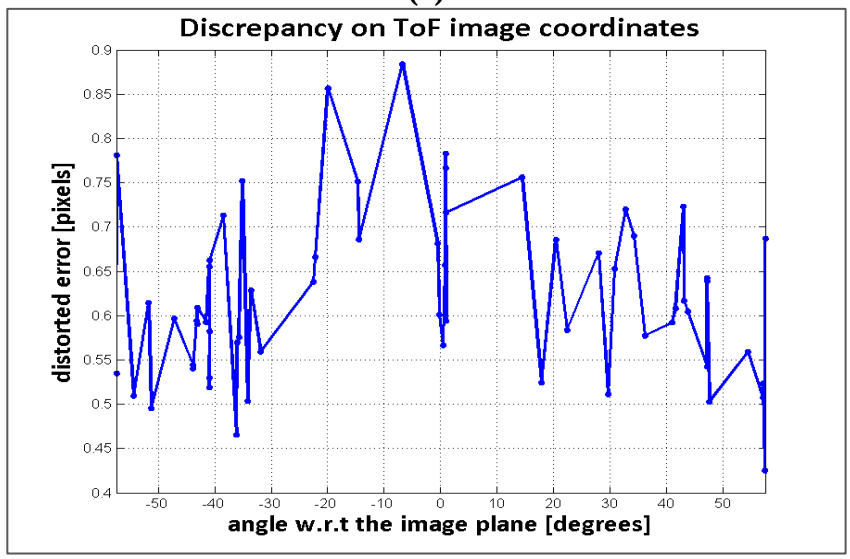



**Figure 4.5** Estimation of the angle of the board plane w.r.t. the image plane – image sample 44.

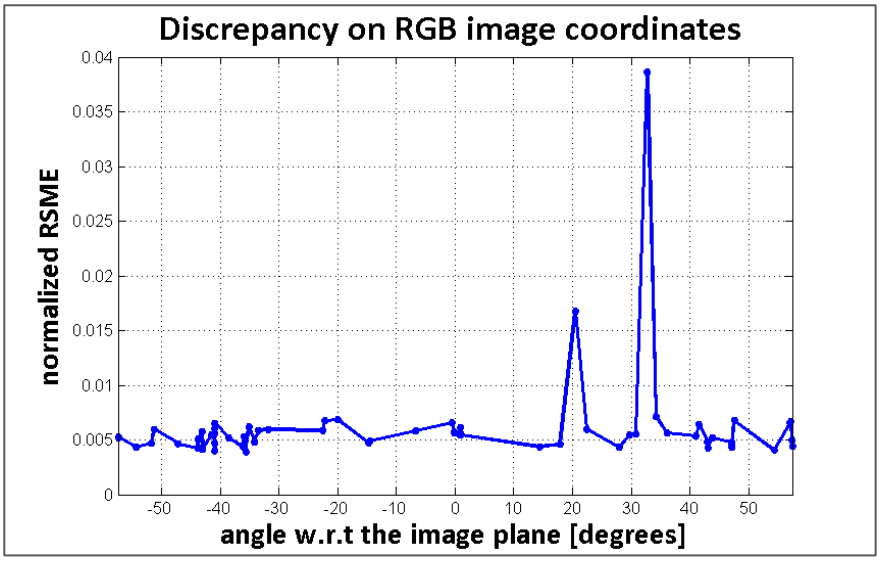



**(a)**

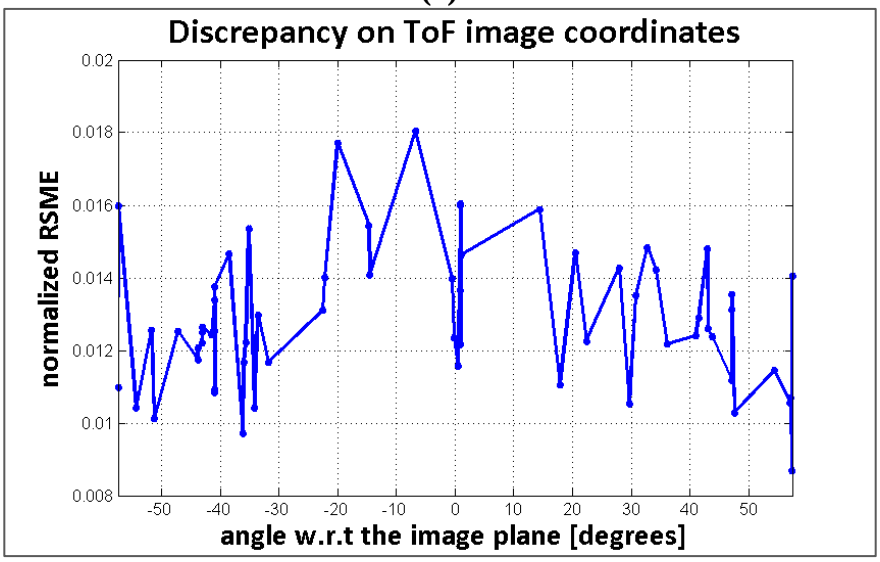



**(b)**

**Figure 4.6** Errors vs the angle of the board plane w.r.t. the image plane. (**a**) RGB camera $\boldsymbol{E_d}$ error. (**b**) ToF camera $\boldsymbol{E_d}$ error.




**(a)**




**(b)**

**Figure 4.7** Errors vs the angle of the board plane w.r.t. the image plane. (**a**) RGB camera $\boldsymbol{E_d}$ error. (**b**) ToF camera $\boldsymbol{E_d}$ error.

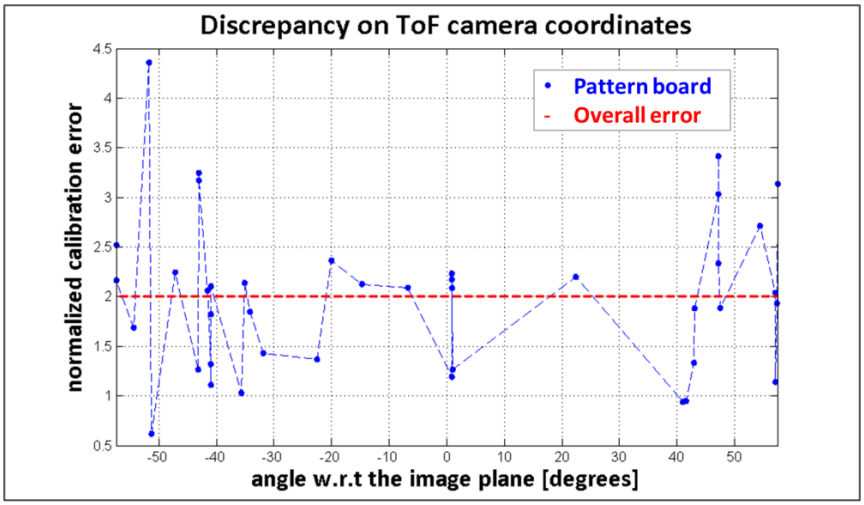
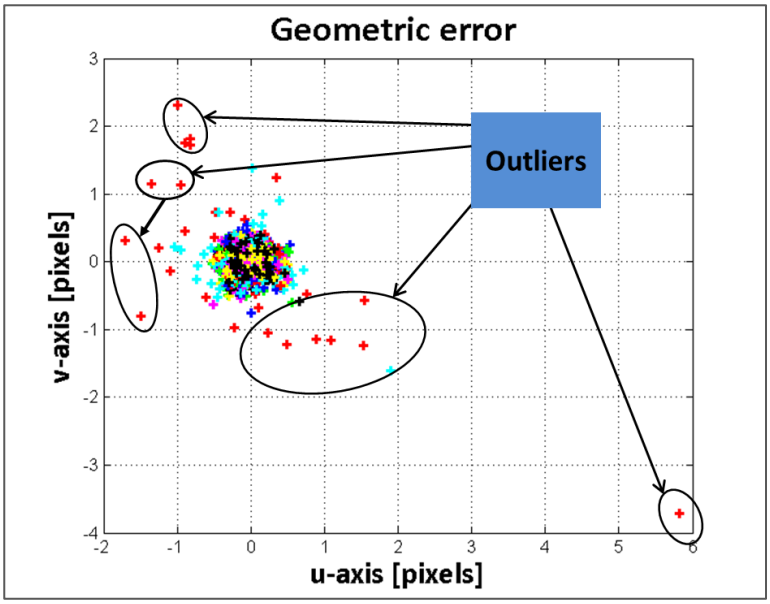


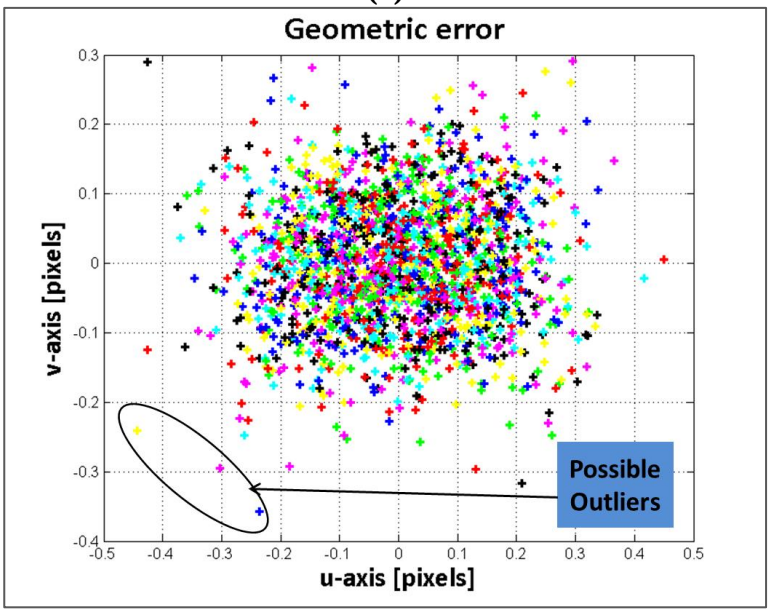

**Figure 4.8** Normalized Calibration Error on the ToF camera coordinates vs the angle of the board plane w.r.t. the image plane.

The obtained results pointed out the sensitivity of the $RMSE_d$ and the $E_d$ to outliers. In Figure 4.2(a) and 4.3(a) two peaks denotes the presence of outliers in the RGB camera error computation. These large errors could be derived from the mismatching on the detection and selection of the control points. According to the data on these two Figures, the image samples 44 and 46 are the pattern boards that produce the outliers. In both cases, the pattern board is posed in a way that its digitalization lays on the left bottom corner on the image coordinates, and the pattern board is rotated 33º and 21º with respect to the image plane, respectively.

The presences of outliers on ToF images are not as clear as in the RGB images, and the geometric error is more likely to a normal distribution. However, the evaluation of the NCE reveals two peaks on the error measurement, which are on image samples 9 and 18. In Figure 4.9 the pixel coordinate error of the 1860 control points is shown. Thus, this graphics illustrates the presence of the outliers. The mismatching of these estimated points on the image coordinates for the RGB and ToF cameras are shown in Figures 4.10 and 4.11 respectively. In some cases, when the outliers are several orders larger than the other samples, the outliers' removal is justified. After the outliers were removed, the errors were computed once again. Table 4.2 presents the results of the calibration parameters accuracy results for the two criteria ($NRME_d$ and $E_d$, and $NCE_{ToF}$) previously described, with the data set before and after the outliers removal.




**(a)**




**(b)**

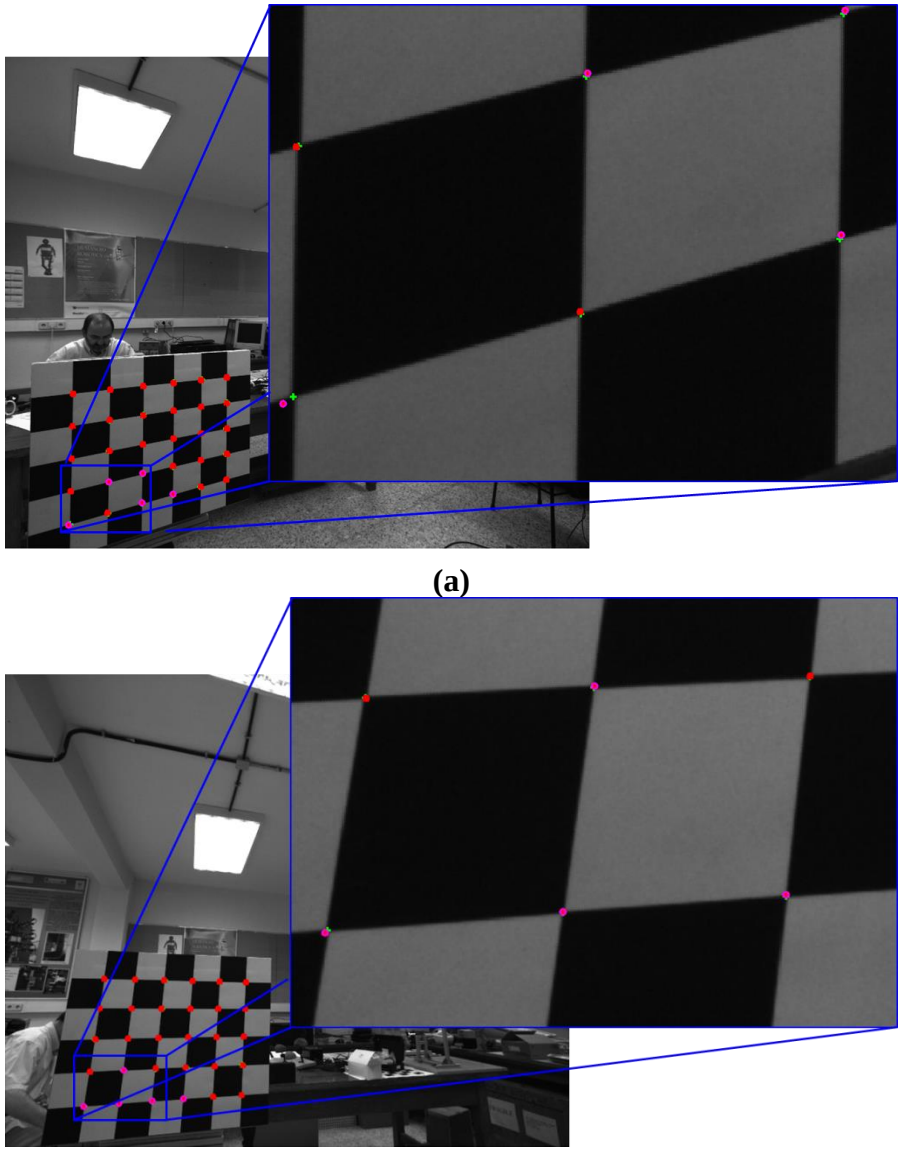
**Figure 4.9** Geometric error on the pixels coordinates. (**a**) Error on RGB images. (**b**) Error on ToF amplitude images.

**Table 4.2** Accuracy of the Standard Calibration Parameters

| **Calibration Error** | | **Camera of/on the Sensory Rig** | |
|---|---|---|---|
| | | ***RGB*** | ***ToF*** |
| 62 image samples | $NCE_{ToF}$ | --- | 1.997 |
| | $E_d$ | 0.2584 | 0.1484 |
| | $NRMSE_d$ | 0.0075 | 0.0130 |
| Outliers removed | $NCE_{ToF}$ | --- | 1.9022 |
| | $E_d$ | 0.2323 | 0.1486 |
| | $NRMSE_d$ | 0.0054 | 0.0130 |

The evaluation of the accuracy of the obtained calibration parameters accuracy for the sensory rig (see Section 3.1.1), analysis demonstrates a satisfactory performance of the computed intrinsic and extrinsic calibration parameters of this sensory system, with an obtained overall error of the discrepancy on the pixel coordinates of $RMSE = 0.0130$ for the ToF image coordinates and $RMSE = 0.0075$ for the RGB image coordinates, and a $NCE_{ToF} = 1.997$ on the ToF camera coordinates system.

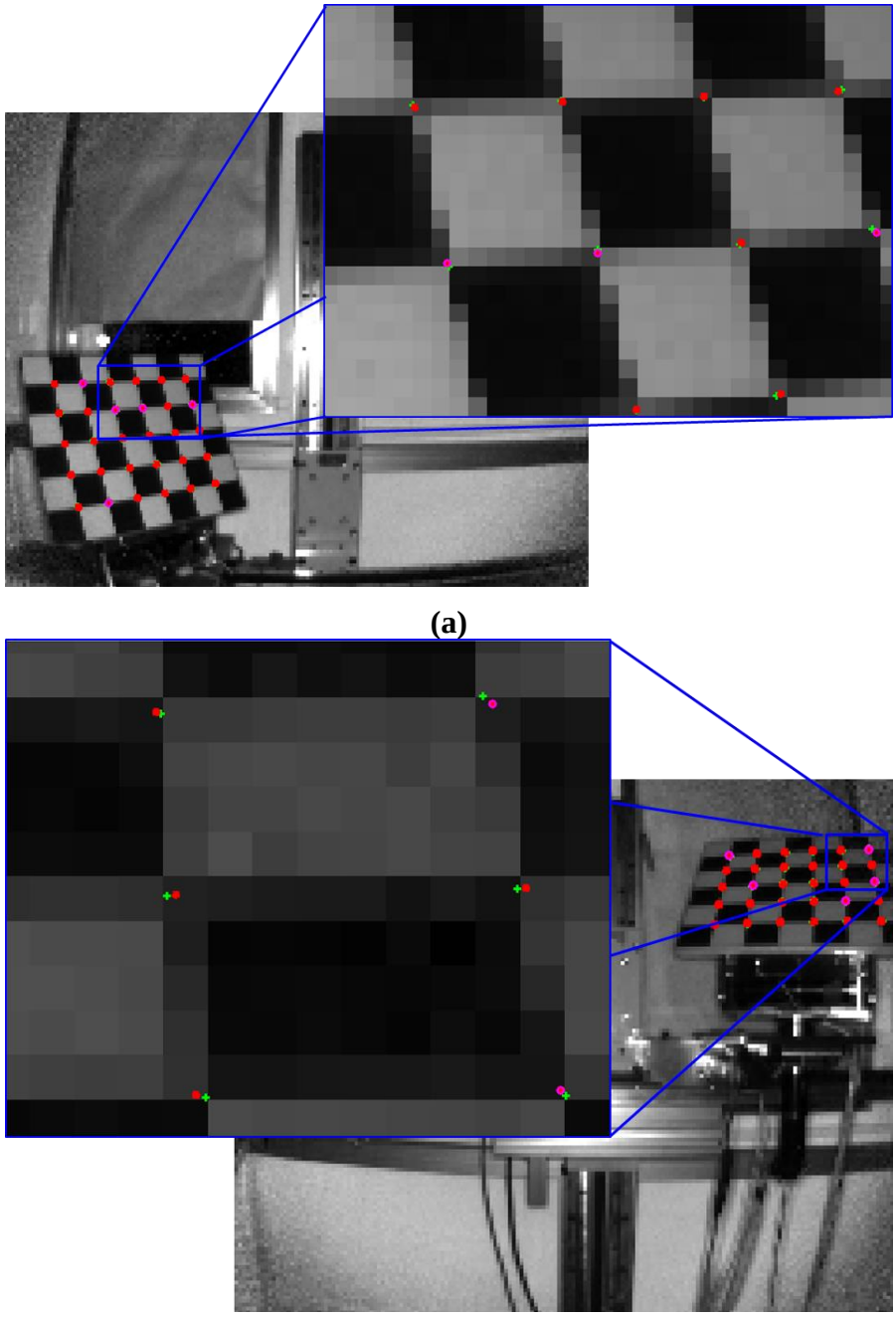
Regarding related investigations on the calibration accuracy issue, in (Wei and Cooperstock 2005) the authors tested the Zhang's algorithm in a casual and in an elaborate setup, with an obtained NCE of 2.56 and 1.67, respectively for each setup. In comparison with the obtained results in this Thesis and in spite of the lack of rigorous elaboration in the calibration setup, the accuracy of the results of this work outperforms their obtained accuracy with a casual setup results in 78%. Concerning their elaborate setup, the obtained results in this section are only 16 % less accurate in comparison with their results. Regarding the accuracy on pixel coordinates, the results of the calibration procedure carried out in this Thesis outperform both, casual and elaborate setup in terms of distorted pixel coordinates error.



**(a)**

**(b)**

**Figure 4.10** RGB camera potential outliers. (**a**) Image sample 44. (**b**) Image sample 46.



**(a)**

**(b)**

**Figure 4.11** ToF camera potential outliers. (**a**) Image sample 9. (**b**) Image sample 18.

On the other hand, the influence of the angle of the plane with respect to the image plane is evident, since the larger the angle the higher the deviation of the 2D or 3D error estimations. In conclusion, the comparison of the results obtained in this research with similar researches of calibration parameters accuracy, indicates that the calibration parameters for the ToF and the RGB camera are capable of back-projecting 2D data and re-projecting 3D points under satisfactory accuracy conditions. Thus, the Homogenous Transformation (Barrientos et al. 2007) between the ToF and the RGB camera can be achieved with sufficient accuracy as well.

## 4.3 Noise-free Data (ideal) Evaluation

Since depth measurements are commonly affected by noise, let assume that the back-projection of the control points on the ToF image plane into the 3D world coordinate are the ground truth depth measurements $\left(X_i^{GT}, Y_i^{GT}, Z_i^{GT}\right)$. Hence, in the standard calibration method validation, these true measurements are used for computing the depth map registration under ideal conditions. In order to carry out a quantitative comparison between the registration methods, the ideal depth measurements $\left(X_i^{GT}, Y_i^{GT}, Z_i^{GT}\right)$ are also applied in the dense map registration by means of the depth-dependent *Hlut* approach. However, since these truth measurements are estimated from the calibration parameters, and the depth-dependent *Hlut* does not consider these calibration parameters for constructing its entries. It would be expected that this method generates more deviations in comparison with the standard calibration.

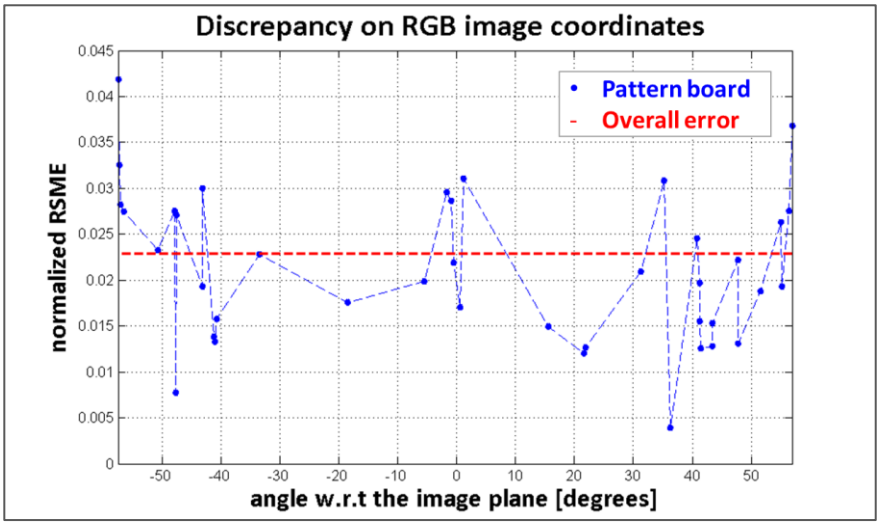
The criterion adopted for the depth map registration accuracy evaluation is the *Accuracy of the Distorted Image coordinates*. This criterion was introduced in Section 4.2. In this case, the normalized RMSE evaluates the geometric distance error between the ground truth control points on the RGB image coordinates $\left(\mathrm{x}_{\mathrm{i}_{RGB}}, \mathrm{y}_{\mathrm{i}_{RGB}}\right)$ and the transferred points $\left(\hat{\mathrm{x}}_{\mathrm{i}_{RGB}}, \hat{\mathrm{y}}_{\mathrm{i}_{RGB}}\right)$, which are computed from the ground truth control points on the ToF image coordinates $\left(\mathrm{x}_{\mathrm{i}_{ToF}}, \mathrm{y}_{\mathrm{i}_{ToF}}\right)$. Similar to the computation of the RMSE in Section 4.2, the residuals of the distance error of the estimates points $\left(\hat{\mathrm{x}}_{\mathrm{i}_{RGB}}, \hat{\mathrm{y}}_{\mathrm{i}_{RGB}}\right)$, and the predicted distance error ($dist_{pred} = 0$) are evaluated (see Equation (4.2)).

For the image registration obtained by means of the standard calibration method, the control points on the ToF amplitude image coordinate

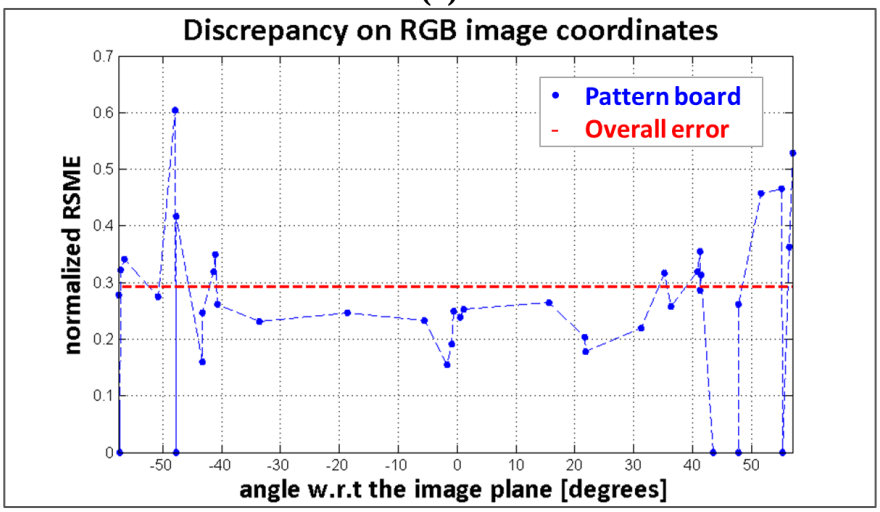
$(\mathrm{x}_{\mathrm{i}_{\mathrm{ToF}}}, \mathrm{y}_{\mathrm{i}_{\mathrm{ToF}}})$ are back-projected to the 3D world coordinates $(X_i^{GT}, Y_i^{GT}, Z_i^{GT})$ and then, these 3D points re-projected to the RGB image coordinates $(\hat{\mathrm{x}}_{\mathrm{i}_{\mathrm{RGB}}}, \hat{y}_{\mathrm{i}_{\mathrm{RGB}}})$. In the case of the depth map registration with the depth-dependent *Hlut*, the 3D ground truth measurements $(X_i^{GT}, Y_i^{GT}, Z_i^{GT})$ of the control points on the ToF image coordinates $(\mathrm{x}_{\mathrm{i}_{ToF}}, \mathrm{y}_{\mathrm{i}_{ToF}})$ were used as the input depth values for implementing the directives listed in Algorithm 3 (see Section 3.2).

Furthermore, complementary information was collected for the error analysis. . For instance, in Table 4.3 the geometric error distribution of the obtained results with the two considered methods is detailed (see page 66). On the other hand, in order to investigate the robustness of the calibration parameters, the influence of the angle of the pattern board with respect to the ToF image plane on the computed RMSE was evaluated. Figure 4.12 shows the results of this analysis. In this case, the influence of the pattern board orientation is not visible, since the control points selection was a procedure manually guided, and therefore, the ambiguities on the corner detection was avoided. Nevertheless, this result pointed out the capability of the obtained calibration parameters to deal with perspective problems.

The obtained results for the standard calibration methods, with a normalized RMSE $= 0.0229$, a mean value in pixel coordinates $Mean_{(u,v)-axis} = [0.45, 0.08]$, a standard deviation $\sigma_{(u,v)-axis} = [0.75, 0.65]$, and a geometric distance error such as the 95.6 % is $\leq 2\ pixels$, show that the estimated standard calibration parameters for Homogenous Transformation provide accurate data fusion between ToF and RGB cameras. On the other hand, the response of the depth-dependent *Hlut* is not as accurate as the results presented in Section 3.3, with a normalized RMSE $= 0.2935$, a mean value in pixel coordinates $Mean_{(u,v)-axis} = [-6.08, 4.33]$ and a standard deviation $\sigma_{(u,v)-axis} = [6.09, 9.78]$.




**(a)**




**(b)**

**Figure 4.12** Normalized RMSE on RGB camera coordinates vs the angle of the board plane w.r.t. the image plane. (**a**) Standard calibration method. (**b**) Depth-dependent *Hlut* approach.

**Table 4.3** Results of the Error Distribution: noise-free (GTD) depth values

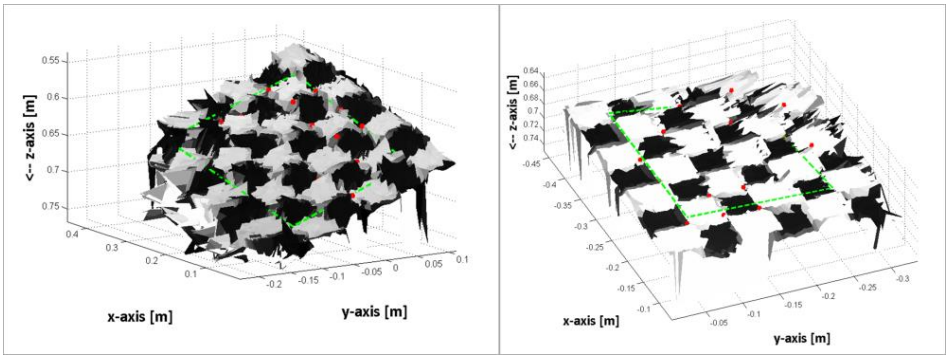
| **Error Distribution [pixels]** | **Error Percentage [%]** | | | | | |
|---|---|---|---|---|---|---|
| | ***Standard Calibration Method*** | | | ***Proposed method (Hlut)*** | | |
| | ***u-axis*** | ***v-axis*** | ***Geometric Dist.*** | ***u-axis*** | ***v-axis*** | ***Geometric Dist.*** |
| $error \leq \|1\|$ | 74.3 | 88.5 | 58.9 | 3.8 | 9.0 | 0.2 |
| $\|1\| < error \leq \|2\|$ | 22.9 | 11.2 | 36.7 | 3.0 | 9.4 | 0.4 |
| $\|2\| < error \leq \|3\|$ | 2.8 | 0.3 | 4.3 | 5.1 | 6.7 | 0.9 |
| $error > \|3\|$ | 0 | 0 | 0.1 | 88.1 | 74.9 | 98.5 |

**Table 4.4** Results of the Error Distribution: raw depth measurements

| **Error Distribution [pixels]** | **Error Percentage [%]** | | | | | |
|---|---|---|---|---|---|---|
| | ***Standard Calibration Method*** | | | ***Proposed method (Hlut)*** | | |
| | ***u-axis*** | ***v-axis*** | ***Geometric Dist.*** | ***u-axis*** | ***v-axis*** | ***Geometric Dist.*** |
| $error \leq \|3\|$ | 12.5 | 23.8 | 2.7 | 13.2 | 39.7 | 3 |
| $\|3\| < error \leq \|6\|$ | 11.5 | 22.5 | 7.2 | 28.6 | 29.7 | 16.6 |
| $\|6\| < error \leq \|9\|$ | 12.6 | 17.3 | 9.2 | 30.1 | 15.1 | 33.9 |
| $error > \|9\|$ | 63.4 | 36.4 | 80.9 | 28.1 | 15.9 | 46.5 |

## 4.4 Raw Depth Measurements Evaluation

For the image registration accuracy evaluation addressed in this section, the acquisitions of the raw depth measurements of the white-black pattern board poses were considered. In Figure 4.13, an illustration of the pattern board depth measurements is shown. As it was previously mentioned, for the depth map registration with the standard calibration parameters, the method described by (Park et al. 2011) was implemented, and for of the depth-dependent *Hlut* approach, the procedure described in Algorithm 3 (see section 3.3) was adopted.



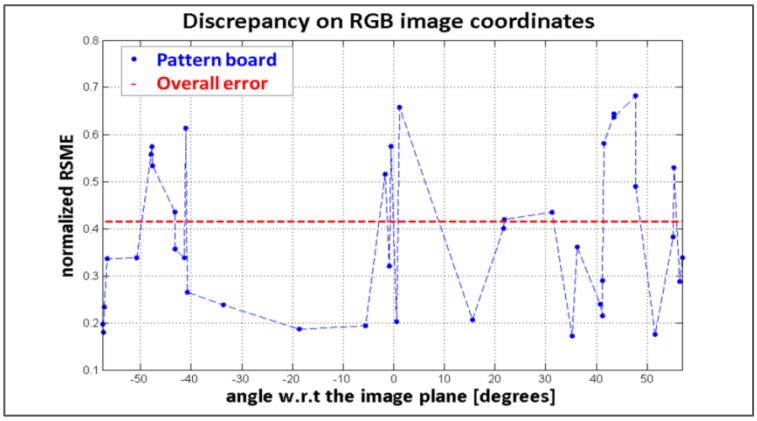

**(a)** **(b)**

**Figure 4.13** Rendering of the depth measurements and amplitude data acquired by the ToF camera. (**a**) Image sample 1. (**b**) Image sample 59.

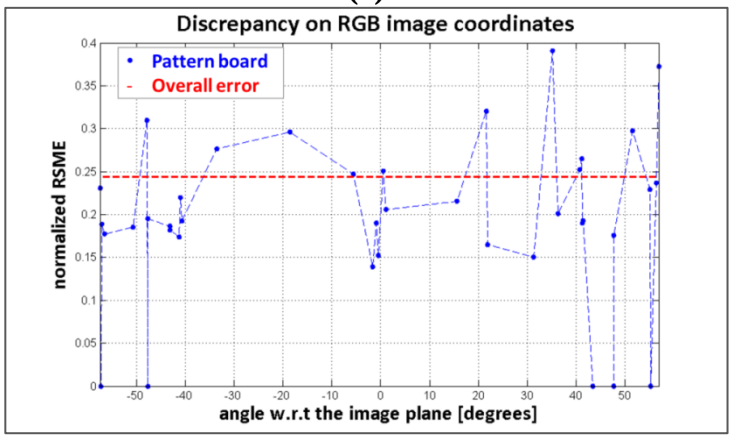
Consistently with the evaluation of the registration accuracy in Section 4.3, in this case the accuracy has also been measured following the criterion of *Accuracy of the Distorted Image coordinates*. For that purpose, the normalized RMSE was used to estimate the discrepancy between the ground control points on the RGB image coordinates $\left(\mathrm{x}_{\mathrm{i}_{RGB}}, \mathrm{y}_{\mathrm{i}_{RGB}}\right)$ and the mapped points on the RGB image coordinate $\left(\hat{\mathrm{x}}_{\mathrm{i}_{RGB}}, \hat{\mathrm{y}}_{\mathrm{i}_{RGB}}\right)$. These mapped points are projected from the control points on the ToF image coordinates $\left(\mathrm{x}_{\mathrm{i}_{ToF}}, \mathrm{y}_{\mathrm{i}_{ToF}}\right)$.

Regarding the numerical results of the accuracy evaluation of raw depth measurements, the method proposed in Thesis, presents an overall RMSE $= 0.2440$, reducing the error in 41 % in comparison with the standard calibration method, which exhibits an overall RMSE $= 0.4150$. In Figure 4.14 the results of the RMSE on each pattern board is shown. The computed

geometric error in pixel coordinates for the proposed approach provides a mean value $Mean_{(u,v)-axis} = [-6.4, 1.8]$ and a standard deviation $\sigma_{(u,v)-axis} = [6.3, 7.13]$, where the 20 % of the geometric distance error is $\leq 6\, pixel$. In contrast, the second method provides a mean value $Mean_{(u,v)-axis} = [-14.3, 5.6]$ and a standard deviation $\sigma_{(u,v)-axis} = [10.7, 7.4]$, where only the 10 % of the geometric distance error is $\leq 6\, pixel$. Therefore, the distribution error analysis also indicates that the depth-dependent *Hlut* approach outperforms the standard calibration method. In Table 4.4 (see page 66), the geometric error distribution is detailed.




**(a)**

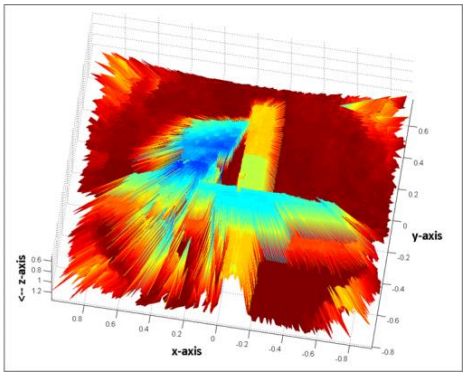



**(b)**

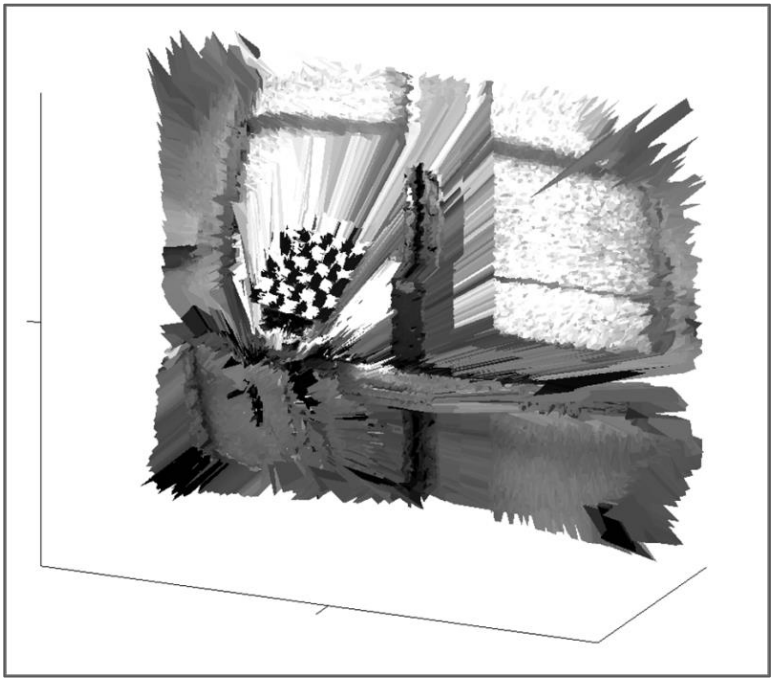
**Figure 4.14** Normalized RMSE on RGB camera coordinates vs the angle of the board plane w.r.t. the image plane. (**a**) Standard calibration method. (**b**) Depth-dependent *Hlut* approach.

The accuracy evaluation in terms of visual results is presented in Figure 4.15. For that purpose, the render of the colour depth map achieved with both methods are shown.

**(a)**

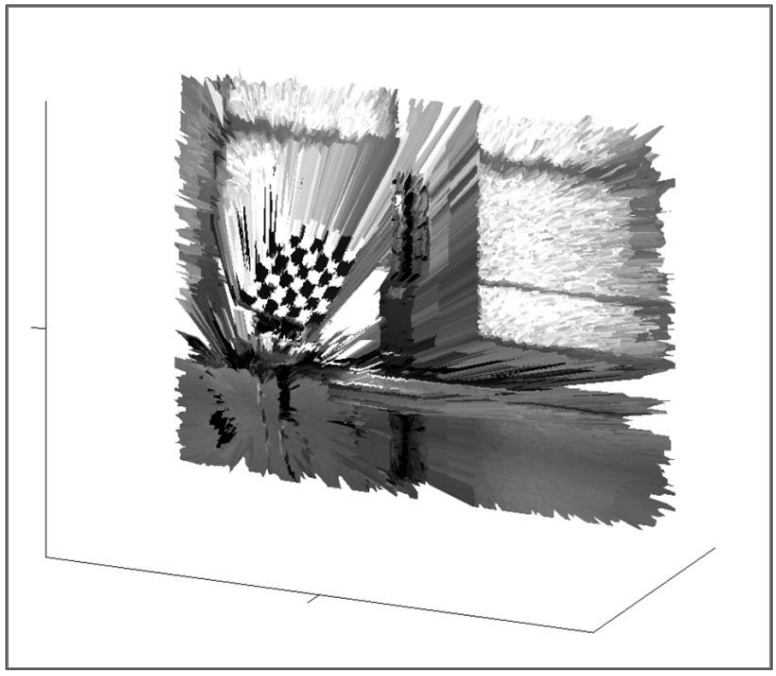



**(b)**



**(c)**

*Figure 4.15 Cont.*



**(d)**

**Figure 4.15** Depth map registration of sample 14. (**a**) RGB image. (**b**) ToF depth measurement. (**c**) Standard calibration result. (d) Depth-dependent Hlut result.

The obtained results show that the proposed method provides fewer oscillations when matching the depth and colour information. Consequently, a more visually homogenous surface on the pattern board is achieved. This is mainly because the proposal of this work considers a range of depth values, thus, slight fluctuations on the measurements are avoided. On the contrary, in the standard calibration method, the depth measurements are directly used for computing data fusion, thus an uneven pattern board surface is produced.

## 4.5 Filtered Depth Measurements Evaluation

The objective of this Section is to evaluate the performance of the registration methods when processing filtered raw depth measurements acquired by the ToF camera. Thus, for the evaluation of filtered data, the raw depth measurements were smoothed by applying the bilateral filtering

(Tomasi and Manduchi 1998) and the non-local means filter (Buades et al. 2005).

Since the evaluation of the filters is not matter of this research, the implementation of these techniques was not fully optimized. In this case, the characteristic behaviour of the denoising filtering for over smoothing edges, did not affect the inner grid on the pattern board that was used for the selection of the control points. However, the global results of the denoising filters modify the depth values on the pattern board. Consequently it is expected that the direct use of these flawed measurements produces misalignment problems. In Figure 4.16 the results of both denoise filtering implementation are illustrated, where the area enclosed in the green corresponds to the inner grid for the ground truth control point's extraction. The offset on the filtering depth values is more noticeable in the data processed with the non-local filter (see Figure 4.16(c)). Consequently the problems of data misalignments are expected to be more evident when implementing the depth map registration procedures with this data.

For the bilateral filtering implementation, the results are consistent with the ones obtained in Section 4.4. The depth-dependent *Hlut* outperforms the standard calibration method in terms of accuracy measured by normalized RMSE and the geometric error distribution, with an overall $\text{RMSE} = 0.2376$, a mean value $Mean_{(u,v)-axis} = [-6.2, 2.1]$ and a standard deviation $\sigma_{(u,v)-axis} = [5.9, 7.0]$. These results show slight improvements in comparison with the raw data processing. On the contrary, the obtained error with the standard calibration method increases when processing filtered data, with an overall $\text{RMSE} = 0.5402$, a mean value $Mean_{(u,v)-axis} = [-15.8, 5.6]$ and a standard deviation $\sigma_{(u,v)-axis} = [17.1, 10.5]$. The obtained result with the proposed approach reduces the error in 56% in comparison with the standard calibration method. Figure 4.17 shows the obtained RSME on each image sample of the pattern board and Table 4.5 summarizes the geometric error distribution.

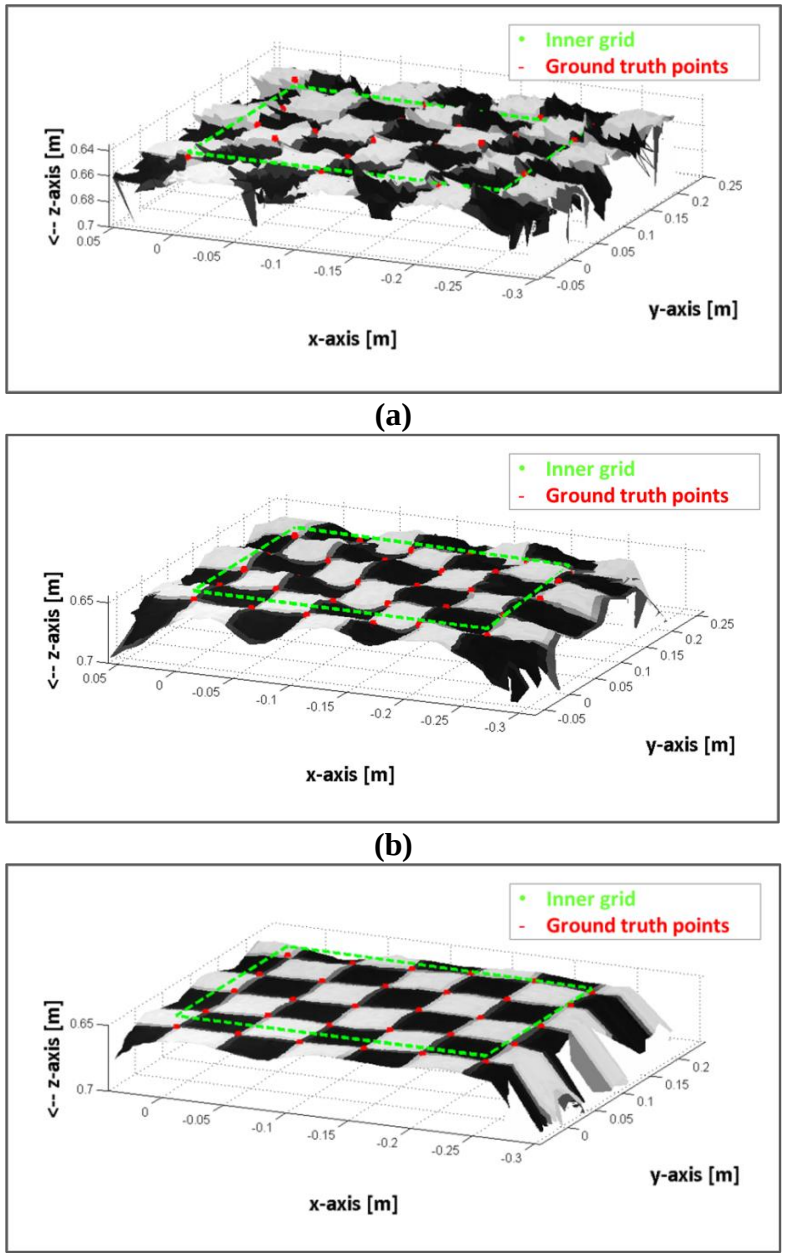
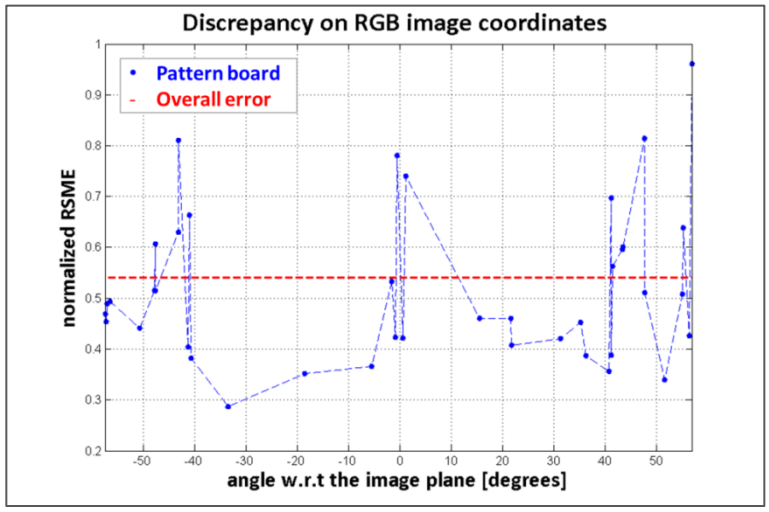


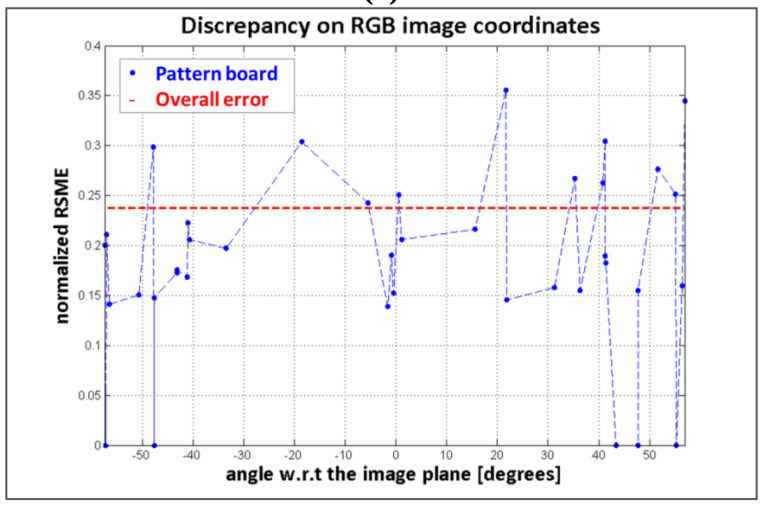

**Figure 4.16** Denoising filtering of sample 41. (**a**) Original data. (**b**) Bilateral filtering. (**c**) Non-local means filter.

In this case, of the data filtering data by means of the non-local filter, the normalized RMSE of the standard calibration method is increased in 216% compared with the error of the depth-dependent *Hlut* approach, where the proposed approach provides an overall RMSE $= 0.2365$, a mean value $Mean_{(u,v)-axis} = [-6.4, 1.8]$ and a standard deviation $\sigma_{(u,v)-axis} = [5.8, 7.0]$, while the standard calibration provides an overall RMSE $= 0.7478$, a mean value $Mean_{(u,v)-axis} = [-29.9, 15.2]$ and a standard deviation $\sigma_{(u,v)-axis} = [8.9, 9.8]$. These results exhibit the low capability of the standard calibration method for processing over smoothed data. This issue is also reflected in the geometric error distribution provided by the standard calibration method. In comparison with the results of the bilateral filtering, the results of the standard calibration method are increased in $[83.5\%, 171\%]$ for the mean values in *(u-v)*-axis respectively, whereas for the depth-dependent *Hlut* approach, the mean errors in *(u-v)*-axis are decreased in $[0\%, 14.2\%]$.

On the other hand, the analysis of the influence of the pattern board poses is illustrated in Figure 4.18, where the normalized RMSE versus the angle of the plane model of the pattern board with respect to the image plane is shown. Lastly, in Table 4.6, the geometric error distribution is detailed.

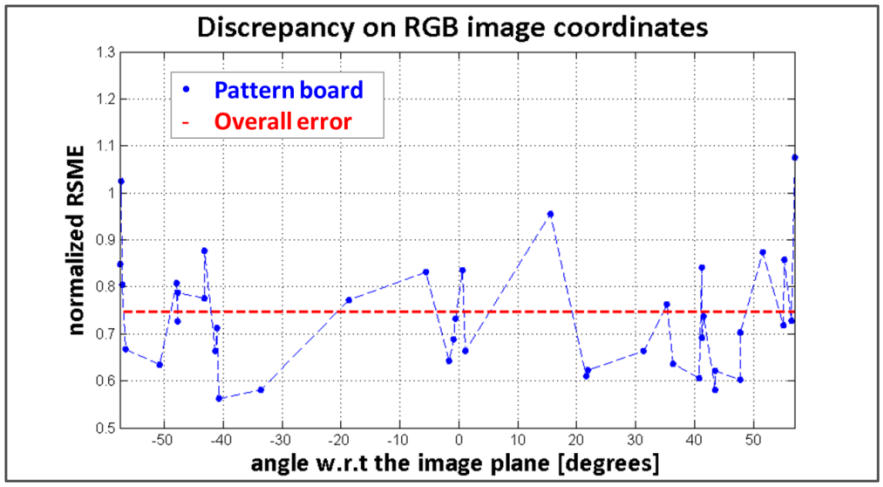



**(a)**

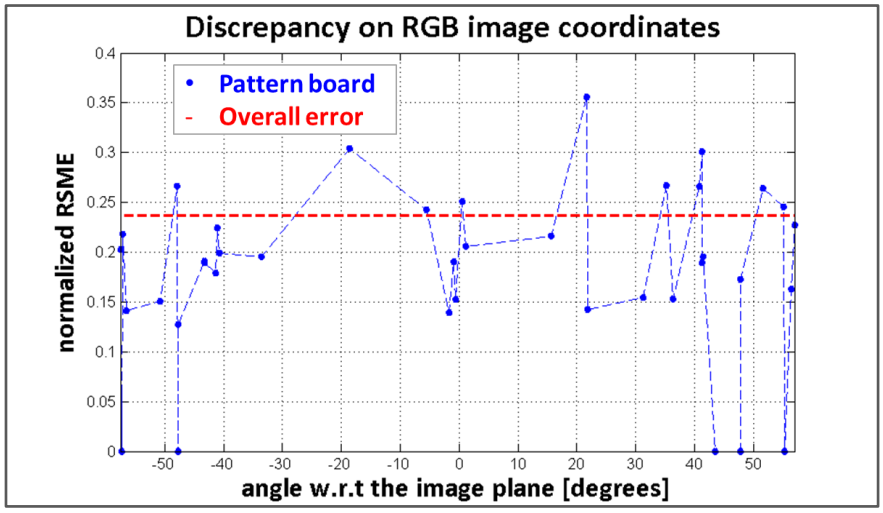



**(b)**

**Figure 4.17** Normalized RMSE on RGB camera coordinates vs the angle of the board plane w.r.t. the image plane. (**a**) Standard calibration method. (**b**) Depth-dependent *Hlut* approach.




**(a)**




**(b)**

**Figure 4.18** Normalized RMSE on RGB camera coordinates vs the angle of the board plane w.r.t. the image plane. (**a**) Standard calibration method. (**b**) Depth-dependent *Hlut* approach.

**Table 4.5** Results of the Error Distribution: filtered depth values - bilateral filtering

| **Error Distribution [pixels]** | **Error Percentage [%]** | | | | | |
|---|---|---|---|---|---|---|
| | ***Standard Calibration Method*** | | | ***Proposed method (Hlut)*** | | |
| | ***u-axis*** | ***v-axis*** | ***Geometric Dist.*** | ***u-axis*** | ***v-axis*** | ***Geometric Dist.*** |
| $error \leq \|3\|$ | 6.2 | 18.1 | 1.0 | 14.4 | 41.0 | 3.3 |
| $\|3\| < error \leq \|6\|$ | 5.3 | 21.6 | 3.0 | 28.5 | 29.2 | 17.3 |
| $\|6\| < error \leq \|9\|$ | 9.3 | 17.2 | 5.3 | 29.7 | 15.0 | 33.8 |
| $error > \|9\|$ | 79.2 | 43.1 | 90.7 | 27.4 | 14.8 | 45.6 |

**Table 4.6** Results of the Error Distribution: filtered depth values - non-local means filter

| **Error Distribution [pixels]** | **Error Percentage [%]** | | | | | |
|---|---|---|---|---|---|---|
| | ***Standard Calibration Method*** | | | ***Proposed method (Hlut)*** | | |
| | ***u-axis*** | ***v-axis*** | ***Geometric Dist.*** | ***u-axis*** | ***v-axis*** | ***Geometric Dist.*** |
| $error \leq \|3\|$ | 0.3 | 2.8 | 0 | 13.8 | 41.6 | 3 |
| $\|3\| < error \leq \|6\|$ | 0.3 | 3.7 | 0 | 28.7 | 29.0 | 17.3 |
| $\|6\| < error \leq \|9\|$ | 0.2 | 9.0 | 0.1 | 29.9 | 14.9 | 34.0 |
| $error > \|9\|$ | 99.2 | 84.5 | 99.9 | 27.6 | 14.5 | 45.7 |

## 4.6 Conclusions

In this Chapter an in-depth evaluation and comparison of the most common method for depth map registration and the proposed method in this Thesis were conducted. Since the two methods rely on depth measurements for implementing the dense map registration, the quality of the depth estimations is a crucial issue for achieving more accurate results. On the other hand, the state-of-the-art presented in Chapter 2 show that the noise in the depth measurements is a persistent problem. Consequently, the investigation of this Chapter was focused on the methods evaluation from the perspective of their response to the noise in the depth measurements.

First, the computation and the validation of the cameras calibration parameters were achieved. Then, three scenarios were considered for the methods comparison: noise-free (ideal) depth information, raw depth information and filtered depth information. For the ideal depth measurements, the standard calibration method evidently outperforms the proposed approach. That is because these true measurements were computed from the calibration results, and the entries of the *Hlut* $\left\{ H_k^{lut} \right\}$ were computed without considering the calibration parameters. On the contrary, when using raw and filtered depth measurements acquired from the ToF camera, the depth-dependent *Hlut* method outperforms the accuracy results of the standard calibration method in all scenarios. For instance, when processing raw data, the proposed approach in this work reduced the error in 41 % with respect the error of the standard calibration method. In the case of the filtered data, the obtained error with the standard calibration method is increased in 127% when using bilateral filtering and in 216% when using non-local filter, compared with corresponding errors of the proposed approach.

The results pointed out the high capability of the proposed method regarding to the standard calibration technique for dealing with slight variation in the depth estimation and for processing non-excessively over-smoothed filtered data.

# Chapter 5

# Experimental Results and Proposed Method Validation

## 5.1 Introduction

In order to evaluate the proposed method, two different and representative scenarios for close range objects detection in robotic applications were considered for the experimental stages of this work. The first experiment was conducted indoors, in a scenario that can be commonly used for robotic tasks such as the mobile robot navigation, the obstacle detection, the fall detection of people, the elderly assistance and others. The experiment was focused on testing the proposal approach in volumetric known objects, objects with large relief with respect to the image extent, objects placed all over the field of view of the system, and at different poses and orientations with respect to the image plane.

The second experiment was addressed toward precision agriculture (PA), one of the most relevant areas in robotics field. Currently, Universities, research groups, and small and large companies, supported by ambitious projects funded by the European Community and other international entities, are joining efforts to investigate and to put in practice the advances on the precision agriculture field. This is the case of the *Intelligent Sensing and Manipulation for Sustainable Production and Harvesting of High Value Crops, Clever Robots for Crops (CROPS)* project funded by the European

Union through the Seventh Framework Program, Grant Agreement Number 246252. The second experiment of this chapter is enclosed under the scope of the CROPS project, where the implementation of a multisensory system, and the depth-dependent *Hlut* approach for the detection and localization of fruits was investigated.

## 5.2 Man-made Indoor Environments

Intelligent service robots are becoming a major interest area for multiple applications in the society. Special attention is been paid to the personal security and assistance of people. These robotic applications are mostly oriented to deliver assistance to elderly people living on their own, and monitoring the children's safety at their homes. In both cases, systems are devoted to identify dangerous situations such as people falling, falling objects or long periods of inactivity. For that purpose, the robotic application needs to be able to track people's motion and detect obstacles. Falls are frequent among elderly people, and are a still-underestimated medical problem with respect to causes and consequences. Falls can have immediate lethal results but also produce many disabling fractures and dramatic psychological problems which reduce elderly people's independence. Research has found that half of those patients with a "long lie" (i.e., those remaining on the floor for more than one hour after a fall) died within six months of the fall, even if there was no direct physical injury (Zambanini and Machajdik 2010). Thus, immediate alarming and helping is essential to reduce the rate of morbidity and mortality (Wild, Nayak and Isaacs 1981). The technical solutions that have been proposed for the detection of falls can be classified into three groups (Noury et al. 2007): wearable device-based, ambience device-based and vision-based methods. Though, in-house vision systems provide several advantages over other sensors: they are able to detect several events simultaneously, do not disturb or interfere in the daily activities of people, provide richer and more accurate data, and report fewer false alarms than other devices.

Most of the related works are based on colour cameras, some of them using static cameras at each room (Foroughi, Aski and Pourreza 2008, Cucchiara et al. 2005). In other cases, the authors propose the implementation of vision system mounted on mobile robots (Di Paola et al. 2008). Stereo vision techniques and multi-cameras approaches have been utilized for acquiring 3D information of the scenes, avoiding occlusions of people and covering large areas (Zambanini and Machajdik 2010). Unfortunately, the configuration of these cameras demands rigorous camera calibration procedure, and involves heavy computational load which restrains real-time operation. On the other hand, simple omnidirectional vision systems offer 360º view angle of the indoor scene in a single image (Ming-Liang, Chi-Chang and Huei-Yung 2006), but preclude the attainment of reliable spatial

data, thus requiring passive methods for the acquisition of the 3D information of the scene, such as the structure from motion or shape. An omnidirectional stereo configuration approach was presented in (Salinas et al. 2011). This work presented a catadioptric rectified configuration for providing 3D information of a scene, by means of a single acquisition of the system in combination with passive triangulation techniques. The system was capable of acquiring large areas, though the spatial resolution was limited as well.

Recent works have introduced the use of ToF cameras for analysing human poses in falling detection (Diraco, Leone and Siciliano 2010) as well as probabilistic methods for motion capturing (Ganapathi et al. 2010). The lack of resolution of the ToF cameras and their incapability of providing contextual information make them less accurate for people's motion detection. Lately, the use of the Kinect® sensor (Microsoft Research 2009) for the skeleton and body motion analysis in fall detection applications has been introduced by (Gasparrini et al. 2014, Rougier et al. 2011). Evidently, object's motion detection is a complex task, in particular in real life environments, where the occlusions and the dynamic changing nature of the scenes take a great deal when processing the data. Therefore, the acquisition of richer and quality information of the 3D world scene is a key stage for achieving proper objects detection.

A solution that fuses high-resolution images and depth information is a promising approach for close range detection of people motion, since this solution provides detailed contextual information, as well as the scene structure at every snapshot of the sensory system. This combination eludes in certain degree, the lack of resolution of ToF cameras and the features mismatching problem of stereo triangulation methods.

In order to evaluate the image registration method proposed in this Thesis in man-made indoor environments, two series of experimental tests were conducted. Details of these experiments are described below:

1. The first group of experiments were focused on the method assessment for registering continuous surfaces angled with respect to the sensory system. For instance, a continuous planar surface which is angled with respect to the image plane might be transformed by several homographies $\{H_k^{lut}\}$. This was achieved by modifying the perspective view of the white-red pattern board. The experimentation setup consisted of the four degrees of freedom robotic platform and the sensory system described in section 3.1, and the 3 × 5 white-red

chessboard of 50 mm each square considered in Section 3.2. The sensory system was mounted on the robotic platform, and the pattern board was positioned in front of the cameras, with the inner grid aligned with respect to the centre of the ToF camera. Then, 25 image samples from different poses of the sensory rig were acquired, and a total of 288 control points were evaluated. The results of the registration procedure of two image samples are illustrated in Figures 5.1 and 5.2. The transformation matrices $\{H_k^{lut}\}$ applied for transferring the points on the ToF pixel coordinates $\{xcp_i^{ToF}\}$ are illustrated with coloured marks (see Figures 5.1(b) and 5.2(b)). Each colour represents an entry *k* of each homography $\{H_k^{lut}\}$, likewise for the mapped points on the RGB pixel coordinates $\{xmap_i^{RGB}\}$ (see Figures 5.1(a) and 5.2(a)).

2. The second group of experiments are focused on the proposed method evaluation for registering images of 3D man-made environments and its implementation on people/object motion detection. These scenes are composed by volumetric objects made of different materials, where the relief of these 3D objects is high enough with respect to the extent of the image view. Hence 3D-space points of an object do not belong to a unique plane and consequently, several homographies $\{H_k^{lut}\}$ should map the object's points. In order to estimate the error of the mapped points, white-red landmarks were attached to some objects among 47 image samples, where a total of 767 control points were evaluated. In this case, the sensory rig was static and the objects were positioned at different distances and poses within the depth of field of the system. In addition, sequences of moving objects and descriptive people's falling postures were also analysed. Figures 5.3 and 5.4 show the results of the registration process for two image samples.

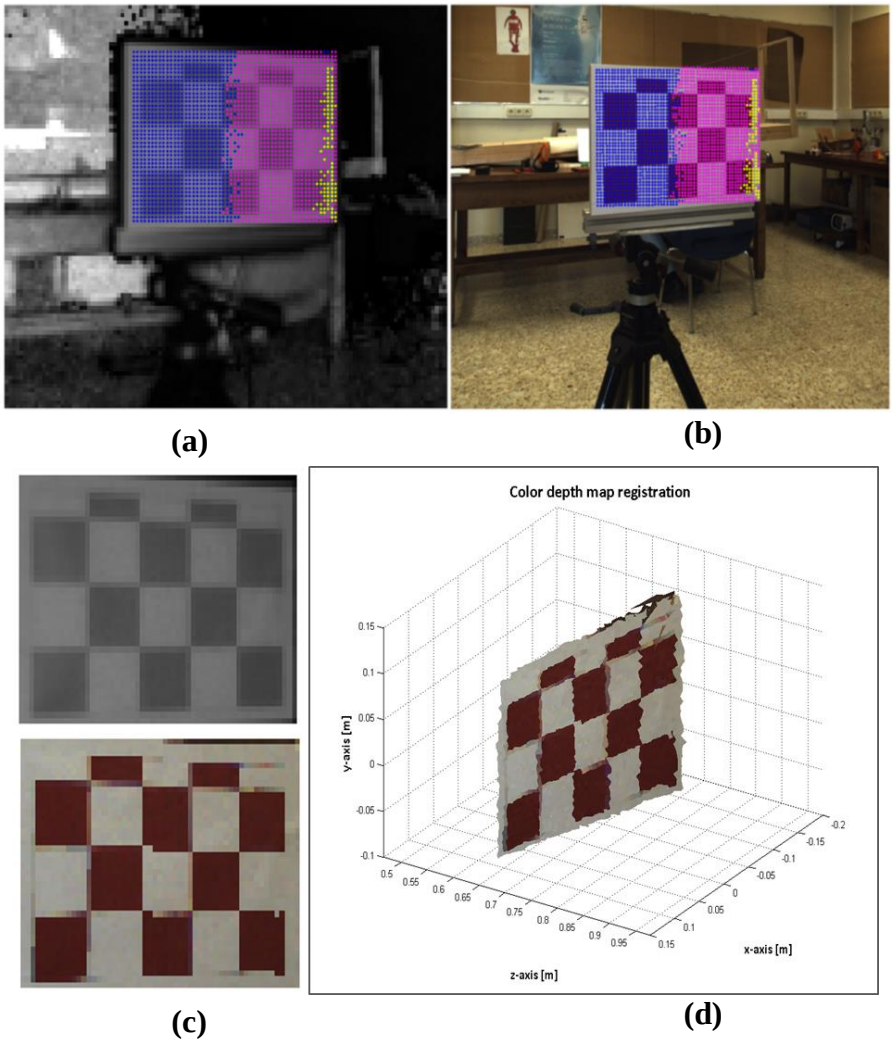



**Figure 5.1** Image sample 19. (**a**) Selected points on the ToF. (**b**) Estimated points on the RGB image. (**c**) Top: selected ToF ROI. Bottom: estimated RGB ROI. (**d**) Colour depth map of the ROI.

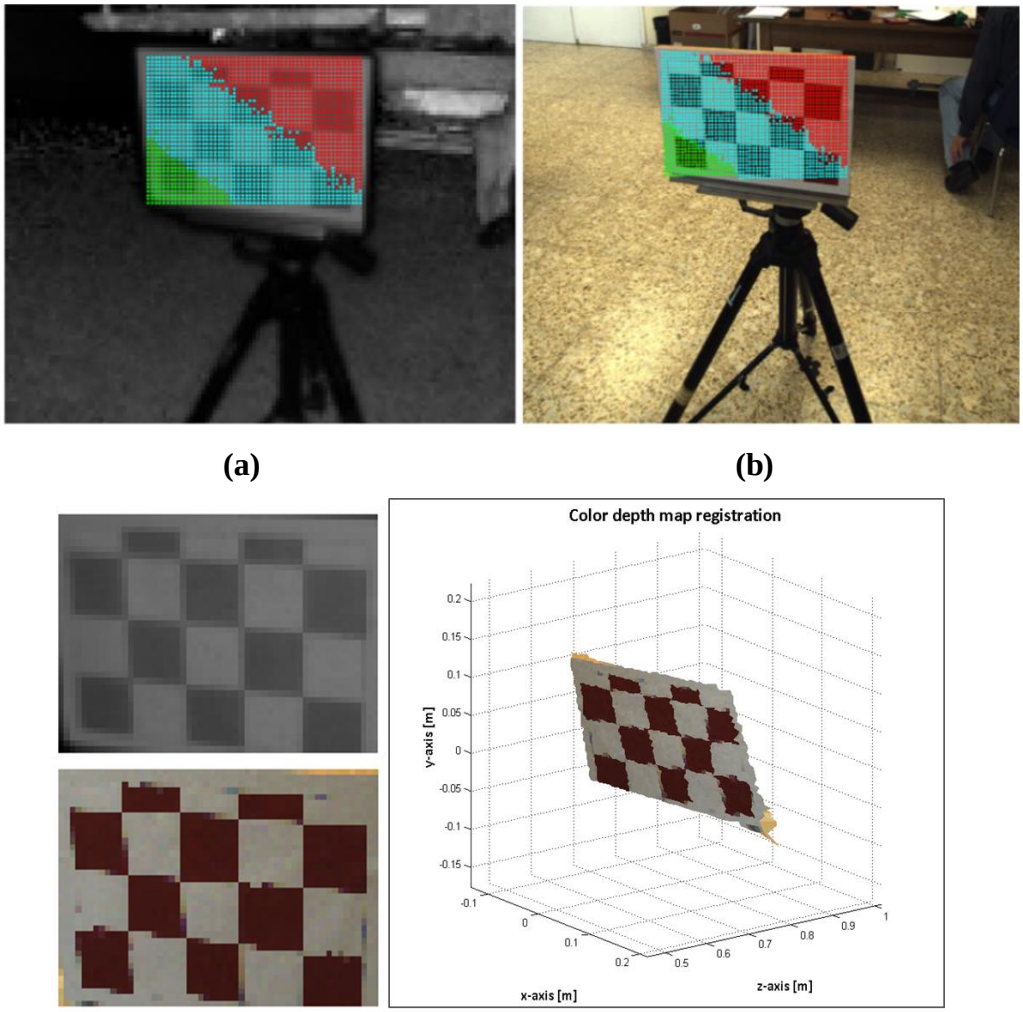
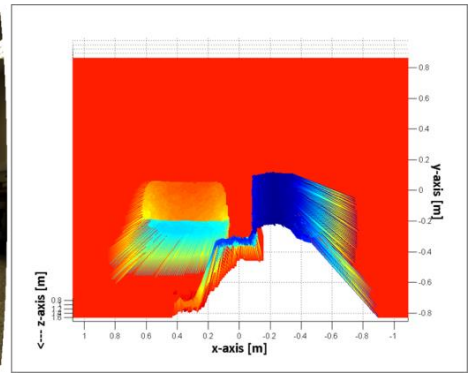
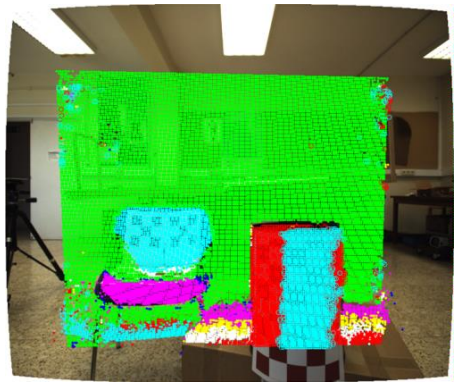



**Figure 5.2** Image sample 25. (**a**) Selected points on the ToF. (**b**) Estimated points on the RGB image. (**c**) Top: selected ToF ROI. Bottom: estimated RGB ROI. (**d**) Colour depth map of the ROI.

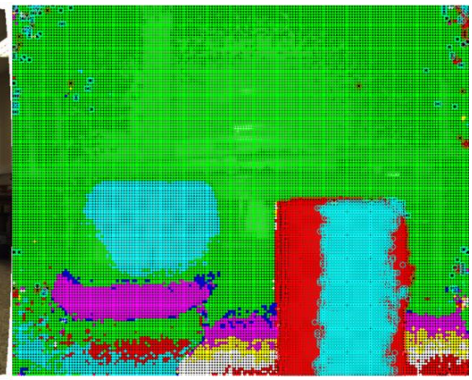
**(a)**




**(b)**

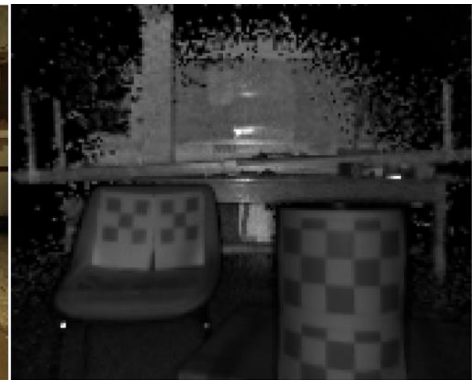


**(c)**



**(d)**

**(e)**



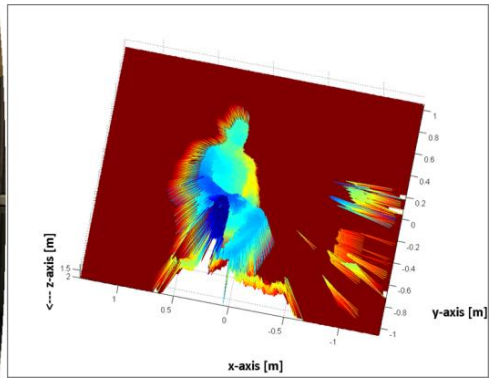
**(f)**

*Figure 5.3 Cont.*

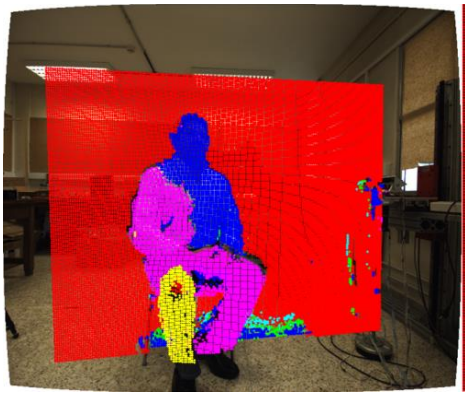
**(g)**

**Figure 5.3** Image sample 19. (**a**) RGB image. (**b**) ToF depth measurements. (**c**) Homography labelled $\boldsymbol{mask_{LRGB}}$ on RGB image coordinates. (**d**) Homography labelled $\boldsymbol{mask_{LRGB}}$ on ToF image coordinates. (**e**) Registered RGB image. (**f**) ToF amplitude image. (**g**) Colour depth map.

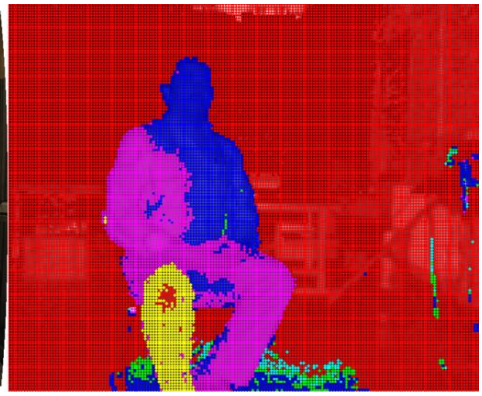
**(a)**




**(b)**



**(c)**

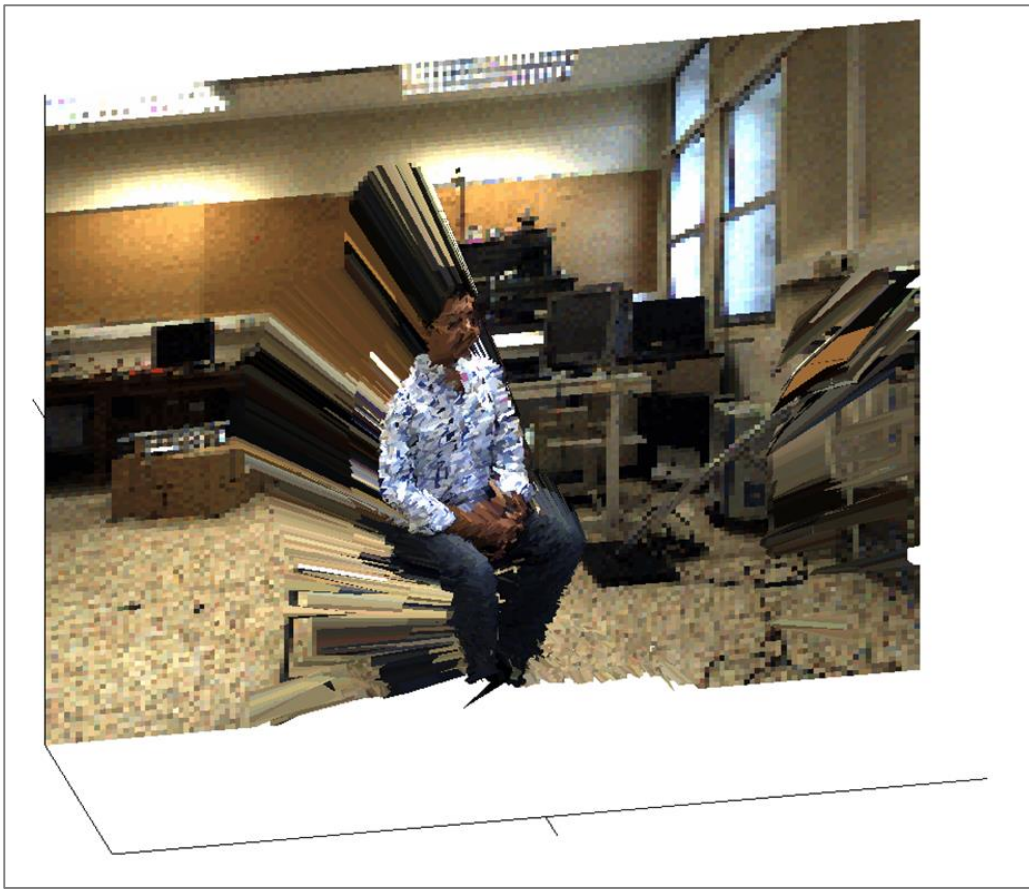


**(d)**

**(e)**

**(f)**

*Figure 5.4 Cont.*



**(g)**

**Figure 5.4** Image sample 8. (**a**) RGB image. (**b**) ToF depth measurements. (**c**) Homography labelled $\boldsymbol{mask_{LRGB}}$ on RGB image coordinates. (**d**) Homography labelled $\boldsymbol{mask_{LRGB}}$on ToF image coordinates. (**e**) Registered RGB image. (**f**) ToF amplitude image. (**g**) Colour depth map.

### *5.2.1 Results and evaluation*

Two series of experimental tests were conducted to evaluate the proposed image registration method in man-made indoor environment. In order to carry out a quantitative assessment of the error, white-red landmarks were attached to the objects of interest. Then, correspondence ground control points were selected on the RGB and ToF image coordinates, and the *Accuracy of the Undistorted Image Coordinates* criterion (Salvi et al. 2002) was evaluated.

For that purpose, the geometric and distance errors (see Equation 3.2) were computed. The results of the error distribution are detailed in Table 5.1, while in Figures 5.5 and 5.6 the results of the geometric and distance error for each group of experimental test are illustrated. The discrepancies for the first group of experiments, in terms of normalized RMSE is 0.1383, and the mean value and the standard deviation are $Mean_{(u,v)-axis} = [-0.4, 0.6]$ and $\sigma_{(u,v)-axis} = [4.3,4.7]$, respectively, where the 70 % of the geometric distance error is $\leq 6\, pixel$.

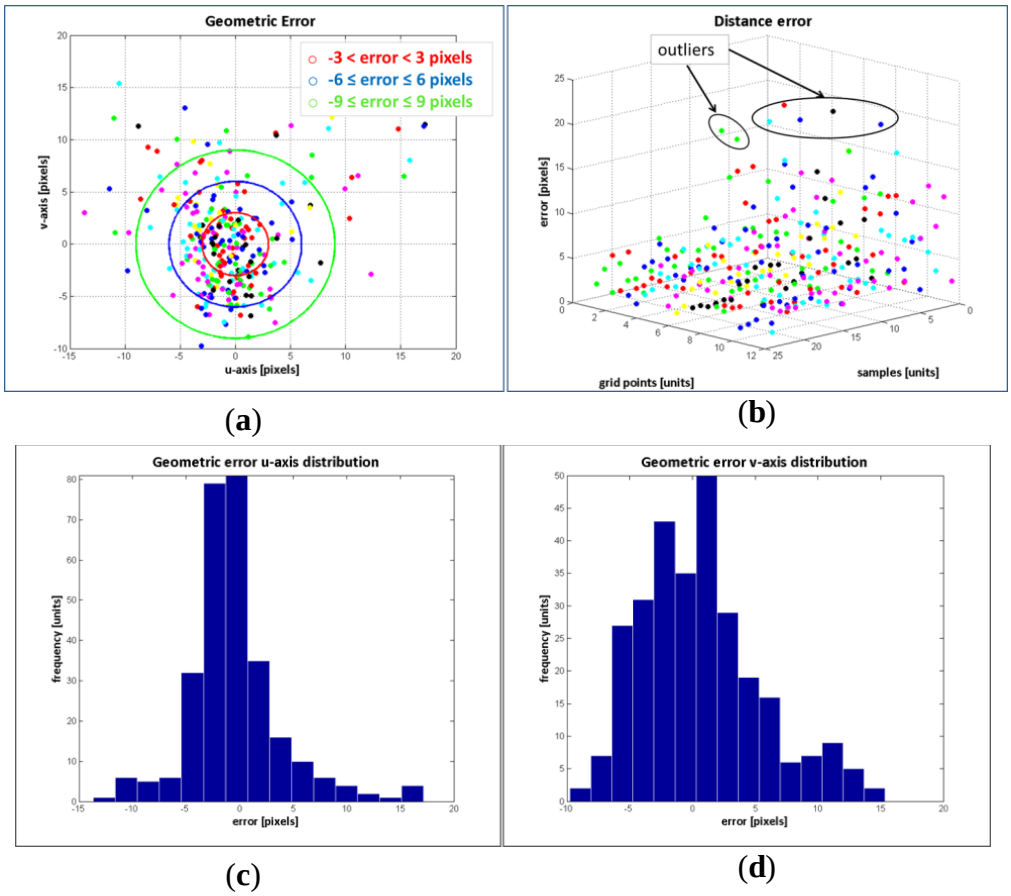



**Figure 5.5** Error evaluation of the experimental tests corresponding to group #1. (**a**) Geometric error. (**b**) Distance error. (**c**) Error distribution in *u*-axis. (**d**) Error distribution in *v*-axis.

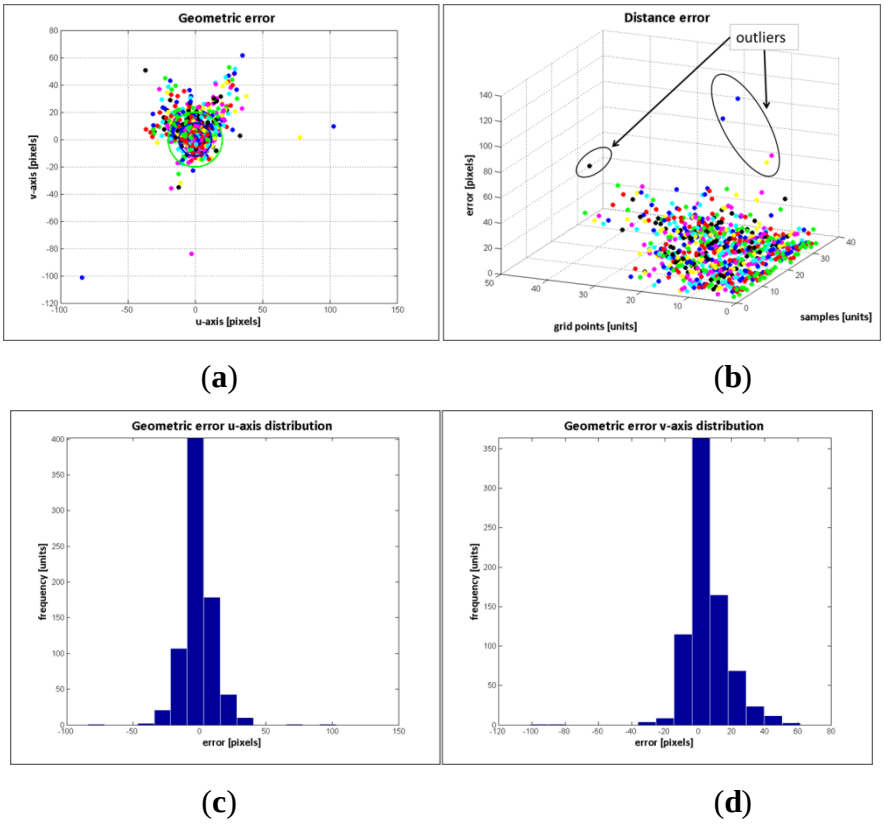



**Figure 5.6** Error evaluation of the experimental tests corresponding to group #2. (**a**) Geometric error. (**b**) Distance error. (**c**) Error distribution in *u*-axis. (**d**) Error distribution in *v*-axis.

The results for evaluating the estimated data on the RGB image coordinates $\{xmap_i^{RGB}\}$, show that the response of the proposed registration method is quite promising, given that a large number of the mapped points have geometric errors less than 3 pixels on the RGB frame and the RMSE is quite low. Most of the erroneous data are due to the depth measurement variations and the flawed data selected as correspondence ground control points. Similar to the accuracy analysis presented in Section 3.3, the outliers might be removed to calculate the accuracy of the depth-dependent *Hlut* approach. However, in this case, only the correspondence points defined as outliers were removed, only 6 points of a set of 288 control points were identified as flawed points, and the accuracy of the test was improved in 8 %. The errors estimations were carry out once again, and the obtained error

distribution was also included in Table 5.1. The results in terms of normalized RMSE are 0.1272, the mean value is RMSE are 0.1272 and the mean value is $Mean_{(u,v)-axis} = [-0.7, -0.3]$ and the standard deviation $\sigma_{(u,v)-axis} = [3.5,4.6]$.

In the second group of experiments, the presence of flawed correspondence data is more visible, mostly because of the object's shape and perspective with respect to the image coordinates. As it shown in the obtained results, with a normalized $RMSE = 0.4005$, a mean value $Mean_{(u,v)-axis} = [-0.16, 5.6]$ and a standard deviation $\sigma_{(u,v)-axis} = [12.2,13.1]$, where only the 20 % of the geometric distance error is $\leq 6\ pixel$. In Figure 5.6(a-b), the identification of the outliers is quite noticeable. Regarding the outliers evaluation, only control points considered as flaws were removed to calculate the errors once again. In this case only 11 points of the 767 sample set were removed and the accuracy was improved 12 %. The results of the error distribution are also detailed in Table 5.1, with a normalized $RMSE = 0.3511$, a mean value $Mean_{(u,v)-axis} = [-0.46, 5.5]$ and a standard deviation $\sigma_{(u,v)-axis} = [10.5,11.3]$.

Regarding mismatching correspondence points, three fundamental factors should be considered: the large difference between the cameras resolution, the foreshortening, which makes less precise the digitalization of some features of the objects, and the noise in the depth measurements, which comprises the intrinsic noise and the one derived from the object albedo. For instance, flying pixels, multi-path interference and features distortion of the depth measurements are expected.

The presence of noise in the depth measurements is a persistent problem when using ToF cameras. The solutions are constricted to filtering techniques, which usually modify or over smooth the original values. In Chapter 4, it has been demonstrated that the depth-dependent method is capable of dealing with small variation in the depth measurements and that this method outperforms the accuracy results obtained with the standard calibration method.

**Table 5.1** Results of the Error Distribution

| Error Distribution (pixels) | | Group #1 | | | Group #2 | | |
|---|---|---|---|---|---|---|---|
| | | *Error Percentage [%]* | | | | | |
| | | ***u-axis*** | ***v-axis*** | ***Geometric Distance*** | ***u-axis*** | ***v-axis*** | ***Geometric Distance*** |
| **Entire samples set** | $error \leq \|3\|$ | 62.2 | 49.3 | 28.8 | 33.0 | 29.1 | 10.4 |
| | $\|3\| < error \leq \|6\|$ | 25.0 | 33.3 | 42.4 | 21.6 | 19.3 | 18.0 |
| | $\|6\| < error \leq \|9\|$ | 6.6 | 9.7 | 14.9 | 12.4 | 13.8 | 16.8 |
| | $error > \|9\|$ | 6.2 | 7.7 | 13.9 | 33.0 | 37.8 | 54.8 |
| **Outliers removed** | $error \leq \|3\|$ | 63.5 | 50.4 | 29.4 | 33.3 | 29.4 | 10.6 |
| | $\|3\| < error \leq \|6\|$ | 25.5 | 34.0 | 43.3 | 22.0 | 19.6 | 18.2 |
| | $\|6\| < error \leq \|9\|$ | 6.7 | 9.6 | 15.3 | 12.6 | 14.0 | 17.1 |
| | $error > \|9\|$ | 4.2 | 6.0 | 12.0 | 32.1 | 37.0 | 54.1 |

For examining the influences of the orientation of the objects with respect to the image plane, the plane model of the pattern grid was estimated from the depth measurements of the ToF camera, and for the volumetric objects, a quadric cylindrical approximation was adopted. Since depth values are slightly noisy, the RANSAC algorithm (Fischler and Bolles 1981) was used to compute the plane model and the quadric cylindrical approximation of the objects of interest. In Figure 5.7 the objects approximation and the computation of its angle $(\beta)$ with respect to the image plane are exemplified. For convenience, the angle is represented such as $\beta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. In Figures 5.8(a) and 5.8(b) the normalized RMSE of each sample and the angle $(\beta)$are illustrated, for the first and second group of experimental tests, respectively

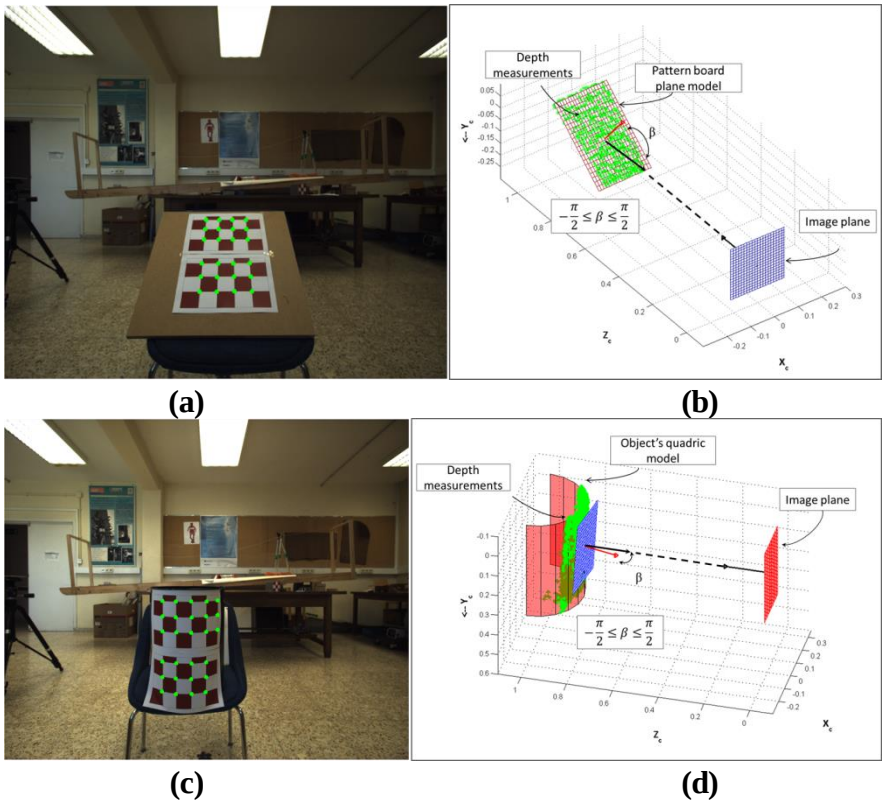
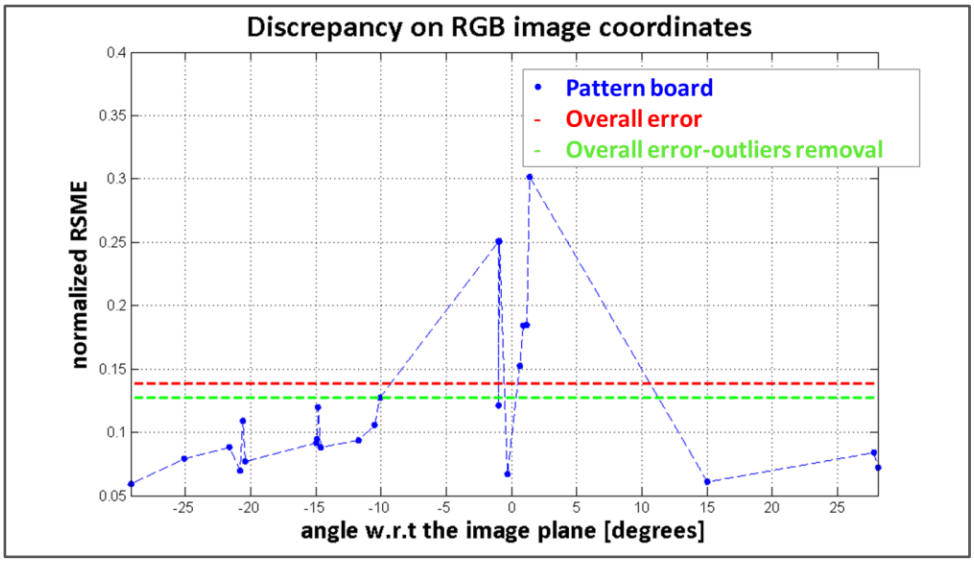
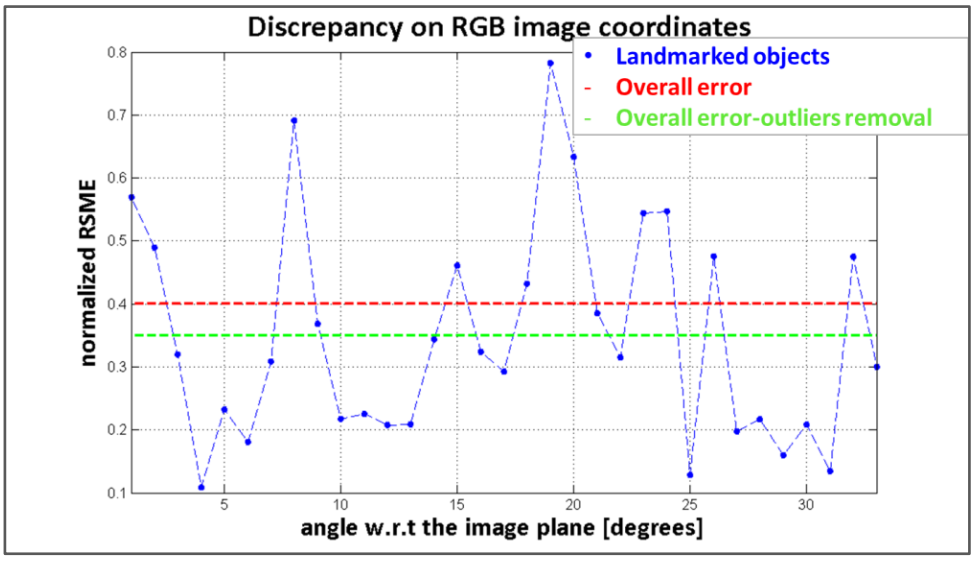



**Figure 5.7** Estimation of the angle $(\boldsymbol{\beta})$ of the objects approximation w.r.t. the image plane. (**a**) RGB image and control points (sample 31). (**b**) Plane model of the pattern board (sample 31). (**c**) RGB image and control points (sample 12). (**d**) Quadric model of an object (sample 12).




**(a)**

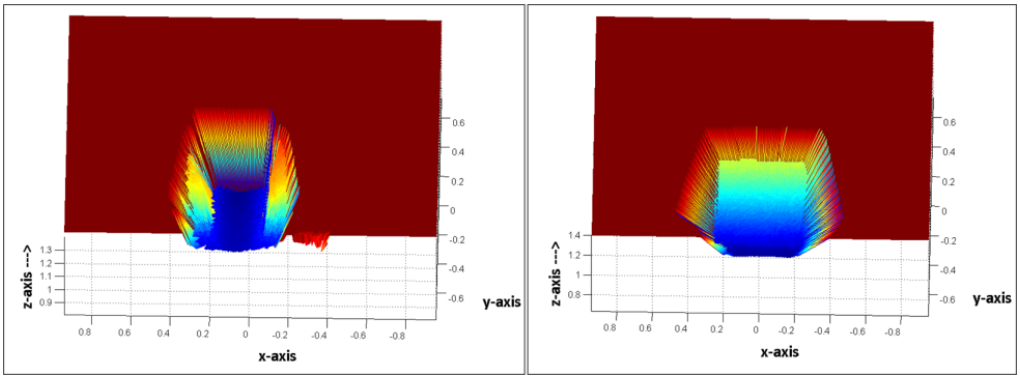



**(b)**

**Figure 5.8** Normalized RMSE on RGB pixel coordinates vs the angle ($\boldsymbol{\beta}$) of the object's approximation w.r.t. the image plane. (**a**) First group of experiments. (**b**) Second group of experiments.

The goodness of the depth-dependent *Hlut* approach in comparison with the standard calibration method is shown in Figures 5.10-5.13, where the visual results of computing depth map registration of image samples 12 and 31 are illustrated. The results of the depth map registration by means of the standard calibration method for the two image samples are shown in Figures 5.10 and 5.12 respectively. While in Figures 5.11 and 5.13, the depth-dependent *Hlut* depth map registration results are presented. Figure 5.7 illustrates the RGB images of the samples 12 and 32, and their corresponding 3D depth maps are in Figure 5.9. For the methods implementation, the raw depth measurements from the ToF camera were used.

The presence of noise in the depth values is denoted by slight variation on the surfaces and edges of the objects. Despite the obtained errors, the visual results of the colour depth map reconstruction show the capability of the proposed registering method for preserving the object's edges and shape. It is possible to notice that in those areas where the data is properly matched, the presence of artifacts or misalignments problems is in general avoided. This contrasts with the standard calibration method, where the object's surfaces are less homogeneous and the object's edges exhibits several misalignment problems.




**(a)** **(b)**

**Figure 5.9** Depth measurements acquired by the ToF camera (in meters). (**a**) Image sample 12. (**b**) Image sample 31.

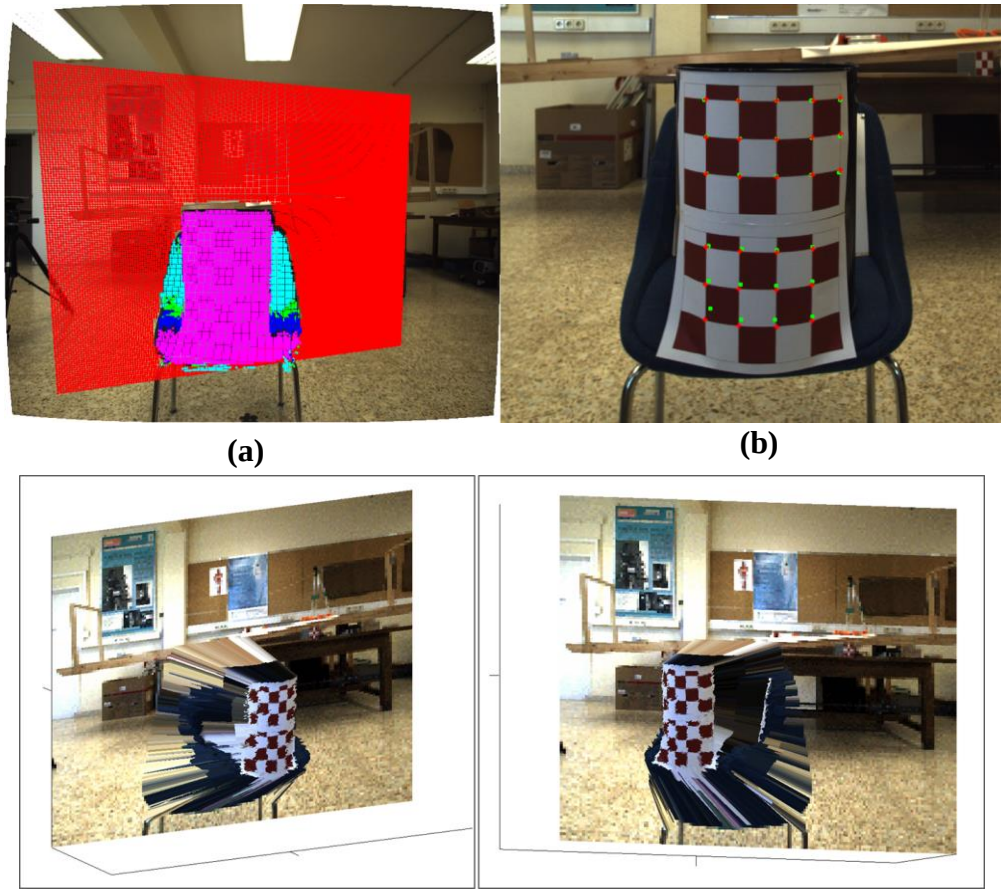
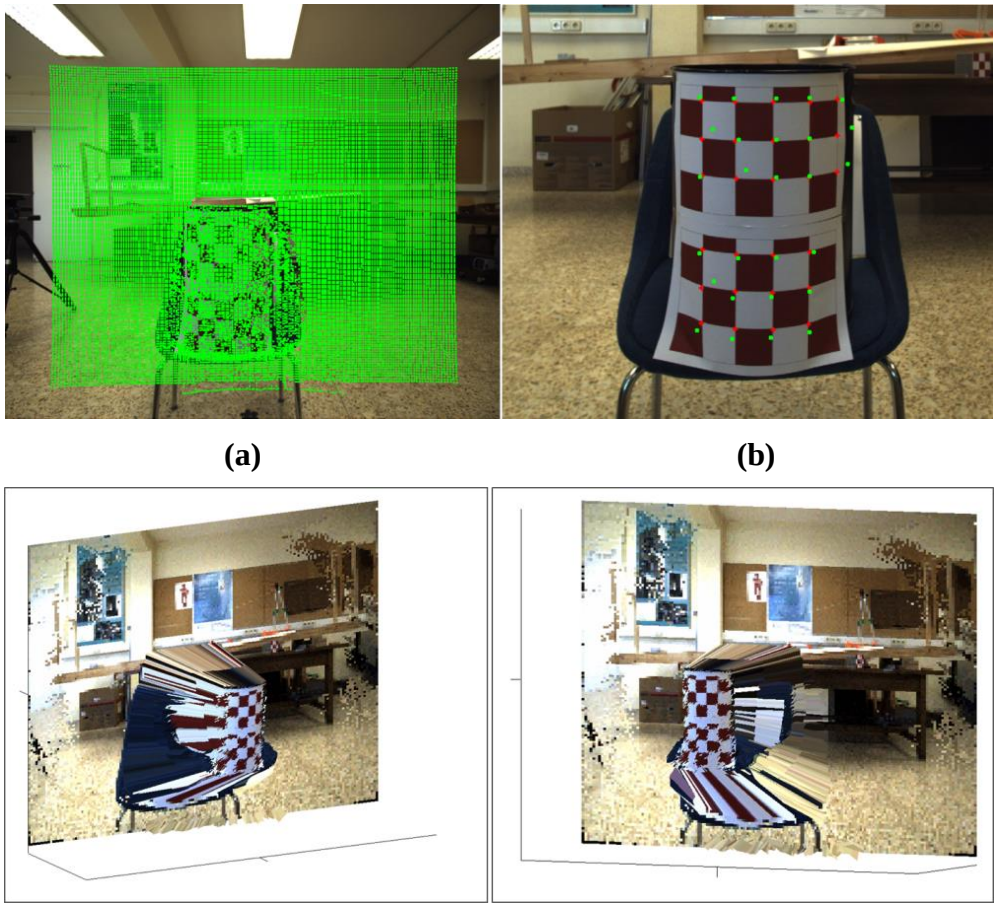


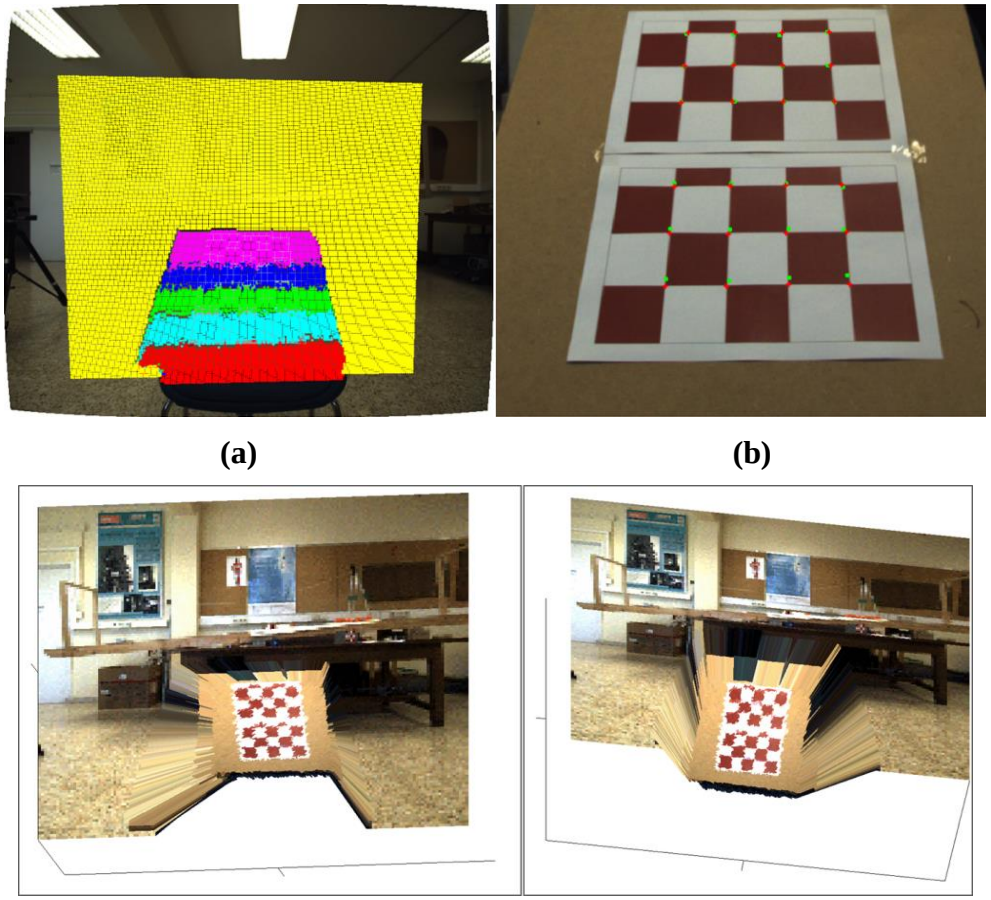
**Figure 5.10** Depth map registration of sample 12 with depth-dependent *Hlut* approach. (**a**) Homography labelled mask, where each colour represents a homography of the *Hlut*. (**b**) Correspondence control points (red) and estimated points (green). (**c**) Low resolution colour depth map – view 1. (**d**) Low resolution colour depth map – view 2.



**(a)** **(b)**

**(c)** **(d)**

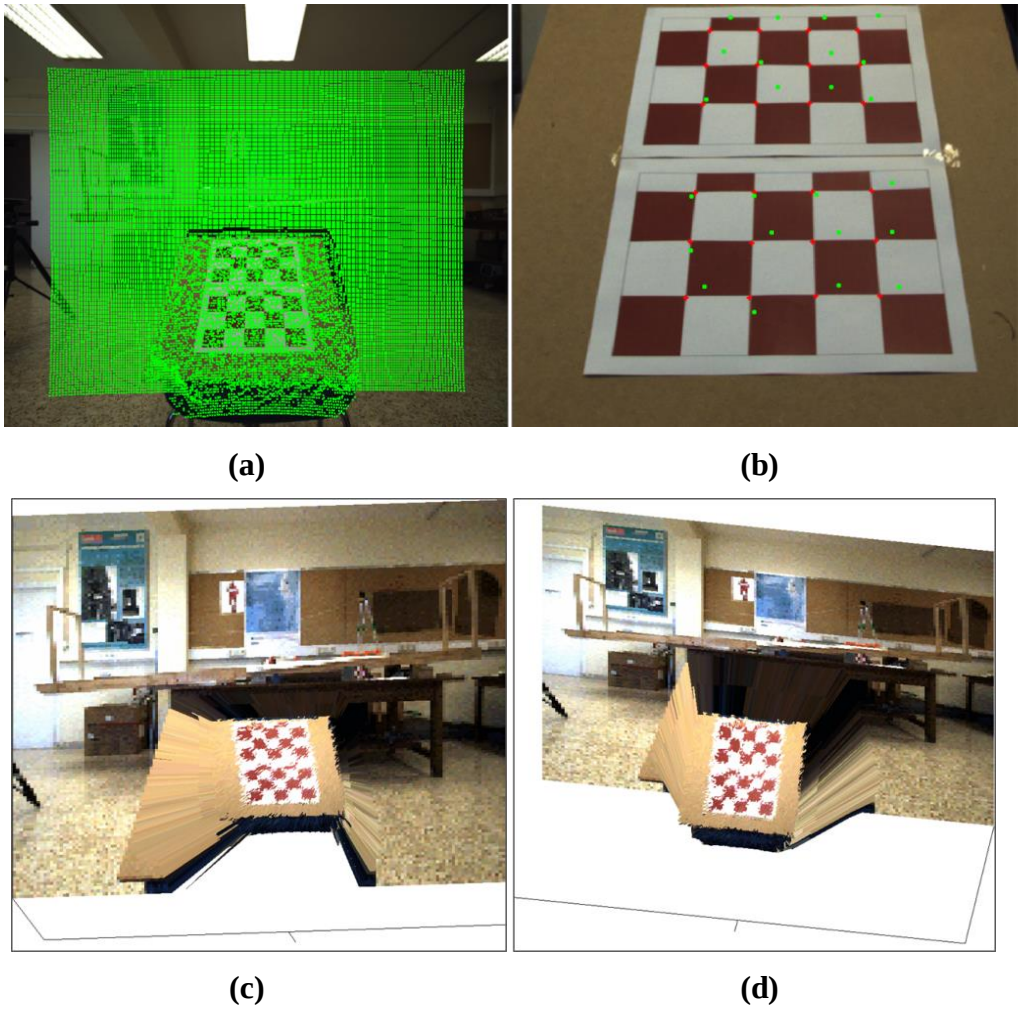
**Figure 5.11** Depth map registration of sample 31 with standard calibration method. (**a**) Registered points on RGB pixel coordinates. (**b**) Correspondence control points (red) and estimated points (green). (**c**) Low resolution colour depth map – view 1. (**d**) Low resolution colour depth map – view 2.



**(a)** **(b)**

**(c)** **(d)**

**Figure 5.12** Depth map registration of sample 31 with depth-dependent *Hlut* approach. (**a**) Homography labelled mask, where each colour represents a homography of the *Hlut*; (**b**) Correspondence control points (red) and estimated points (green); (**c**) Low resolution colour depth map – view 1; (**d**) Low resolution colour depth map – view 2.



**(a)** **(b)**

**(c)** **(d)**

**Figure 5.13** Depth map registration of sample 31 with standard calibration method. (**a**) Registered points on RGB pixel coordinates. (**b**) Correspondence control points (red) and estimated points (green). (**c**) Low resolution colour depth map – view 1. (**d**) Low resolution colour depth map – view 2.

Evidently noise in depth measurements is a drawback for computing image registration. Nonetheless, the proposed method has shown to be robust enough to deal with variations within ±12 mm. The object distance is also a significant issue, since the sensing of the object's edges, borders and dimensions are altered because of the low ToF camera resolution, and

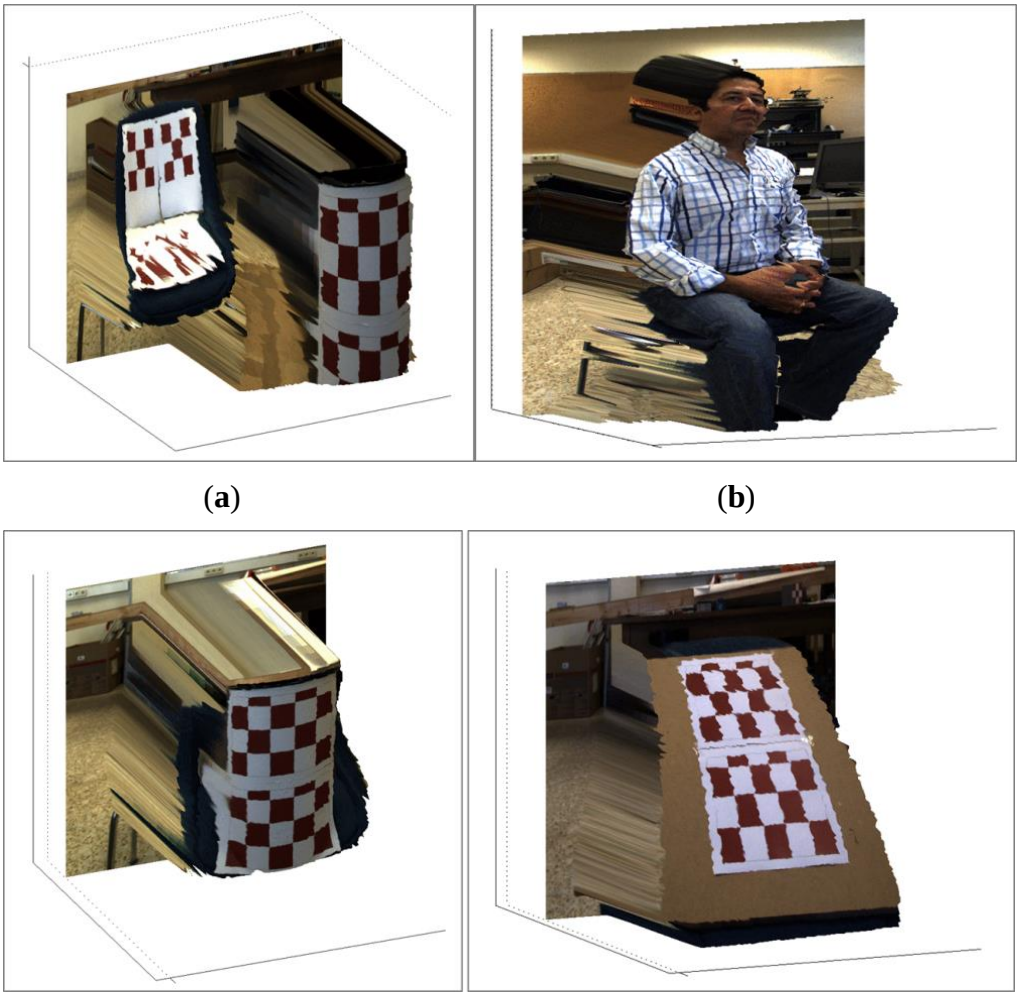
because of the working principle of these cameras. In addition due to the large difference in the camera resolution, when mapping points from the ToF to the RGB image, there are approximately 10 pixels of unmapped points between adjacent mapped points.

Regarding the ToF data improvement and enlargement, the proposed method provides a labelled homography mask $mask_{LRGB}$ on the RGB pixel coordinates, which corresponds to the entries $k$ of the homographies $H_k^{lut}$ used to transfer the data. For the ToF camera resolution enlargement, it is possible to classify the unmapped points on the RGB image coordinates by assigning them a $H_k^{lut}$. Hitherto, the proposed method was able to put colour on the depth map while dealing with slight variation on the depth measurements. Now by using entire classified homography labelled mask $mask_{LRGB}$, the method is also capable of assigning depth to the colour information. Thus, a high-resolution colour depth map is computed by transferring the RGB data to the ToF frame such that:

$$xmap_i^{ToF} = H_k^{lut^{-1}} x_i^{RGB} \tag{5.1}$$

The obtained rendering results of the high resolution colour depth map of 5 Megapixels in size are displayed in Figure 5.14. The high resolution colour depth maps were computed from the images samples 8, 12, 19 and 31, depicted in Figures 5.4, 5.3 and 5.7, respectively. The visual results show a satisfactory performance of the proposed method. For instance, a cylinder bucket, a chair, or the angled board, which are continuous surfaces, were properly mapped, without any presence of discontinuity on their surfaces, as well as the person who is sitting on the chair. All of these objects represent typical targets in close range detection. The shape and edges of objects presented almost no difficulties or artifacts, and so far, no enhancement algorithm has been implemented yet.

Regarding noise filtering and data enhancement, once the registration has been done, the depth along with the labelled homography mask $mask_{LRGB}$ and the corresponding RGB colour information could be used to extrapolate the depth measurements by means of guided or weight-based methods, whereas the values on $mask_{LRGB}$ might be used as weights. In this way, boundaries could be properly classified and edges enhancement procedure could be computed and, finally, high quality maps could be achieved.

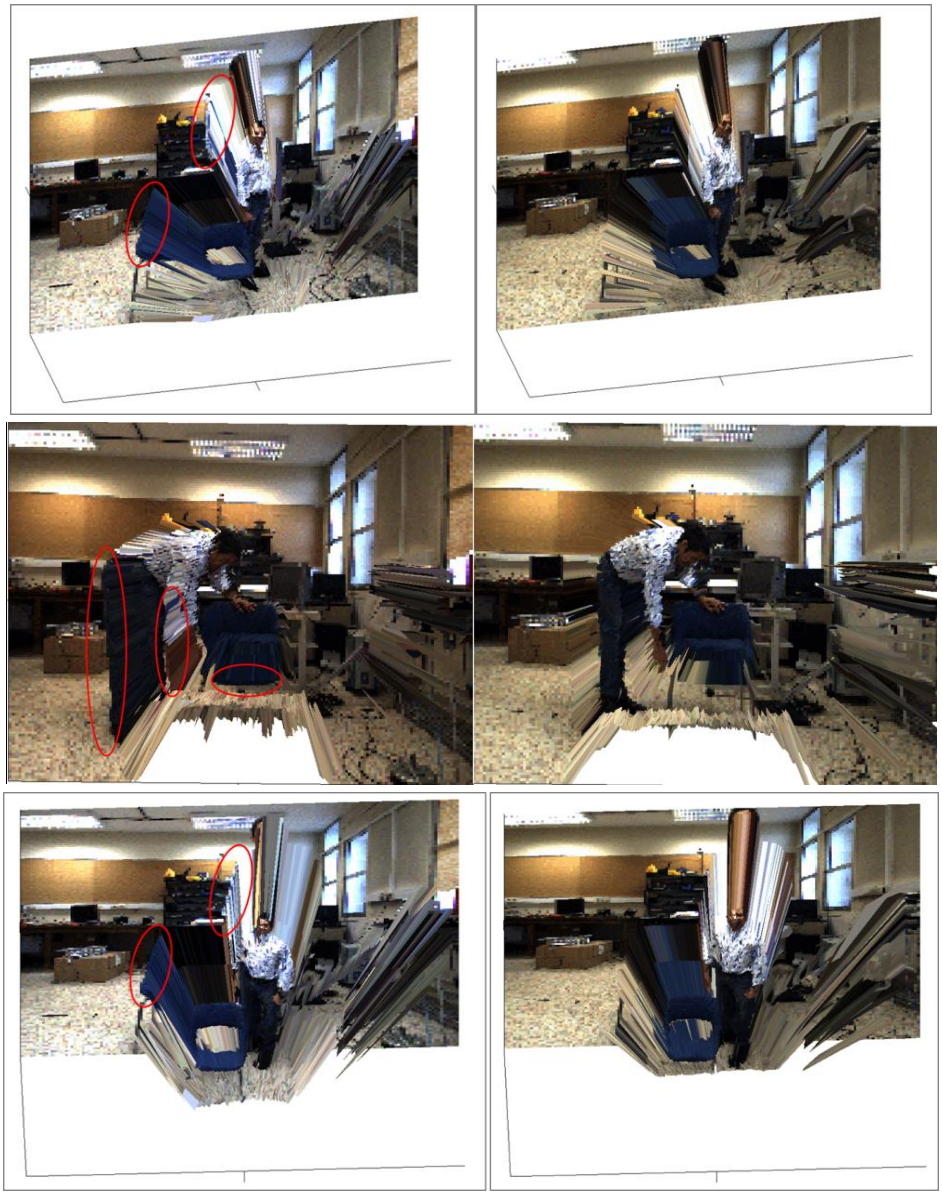


(**a**) (**b**)

(**c**) (**d**)

**Figure 5.14** High resolution colour depth map reconstruction. (**a**) Two volumetric objects placed at different distances from the sensory system (sample 19). (**b**) An object with a large relief with respect to the image extent (sample 8). (**c**) A curved object (sample 12). (**d**) A continuous surface which is slanted with respect to the cameras axis (sample 31).

Additionally, a comparative visual assessment between the image registration results achieved with the standard calibration method and with the proposal approach is presented in Figure 5.15.



**(a)** **(b)**

**Figure 5.15** Colour depth map registration results of image samples 11, 12 and 17. (**a**) Standard calibration method results. (**b**) Depth-dependent *Hlut* approach results.

In this comparison, two different postures of a person were computed. The colour depth maps obtained by means of the standard calibration exhibit several artifacts in the objects edges. For instance, the depth information of the chair boundaries are not aligned and are inconsistent with of the colour information of the chair boundaries, as it is illustrated in Figure 5.15(a) and highlighted with the help of the red marks. Several pixels that belong to the person (the arm, the leg, the back) and to the chair are mapped as part of the background.

On the contrary, the obtained results with the depth-dependent *Hlut* approach, which are depicted in Figures 5.15(b), show a satisfactory alignment between depth and colour information of the objects boundaries. Once again, the proposed approach outperforms the standard calibration method in terms of accuracy and proper alignment of the mapped data.

For the method evaluation oriented to in-house video surveillance and people motion detection task, sequences of people movements were analysed. In the acquired sequences, common occlusions caused by objects in the room are represented, along with illustrative people's falling postures. The experimental test aims the evaluation of the capabilities of the depth-dependent *Hlut* approach in tracking objects applications. The proposed procedure for detecting variance in image regions is based on a robust multiple objects motion detection algorithm, introduced by Black and Anandan (1996), with the combination of a quadric surface approximation of the 3D objects structure of the computed inlier motions. The proposed procedure is depicted in Figure 5.16, and in Figures 5.20-5.23, the visual results of the image registration procedure and the motion detection task are shown.

The proposed procedure for objects motion detection, computes the structure from motion analysis on the amplitude images acquired by the ToF camera instead of using the RGB images, in order to avoid false inliers due to the illumination conditions and shadows

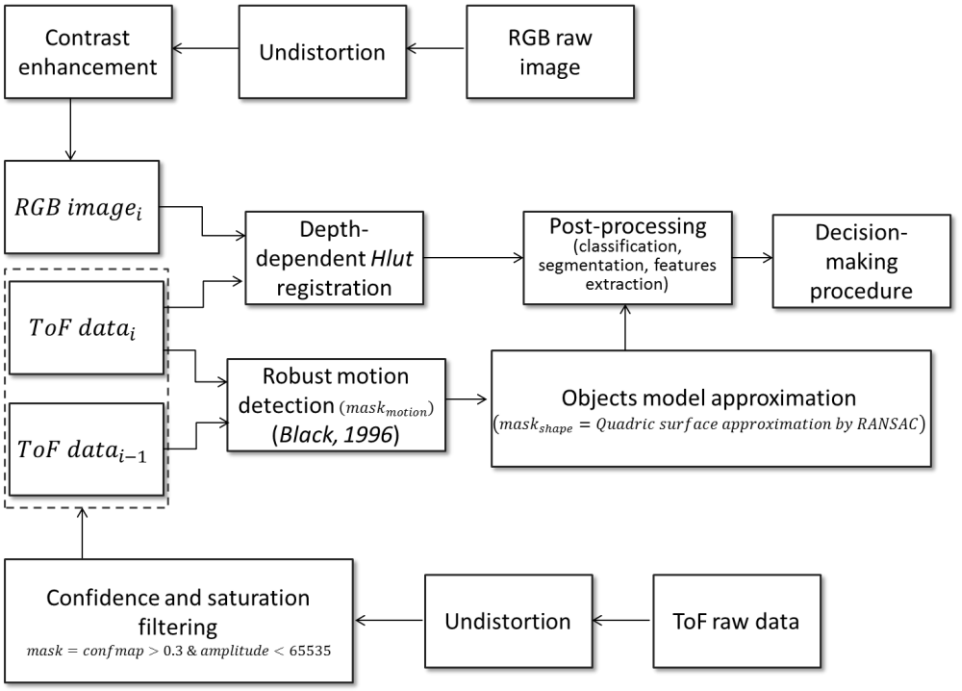



**Figure 5.16** Procedure for objects motion detection.

Figures 5.18 and 5.19 show a comparison of the visual results of structure from motion computation on RGB registered images (176×144 pixels) and on amplitude images (176×144 pixels), respectively. While the input image pair for the motion detection analysis is presented in Figure 5.17.

**Figure 5.17** Image pair of a person's falling sequence. (**a**) Image sample $\boldsymbol{sample_t}$. (**b**) Image sample $\boldsymbol{sample_{t+1}}$.

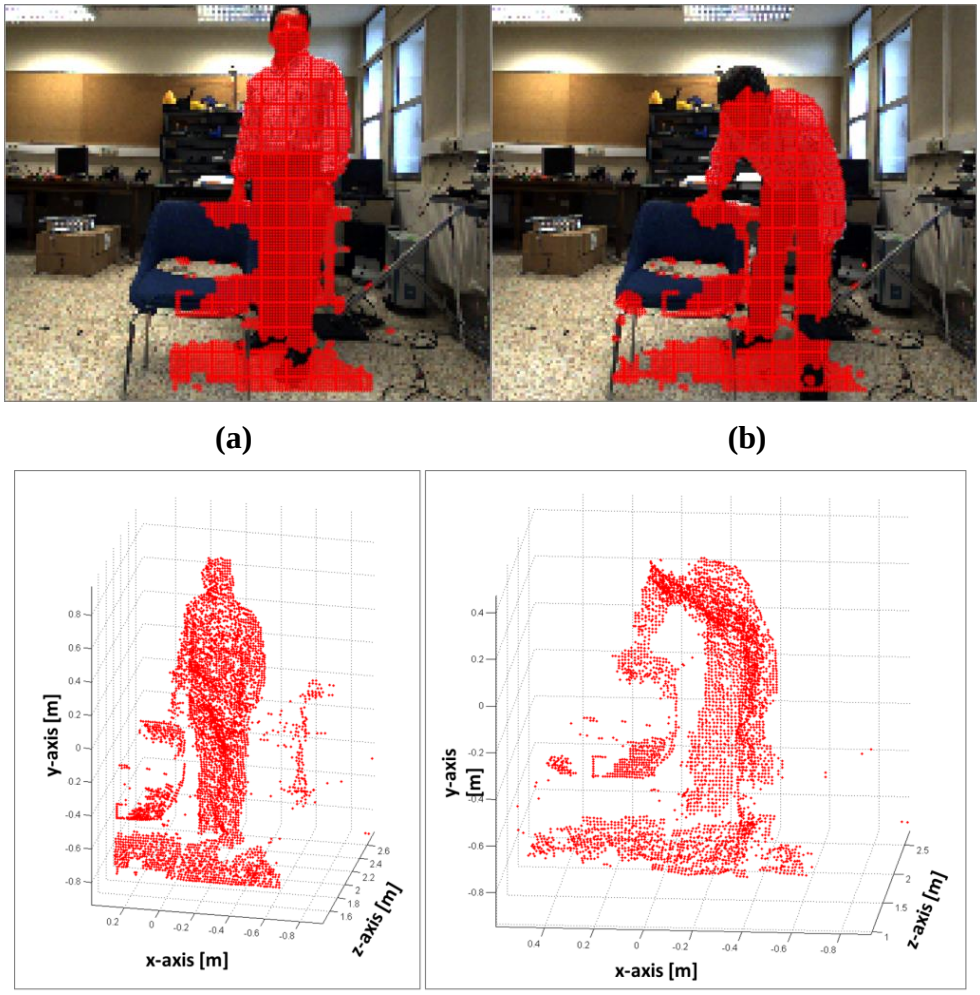



**Figure 5.18** Results of the robust structure from motion algorithm implementation on RGB registered images. (**a**) Inliers of motion detection in $\boldsymbol{sample_t}$. (**b**) Inliers of motion detection in $\boldsymbol{sample_{t+1}}$. (**c**) Depth measurements of the inlier motion region of $\boldsymbol{sample_t}$. (**d**) Depth measurements of the inlier motion region of $\boldsymbol{sample_{t+1}}$.

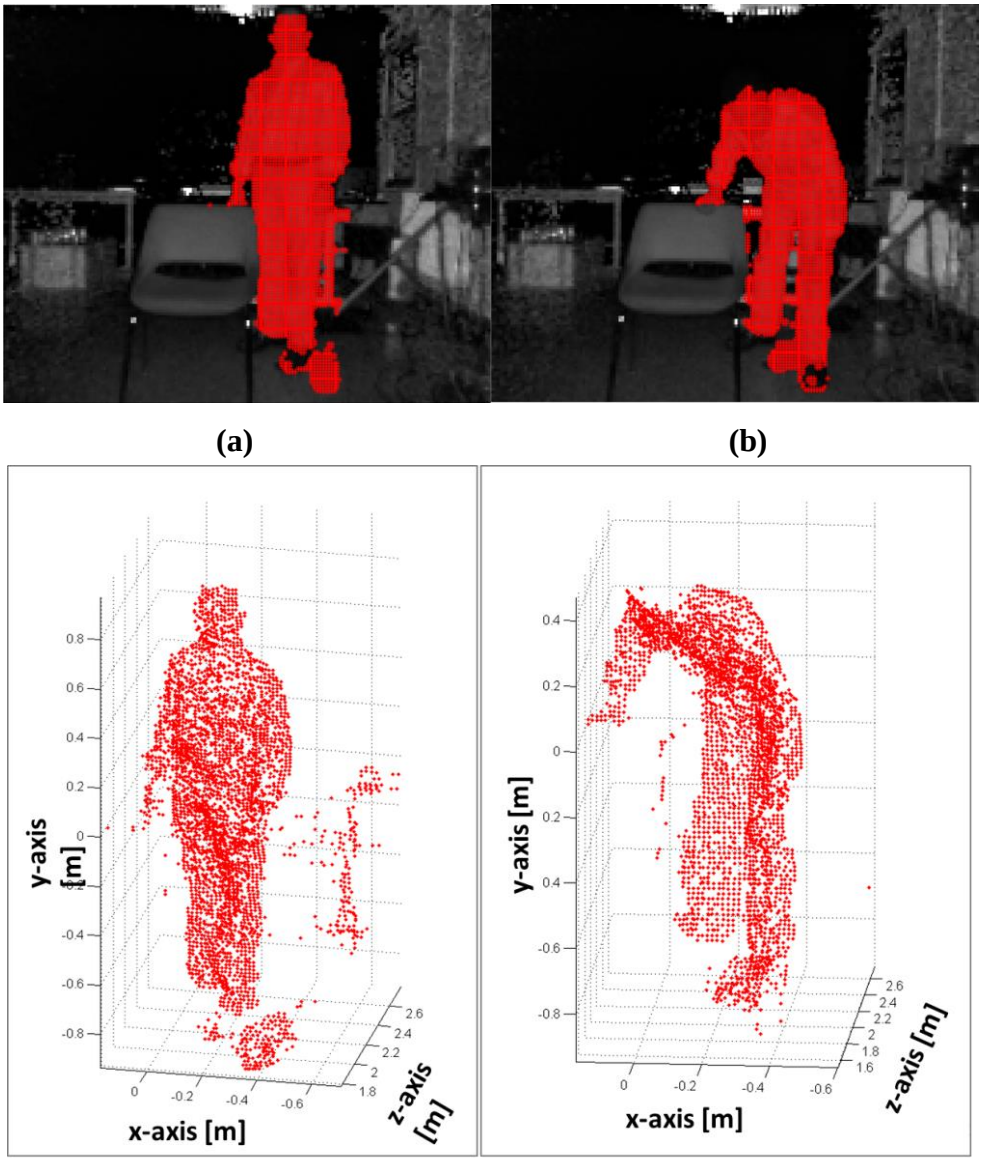



**Figure 5.19** Results of the robust structure from motion algorithm implementation on ToF amplitude images. (**a**) Inliers of motion detection in $\boldsymbol{sample_t}$. (**b**) Inliers of motion detection in $\boldsymbol{sample_{t+1}}$. (**c**) Depth measurements of the inlier motion region of $\boldsymbol{sample_t}$. (**d**) Depth measurements of the inlier motion region of $\boldsymbol{sample_{t+1}}$.

The results in Figure 5.18, of implementing the robust motion detection algorithm in colour information, illustrate the erroneous identification of the man's shadow as inlier. Given that the shadow covers part of the chair and a part of the floor, the intensities variation produced due to the combination of the textured floor and the man's shadow are computed as intensity gradients. This is a common problem when computing motion detection in indoor environments. The results in Figure 5.19 present a solution for reducing this problem, in which amplitude images are used for the motion detection implementation. Consequently, the uncertainties in the features extraction stage are also reduced. For instance, the person's postures shown in Figures 5.19(c-d) are free of outliers

For implementing the proposal procedure described in Figure 5.16, two different image pairs of sequences of person's postures were selected. The visual results are illustrated in Figures 5.20-5.23. Figure 5.20 and 5.22 show the results of the structure from motion algorithm for the first and the second image pair respectively. Figures 5.21 and 5.23, display the results of the registration procedure from the inlier motion regions, for the first and the second image pair respectively. As for the results of the method implementation, 3D structures of the man's body postures are provided, along with high resolution colour information of the obtained 3D structure. Consequently, an ellipsoidal approximation of the body structure could be achieved as well as the fall detection task.

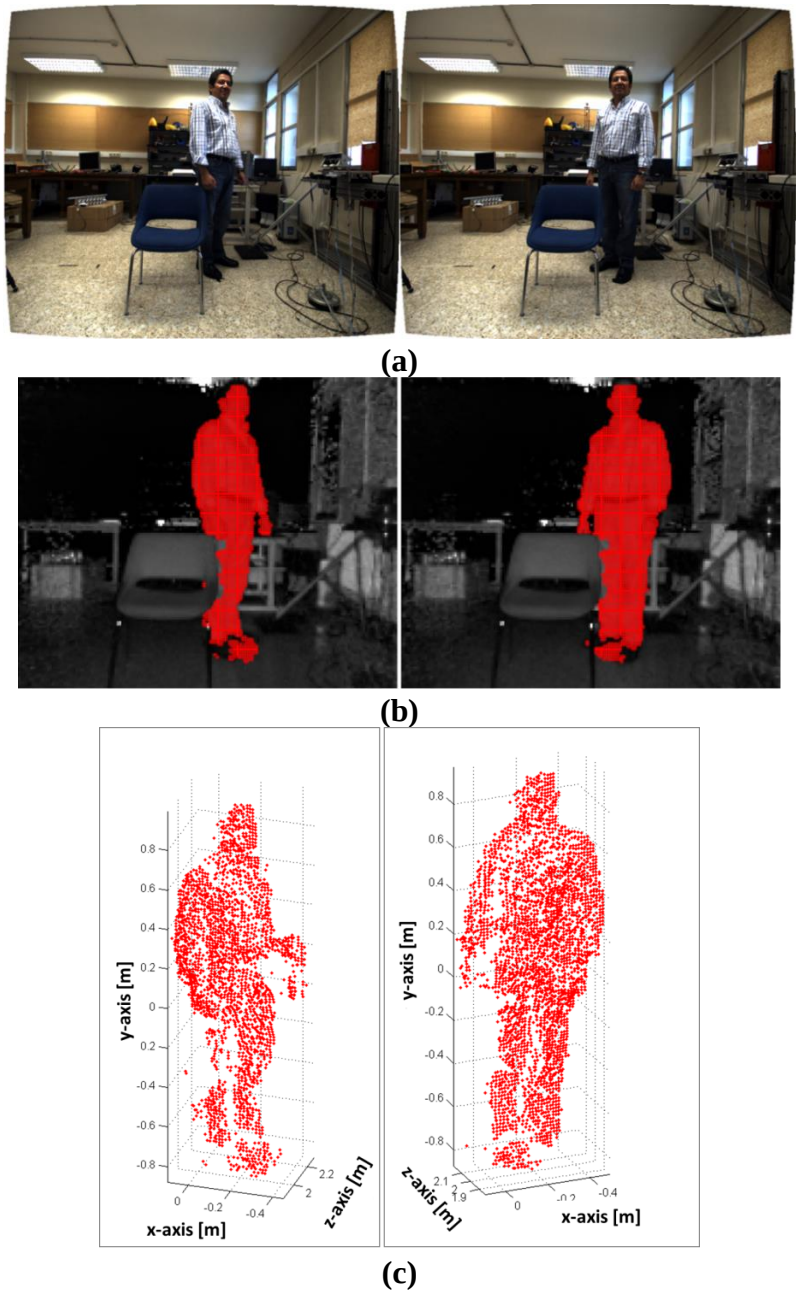



**Figure 5.20** Motion detection results of image pair $(\boldsymbol{sample}_{t=11}, \boldsymbol{sample}_{t=12})$. (a) Image pair of a sequence of a person's postures. (b) Inliers of motion detection in amplitude images. (c) Depth measurements of the inlier motion region.

**(a)**

**(b)**

**Figure 5.21** Image registration results of the motion inliers regions of image pair ($\boldsymbol{sample_{t=11}}, \boldsymbol{sample_{t=12}}$). (**a**) High resolution colour depth map. (**b**) High resolution colour information of the inlier motion region.

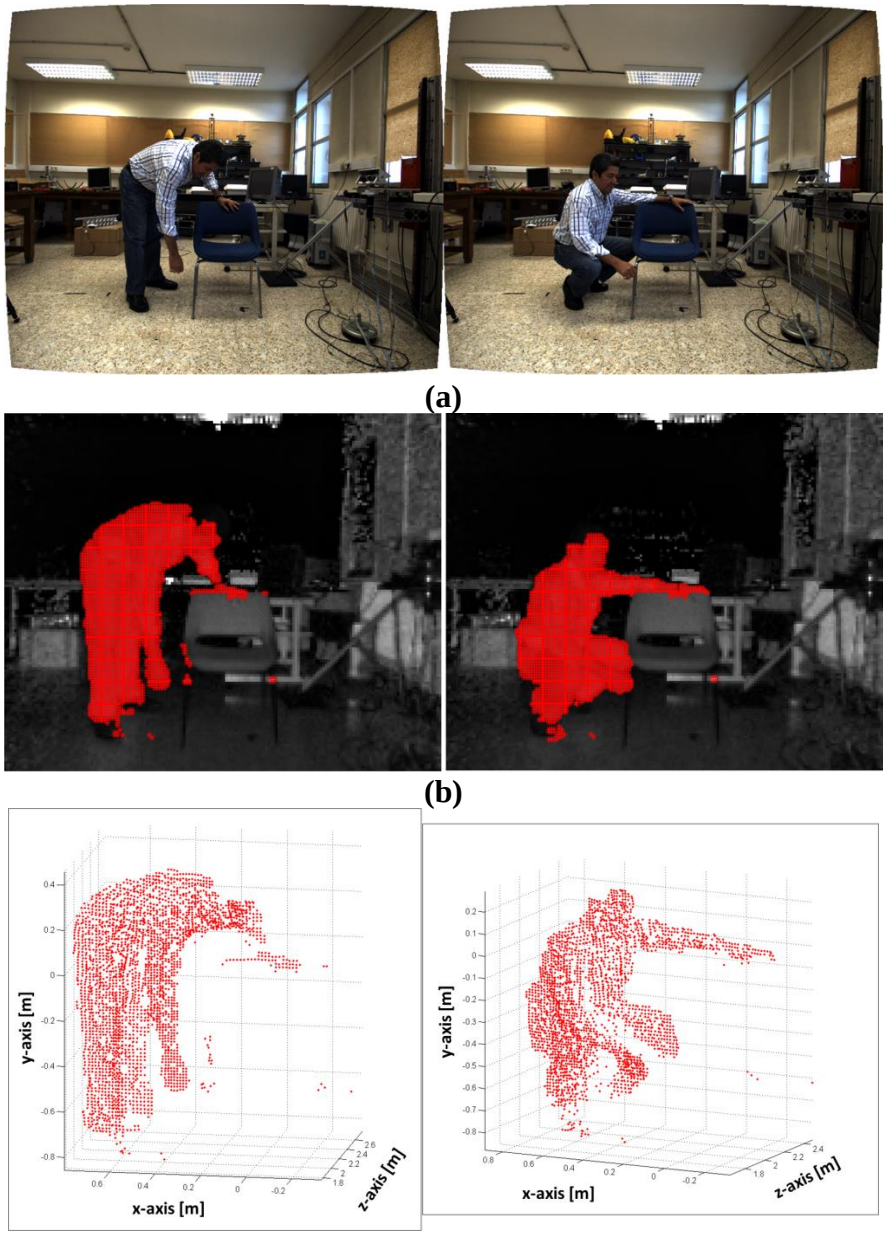
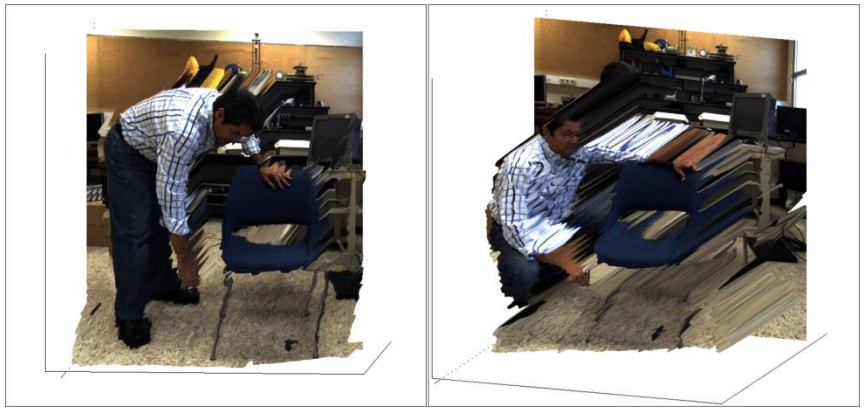



**Figure 5.22** Motion detection results of an image pair ($\boldsymbol{sample_{t=17}}, \boldsymbol{sample_{t=18}}$). (**a**) Image pair of a sequence of a person's postures. (**b**) Inliers of motion detection in amplitude images. (**c**) Depth measurements of the inlier motion region.



**(a)**

**(b)**

**Figure 5.23** Image registration results of the motion inliers regions of image pair ($\boldsymbol{sample_{t=17}}, \boldsymbol{sample_{t=18}}$). (**a**) High resolution colour depth map. (**b**) High resolution colour information of the inlier motion region.

### *5.2.2 Conclusions*

The visual results show a satisfactory performance of the proposed method. In order to evaluate the method, two experimental tests were conducted. The scenes of these experiments comprise non-planar objects, scenes which are clearly explained with several planes, meaning that the image registration procedure by means of the depth-dependent *Hlut* approach should be done with several homographies as well. In order to perform a

quantitative assessment, in the first group of experiments, the white-red pattern board was used as target, placing it at several positions and orientations with respect to the image planes, and at several distances from the sensory system. In the second group, white-red landmarks were attached to some of the objects into the scene. These objects were also placed at several positions, orientation and distances with respect to the sensory rig.

The obtained results do not show any presence of misalignment problems between depth and colour information, likewise the presence of discontinuities on the objects surfaces. Discontinuities could appear within the homographies transition regions. The proposed method has proven to be robust enough for evading this issue. For instance, a cylinder bucket, a chair, or the angled board, which are continuous surface, are properly mapped, without any presence of discontinuity on their surfaces. The shape and edges of objects presented almost no difficulties or artifacts, and so far, no enhancement algorithm has been implemented.

Additionally, the visual results of the depth-dependent *Hlut* method were compared with the results of the standard calibration method. As it was demonstrated in Chapter 4, the proposal approach outperforms the standard calibration method, in terms of accuracy and data alignment when computing raw depth measurements.

In order to evaluate the capability of the proposed method oriented to in-house video surveillance and motion detection task, sequences of people postures were analysed. For that purpose, a motion detection procedure was introduced. The procedure computes a robust structure from motion algorithm on the amplitude images acquired by the ToF camera. Then, a motion mask is used to provide 3D body structures and its corresponding high resolution colour information from the RGB registered data. The proposed procedure reduced the problems of false inliers produced by shadows and the illumination condition. The output of the process provides valuable information for the decision-making stage, since data quadric approximation of the 3D body structure delivers the characteristic of an ellipsoid, and the colour information could be used for a person feature extraction.

Nevertheless, in future investigation more sophisticated motion detection algorithm should be considered. For instance, algorithms that combine depth and colour information for computing motion detection. In such a case, the low confidence regions of the amplitude image could be detected. Normally,

these regions correspond to sensing of dark objects. As it occurs with the man's hair, which is black and consequently is estimated as zero.

## 5.3 Precision Agriculture: Detection and Localization of Fruits for Automatic Harvesting

This experimental section is enclosed in the scope of the Project entitled *Intelligent Sensing and Manipulation for Sustainable Production and Harvesting of High Value Crops, Clever Robots for Crops (CROPS)*, which was funded by the European Union through the Seventh Framework Program, Grant Agreement Number 246252. The Project is framed within the topic *Automation and robotics for sustainable crop and forestry management*. In summary, CROPS project was intended for the development of scientific know-how for a highly configurable, modular and clever carrier platform that includes modular parallel manipulators and "intelligent tools" (sensors, algorithms, sprayers, grippers) that can be easily installed onto the carrier and are capable of adapting to new tasks and conditions. Several technological demonstrators were developed for high value crops like greenhouse vegetables, fruits in orchards, and grapes for premium wines. The CROPS robotic platform should be capable of site-specific spraying (spray applied only towards foliage and selected targets) and selective harvesting of fruit (detects the fruit, determines its ripeness, moves towards the fruit, grasps it and softly detaches it). Another objective of CROPS was to develop techniques for reliable detection and classification of obstacles and other objects to enable successful autonomous navigation and operation in plantations and forests. The agricultural and forestry applications share many research areas, primarily regarding sensing and learning capabilities.

Precision agriculture (PA) is a continuously growing research area, where service robots are becoming an important part for improving competitiveness and sustainable production (Aracil, Balaguer and Armada 2008). PA oriented to the automatic harvesting of fruits requires the investigation of non-destructive sensors capable of collecting precise and unambiguous information for an efficient detection and localization of fruits. This task of detection and localization in natural scenes is quite challenging, since most fruits are partially occluded by leaves, branches or overlapped with other fruits(Plebe and Grasso 2001). These occlusions eliminate the direct correspondence between visible areas of fruits and the fruits themselves by introducing ambiguity in the interpretation of the shape of the occluded fruit (Kelso 2009). In addition, colours of fruits cannot be rigidly defined because the high variability exhibited

among the different cultivars within a same species and the different levels of ripeness. Moreover, fruits can be found in random positions and orientations on trees which can be of various sizes, volumes and limb structures. Environmental conditions such as wind, rain, dust, moisture and lighting also increase the technical challenge imposed to the sensory system (Sarig 1990).

Given the strong dependence of the fruit harvesting robots on sensorial information, and the numerous problems to be solved in this area due to the application requirements, there has been an intensive research effort during the last four decades, aiming to provide automatic detection and localisation of fruits. Most of the related studies reported in the literature are based on the use of computer vision and other image processing techniques. One of the first studies was presented by Schertz and Brown (1968), whose identified from their measurements that the surface of oranges reflected ten times more light than the leaves. In (Parrish and Goksel 1977) the first computer vision system for detecting apples and guiding a harvesting robot was implemented. The proposed system was based on a monochrome camera and a red optical filter to increase the contrast between red apples and green-coloured leaves.

In (Buemi, Massa and Sandini 1985) a vision system based on a single colour camera was proposed for the tomato harvesting Agrobot robotic system. Hue and saturation histograms were employed to perform thresholding to segment the image whereas the 3D information was obtained by stereo-matching of two different images of the same scene. Two approaches based on colour information to solve the fruit recognition problem for a citrus picking robot were presented in (Slaughter and Harrel 1987, Slaughter and Harrel 1989). A system based on a monochrome camera to detect and locate tomatoes in natural settings was also developed in (Whittaker et al. 1987). Each acquired image was processed in order to find circular arcs that could correspond to tomato contours. The automatic detection of apples by using a stereo vision system which provided the 3D-dimensional position of each detected fruit was addressed in (Kassay 1992). A sensory system based on an infrared laser range-finder sensor that provided range and reflectance images, capable of detecting spherical fruits in non-structured environments was designed and implemented in (Jiménez, Ceres and Pons 2000b). Some comprehensive reviews like (Sarig 1990, Jiménez, Ceres and Pons 2000a) cover several aspects of these and other not-mentioned-systems.

More recently, Van Henten et al. (2002) achieved a high detection rate of cucumber fruits by combining the images acquired by two cameras, one equipped with an 850 nm filter and the other with a filter in the 970 nm band.

In (Bulanon et al. 2004) authors used a real time machine vision system based on a CCD colour camera to determine the location of the apples centres and the abscission layer of the peduncles. In a later approach, Bulanon and Kataoka (2010) extended their earlier study by combining the machine vision system based on a CCD colour camera with a laser ranging sensor to determine the distance to the fruit. Tanigaki et al. (2008) designed and manufactured a 3D vision system that has two laser diodes for a cherry-harvesting robot. One of these laser diodes emits a red beam and the other an infrared beam. The 3D shape of the cherries was measured by scanning the laser beams, and the red fruits were distinguished from other objects by the difference in the spectral-reflection characteristics between the red and infrared laser beams. A multispectral analysis was also carried out in (Bulanon, Burks and Alchanatis 2010) to enhance citrus fruit detection in the field. In (Hayashi et al. 2010, Hayashi et al. 2012) authors proposed a machine vision unit that consists of three aligned CCD cameras for guiding a strawberry-harvesting robot. In this case, the two side cameras were used to provide stereo vision to determine the fruit position in the 3D space, while a camera located in the centre was used to detect the peduncle and to calculate its inclination.

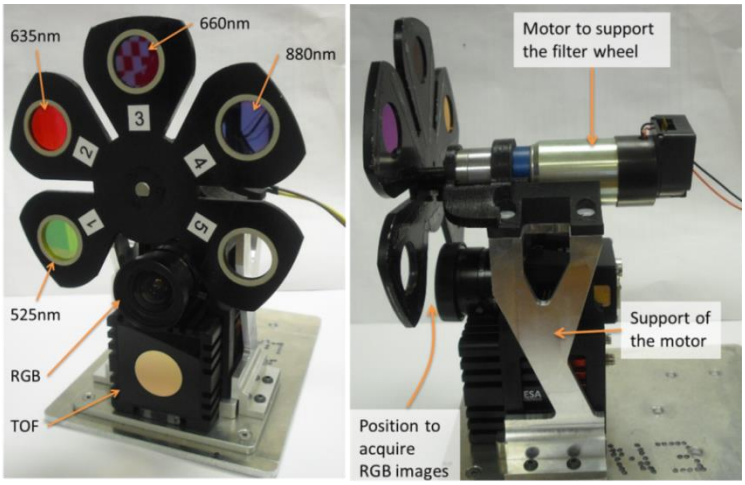
All the studies mentioned above are limited to fruit detection. Nevertheless, for the harvesting task, it would be advantageous to detect and localise other plant elements (e.g., branches, leaves, cables, etc.) that could interference in the free motion of the robotic manipulator. In (Fernández et al. 2013a) Cabernet Sauvignon grapevine elements are discriminated for precision viticulture tasks such as harvesting, whereas in (Bac, Hemming and Henten 2013) the problem of plant parts detection is addressed for the motion planning of a sweet-pepper harvesting robot. Also worthy of mention are the researches carried out by (Berestein et al. 2010, Dey and Mummert 2012). Although the sensory systems proposed in these studies have not been designed for harvesting robots, they addressed the detection and localization of plant elements for other precision agriculture tasks as selective spraying and yield estimation.

The objectives of the research of this experimental section are twofold. The first objective is to evaluate and validate the capabilities of the image registration method proposed in this Thesis, which is the depth-dependent *Hlut* approach for combining RGB high resolution cameras and ToF cameras. The second objective is to assess the feasibility of detecting, discriminating and locating fruits and other plant elements in natural environments by

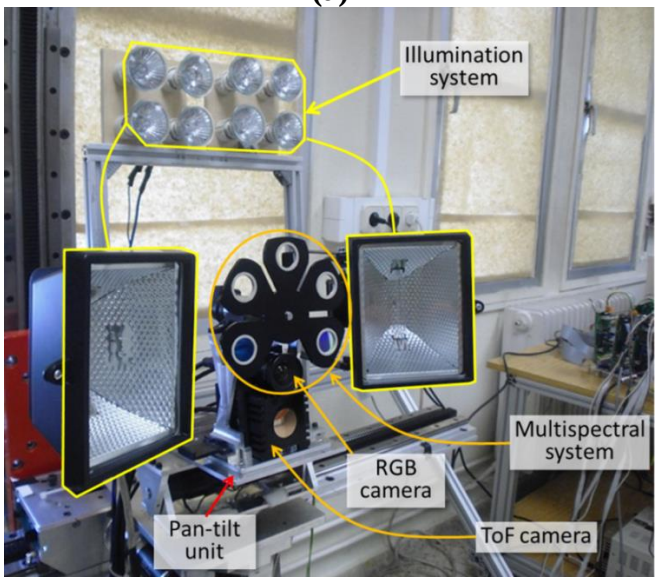
utilising a unique modular and easily adaptable multisensory system and a set of associated pre-processing algorithms. The proposed solution is intended to be used in autonomous harvesting robotic systems, without requiring previous preparation of the crops or previous knowledge of the environment. The proposed multisensory system consists of an AVT Prosilica GC2450C high resolution CCD colour camera (Allied Vision 2011), a multispectral imaging system and a Mesa SwissRanger SR-400011 ToF 3D camera (MESA Imaging 2011) (see Figure 5.24 for a graphical description), similar to the sensory rig introduced in Section 3.1. Though, in this case, the high resolution colour camera is not only utilised for the acquisition of RGB images, but also as part of the multispectral system, in which case it is set in the monochrome mode. The multispectral system is completed with a custom-made filter wheel and a servomotor that is responsible for the accurate positioning of the filter wheel. This positioning can be achieved with a maximum angular velocity of 210 rpm and a position error if ±0.01285°. The filter wheel allows interchanging up to five optical filters, facilitating the adaptation of the system for the detection of different kinds of crops. Since correct illumination could be critical in some scenarios, the system also includes two different light sources, an array of xenon lamps and two halogen spots, located above and at both sides of the sensory system, respectively. This lighting system is connected to a control unit that enables the independent power on and off of the lamps, and the control of their intensities. A visual description of the proposed system is shown in Figure 5.24

The RGB camera and the multispectral imaging system will provide the input data required for the detection and characterization of areas of interest that could belong to fruits, whereas the ToF 3D camera will supply simultaneously fast acquisition of accurate distances and intensity images of targets, enabling the localization of fruits in the coordinate space.

In order to confer versatility to the set-up, the whole proposed multisensory system is installed on a pan-tilt unit that facilitates the data acquisition of different viewpoints. The tilt movement has a limited angular displacement of $\alpha = \pm 30°$ relative to the horizontal axis due to mechanical constraints. The yaw movement has no mechanical constraint, so it could rotate 360° around the vertical axis. However, for the stated application, the automatic yaw movement will be restricted for azimuthal angles within the range given by $0° \leq \beta \leq 180°$.




**(a)**




**(b)**

**Figure 5.24** Close-up views of the multisensory system for fruit harvesting. (**a**) Multisensory rig and filter wheel. (**b**) Complete view of the system.

The control architecture for the proposed multisensory system consists of two main parts, a unit implemented in Robot Operating System (ROS (2007), http://www.ros.org/) responsible for managing the sensing devices and the high level control of the hardware elements, and a second unit implemented

in QNX RTOS (http://www.qnx.com) for the low level control of the hardware elements, which are the motorised filter wheel, the illumination system and the pan-tilt unit (see Figure 5.25). A general description of the software implementation based on experiences from software development within the CROPS project (Crops-project 2010) is presented in (Barth et al. 2014), Thus, the principal functions of the first unit are the initialisation and setting of the CCD and ToF cameras according to the working conditions (acquisition mode, pixel format, exposure and integration time), and the control of the image acquisition procedure. Synchronous acquisition of the CDD and ToF camera is achieved when the sensory system controller publishes a trigger message that is sent when the filter wheel reaches a requested position. Immediately after the frame data acquisition is successfully completed, the sensory system controller node sends a command to the second unit implemented in QNX in order to initiate the motion of the filter wheel to the next target position. This node also sends commands for controlling the lights and the pan-tilt unit when required.

The second unit is in charge of the low level control for the high accurate positioning of the filter wheel (with a position error of ±0.01285° and a maximum time delay of 50 ms for the positioning of each filter), switching on/off and intensity variation of the illumination system, as well as the high accurate positioning of the pan-tilt unit, being the PID controller the selected option for this purpose. First and second unit communicate between them via TCP messages. These messages contain the parameters and commands required for controlling and monitoring the motion and the data acquisition tasks of the multisensory system. The entire process of acquiring and registering a pair of images from the RGB and the ToF camera takes 300 ms running in a x64 bits Intel(R) Xenon(R) CPU @ 2.66 GHz and 6 GB RAM.

Before investigating methodologies and techniques for detecting and locating fruits with high accuracy, it is necessary to count with appropriate pre-processing algorithms that allow taking full advantage of the data acquired with the designed multisensory system. For that purpose, two complementary pre-processing algorithms are proposed: a pixel-based classification algorithm that labels areas of interest that are candidates for belonging to fruits and the depth-dependent *Hlut* registration algorithm proposed in this Thesis, which listed in Algorithm 3 in Section 3.3. This algorithm combines the results of the aforementioned classification algorithm with the data provided by the ToF camera for the 3D reconstruction of the desired regions.

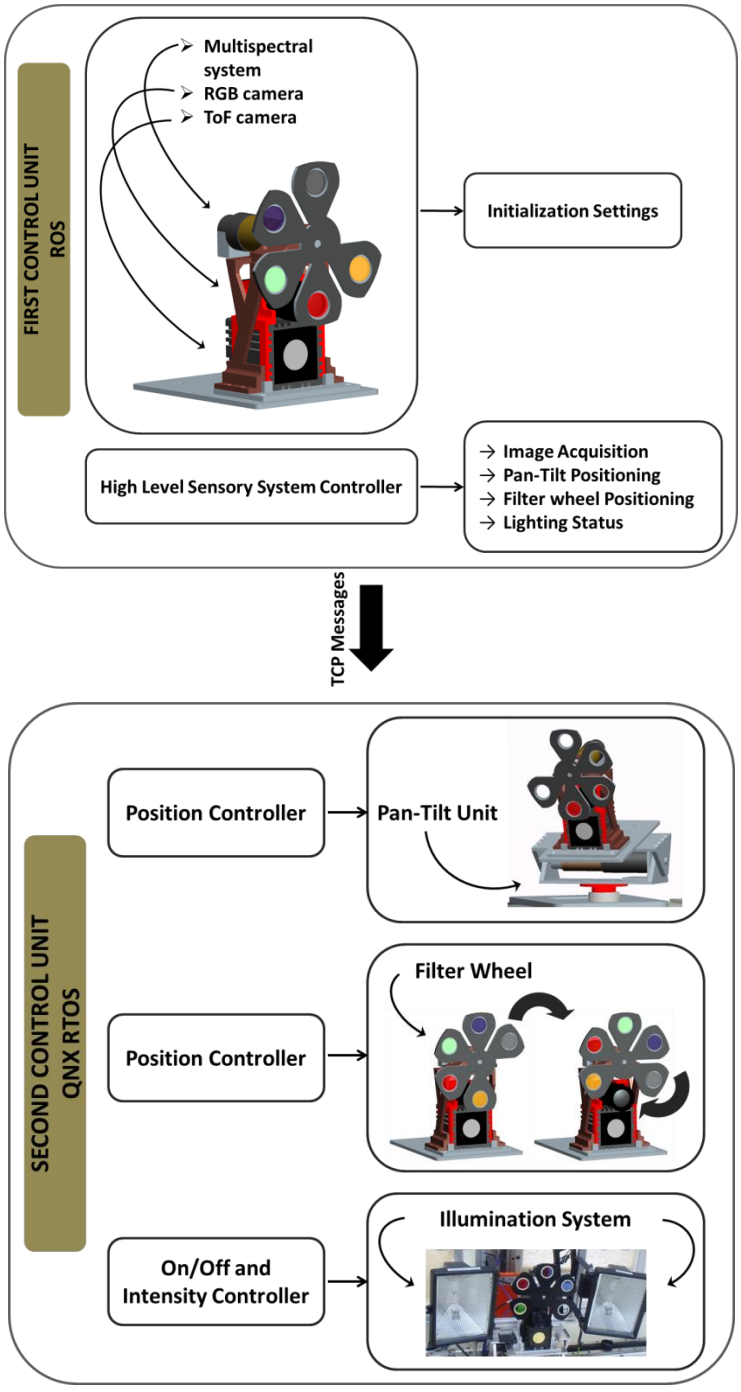



**Figure 5.25** Multisensory system structure.

Several studies have demonstrated that different targets with a similar appearance when they are captured by an RGB camera can exhibit distinctive properties if they are examined with spectral systems capable of acquiring several separated wavelengths (Namin and Petersson 2012). For this reason, the first algorithm deals with the combination of RGB and filtered images acquired with the proposed multisensory system in order to achieve a classification system capable of distinguishing the different elements of the scene (Fernández et al. 2013a). The algorithm, based on Support Vector Machines (SVMs), is capable of labelling each pixel of the image into four classes that are: stems and branches, fruits, leaves, and background. SVM is a supervised learning method utilized for classifying set of samples into two disjoint classes, which are separated by a hyperplane defined on the basis of the information contained in a training set (Mucherino, Papajorgji and Pardalos 2009). In the case at hand, four SVMs are utilized sequentially, each one for detecting a class against the rest. Therefore, after the first SVM is applied, pixels identified as belonging to fruit class are labelled and a mask is generated in such a way that only the remaining pixels are considered for the following SVMs. This step is then repeated for the rest of the classes in the following order: leaves, stems and branches, and finally background. The SVM classifiers are trained by selecting a random subset of samples from the RGB and filtered images and manually labelling the regions of interest from these images into the four semantic classes mentioned above. The algorithm was implemented in C++ with the aid of the Open Source Computer Vision Library (OpenCV) (OpenCV Developers Team 2000, Bradski and Kaehler 2008).

Once regions of interest into the scene have been detected and classified, it is necessary to locate them spatially. While ToF depth measurements are fundamental for localisation purposes, it is still necessary to automatically match this information with the classification map obtained from the previous step in a common reference frame. Thus, the depth-dependent *Hlut* approach is implemented for accomplishing this task. The procedure structure for the pre-processing algorithms is summarized in Figure 5.26.

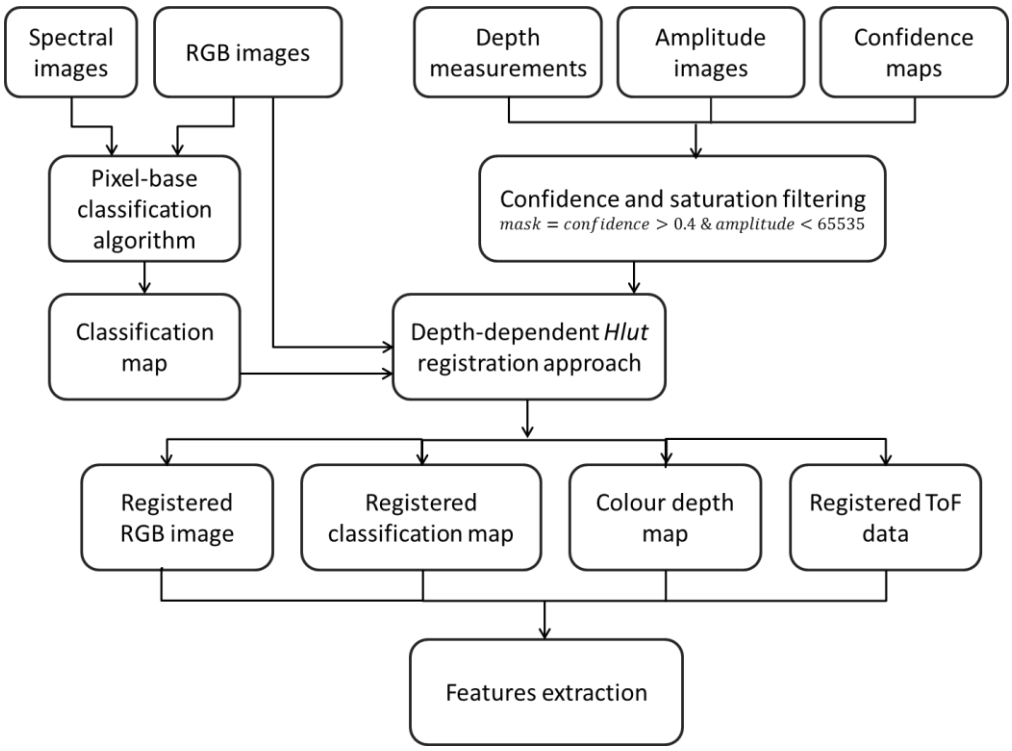
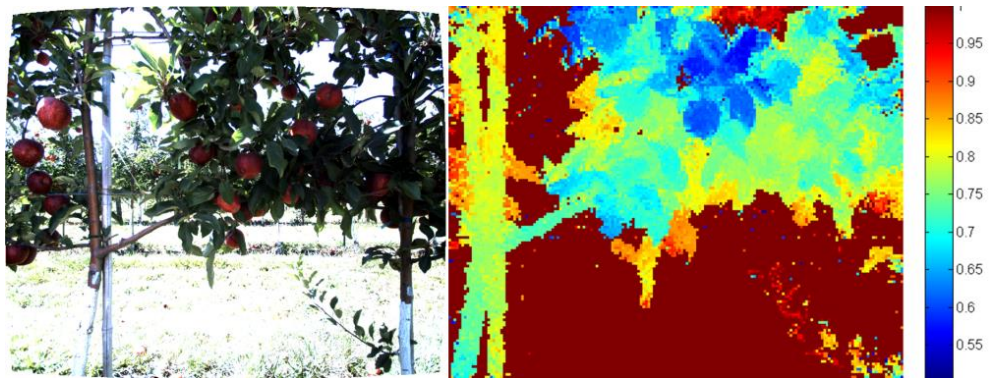


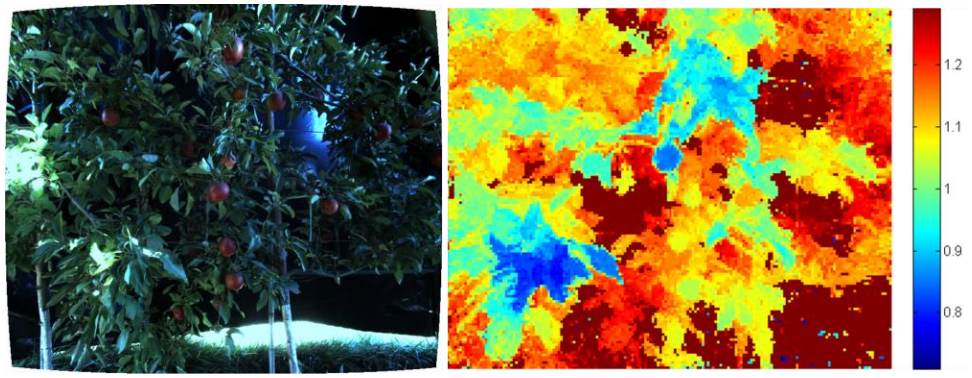

**Figure 5.26** Structure of the multisensory system pre-processing procedure.

### *5.3.1 Results and validation*

In order to evaluate the feasibility of the multisensory system and the associated set of pre-processing algorithms for detecting and locating fruits in natural scenarios, an extensive experimental campaign has been conducted in both laboratory and on the field conditions. Details of these experiments are described below. In Figure 5.27, examples of apple orchards acquired in natural condition are illustrated.




**(a)**




**(b)**

**Figure 5.27** Scenes of apple orchard on the field. (**a**) RGB image and ToF range data. (**b**) RGB image and ToF range data.

### *5.3.1.1 Laboratory conditions*

Due to the complexity of natural scenes illustrated in Figure 5.27, an initial experimental set in laboratory conditions was conducted in order to evaluate the feasibility of combining an RGB and a ToF camera for detecting and locating fruits in apple orchard.

This experimental stage is twofold. The first part is oriented to the validation of the depth-dependent *Hlut* approach for computing depth map registration of apple orchard scenes. The second part is focused on the evaluation of the registered data for extracting spatial features from the classified regions of fruits. In its important to mention that the research related to the attainment of multispectral classification map escapes the scope of this Thesis, and further detailed information about the multispectral

classification algorithm results can be found in (Fernández et al. 2014). For the proposed method validation, three artificial apples were used as targets for computing depth map registration with the depth-dependent *Hlut* approach. These apples were custom-made manufactured by means of a 3D prototype printing machine and they were attached to a panel board. In order to evaluate the capabilities of the proposed registration method, the targets were placed at several positions, orientations and distances with respect to the multisensory system.

In Figure 5.28 some views of the image samples acquisitions are shown, whereas the results of the implementation of the Algorithm 3 (see Section 3.3), for the registration of the images samples displayed on Figures 5.28(a) and 5.28(b) are illustrated in Figures 5.29 and 5.30, respectively.

**(a)** **(b)**

**Figure 5.28** Images of artificial apples acquired in laboratory conditions. (**a**) Occlusions free scene (sample 33). (**b**) Scene with occlusions (sample 50).

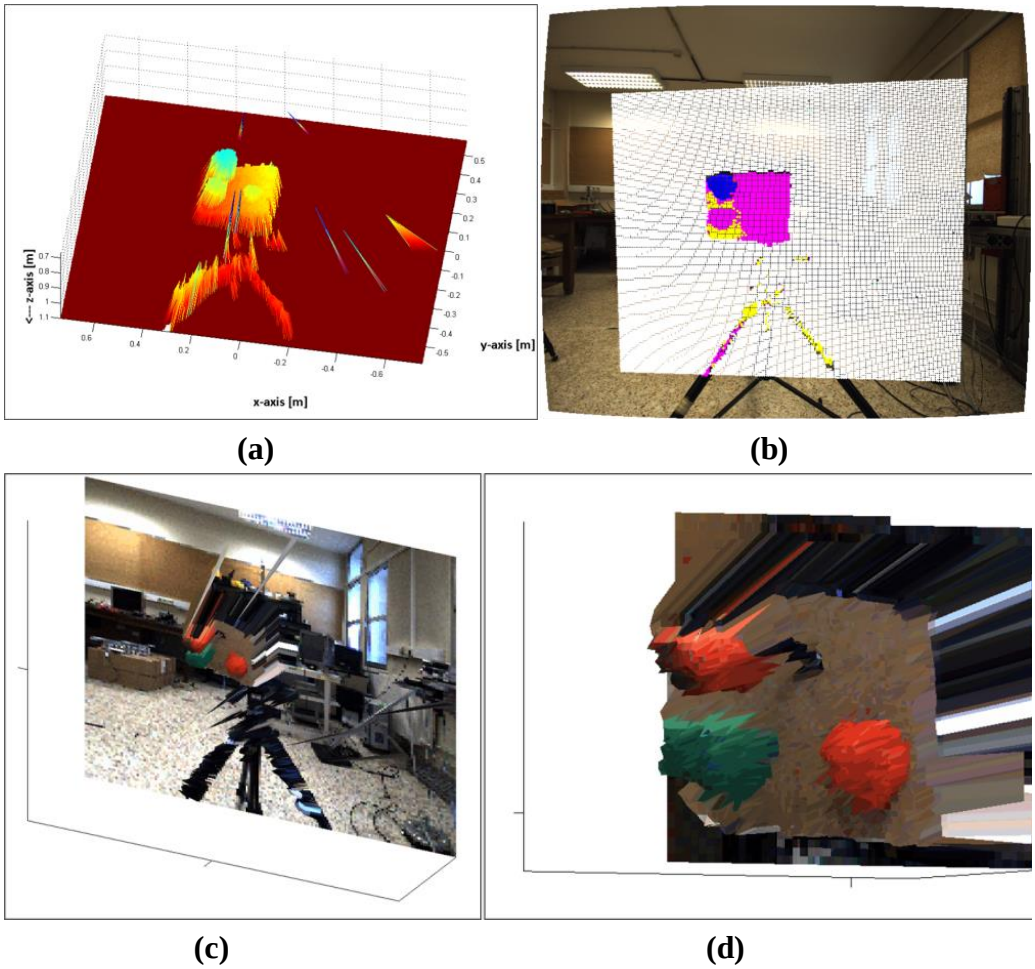



**Figure 5.29** Results of image registration of image sample 33. (**a**) Depth measurements. (**b**) Homography labelled mask, each colour represents a homography of the *Hlut*. (**c**) Low resolution colour depth map. (**d**) Close-up of the high resolution colour depth map.

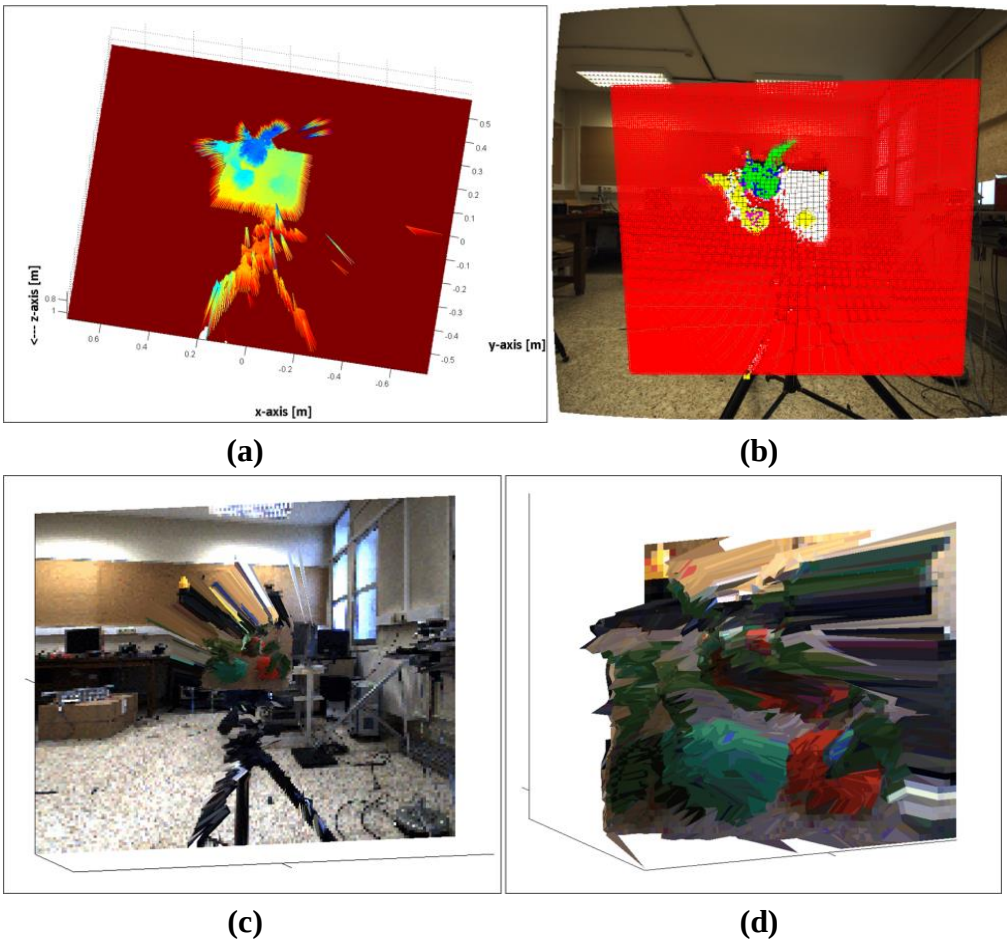
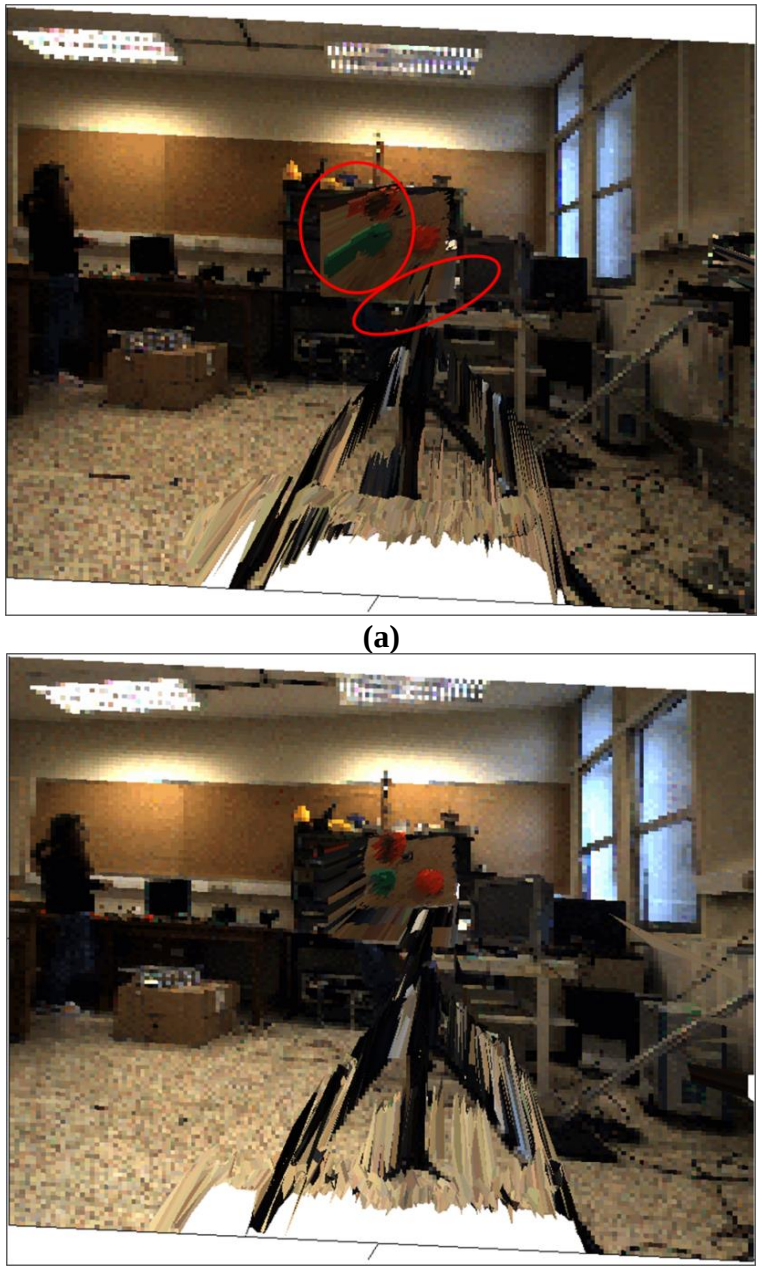



**Figure 5.30** Results of image registration of image sample 50. (**a**) Depth measurements. (**b**) Homography labelled mask, each colour represents a homography of the *Hlut*. (**c**) Low resolution colour depth map. (**d**) Close-up of the high resolution colour depth map.

In addition, Figure 5.31 shows the visual results of the image registration procedure of a scene by implementing the depth-dependent *Hlut* approach, along with the obtained results by means of the standard calibration method. The comparison of the results of the two methods, in terms of data alignment, shows the proposal of this work outperforms the standard calibration method. More accurate matched edges are provided when applying the proposed approach. The red markers in Figure 5.31(a) illustrate this issue.



**(a)**

**(b)**

**Figure 5.31** Results of image registration of image sample 52. (**a**) Standard calibration method. (**b**) Depth-dependent *Hlut* approach.

Since the attainments of the proposed depth map registration approach has been demonstrated and the step of fruits classification has been previously achieved, the spatial features extraction from classified regions of fruits is investigated. Figure 5.32 summarizes the implemented algorithm for the features extraction procedure.

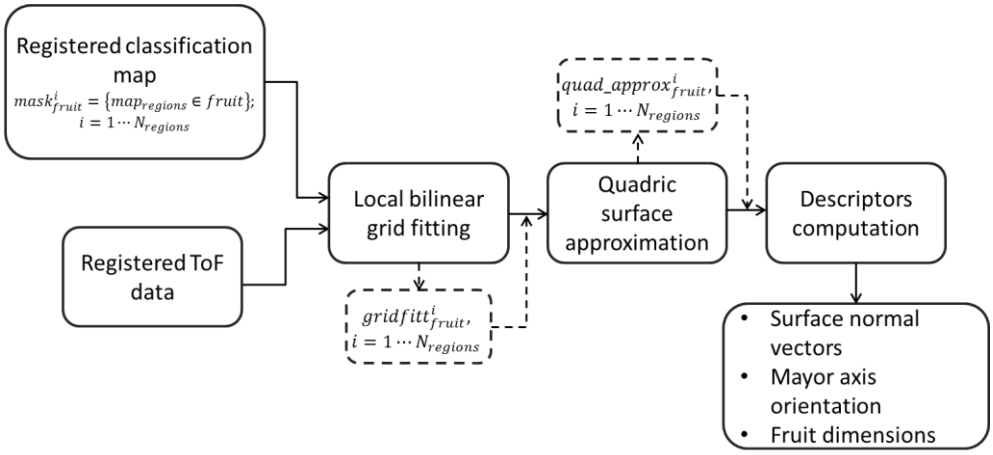



**Figure 5.32** Features extraction procedure.

The proposal is based on a bilinear interpolation for the grid fitting of the depth measurements registered as fruits, $mask^i_{fruit}; i = 1 \cdots N_{regions}$, where $mask^i_{fruit} = \{map \ni fruit\}$, combined with a quadric surface approximation $\left(quad\_approx^i_{fruit}\right)$ of the resulting depth estimations from the grid fitting process $\left(gridfitt^i_{fruit}\right)$, where the inputs of the process corresponds to the registered classification map and the depth values. Once the approximation is successfully accomplished, the extraction of relevant descriptors is possible, such as: fruits main axis orientation with respect to the ToF image plane, fruits position with respect to the ToF camera coordinates, fruits dimensions and others. In Figures 5.33 and 5.34, the visual results of the proposed procedure for extracting object descriptors of the three fruits on the image samples 33 and 50 are shown, respectively. This first set of results was computed from the low resolution colour depth maps, while on the second set of results, the proposed procedure was implemented on high resolution colour depth maps. In Figure 5.35, the result of this implementation on image sample 33 is shown.

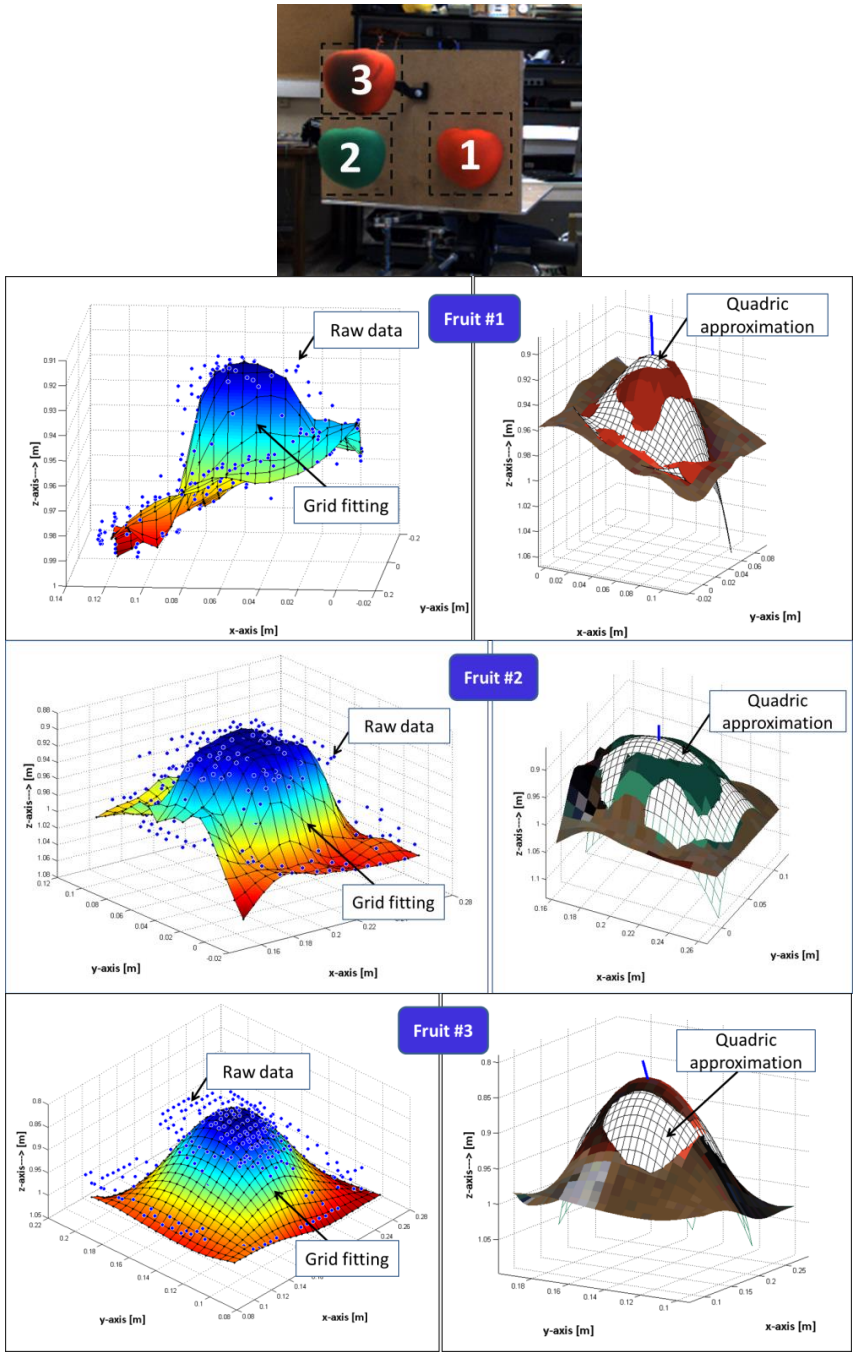



**Figure 5.33** Low resolution results on the feature extraction procedure of image sample 33 (see Figures 5.28(a) and 5.29).

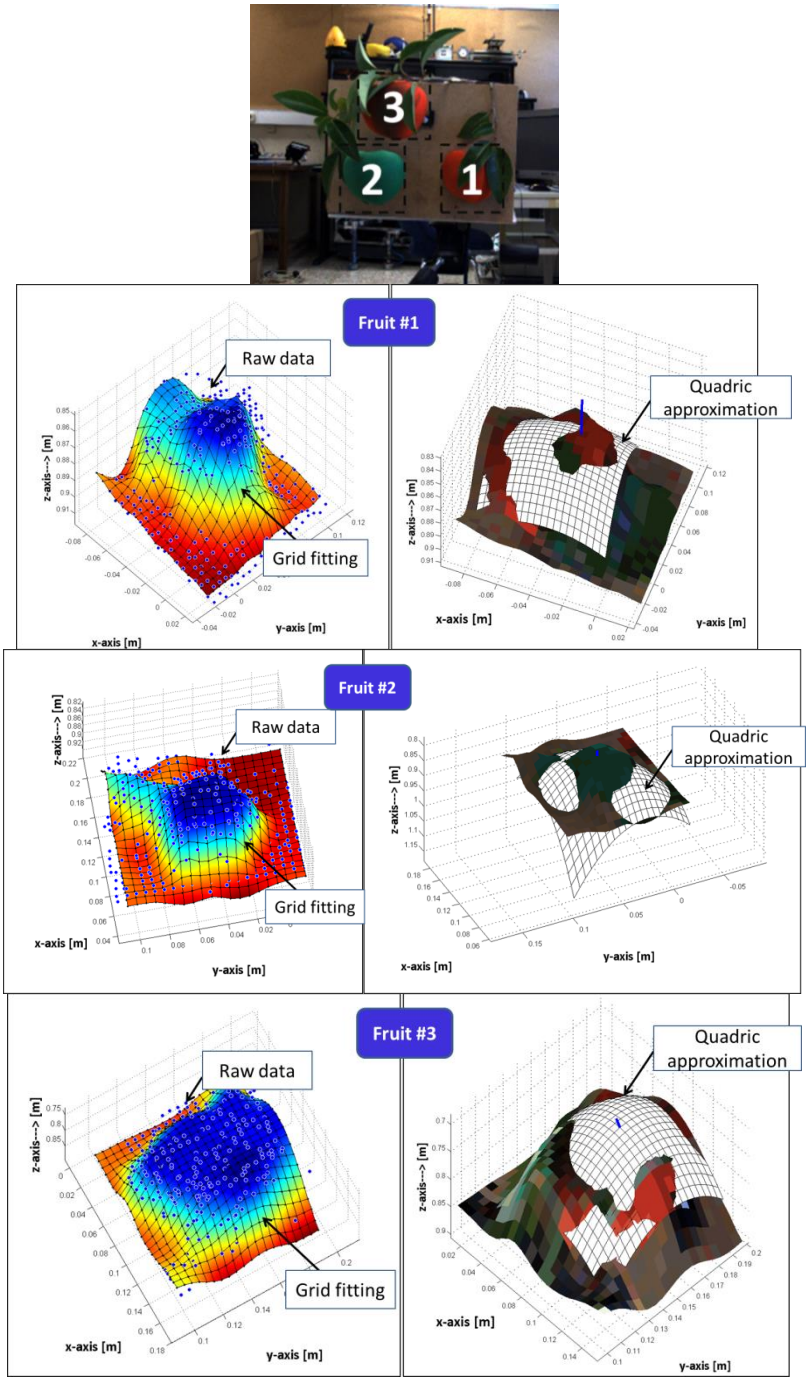



**Figure 5.34** Low resolution results of the feature extraction procedure on image sample 50 (see Figures 5.28(b) and 5.30).

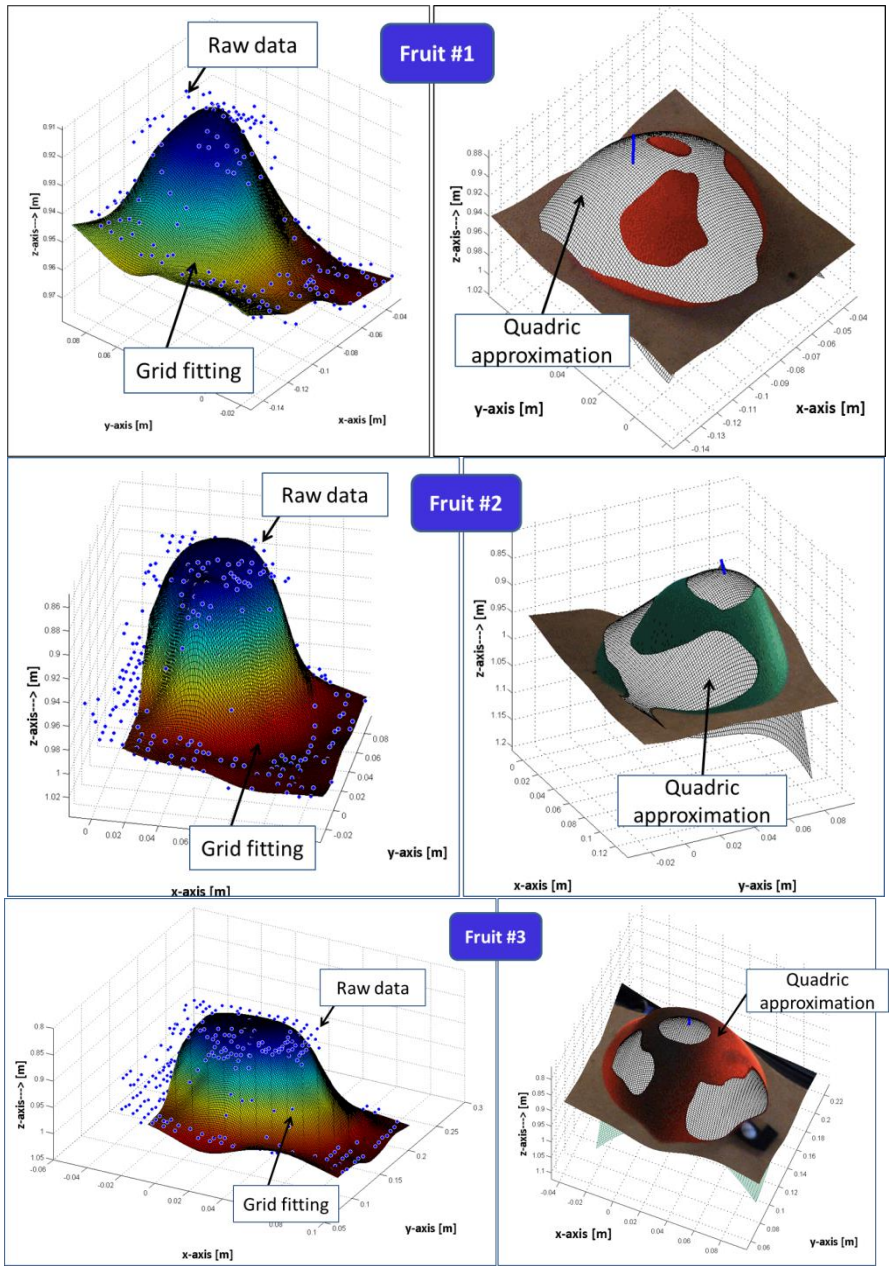



**Figure 5.35** Results of the feature extraction procedure of image sample 33, implemented on high resolution colour depth information.

The evaluation of the features extraction is addressed by considering the perspective view of the fruits with respect to the images plane. Then, the fruits dimensions were only computed on image samples on which the fruits are completely visible. Regarding the fruits main axis orientation with respect to the ToF image plane, a comparison between the orientation of the fruits and the orientation of the panel board on each image sample was conducted. Since the apples are fixed to the panel board, it is expected that all pieces of the set have the same orientation. Consequently, the discrepancy between these orientations should be small. The obtained Mean errors are detailed in Table 5.2.

**Table 5.2** Results of the Feature Extraction Error in Terms of Mean Error

| **Fruit #** | **Orientation with respect to the panel board [degrees]** | **Fruit Dimensions [mm]** | | |
|---|---|---|---|---|
| | | ***Width*** | ***Height*** | ***Depth*** |
| **1** | 3.5 | 9 | 8 | 13 |
| **2** | 5.6 | 6 | 9 | 1.1 |
| **3** | 4.8 | 7 | 8 | 9 |

A very relevant issue for satisfactory features retrieval from the apples scenes is the avoidance of pre-filtering algorithms. Usually, filtering algorithm over-smooth the depth values, in this in this case, could cause loss of crucial information of the objects features. In the particular considered scenario, only tens of points on the ToF image coordinates are able of sensing the apples. Therefore, any loss of this information should be avoided. Conventionally, the proposed registration approach does not strictly require a pre-filtering algorithm for achieving accurate depth map registration. Consequently, complete raw data of the sensed apples is available for post-processing algorithm. Figure 5.36 illustrates the use of this low resolution raw information in combination with the high resolution colour information for improving the registered regions with fruits. The procedure is based on a local 3D depth values grid fitting algorithm, and a quadric surface approximation, which is constrained with the colour information mask that corresponds to apples.

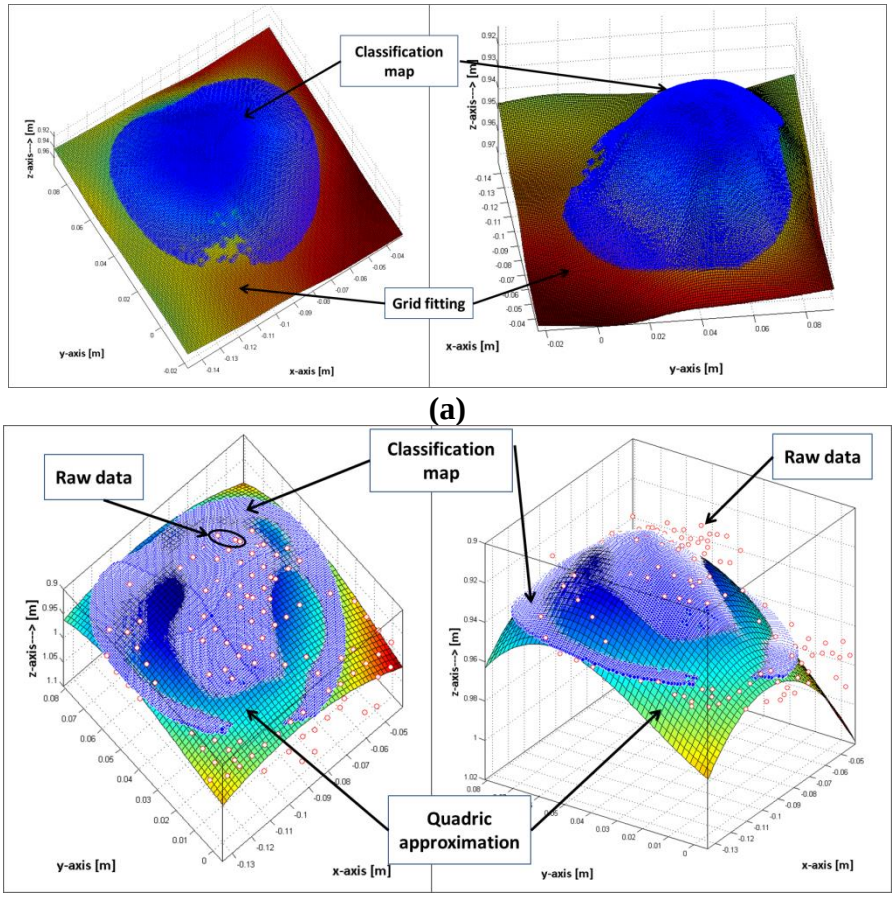


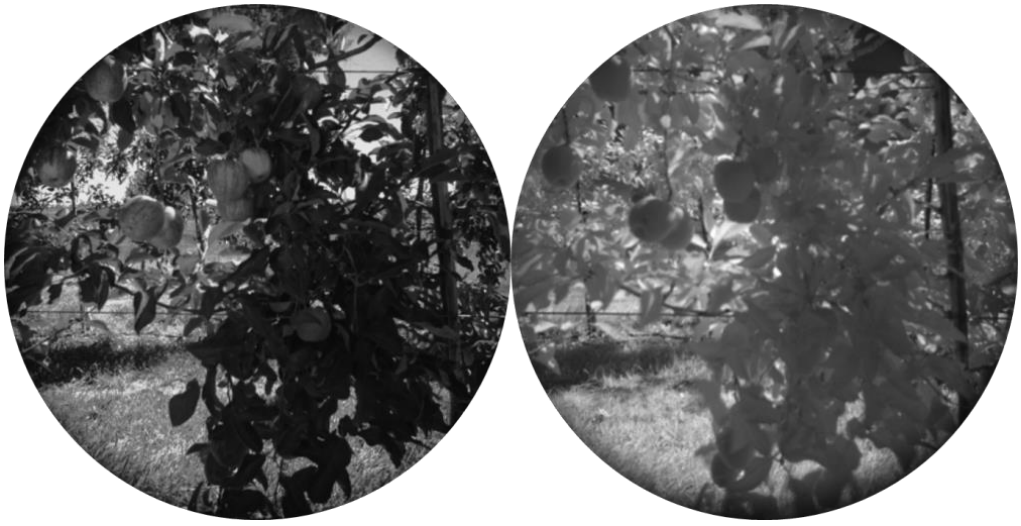

**Figure 5.36** Details of the registration data improvement in the high resolution context. (**a**) Guided local data grid fitting $\left(\boldsymbol{gridfitt}^{\boldsymbol{i}}_{\boldsymbol{fruit}}\right)$. (**b**) Guided quadric surface approximation $\left(\boldsymbol{quad\_approx}^{\boldsymbol{i}}_{\boldsymbol{fruit}}\right)$.

### *5.3.1.2 Field conditions*

The extensive experimentation carried out in field conditions was conducted in an apple orchard located in Chillan, Chile, and it involved the data acquisition process by means of the proposed multisensory system and a ground truth data collection of the acquired scenes. This experimental set is composed with two phases. The first phase of the experimental campaign was devoted to the acquisition of training data for the design of the pixel-based classification algorithm (Fernández et al. 2014). In this case the acquired

dataset included RGB and monochrome images with band-pass filters that have centre wavelengths of 635 nm and 880 nm (Fernández et al. 2013a). Since the aforementioned algorithm deals with the classification of each image pixel, each testing set consists of 5,018,400 samples (2448 × 2050 pixels on the image). In order to train the SVMs of the proposed classification algorithm, four acquired datasets were randomly selected. From these RGB and filtered images, representative regions of interest of different sizes were selected for each desired class. Then, the mean reflectance values of these regions were treated as training samples and were manually labelled in four semantic classes: fruits (apples), stems, leaves and background. With the obtained set of 40 samples per class, the SVMs of the proposed pre-processing algorithm were trained to classify the pixels of the images. The sampling approach for training data could be then considered as a stratified random sampling method, since the population is divided into smaller groups known as strata, which are formed based on members' shared features (Waske and Benediktsson 2007). Random samples from each stratum are taken, and these subsets are then combined to form the random training sample.

For the second phase of the experimental campaign, aimed at evaluating the proposed system, the acquired dataset included not only RGB and monochrome filtered images, but also range data. Outputs provided by the proposed system consist of a pixel-based classification map and the ToF depth measurements registered data. While the first phase concerning the attainment of the multispectral classification map escapes the range of this research, the second phase related to the registration of the classification map with the depth information, is part of the scope of the investigations of this Thesis. Figures 5.37 and 5.38 show the RGB and the filtered images acquired with the multisensory system, as well as the resulting classification map for an apple crop scene, while the corresponding raw depth measurements and the amplitude image acquired by the ToF camera of the scene is shown in Figure 5.39.

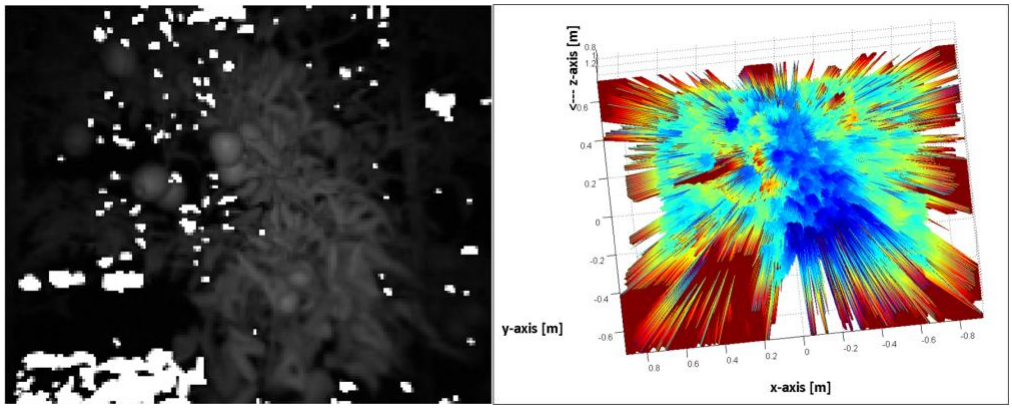


**(a)** **(b)**

**Figure 5.37** Spectral images of the apple orchard. (**a**) 635 nm image.(**b**) 880 nm image.

**(a)** **(b)**

**Figure 5.38** Apple orchard images. (**a**) RGB image. (**b**) Classification map.




**(a)** **(b)**

**Figure 5.39** ToF data of the apple orchard. (**a**) Amplitude image. (**b**) Depth measurements.

For validation purposes, a total of 12 scenes from the apple crop were processed and evaluated. Ground truth data was carefully collected and produced for each scene in order to carry out a quantitative assessment of the proposed solution. This process involved as first step the manual labelling of some fruits of the scenes acquired and processed during the experimental campaign, as well as the manual measurement of the distance from the frontal plane of the ToF camera to the centre of the visible outer surface of each labelled fruit. Horizontal and vertical distances from a defined reference frame to the centre of the visible outer surface of each labelled fruit were also measured manually. For instance, Figure 5.40 shows an example of the labelling of one of the scenes acquired in the apple orchard. Note that these images have been acquired with an external camera, different from the RGB camera included in the multisensory rig, only for illustration purposes, and consequently, as can be observed, the point of view is different if they are compared with Figures 5.38(a).

Table 5.3 summarises the ground truth measurements collected for this scene, where X and Y correspond to the horizontal and vertical distances measured from the origin of the reference frame defined on the image to the centre of the visible outer surface of each labelled fruit, and Z represents the orthogonal distance measured from the frontal plane of the ToF camera to the centre of the visible outer surface of each labelled fruit. The reference frame defined on each image for the ground truth data collection is the centre of the fruit labelled as 1. Thus a transformation of these measurements is required in

order to compare them to the data provided by the ToF camera. This transformation only affects to the X and Y coordinates, since z coordinate is always referenced to the ToF camera.

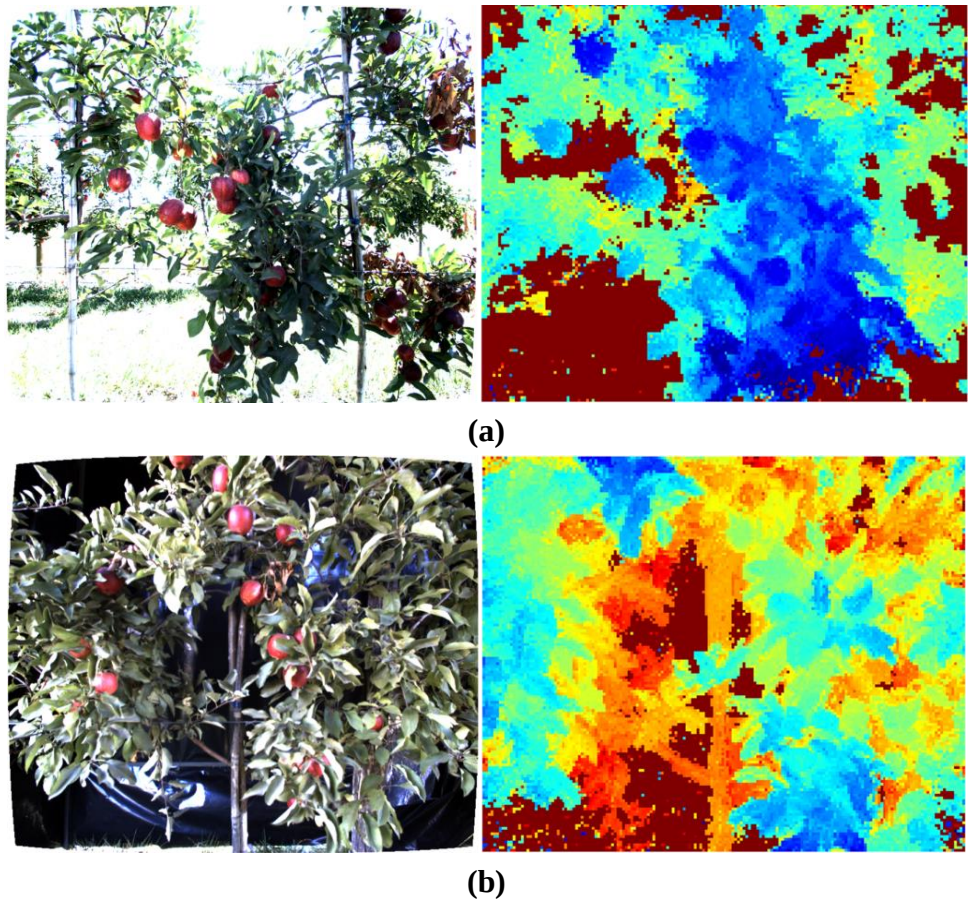

**Figure 5.40** Ground truth data acquisition of an apple orchard.

**Table 5.3** Ground truth measurements of the scene presented in Figure 5.40 (scene 5 of Day 1)**.**

| **Reference Frame – Centre of the Fruit 1** | | | |
|---|---|---|---|
| **Fruit** | **X[mm]** | **Y[mm]** | **Z[mm]** |
| **1** | 0 | 0 | 794,4 |
| **2** | -180 | 95 | 862 |
| **3** | 160 | 60 | 758 |
| **4** | 170 | 30 | 819 |
| **5** | -50 | 230 | 754 |
| | | Mean distance | 797.48 |

After evaluating the data registered from the ToF camera with the collected ground truth, it was obtained that the position error ranges from 0 to 4.5 cm in the x-axis, from 0 to 6.1 cm in the y-axis and from 1 to 7.6 cm in the z-axis, with a mean error of 0.8 cm in the x-axis, 1.5 cm in the y-axis and 2.3 cm in the z-axis.

The obtained visual results are also promising, since in spite of the complexity of the scenes, the depth-dependent *Hlut* method is capable of dealing with noise in depth measurements and the results show a satisfactory alignment between colour and depth measurements. In Figures 5.42 and 5.43, the depth map registration results of two natural scenes from apple orchard are shown, while in Figures 5.44 and 5.45, the results of the features extraction process of some of the fruits are illustrated. For the computing the features extraction image processing, the procedure described in Figure 5.32 was implemented. The two input scenes comprises: the scene 5 (Day 1) and the scene 12 (Day 3), which are illustrated in Figure 5.41



**(a)**

**(b)**

**Figure 5.41** Scenes of apple orchard on the field. (**a**) RGB image and ToF range data of scene 5 (Day 1). (**b**) RGB image and ToF range data of scene 12 (Day 3).

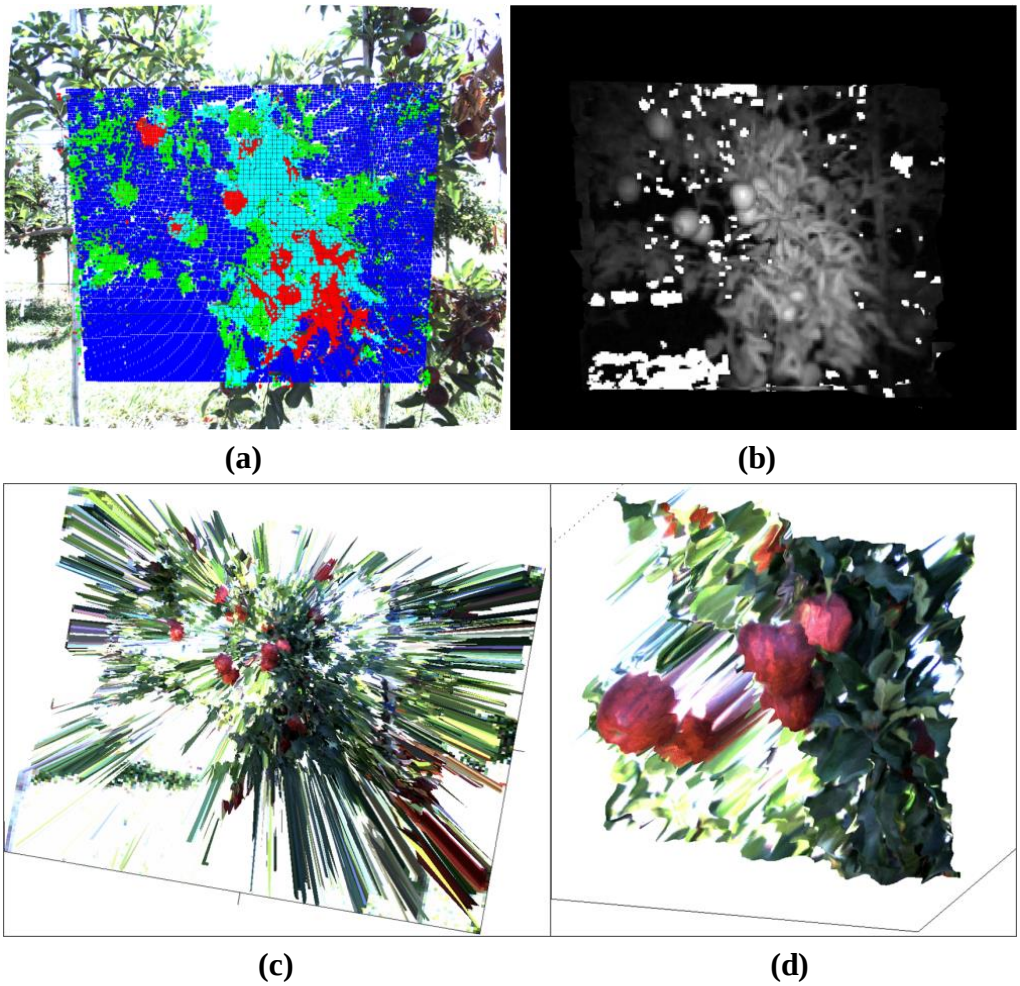



**Figure 5.42** Results of image registration of scene5 (Day 1). (**a**) Depth measurements. (**b**) Homography labelled mask, where each colour represents a homography of the *Hlut*. (**c**) Low resolution colour depth map. (**d**) Close-up of the high resolution colour depth map.

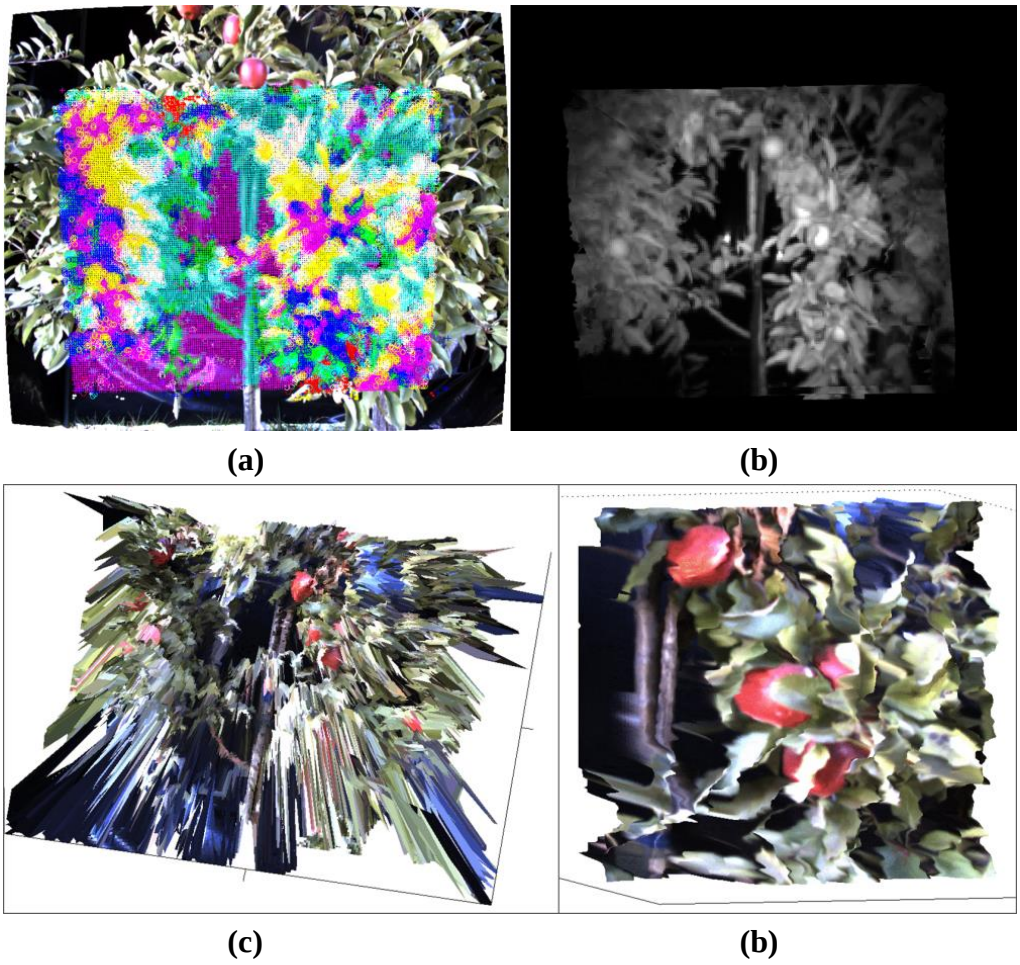



**Figure 5.43** Results of image registration of scene12 (Day 2). (**a**) Depth measurements. (**b**) Homography labelled mask, where each colour represents a homography of the *Hlut*. (**c**) Low resolution colour depth map. (**d**) Close-up of the high resolution colour depth map.

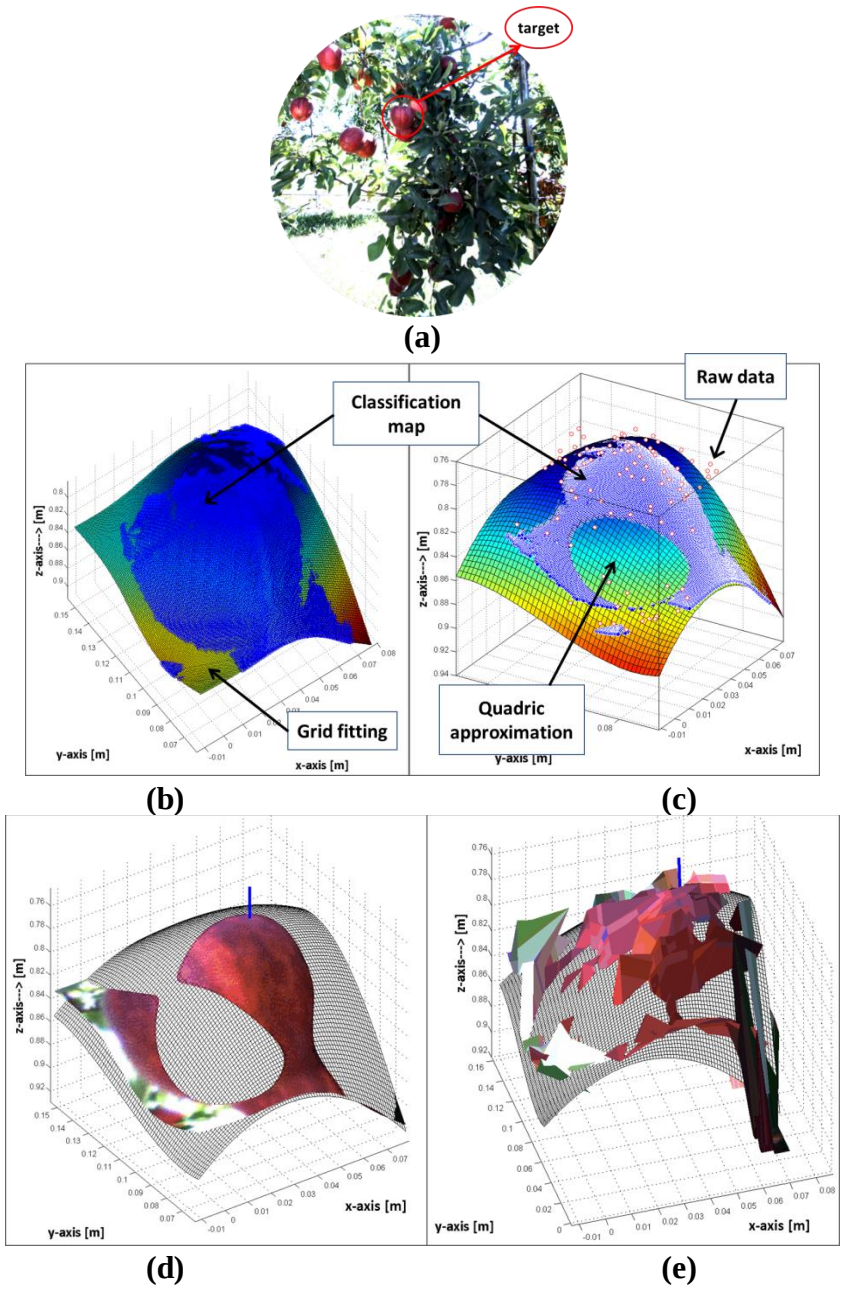



**Figure 5.44** Results of the feature extraction procedure of a fruit in scene 5 (Day 1). (**a**) Fruit position in the RGB image coordinates. (b) Guided local data grid fitting. (**c**) Guided quadric surface approximation. (**c**) Depth and colour model of the fruit. (**e**) Raw depth and colour model of the fruit.

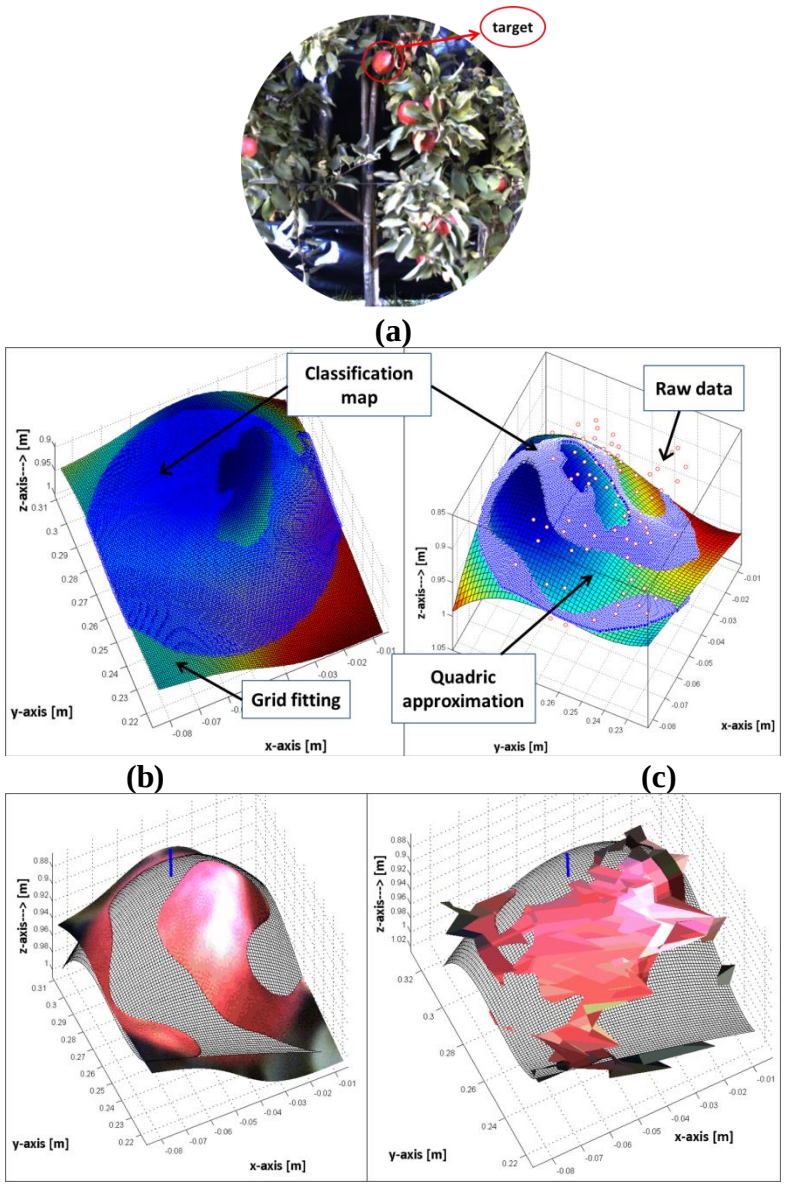



**Figure 5.45** Results of the feature extraction procedure of a fruit in scene 12 (Day 3). (**a**) Fruit position in the RGB image coordinates. (b) Guided local data grid fitting. (**c**) Guided quadric surface approximation. (**c**) Depth and colour model of the fruit. (**e**) Raw depth and colour model of the fruit.

In natural scenes it is quite easy to find a great number of elements that can affect the response of the ToF camera, which is characterised by suffering from flying pixels, noise and incorrect depth measurements due to the scene geometry and material properties. For instance, the modulated light used by the ToF camera is frequently reflected by multiple surfaces inside the scene before reaching the camera sensor. Border of fruits and leaves produces commonly this kind of multi-path interferences, affecting the range data measurements and consequently the fruits properties. Plants elements can also be moved by the wind during the acquisition process, producing erroneous measurements. It has to be considered also that the registration algorithm is dealing with a correspondence between images of 144 × 176 pixels from the ToF camera and images of 2050 × 2480 pixels from the classification maps. Moreover, manual measurement of distances for ground truth data is not exempt from errors, which could explain the appearance of some isolated maximum errors, far from the mean values. Therefore, the mean position errors obtained during the experimental test are quite acceptable bearing in mind the high complexity of the studied scenes and the large difference in the resolution of the ToF images and the classification maps.

Regarding the problems caused by the occlusions of leaves and branches, and the overlapped fruits, the preliminary results of the proposed method, which is based on the depth-dependent *Hlu*t registration approach and the guided fruits approximation procedure, shown the capabilities of the proposal to deal with these issues. Nevertheless, more robust methods for surfaces approximation should be evaluated. For instance, in Figures 5.46, the results of the feature extraction of an occluded apple are presented, while in Figure 5.47, the problem of overlapped fruits is illustrated.

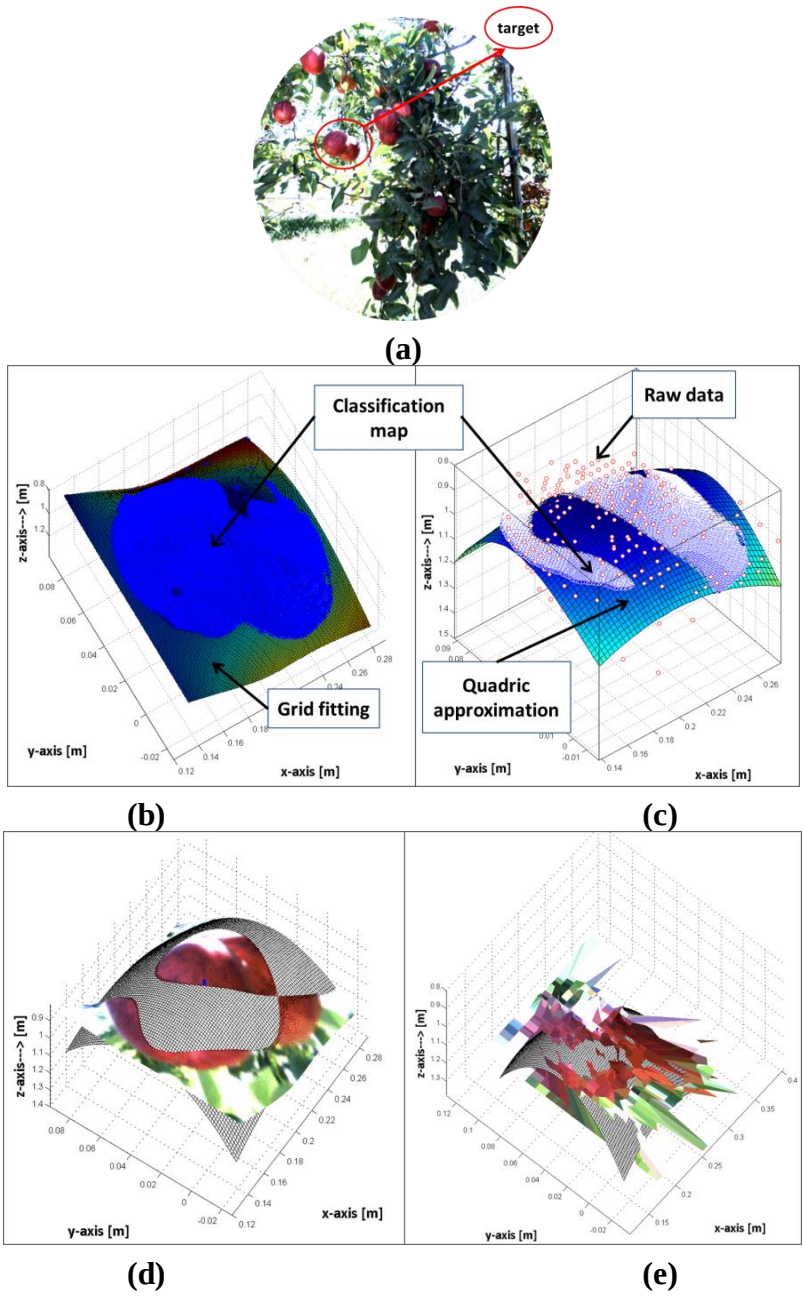



**Figure 5.46** Results of the feature extraction procedure of a fruit in scene 12 (Day 3). (**a**) Fruit position in the RGB image coordinates. (b) Guided local data grid fitting. (**c**) Guided quadric surface approximation. (**c**) Depth and colour model of the fruit. (**e**) Raw depth and colour model of the fruit.

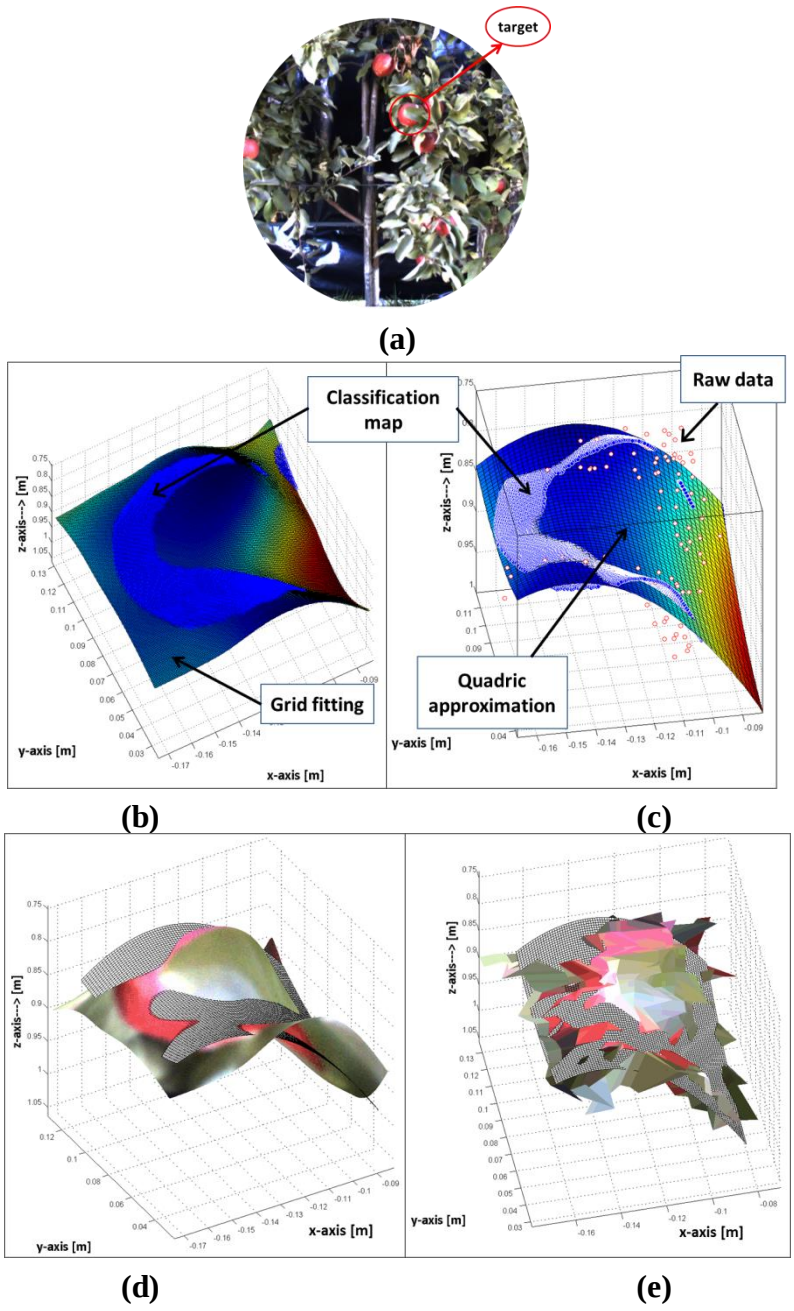



**Figure 5.47** Results of the feature extraction procedure of a fruit in scene 12 (Day 3). (**a**) Fruit position in the RGB image coordinates. (b) Guided local

data grid fitting. (**c**) Guided quadric surface approximation. (**c**) Depth and colour model of the fruit. (**e**) Raw depth and colour model of the fruit.

### *5.3.2 Conclusions*

This experimental section proposes a modular and easily adaptable multisensory system and a set of associated pre-processing algorithms for the detection and localisation of fruits. The solution includes a colour camera and a multispectral system for acquiring reflectance measurements in the visible and NIR regions that are used for finding areas of interest that belong to the fruits, and a ToF camera that provides fast acquisition of distances enabling the localisation of the targets in the coordinate space.

The pre-processing algorithms designed for the proposed multisensory system include a classification algorithm based on SVMs that identifies pixels that belong to fruits and a registration algorithm that combines the results of the aforementioned classification algorithm with the data provided by the ToF camera in order to obtain a direct correspondence among their pixels, so range data can be associated to pixels labelled as fruit. An extensive experimental campaign was carried out in order to assess the proposed solution, including the acquisition of not only test data but also training and ground truth data. The experimental analysis of image registration in laboratory conditions indicates that the proposal is accurate enough for computing feature extraction of fruits, taking into account the complexity of the scene. For instance, the small dimensions of the objects compared with the extant of the image view, as well as the spherical shape of the objects are additional difficulties when using ToF cameras.

On the other hand, in spite of the challenging scenarios found in natural environments, the proposed solution exhibited a satisfactory performance. Multisensory system provides all the data required for detecting and locating fruits, showing a great versatility in dealing with different crops. The pre-processing algorithm based on SVM classifiers affords an accurate enough discrimination of apple tree elements, without any pre-treatment of the images, and without any preparation of the crops. Finally, registration algorithm allows the spatial localisation of the regions of interest classified as fruits with enough accuracy.

Finally, it is important to remark that the proposed method is characterized by not requiring strictly a pre-filtering of the depth data for providing accurate enough registered data, in terms of alignment between the depth and colour information, which is a significant advantage. Filtering algorithms

usually over smooth the data, which could be very problematic due to the dimension and the shape of the fruits. However, in the proposed post-processing procedure the data fitting and filtering is addressed from a local perspective. Consequently, the shape of the apple could be recovered from the combination of the little depth information and the large colour information, while avoiding the previous loss of depth values caused by the filtering process.

# Chapter 6

## Conclusions, Contributions and Future Research Directions

In the research of this Thesis, a non-feature-based methods for automatic image registration has been presented, which is relied on depth dependant planar projective transformation. For this purpose, a framework for automatic image registration of low resolution ToF and high resolution RGB images was designed and implemented. The implemented framework is capable of registering images with non-common features and dealing with moderate noisy depth measurements. The method is based on a depth-dependent homography lookup table (*Hlut)*. By this means, the 3D world is parametrized in *n*-planes which correspond to the entries of the *Hlut*. Hence, points transformation between views is reliant on the distance from the objects to the sensory system, this being a non-feature-based method. Since the method relies on planar projective transformation, the computational load is very low, making it suitable for near real-time applications.

The conclusions for each of the two objectives proposed in this Thesis are presented in detail below.

## 6.1 Design, Implementation and Validation of the Proposed Image Registration Method

The results of the depth-dependent *Hlut* approach validation, which were presented in Chapter 3, show that the proposed solution exhibits a satisfactory performance in terms of both visual quality and RMSE. The method normally maps points with an error of less than 4 pixels, measured on the RGB frame, which is a small error considering the RGB camera resolution (2448 $\times$ 2050 pixels). These errors represent slight distortions of the mapped points at working distances within 400–2300 mm. The procedure is capable of computing a low resolution colour depth map together with a labelled homography mask $mask_{LRGB}$ on the RGB image coordinates. The values of the $mask_{LRGB}$ correspond to the homographies $\{H_k^{lut}\}$ used for transferring the data. Since there is a large difference between the cameras resolution, within adjacent estimated points on the RGB pixel coordinates, several coloured points remain unmapped. This labelled mask $mask_{LRGB}$ is intended to be used for matching the unmapped points on the RGB image frame. This work presents an initial approach for this procedure, where a nearest neighbourhood algorithm was applied to create a entire mask of $\{H_k^{lut}\}$ on the RGB pixel coordinates. Then, the high resolution colour depth map was straightforward computed by mapping points from the RGB to the ToF with the homographies $\left\{H_k^{lut^{-1}}\right\}$.

The next step in the proposed method evaluation comprises an in-depth comparison between the standard calibration method and the depth-dependent *Hlut* approach, presented in Chapter 4. The standard calibration method is the most commonly implemented method for computing depth map registration in the literature, and for this reason it was selected for the comparative benchmarking. On the other hand, according to the state-of-the-art studied in Chapter 2, the depth estimation acquired with the ToF cameras are affected by systematic (related to the camera's internal configuration, hardware, etc.) and non-systematic (external factor, such as lighting conditions, motion blurring, etc.) errors, being some of them reduced by the camera calibration process and others by data filtering. Since the two registration methods are reliant on the accuracy of the depth measurements, the methods evaluation and comparison was addressed from the perspective of their response to the noise in the depth estimations.

First, the calibration parameters were computed as much accurate as possible. The obtained extrinsic and intrinsic parameters were compared with similar researches of the field, and they have proven to surpass the accuracy of these researches results. Next, the investigation for the methods evaluation was distributed in three situations of data processing: noise-free data input (ideal), raw data input and filtered data input. The obtained results indicate that the method proposed in this Thesis outperforms the accuracy results of the standard calibration method. For instance, when processing raw data, the proposed approach reduced the error in 41%, with an obtained $\mathrm{RMSE} = 0.2440$, in comparison with the error of the standard calibration method, with an $\mathrm{RMSE} = 0.4150$. Similarly for the filtered data, the obtained error with the standard calibration method is increased in 127% when using bilateral filtering and in 216% when using non-local filter, compared with corresponding errors of the proposed approach. The obtained errors with the proposed approach are $\mathrm{RMSE} = 0.2376$ and $\mathrm{RMSE} = 0.2365$, respectively for each filtering technique, and regarding the results of the standard calibration method, the obtained errors are $\mathrm{RMSE} = 0.5402$ and $\mathrm{RMSE} = 0.7478$, respectively.

In conclusion, these results point out the capability and flexibility of the proposed method for dealing with slight variation in the depth estimation and for processing non-extremely smoothed filtered data. Since the depth-dependent *Hlut* method is reliant on a range of depth measurements instead of the exact value of each measure, moderate variations in the depth estimation could be avoided for the data registration.

## 6.2 Experimental Testing and Validation of the Proposed Method in Indoors and Outdoors Robotic Applications

For the experimentation and the validation of the proposed registration approach in combination with image fusion (pixel-based) algorithms, which are presented in Chapter 5, two relevant robotic applications were considered for this purpose. In following subsections, the specific conclusions of these applications are presented.

### *6.2.1 Experimental stage in indoors robotic application*

The first experimental test was oriented to indoors applications, such as the in-house video surveillance and the people falling monitoring. In which the implementation of the proposed approach in people motion detection

tasks was investigated. Besides that, this experimental stage was also comprises the validation of the proposed method accuracy, and its capability for an adequate image registration of large surfaces by means of several homographies $\{H_k^{lut}\}$, whereas the presence of discontinuities within the homographies transitions are avoided. For that purpose, this experimental stage was composed of two series of experiments, whose main results are summarizes below.

#### *6.2.1.1 First series of experiments (indoors): validation of the accuracy and satisfactory image registration of large surfaces*

First, a planar surface was used as target (the pattern board), and several images of it were acquired at different positions, orientation and distances with respect the sensory system. Then, this process was repeated but using various volumetric and non-uniformed objects such as a chair, a cylinder bucket and a person. The numerical and the visual results exhibit a satisfactory performance of the proposal, with an obtained normalized RMSE of 0.1272, a mean value given by $Mean_{(u,v)-axis} = [-0.7, -0.3]$ and a standard deviation $\sigma_{(u,v)-axis} = [3.5,4.6]$, for the planar objects, and a $RMSE = 0.3511$, a mean value $Mean_{(u,v)-axis} = [-0.46, 5.5]$ and a standard deviation $\sigma_{(u,v)-axis} = [10.5,11.3]$ for the volumetric and non-uniform objects. In addition, the visual results demonstrate the capability of the proposed registration method for avoiding discontinuities on the mapped surfaces, preserving the edges and shape of objects and, providing accurate enough alignment of depth and colour information. This contrasts with the results obtained with the standard calibration method, where the object's surfaces are less homogeneous and the object's edges exhibit several misalignment problems.

#### *6.2.1.2 Second series of experiments (indoors): evaluation of the method capability for motion detection*

For the purpose of evaluation of the proposed method capability for the people motion detection task, a motion detection procedure was introduced. The procedure computes a robust structure from motion algorithm on the amplitude images acquired by the ToF camera. Then, the resulting motion mask in combination with the depth measurements and the registered RGB image, are used to provide the 3D structure of the person's body and its

corresponding high resolution colour information. The proposed procedure reduced the problems of false inliers produced by shadows and the varying illumination conditions. The output of the process provides valuable information for the decision-making stage, since data quadric surface approximation of the 3D body structure, might delivers the characteristics of an ellipsoid, often used in people's falling detection investigations, and the high resolution colour information could be used for a person feature extraction. On the other hand, the results demonstrate that the proposed method is capable of computing high resolution colour dense map, while the loss of colour information is avoided.

### *6.2.2 Experimental stage in outdoors robotic application*

The second application considered for testing and validating the proposed approach was framed within the European Project entitled *Intelligent Sensing and Manipulation for Sustainable Production and Harvesting of High Value Crops, Clever Robots for Crops (CROPS)*. The general objective of this experimental stage is to assess the feasibility of detecting and locating fruits (apples) and other plant elements in natural environments by utilising a unique modular and easily adaptable multisensory system in combination with the proposed depth-dependent *Hlut* approach. For that purpose, two experiments were conducted. An initial experimental setup in laboratory conditions was adopted, where scenes of artificial apples were analysed. The obtained visual results show a satisfactory performance of the image registration procedure, in spite of the complexity of the scenes, because of the small size of the fruits with respect of the images view and the rounded shape of the fruits. These issues become more relevant when the targets are angled with respect to the image plane. In all cases evaluated, the presence of misalignment problems is almost imperceptible and the shape and edges of objects are preserved. Additionally, a feature extraction procedure was proposed and implemented. The results illustrate the capability of the proposal for detecting and locating fruit.

A second experimental step was carried out in natural conditions, where an extensive campaign for collecting data from an apple orchard was conducted. The complexity of these scenes is increased due to the varying illumination conditions, the random position of the apples on the trees, the natural elements of the plants and the dynamic nature of the environment, such as the presence of wind. Despite these difficulties, the visual results are also promising, since high resolution colour depth information was achieved

with enough accuracy, and the proposed feature extraction was also successfully implemented.

An important advantage of the proposed approach presented in this Thesis, was illustrated in Chapter 5 within the results of the experiments concerning to the apples detection and localisation. The proposed method does not strictly require a pre-filtering process, which usually over-smooths the depth information in an arbitrary way, and nevertheless, this method is capable of providing accurate enough registered data in terms of alignment between the depth and colour information. Hence, small objects or features of the objects are not removed, and in the post-processing procedure the data fitting and filtering might be addressed from a local perspective, by means of the combination of the little depth information and the large colour information.

## 6.3 Main Contributions

The main contributions of this Thesis are highlighted as follows:

- The design, implementation and evaluation of a novel approach for the automatic registration of images acquired with a ToF and RGB camera has been introduced. The detailed procedure for the automatic computation of the depth-dependent *Hlut* approach has been listed in the pseudocode gathered in Algorithm 1. This contribution has allowed the dissemination in a journal paper (Salinas et al. 2015).
- A comprehensive comparative comparison between the proposed method and the standard calibration method has been conducted, where three relevant input data scenarios were considered. The visual and numerical results have been analysed, in which the proposed method outperforms the accuracy results of the standard calibration method and, it also has demonstrated to be capable of dealing with slight variation in the depth measurements. An initial comparison process was also included in the journal paper (Salinas et al. 2015).
- An in-depth accuracy and response validation of the method has been carried out. The proposal has shown a satisfactory performance in the large surface registration, providing uniform registered colour depth regions without the presence of discontinuities.
- The method has been implemented in motion detection tasks. These results exhibit the capability of the proposal for people falling detections tasks. The procedure provides 3D information and the contextual information of the body structure, which corresponds to the

motion inliers. The preliminary study of this algorithm was reported in the journal paper (Salinas et al. 2012).

- The problems of shadows and variation in the illumination conditions, which usually produce the false inlier in the motion detection algorithm, have been addressed by using the ToF amplitude images instead of the RGB images in the motion analysis algorithm.
- In this Thesis a new approach for the detection and localisation of fruits in natural environments for harvesting robots have been introduced. This approach presents a unique multisensory system and the combination of the image registration and fusion algorithms, based on the depth-dependent *Hlut* method. The conceptual idea of this proposal was reported in the journal paper (Fernández et al. 2014, Fernández et al. 2013a) and in the international conferences proceeding (Barth et al. 2014, Montes et al. 2012). In this Thesis, an extensive and complementary research on this field has been presented.
- Finally, the proposed approach has demonstrated its capability for the feature extraction of small and rounded objects which have been acquired under very complex conditions.

## 6.4 Future Research Directions

Although for some robotic applications the results presented in this Thesis are accurate enough, other applications might require high-quality and high-accuracy colour depth maps. In future researches, more sophisticated algorithms for edge and depth measurement enhancement, as well as for the detection and removal of the outliers, should be investigated. Since the labelled homography mask $mask_{LRGB}$ was created for further implementations of these algorithms, smart and guided algorithms should be adopted to use the combination of the labelled information, the depth values and the texture of the RGB images.

The proposed method has been conceived as a flexible and adaptable approach. Hence, ongoing investigations with the proposed framework attempts to apply this methodology in other multisensor configurations, which are composed with two or more sensors such as thermal cameras, SWIR cameras, multispectral systems, and so on. For that purpose, an extensive and detailed description of the procedure for the automatic depth-dependent *Hlut* method computation is presented in this Thesis and listed in Algorithm 1.

# Part II: Método para el registro automático de imágenes basado en trasformaciones proyectivas planas dependientes de las distancia y orientado a imágenes sin características comunes

# Resumen

## 1. Fusión de imágenes multisensorial

La fusión de imágenes es una de las operaciones más importantes en el procesamiento de imágenes, que tiene como objetivo el mejoramiento del conocimiento y de la representación de entornos tridimensionales, mediante la adquisición de imágenes de una escena con múltiples y diferentes sensores, y capturadas en tiempos diferentes. La fusión sensorial ha sido ampliamente utilizada en la mayoría de campos de investigación, donde se requiere el análisis de imágenes. Dichos campos incluyen el análisis de imágenes médicas (James and Dasarathy 2014, Wyawahare et al. 2009), detección remota (Inglada and Giros 2004, Fonseca and Manjunath 1996), visión por computador (Salvi et al. 2007), robótica (Hines et al. 2003, Luo et al. 2002). Dada la gran diversidad y cantidad de aplicaciones, y el creciente número y diversidad de sensores para la captura de datos, es casi imposible, que una única metodología de fusión multisensorial tenga la capacidad de satisfacer a todos los campos de investigación previamente mencionados. Por lo tanto, la decisión de adoptar una determinada solución en temas de fusión sensorial, está directamente relacionada con la naturaleza de la aplicación y con la información que es considerada relevante.

Por lo general, la fusión de imágenes puede clasificarse en tres grupos de algoritmos, los cuales son, a nivel de píxeles, a nivel de características y a nivel simbólico. Los algoritmos a nivel de píxeles, han sido ampliamente investigados en comparación con los otros dos grupos de algoritmos. Como ejemplo de ello, en (Sahu and Parsai 2012), los autores presentan un estudio y

revisión del estado del arte de los algoritmos a nivel de píxeles. Estos algoritmos se basan en la variación de la intensidad de los píxeles, y pueden trabajar tanto en el domino del espacio, como en el de la frecuencia.

La estructura del procesamiento de la fusión de imágenes, normalmente se compone de cuatro pasos: pre-procesamiento (eliminación del ruido), el registro de imágenes (alineación de las imágenes), la fusión de imágenes (a nivel de píxeles) y el post-procesamiento (clasificación, segmentación y extracción de características). Los principales pasos del concepto general para para la fusión de imágenes, se muestran en la Figura 1.1.

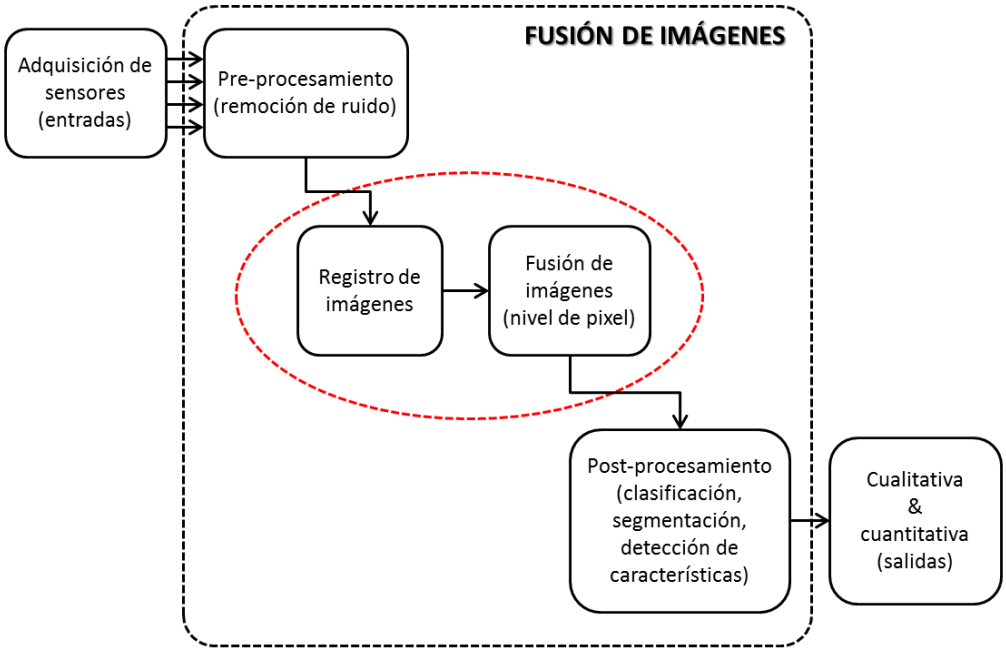



**Figura R.1** Diagrama de flujo para el procedimiento para la fusión de imágenes.

En la mayoría de los casos, para la implementación del algoritmo que permite la fusión de imágenes, se asume que los datos de entrada se encuentran perfectamente alineados. Sin embargo, en la práctica estas situaciones son difíciles de encontrar. Únicamente se encuentran en aquellas situaciones, en las que no se modifican los parámetros intrínsecos y extrínsecos de las cámaras, se obtienen imágenes que se encuentran geométricamente alineadas (Hall and Llinas 1997). Los demás casos

requieren la implementación previa de métodos de registro de imágenes. Mediante estos métodos se obtienen imágenes geométricamente alineadas, y dependiendo de la naturaleza tanto de las fuentes de captura de datos, como de las aplicaciones, dichos métodos se categorizan en aquellos que están basados en el análisis de características y en los de análisis de segmentos. Para ambas aproximaciones, el procedimiento de registro de imágenes se compone de cuatro pasos: la detección de características, la búsqueda de patrones comunes, la estimación de un modelo de transformación y la transformación entre las imágenes (Zitová and Flusser 2003).

El registro de imágenes es un paso fundamental en el procedimiento de fusión de imágenes, ya sea empleando para ello los parámetros de calibración de las cámaras o métodos más comunes, como los basados en el análisis de características y de segmentos. Este paso es primordial porque los errores obtenidos en este proceso, son posteriormente transferidos a cualquier clase de algoritmo diseñado para la fusión de imágenes (a nivel de píxeles, características o simbólico). Esta relación tan compleja entre el registro y la fusión de imágenes, puede enfocarse desde el punto de vista en el que se consideran dos diferencias entre las imágenes de entrada: las diferencias en el espacio y las diferencias que no tiene relación con el espacio (Zhang and Blum 2001). La primera de ellas tiene relación con el des-alineamiento espacial entre las imágenes, que puede deberse a las transformaciones geométricas entre las mismas (rotación, traslación, escalado, etc.). A la segunda se le atribuyen parámetros relacionados con el entorno, tales como las cambios en la iluminación, las escenas dinámicas, el uso de fuentes de captura de datos diferentes, así como el uso de sensores similares pero con configuraciones distintas. Por lo que, en lo referente a las diferencias en el espacio, el registro de imágenes es el encargado de solventar dichas diferencias, mientras que el resto de diferencias están a cargo de los algoritmos de fusión de imágenes.

El registro de imágenes, tal como ha sido mencionado anteriormente, es un método basado en el análisis de características, que por lo tanto, depende de la robustez de las características comunes que pueden encontrarse entre las imágenes. En casos donde estas características (intensidad de píxeles o similitud de regiones) no se encuentran disponibles o no son lo bastante robustas, en todas ellos se deben adoptar soluciones específicas. Un ejemplo de ello se observa en la combinación de cámaras de Tiempo de Vuelo (Time-of-Flight (ToF)) con cámaras de color (RGB), en el cual, las cámaras ToF no son capaces de detectar la información contextual del entorno, al que las

cámaras RGB. Las cámaras ToF proporcionan imágenes en escala de grises de baja resolución, que representan la respuesta de la señal de amplitud, y las correspondientes medidas de profundidad. Esto se consigue mediante la medición del tiempo de vuelo que le toma al haz de luz infrarroja, emitido por la ToF, viajar hasta el objeto y volver al sensor de la cámara. Por lo general, las medidas de profundidad proporcionadas por las cámaras ToF muestran la presencia de ruido, esto se debe a la configuración interna del hardware de los dispositivos y a diversas condiciones del entorno. Por ello, para la captura de medidas de profundidad fiables se requiere la aplicación de técnicas de calibración de los sensores y de filtrado de datos (Reynolds et al. 2011). Sin embargo, este problema de registro de imágenes, puede solventarse mediante la calibración de las cámaras, dado que los parámetros internos y externos de estas cámaras pueden utilizarse para la transformación homogénea entre las coordenadas de ambos dispositivos. Una de los primeros trabajos que estudió la fusión entre cámaras TOF y RGB está en Reulke (2006). Esta estrategia de percepción, basada en el uso de cámaras ToF y RGB, permite la captura de la información relativa a entornos tridimensionales, la cual incluye información de alta resolución del contexto de una escena y de la estructura 3D de la misma, todo ello alcanzando elevadas frecuencias de captura de datos. Estas características suponen una gran ventaja en aplicaciones de sistemas robóticos, en especial aquellas que se ejecutan en condiciones de tiempo real. Por otro lado, esta estrategia de fusión de datos no depende de las características comunes entre imágenes, tal como ocurre en la mayoría de las técnicas pasivas para la adquisición de información tridimensional. En comparación con otras técnicas activas para la percepción visual 3D, las cámaras ToF no requieren partes móviles para realizar medidas de profundidad, como sucede con los escáneres de láser, ni tampoco requieren de entornos con sistemas de iluminación altamente controlados, como es el caso de los métodos de iluminación estructurada. Una revisión exhaustiva de los métodos para la percepción visual 3D, ha sido presentada por Sansoni et al. (2009).

El propósito del estudio presentado en esta tesis, es la investigación de técnicas que permitan la fusión entre imágenes adquiridas por cámaras ToF y RGB, mediante soluciones de registro de imágenes que tengan alta exactitud y sean flexibles, y de igual manera, puedan ser adaptadas a aplicaciones en tiempo real. Esta solución debe ser capaz de resolver la problemática de fusionar sensores que no proporcionan imágenes con características similares. Por consiguiente, esta tesis se enfoca en el diseño y la implementación de

métodos de registro y fusión de imágenes, así como en la experimentación y validación de estos métodos en aplicaciones robóticas, tanto en entornos interiores como exteriores. Las aplicaciones tratadas en esta tesis son la seguridad y vigilancia para la monitorización de personas en hogares, y la detección y localización de frutas para tareas de cosechado, relativas a la agricultura de precisión. En ambos casos, dichas aplicaciones de sistemas robóticos, deben satisfacer condiciones de ejecución en tiempo real.

## 2. Motivación y alcance

La fusión de imágenes es una de las técnicas más importantes en la percepción visual tridimensional orientada a aplicaciones de sistemas robóticos, que permiten la captura de la información del contexto y de la estructura de una escena. Tal como se ha mencionado anteriormente, son dos los pasos más importantes en la fusión de imágenes, que permiten conseguir una representación más exacta y de alta calidad de entornos tridimensionales. Dichos pasos comprenden al registro y fusión (a nivel de píxeles) de imágenes, y se muestran enmarcados en rojo en la Figura 1.1, la cual ilustra el diagrama de flujo operaciones para la fusión de imágenes. El éxito o fracaso del proceso de registro de imágenes, condicionará los resultados del proceso de fusión de datos. Por consiguiente, se debe prestar especial atención al proceso de registro de datos, para poder garantizar la exactitud en la alineación espacial de las imágenes, lo cual forma una parte muy importante del trabajo de investigación de esta tesis doctoral.

Por tal razón, se ha investigado la metodología de fusión de datos entre cámaras ToF y RGB. La compromiso entre las imágenes de color de alta resolución obtenidas por las cámaras RGB y las medidas de profundidad adquiridas por las cámaras ToF, convierten a esta estrategia, en método capaz de proporcionar información de objetos definidos de forma apropiada, tanto en el contexto como en la estructura, con una descripción exacta de los bordes y de las formas de los objetos, y que además, es adecuada para aplicaciones en tiempo real.

Existen diversas problemáticas relacionadas con el registro de estos dos tipos de cámaras. Por un lado, la baja resolución de las cámaras ToF hacen que el cálculo de parámetros exactos de calibración, no sea una tarea sencilla de conseguir. Aún, si se obtienen parámetros de calibración con bastante exactitud, el ruido presente en las medidas de profundidad, introduce errores durante la transformación homogénea entre las cámaras. Con el objetivo de reducir el ruido en las medidas de la cámara ToF, normalmente se aplican

técnicas de filtrado de datos, que en este caso en particular, debido a la baja resolución de estas cámaras, hace que los datos sean aún más susceptibles al efecto de sobre-alisado de bordes y superficies.

Por otro lado, dado que existe una gran diferencia entre las resoluciones de ambas cámaras, únicamente se pueden transformar centenas de píxeles en las imágenes ToF y RGB, que se corresponden con la resolución de las imágenes ToF (144×176 píxeles). Con el objetivo de aprovechar toda la información de color proporcionada por las imágenes de alta resolución de las cámaras RGB, el estudio de métodos que proporcionan mapas de color y de profundidad de alta resolución, es de gran relevancia para este trabajo de investigación.

Otro reto al que se debe enfrentar este trabajo de investigación, es la validación de los resultados obtenidos mediante la fusión sensorial de imágenes. Con el objetivo de validar el método de fusión de imágenes propuesto, se han seleccionado dos aplicaciones de sistemas robóticos, una de ellas enfocada a entornos interiores y la otra a exteriores, en los cuales se deben probar las capacidades del método propuesto en esta tesis.

## 3. Objetivos de la investigación

El primer y principal objetivo de este trabajo de investigación es el de diseñar, implementar y validar un método para el registro de imágenes obtenidas mediante la fusión sensorial de una cámara ToF y una RGB. Esta solución debe ser capaz de resolver problemas como la carencia de características comunes entre las imágenes capturadas por dichos sensores, la presencia de ruido en las medidas de profundidad de la cámara ToF y la gran diferencia entre las resoluciones de estas cámaras (ToF y RGB). Adicionalmente, el registro de imágenes debe poder ser adaptado para aplicaciones en tiempo real. Como parte del proceso de validación del método propuesto, se deben evaluar resultados tanto visuales, como numéricos. Dado que existen otras metodologías para el registro de las imágenes, se debe realizar una evaluación comparación entre el método propuesto y el método de mayor relevancia, que en este caso es el método basado en la calibración estándar de cámaras.

El segundo objetivo en esta tesis, es el desarrollo de una validación experimental que comprende dos aplicaciones de sistemas robóticos: la seguridad y vigilancia para la monitorización de caída de personas en hogares, y la detección y localización de frutas en cultivos para aplicaciones con robots cosechadores, utilizados en agricultura de precisión. En dichas aplicaciones se deben evaluar las capacidades del método propuesto en esta

tesis, en combinación con algoritmos de fusión de imágenes. El propósito de estas evaluaciones es el de demostrar que la metodología propuesta en este trabajo de investigación, tiene la capacidad de proporcionar, de forma satisfactoria, información exacta y de calidad con referencia a la clasificación y segmentación de los datos y la extracción de las características de los objetos de interés. Todo ello, orientado a tareas específicas de aplicaciones de sistemas robóticos de servicio y de agricultura de precisión.

## 4. Organización de la tesis

Con el propósito de abordar los objetivos, la memoria de la tesis está organizada de la siguiente forma:

El Capítulo 2 presenta el estado del arte de las técnicas de fusión sensorial de imágenes, las cuales son fundamentales para diversos métodos de percepción visual tridimensional. De estas técnicas, en este estudio se presenta una investigación más exhaustiva enfocada a las técnicas de fusión de cámaras ToF y RGB.

El Capítulo 3 se dedica al diseño, implementación y validación de la metodología para el registro de imágenes obtenidas mediante la fusión de una cámara ToF y una RGB. Se presentan los conceptos fundamentales para el diseño de la propuesta, así como la información detallada para el cómputo del nuevo enfoque denominado *depth-dependent Hlut*. También se presenta el análisis preliminar de los resultados de la evaluación de la precisión del método. Finalmente, se propone un procedimiento para la obtención de mapas de color y profundidad de alta resolución.

En el Capítulo 4 se presentan los resultados de la comparación entre el método propuesto *Hlut* y el método de calibración estándar de las cámaras. Esta comparación aborda tres escenarios con distintas medidas de la profundidad: sin ruido (ideales), sin procesar y filtradas. El primer grupo de datos, se obtiene de los pasos del procedimiento de calibración de las cámaras. El segundo grupo, corresponde a las medidas de profundidad adquiridas directamente por la cámara ToF. Por último, para el tercer escenario, se aplican dos técnicas de filtrado de datos en las medidas en crudo de la cámara ToF, que son el filtrado bilateral y el filtrado no local de medias.

El Capítulo 5 aborda la sección de experimentación, la cual comprende la validación del método en dos aplicaciones de sistemas robóticos relevantes (en entornos de interiores y de exteriores). El primer conjunto de experimentos enmarcados en la primera aplicación robótica, está enfocado a

la seguridad y vigilancia en hogares, para la monitorización del movimiento de personas. Para este caso, se han diseñado dos series de experimentos. La primera de ellas, se centra en la evaluación de la exactitud del método y de su capacidad para el registro correcto de superficies grandes e inclinadas. La segunda serie, está enfocada en la validación del método en tareas de detección del movimiento. El conjunto de experimentos enmarcados en la segunda aplicación, está orientado a la detección y localización de frutas (manzanas) en árboles para aplicaciones que utilizan sistemas robóticos cosechadores. Con este propósito, se realizan experimentos tanto en laboratorio como en campos de cultivos. De forma adicional, se propone un procedimiento para la extracción de características de los objetos de interés. Dicho proceso combina el método *Hlut* y técnicas de fusión de imágenes a nivel de píxeles.

Finalmente, en el Capítulo 6, se resumen los resultados más importantes obtenidos en este trabajo de investigación, además de las contribuciones más relevantes de esta tesis doctoral y finalmente se presentan posibles líneas futuras de investigación.

## 5. Conclusiones, aportaciones principales y trabajos futuros

En el trabajo de investigación realizado en esta tesis, se ha presentado un método para el registro automático de imágenes, el cual no depende del análisis de características similares entre imágenes, el cual se fundamenta en transformaciones proyectivas planas dependientes de la distancia. Para ello, se ha diseñado e implementado una un sistema que consiste en una cámara ToF de baja resolución y por una cámara de color de alta resolución, para efectuar el registro y la fusión de imágenes adquiridas por ambos dispositivos. Dicho sistema, tiene la capacidad de registrar imágenes que no proporcionan características comunes entre sí, y, de solventar la presencia de ruido moderado en las medidas de profundidad obtenidas por la cámara ToF.

El método propuesto en esta tesis está basado en una tabla de búsqueda de homografías planas, las cuales dependen de las medidas de profundidad, denominada en inglés *depth-dependent homography lookup table (Hlut)*. Los elementos de esta tabla se obtienen mediante la discretización virtual del espacio tridimensional en {*n*-planos}, los cuales se encuentran dispuestos frente al sistema y paralelos a éste. A partir de éstos, la transformación de puntos entre las vistas depende de la distancia entre los objetos y el sistema multisensorial. Estas características convierten a esta solución, en una técnica no basada en el análisis de las características similares entre imágenes. Dado que este método se

fundamenta en transformaciones proyectivas planas, los requerimientos de carga computacional son bajos, haciendo posible la implementación del método en aplicaciones de tiempo real.

Las conclusiones para cada uno de los objetivos propuestos en esta tesis, se detallan a continuación.

## 5.1 Diseño, implementación y validación del método propuesto para el registro de imágenes

La validación de los resultados obtenidos por el método *Hlut*, los cuales se han presentado en el Capítulo 3, muestran que la solución propuesta presenta un rendimiento satisfactorio, tanto en términos de calidad de los resultados visuales, como en términos numéricos evaluados por medio de la raíz cuadrada del error cuadrático medio (RMSE). Por otro lado, mediante este método se consigue la transformación de puntos con un error inferior a 4 píxeles, medidos en el plano imagen de la cámara RGB, lo cual se considera un valor pequeño en comparación con la resolución de la imagen RGB que es de $2448 \times 2050$ píxeles. Dichos errores representan pequeñas distorsiones, dadas las distancias de trabajo del sistema, que se encuentran entre los 400 y 2300 mm. Mediante este procedimiento, se consiguen mapas de color y de profundidad de baja resolución, y además de ello, se proporciona una máscara de etiquetas de homografías $mask_{LRGB}$ en el plano imagen de la cámara RGB. Los valores de dicha máscara se corresponden con las homografías $\{ H_k^{lut} \}$ utilizadas para la transformación de los puntos. Dado que existe una gran diferencia entre las resoluciones de las cámaras, entre cada par de puntos adyacentes estimados en la imagen RGB, existen varios puntos de color sin ser mapeados. Por este motivo, la máscara de etiquetas $mask_{LRGB}$, ha sido creada con el objetivo de ser implementada como una herramienta en la resolución de dicho problema. En este trabajo de investigación se presenta una primera propuesta para el uso de esta máscara, la cual está basada en el algoritmo de vecinos más próximos, obteniendo como resultado una máscara completa de etiquetas de homografías $\{ H_k^{lut} \}$ en el plano imagen de color. De esta manera se obtiene un mapa de color y de profundidad de alta resolución, mediante la transformación inversa de las homografías $\left\{ H_k^{lut^{-1}} \right\}$, que permiten transferir puntos desde el plano imagen RGB al plano imagen ToF.

El siguiente paso en la evaluación del método propuesto, consiste en la comparación exhaustiva entre el método *Hlut* y el método de la calibración

estándar de las cámaras, presentada en el Capítulo 4 de la tesis. De acuerdo la revisión del estado del arte, el método de calibración estándar es el método más utilizado para el cómputo del registro de mapas de profundidad y, por esta razón, es el elegido como base para esta evaluación comparativa. Por otro lado, de acuerdo al estudio del estado del arte realizado en el Capítulo 2, existen errores sistemáticos (relativos a los parámetros internos de configuración, el hardware, etc.) y no sistemáticos (factores externos, tales como las condiciones de iluminación, desenfoque del movimiento, etc.), muchos de los cuales son reducidos mediante la calibración de estos dispositivos y el filtrado de los datos. Dado que ambos métodos de registro de imágenes dependen de las medidas de profundidad adquiridas por la cámara ToF, esta evaluación comparativa se ha llevado a cabo desde un punto de vista tal, que se evalúa la respuesta de estos métodos bajo la presencia de ruido en dichas medidas.

Lo primero fue la realización del cómputo, lo más exacto posible, de los parámetros de calibración de las cámaras. Con este objetivo se han comparado los resultados obtenidos tanto, de los parámetros intrínsecos como extrínsecos, con los resultados de exactitud obtenidos en estudios similares de este campo. A continuación, se ha realizado la evaluación de los métodos en tres situaciones, con datos de entrada: sin ruido (ideales), sin procesar y filtrados. Los resultados obtenidos durante la evaluación indican que, el método propuesto supera los valores de exactitud obtenidos por el método de calibración estándar. Como ejemplo de ello, en el procesamiento de datos sin filtrar, el método *Hlut* reduce el error en un 41%, con un error de $\mathrm{RMSE} = 0.2440$, comparándolo con el error del método de calibración estándar que posee un $\mathrm{RMSE} = 0.4150$. De igual forma ocurre con los datos filtrados, en los que los errores obtenidos por el método de calibración estándar se incrementan en un 127% con el filtrado bilateral y en un 216% con el filtro de medias no locales, ambos comparados con los datos obtenidos con el método propuesto. Los errores obtenidos con el método *Hlut* correspondientes al filtrado bilateral y al filtro de medias no locales son de $\mathrm{RMSE} = 0.2376$ y $\mathrm{RMSE} = 0.2365$, respectivamente. De forma similar, los errores obtenidos con el método de calibración estándar son de $\mathrm{RMSE} = 0.5402$ y $\mathrm{RMSE} = 0.7478$, respectivamente, para cada técnica de filtrado de datos.

Se puede concluir que estos resultados destacan la capacidad y flexibilidad del método propuesto para lidiar con variaciones moderadas en las medidas de profundidad y con los datos filtrados que han sido extremadamente

alisados. Dado que el método *Hlut* depende de un rango de medidas de profundidad, en lugar de utilizar los valores exactos de cada una de las medidas de profundidad, permite eludir las desviaciones en las medidas de profundidad.

## 5.2. Análisis experimental y validación del método propuesto en aplicaciones de sistemas robóticos en entornos de interiores y de exteriores

Para el proceso de experimentación y validación de la solución propuesta en esta tesis, se han elegido dos aplicaciones de sistemas robóticos presentadas, que son presentador en el Capítulo 5. Dicha solución consigue la fusión de los datos mediante el registro de imágenes con el método *Hlut* y algoritmos de fusión de imágenes a nivel de píxeles. En las siguientes subsecciones, se detallan las conclusiones para cada una de las aplicaciones.

### *5.2.1 Etapa experimental en aplicaciones robóticas en entornos de interiores*

La primera prueba experimental está enfocada a las aplicaciones en entornos de interiores, tales como la seguridad y vigilancia en hogares y la monitorización de caídas de personas. Para estas aplicaciones, se ha analizado la implementación de la estructura propuesta en tareas de análisis y detección del movimiento. Además de esto, esta etapa de experimentación también comprende, la validación de la exactitud del método propuesto y de sus capacidades para registrar superficies grandes mediante varias homografías $\{H_k^{lut}\}$, considerando que se evita la presencia de discontinuidades entre las transiciones de dichas homografías. Con este objetivo se han realizado dos series de ensayos y los resultados de mayor relevancia se detallan a continuación.

#### *5.2.1.1 Primera serie de ensayos (interiores): validación de la exactitud y el correcto registro de imágenes de grandes superficies*

En primer lugar, se realizó la adquisición de varias imágenes de una superficie plana (tablero de patrones), la cual se colocó en diversas posiciones, orientaciones y distancias con respecto al sistema sensorial. A continuación, se repitió este proceso, pero esta vez, utilizando varios objetos volumétricos y no uniformes, tales como una silla, un cubo cilíndrico y una

persona. Los resultados visuales y numéricos, resaltan el buen desempeño de la propuesta, obteniendo un error normalizado RSME de 0.1272, un error medio de $Mean_{(u,v)-axis} = [-0.7, -0.3]$ y con una desviación estándar $\sigma_{(u,v)-axis} = [3.5,4.6]$ para los objetos planos y, para los objetos volumétricos, un valor de $RMSE = 0.3511$, un error medio de $Mean_{(u,v)-axis} = [-0.46, 5.5]$ y una desviación estándar de $\sigma_{(u,v)-axis} = [10.5,11.3]$. Además de estos datos prometedores, los resultados visuales demuestran la capacidad del método de evitar la presencia de discontinuidades en el mapeo de datos de superficies, en las cuales se preservan los bordes y las formas, proporcionando la alineación lo bastante exacta de la información de color y de profundidad. Esto contrasta con los resultados obtenidos por medio del método de calibración estándar, donde las superficies de los objetos son menos homogéneas y los bordes de los objetos exhiben varios problemas de desalineación.

#### *5.2.1.2 Segunda serie de ensayos (interiores): evaluación de las capacidades del método propuesto en tareas de análisis de movimiento*

Con el propósito de evaluar el potencial del método propuesto en tareas de análisis del movimiento, se ha presentado un procedimiento para la detección del movimiento. Este procedimiento consiste en un algoritmo robusto de análisis afín del movimiento, implementado en las imágenes de amplitud adquiridas por la cámara ToF. Luego, la máscara de los movimientos detectados en el paso anterior, se combina con las medidas de profundidad y con la imagen registrada de color, para ser utilizados en el cómputo de la estructura tridimensional de los cuerpos de las personas y su correspondiente información de color de alta resolución. Este procedimiento reduce los falsos positivos en la detección del movimiento, normalmente producidos por sombras y variaciones en el sistema de iluminación. La información resultante de este proceso tiene un gran valor para el procedimiento de toma de decisiones, dado que los datos de la aproximación cuadrática de la superficie de la estructura 3D de un cuerpo, pueden ser utilizadas para la extracción de características de una elipsoide, a menudo utilizada en estudios de detección de caídas de personas. En este caso, la información de color de alta resolución, puede ser empleada para la extracción de características referentes de las personas. Por otro lado, los resultados obtenidos, demuestran que el método propuesto es capaz de proporcionar mapas de color y de

profundidad de alta resolución, evitando así, la perdida de la información de color.

### *5.2.2 Etapa experimental en aplicaciones robóticas en entornos de exteriores*

La segunda aplicación elegida para la evaluación y la experimentación del método propuesto se encuentra enmarcada dentro de un proyecto europeo, denominado por el acrónimo CROPS y cuyo título en inglés es *Intelligent Sensing and Manipulation for Sustainable Production and Harvesting of High Value Crops, Clever Robots for Crops*. El objetivo general de esta etapa es la validación de la implementación del método propuesto *Hlut* en combinación con un sistema multiespectral, en las tareas de detección y localización de frutas (manzanas) y otros elementos de las plantas, todo ello en entornos naturales. Con este objetivo se han realizado dos grupos de experimentos. El primero de ellos es un ensayo inicial en condiciones de laboratorio, en el que se han analizado escenas compuestas por un conjunto de manzanas artificiales. Los resultados obtenidos de dichos análisis muestran una respuesta satisfactoria en el proceso de registro de imágenes, tomando en cuenta la complejidad de las escenas debido al pequeño tamaño de las frutas, en relación al campo de visión de las imágenes y la forma redondeada de las frutas. Estas dificultades se hacen más relevantes cuando los objetos están inclinados con respecto al plano de las imágenes. En todos los casos que han sido evaluados, los problemas de desalineación son casi imperceptibles y se preservan tanto los bordes, como la forma de los objetos.

El segundo grupo ha sido desarrollo en entornos naturales, para lo cuales, se ha llevado a cabo una campaña de adquisición de datos en campos de cultivos de manzanas. En este caso, la complejidad de las escenas es aún mayor, debido a la variación en las condiciones de la iluminación, la posición aleatoria de las manzanas, los elementos naturales de las plantas y la naturaleza cambiante en estos entornos, tales como la presencia del viento. A pesar de estas dificultades, los resultados visuales y numéricos obtenidos son muy prometedores, dado que se ha conseguido el cómputo de mapas de color y de profundidad de alta resolución con una exactitud aceptable y se ha procedido a la implementación exitosa de un procedimiento para la extracción de características de las frutas.

Una ventaja importante del método propuesto en esta tesis, se ilustra en el Capítulo 5, con los resultados de la detección y localización de frutas. El método propuesto *Hlut*, no requiere a priori de un proceso de filtrado de

datos, que normalmente produce el sobre-alisado de los datos de profundidad. Sin embargo, este método es capaz de proporciona mapas de color y de profundidad de alta resolución y elevada exactitud, en términos de alineamiento de datos. Con esto se consigue que los objetos muy pequeños o las características de éstos no sean eliminados y, en la etapa de post-procesamiento, el filtrado y ajuste de datos podrían abordarse desde una perspectiva local, empleando para ello, la combinación de los pocos datos de profundidad de los que se dispone y la gran cantidad de información de color.

## 5.3 Aportaciones Principales

Las principales aportaciones de la tesis se indican a continuación:

- Se ha presentado el diseño, implementación y evaluación de un método novedoso para el registro automático de imágenes adquiridas con una cámara ToF y una cámara RGB. Los detalles del procedimiento para el cómputo automático de método propuesto, denominada en inglés como *depth-dependent Hlut*, están descritos en el seudocódigo del Algoritmo 1. Esta contribución de la tesis ha dado lugar a una publicación en una revista SCI (Salinas et al. 2015).
- Se ha realizado una exhaustiva evaluación comparativa entre el método propuesto y el método de la calibración estándar, en la que se han considerado tres escenarios diferentes y con datos de entrada relevantes: datos sin ruido (ideales), datos sin procesar y datos filtrados. Se han analizado los resultados visuales y numéricos, en los cuales el método propuesto supera a los resultados de exactitud obtenidos por el método de calibración estándar. También se ha demostrado el método propuesto tiene la capacidad de lidiar con la presencia de ruido moderado en las medidas de profundidad.
- Se ha realizado una validación exhaustiva de la exactitud y del rendimiento del método. Esta propuesta ha mostrado un rendimiento satisfactorio en el registro de superficies grandes, para las que se proporcionan regiones registradas de color y de profundidad que son uniformes y que no presentan discontinuidades.
- Este método ha sido implementado en tareas de detección del movimiento. Los resultados obtenidos resaltan la capacidad de la

propuesta para ser utilizados en tareas de monitorización de caídas de personas. Este procedimiento proporciona información 3D y contextual de la estructura de un cuerpo, la cual se corresponde con las regiones detectadas como áreas de movimiento. Estudios preliminares del algoritmo de detección del movimiento fueron presentados en un artículo de revista (Salinas et al. 2012).

- Los problemas ocasionados por las sombras y los cambios en las condiciones de iluminación, los cuales provocan la detección de falsos positivos del movimiento, se solucionaron mediante la implementación de dicho algoritmo sobre las imágenes de amplitud adquiridas por la cámara ToF, en lugar de utilizar las imágenes de color.
- En esta tesis se ha presentado un nuevo enfoque para la detección y localización de frutas para que sean recolectadas por medio de robots cosechadores. Este enfoque presenta un sistema multisensorial y la combinación de algoritmos de registro y de fusión imágenes, basados en el método propuesto *depth-dependent Hlut*. La idea conceptual de este enfoque ha sido presentado en revistas SCI (Fernández et al. 2014, Fernández et al. 2013a) y en conferencias internacionales publicadas en capítulos de artículos (Barth et al. 2014, Montes et al. 2012). En la memoria de tesis se ha presentado un estudio más detallado y complementario de este tema.
- Se ha demostrado la capacidad del método propuesto para la extracción de características de objetos pequeños y redondeados, los cuales han sido adquiridos en circunstancias de complejidad elevada.

## 5.4 Trabajos futuros

Aunque es cierto que para muchas aplicaciones robóticas, los resultados presentados en esta tesis son lo bastante exactos, para otras aplicaciones pueden ser requeridos mapas de color y de profundidad de alta resolución y también de alta calidad. Por lo tanto, considerando lo anteriormente expuesto, en trabajos futuros, se deben investigar algoritmos más sofisticados para la mejora de los bordes de los objetos y de sus medidas de profundidad, así como métodos para la eliminación de datos atípicos. Dado que la máscara de etiquetas de homografías $mask_{LRGB}$, ha sido creada con el objetivo de la

implementación de dichos algoritmos, se deberían adoptar algoritmos inteligentes y guiados, mediante el uso de la información de dichas etiquetas, en combinación con las medidas de profundidad y del color.

Por otro lado, esta propuesta ha sido concebida como una metodología flexible y adaptable. Por consiguiente, en investigaciones futuras, se pretende aplicar esta metodología en la fusión multisensorial compuesta por otras configuraciones de dos o más sensores, de diversos tipos, tales como cámaras térmicas LWIR, cámaras SWIR, sistemas multiespectrales. Con este objetivo, en esta tesis se ha presentado una descripción detallada del procedimiento para el diseño e implementación del método *Hlut* (seudocódigo del Algoritmo 1 y del Algoritmo 2).

# References


Allied Vision, T. 2011. AVT Prosilica GC2450.

Aracil, R., C. Balaguer & M. Armada (2008) Robots de Servicio. *Rev. Iberoam. Autom. Inf. Ind,* 5**,** 6-13.

Ayache, N. & P. T. Sander. 1991. *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception*. MIT Press.

Bac, C. W., J. Hemming & E. J. v. Henten (2013) Robust pixel-based classification of obstacles for robotic harvesting of sweet-pepper. *Computers and Electronics in Agriculture* 96**,** 148-162.

Baker, S. & S. Nayar (1999) A Theory of Single-Viewpoint Catadioptric Image Formation. *International Journal of Computer Vision,* 35**,** 175-196.

Barrientos, A., L. F. Peñin, C. Balaguer & R. Aracil. 2007. *Fundamentos de robótica*. McGraw-Hill.

Barth, R., J. Baur, T. Buschmann, Y. Edan, T. Hellström, T. Nguyen, O. Ringdahl, W. Saeys, C. Salinas & E. Vitzrabin. 2014. Using ROS for agricultural robotics - Design considerations and experiences. In *Proceedings of the Second International Conference on Robotics and associated High-technologies and Equipment for Agriculture and forestry (RHEA-2014),* 509-518. Madrid, Spain.

Beder, C., B. Bartczak & R. Koch. 2007. A Comparison of PMD-Cameras and Stereo-Vision for the Task of Surface Reconstruction using Patchlets. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 1-8.

Beraldin, J.-A. & M. Gaiani. 2005. Evaluating the performance of close-range 3D active vision systems for industrial design applications. In

*Electronic Imaging 2005*, 67-77. International Society for Optics and Photonics.

Berestein, R., O. Ben-Shahar, A. Shapiro & Y. Edan (2010) Grape clusters and foliage detection algorithms for autonomous selective vineyard sprayer. *Intell. Serv. Robot,* 3**,** 233-243.

Bernardini, F. & H. Rushmeier. 2002. The 3D model acquisition pipeline. In *Computer graphics forum*, 149-172. Wiley Online Library.

Besl, P. (1988) Active, optical range imaging sensors. *Machine Vision and Applications,* 1**,** 127-152.

Black, M. J. & P. Anandan (1996) The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding,* 63**,** 75-104.

Blais, F. (2004) Review of 20 years of range sensor development. *Journal of Electronic Imaging,* 13**,** 231-243.

Bouguet, J. Y. 2008. Camera calibration toolbox for Matlab.

Bradski, G. & A. Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.".

Buades, A., B. Coll & J. M. Morel. 2005. A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 60-65 vol. 2.

Buemi, F., M. Massa & G. Sandini. 1985. Agrobot: a robotic system for greenhouse operations. In *4th Workshop on robotics in Agriculture, IARP*, 172-184 Toulouse.

Bulanon, D. M., T. F. Burks & V. Alchanatis (2010) A multispectral imaging analysis for enhancing citrus fruit detection. *Environ. Control Biol,* 48**,** 81-91.

Bulanon, D. M. & T. Kataoka (2010) A fruit Detection System and an End Effector for Robotic Harvesting of Fuji Apples. *Agricultural Engineering International: the CIGR Ejournal. Manuscript,* XII**,** 203-210.

Bulanon, D. M., T. Kataoka, H. Ukamoto & S. Hata. 2004. Development of a real-time machine vision system for the apple harvesting robot. In *In SICE Annual conference in Sapporo*, 595-598. Hokkaido Institute of Technology, Japan

Cabral, E. L., J. C. de Souza & M. C. Hunold. 2004. Omnidirectional stereo vision with a hyperbolic double lobed mirror. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 1-9. IEEE.

Crops-project. 2010. Intelligent Sensing and Manipulation for Sustainable Production and Harvesting of High Value Crops, Clever Robots for Crops (CROPS)

Cucchiara, R., C. Grana, A. Prati & R. Vezzani (2005) Computer vision system for in-house video surveillance. *Vision, Image and Signal Processing, IEE Proceedings -,* 152**,** 242-249.

Chai, T. & R. R. Draxler (2014) Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development,* 7**,** 1247-1250.

Chan, D., H. Buisman, C. Theobalt & S. Thrun. 2008. A noise-aware filter for real-time depth upsampling. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008*.

Chiabrando, F., R. Chiabrando, D. Piatti & F. Rinaudo (2009) Sensors for 3D Imaging: Metric Evaluation and Calibration of a CCD/CMOS Time-of-Flight Camera. *Sensors,* 9**,** 10080-10096.

DePiero, F. W. & M. M. Trivedi. 1996. 3-D Computer Vision Using Structured Light: Design, Calibration, and Implementation Issues. In *Advances in Computers,* ed. V. Z. Marvin, 243-278. Elsevier.

Deshmukh, M. & U. Bhosle (2011) A survey of image registration. *International Journal of Image Processing,* 5**,** 245-269.

Dey, D. & L. Mummert. 2012. Classification of Plant Structures from Uncalibrated Image Sequences. In *IEEE Workshop on Applications of Computer Vision (WACV)*, 329-336. Breckenridge, CO, USA.

Di Paola, D., D. Naso, A. Milella, G. Cicirelli & A. Distante. 2008. Multi-Sensor Surveillance of Indoor Environments by an Autonomous Mobile Robot. In *Mechatronics and Machine Vision in Practice, 2008. M2VIP 2008. 15th International Conference on*, 23-28.

Diraco, G., A. Leone & P. Siciliano. 2010. An active vision system for fall detection and posture recognition in elderly healthcare. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010,* 1536-1541.

Fernández, R., H. Montes & C. Salinas (2015) VIS-NIR, SWIR and LWIR Imagery for Estimation of Ground Bearing Capacity. *Sensors,* 15**,** 13994.

Fernández, R., H. Montes, C. Salinas, P. González de Santos & M. Armada (2012) Design of a training tool for improving the use of hand-held detectors in humanitarian demining. *Industrial Robot: An International Journal,* 39**,** 450-463.

Fernández, R., H. Montes, C. Salinas, J. Sarria & M. Armada (2013a) Combination of RGB and multispectral Imagery for Discrimination of Cabernet Sauvignon Grapevine Elements. *Sensors,* 13**,** 7838-7859.

Fernández, R., C. Salinas, H. Montes & J. Sarria (2014) Multisensory System for Fruit Harvesting Robots. Experimental Testing in Natural Scenarios and with Different Kinds of Crops *Sensors* 14**,** 23885-23904.

Fernández, R., C. Salinas, H. Montes, J. Sarria & M. Armada. 2013b. Validation of a Multisensory System for Fruit Harvesting Robots in Lab Conditions. In *ROBOT2013: First Iberian Robotics Conference* eds. M. Armada, A. Sanfeliu & M. Ferre. Madrid: Springer.

Fischler, M. A. & R. C. Bolles (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM,* 24**,** 381-395.

Foix, S., G. Alenya & C. Torras (2011) Lock-in Time-of-Flight (ToF) Cameras: A Survey. *Sensors Journal, IEEE,* 11**,** 1917-1926.

Fonseca, L. M. G. & B. S. Manjunath (1996) Registration Techniques for Multisensor Remotely Sensed Imagery. *Journal of Photogrammetry Engineering and Remote Sensing,* 62**,** 1049-1056.

Forest Collado, J. 2004. *New methods for triangulation-based shape acquisition using laser scanners*. Universitat de Girona.

Foroughi, H., B. S. Aski & H. Pourreza. 2008. Intelligent video surveillance for monitoring fall detection of elderly in home environments. In *Computer and Information Technology, 2008. ICCIT 2008. 11th International Conference on*, 219-224.

Ganapathi, V., C. Plagemann, D. Koller & S. Thrun. 2010. Real time motion capture using a single time-of-flight camera. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 755-762.

Gasparrini, S., E. Cippitelli, S. Spinsante & E. Gambi (2014) A depth-based fall detection system using a Kinect® sensor. *Sensors,* 14**,** 2756-2775.

Geyer, C. & K. Daniilidis (2001) Catadioptric Projective Geometry. *International Journal of Computer Vision,* 45**,** 223-243.

Gilvydis, J. B. 1985. Observation system for military vehicles. Google Patents.

Gluckman, J., S. K. Nayar & K. J. Thoresz. 1998. Real-time omnidirectional and panoramic stereo. In *Image Understanding Workshop*, 299-303. Citeseer.

Gonzalez de Santos, P., J. A. Cobano, E. Garcia, J. Estremera & M. Armada (2007) A six-legged robot-based system for humanitarian demining missions. *Mechatronics,* 17**,** 417-430.

Goshtasby, A. A. & S. Nikolov (2007) Guest editorial: Image fusion: Advances in the state of the art. *Inf. Fusion,* 8**,** 114-118.

Govender, M., K. Chetty & H. Bulcock (2007) A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa,* 33.

Guðmundsson, S. Á., H. Aanæs & R. Larsen (2008) Fusion of Stereo Vision and Time-of-Flight Imaging for Improved 3D Estimation. *International Journal of Intelligent Systems Technologies and Applications,* 5**,** 425-433.

Guðmundsson, S. Á., M. Pardas, J. R. Casas, J. R. Sveinsson, H. Aanæs & R. Larsen (2010) Improved 3D reconstruction in smart-room environments using ToF imaging. *Computer Vision and Image Understanding,* 114**,** 1376-1384.

Guomundsson, S. A., H. Aanaes & R. Larsen. 2007. Environmental Effects on Measurement Uncertainties of Time-of-Flight Cameras. In *Signals, Circuits and Systems, 2007. ISSCS 2007. International Symposium on*, 1-4.

Hahne, U. 2009. Depth Imaging by Combining Time-of-Flight and On-Demand Stereo. In *Dynamic 3D Imaging. Lecture Notes in Computer Science*, 70-83.

Hahne, U. & M. Alexa (2008) Combining Time Flight depth and stereo images without accurate extrinsic calibration. *Int. J. Intell. Syst. Technol. Appl.,* 5**,** 325-333.

Hall, D. L. & J. Llinas (1997) An introduction to multisensor data fusion. *Proceedings of the IEEE,* 85**,** 6-23.

Hartley, R. & A. Zisserman. 2003. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Hayashi, S., K. Shigematsu, S. Yamamoto, K. Kobayashi, Y. Kohno, J. Kamata & M. Kurita (2010) Evaluation of a strawberry-harvesting robot in a field test. *Biosystems Engineering,* 105**,** 160-171.

Hayashi, S., S. Yamamoto, S. Sarito, Y. Ochiai, Y. Kohno, K. Yamamoti, J. Kamata & M. Kurita. 2012. Development of a movable strawberry-Harvesting Robot Using a Travelling Platform. In *International Conference of Agricultural Engineering*. Valencia, Spain.

Hebert, M. 2000. Active and passive range sensing for robotics. In *Robotics and Automation, 2000. Proceedings. ICRA '00. IEEE International Conference on,* 102-110 vol.1.

Hines, G. D., Z.-u. Rahman, D. J. Jobson & G. A. Woodell. 2003. Multi-image registration for an enhanced vision system. In *SPIE Visual Information and Processing*, 231-241.

Hong, J., X. Tan, B. Pinette, R. Weiss & E. M. Riseman (1992) Image-based homing. *Control Systems, IEEE,* 12**,** 38-45.

Huhle, B., T. Schairer, P. Jenke & W. Straßer (2010) Fusion of range and color images for denoising and resolution enhancement with a non-local filter. *Computer Vision and Image Understanding,* 114**,** 1336-1345.

Inglada, J. & A. Giros (2004) On the possibility of automatic multisensor image registration. *Geoscience and Remote Sensing, IEEE Transactions on,* 42**,** 2104-2120.

James, A. P. & B. V. Dasarathy (2014) Medical image fusion: A survey of the state of the art. *Information Fusion,* 19**,** 4-19.

Janesick, J. R., T. Elliott, S. Collins, M. M. Blouke & J. Freeman (1987) Scientific Charge-Coupled Devices. *Optical Engineering,* 26**,** 268692-268692-.

Jiménez, A., R. Ceres & J. Pons (2000a) A survey of computer vision methods for locating fruits on trees. *Transactions of the ASAE,* 43**,** 1911-1920.

Jiménez, A., R. Ceres & J. Pons (2000b) A vision system based on a laser range-finder applied to robotic fruit harvesting. *Machine Vision and Applications,* 11**,** 321-329.

Kassay, L. 1992. Hungarian robotic apple harvester. In *ASAE annual meeting papers*, 1-14. St. Joseph, MI 49085.

Kelso, C. R. 2009. Direct occlusion handling for high level image processing algorithms. Rochester Institute of Technology.

Linarth, A., J. Penne, B. Liu, O. Jesorsky & R. Kompe. 2007. Fast Fusion of Range and Video Sensor Data. In *Advanced Microsystems for Automotive Applications 2007,* eds. J. Valldorf & W. Gessner, 119-134. Springer Berlin Heidelberg.

Lindner, M., M. Lambers & A. Kolb (2008) Sub-pixel data fusion and edge-enhanced distance refinement for 2D/3D images. *Int. J. Intell. Syst. Technol. Appl.,* 5**,** 344-354.

Litwiller, D. 2001. CCD vs CMOS: Facts and fiction. In *Photonics Spectra*. Laurin Publishing Co. Inc.

López-Orozco, J. A. (1999) Integración y fusión multisensorial de robots móviles autónomos. PhD Thesis. Universidad Complutense de Madrid, Spain.

Luo, R. C., Y. Chih-Chen & S. Kuo Lan (2002) Multisensor fusion and integration: approaches, applications, and future research directions. *Sensors Journal, IEEE,* 2**,** 107-119.

MESA Imaging, A. 2011. SwisRanger SR4000.

Microsoft Research. 2009. Meet Kinect for Windows.

Ming-Liang, W., H. Chi-Chang & L. Huei-Yung. 2006. An Intelligent Surveillance System Based on an Omnidirectional Vision Sensor. In *Cybernetics and Intelligent Systems, 2006 IEEE Conference on*, 1-6.

Montes, H., R. Fernández, C. Salinas & M. Armada. 2012. Robotic multisensory system for precision agriculture applications. In *The 15th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines*, eds. A. K. M. Azad, N. J. Cowan, M. O. Tokhi, G. S. Virk & R. D. Eastman, 731-738. World Scientific Publishing Co. Pte. Ltd.

Mucherino, A., P. J. Papajorgji & P. Pardalos. 2009. *Data Mining in Agriculture*.

Namin, S. T. & L. Petersson. 2012. Classification of Materials in Natural Scenes using Multi-Spectral Images. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1393-1398. Vilamoura, Algarve, Portugal: IEEE.

Nene, S. A. & S. K. Nayar. 1998. Stereo with mirrors. In *Computer Vision, 1998. Sixth International Conference on*, 1087-1094.

Noury, N., A. Fleury, P. Rumeau, A. K. Bourke, G. O. Laighin, V. Rialle & J. E. Lundy. 2007. Fall detection - Principles and Methods. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 1663-1666.

Open Source Computer Vsion Library.

Osborne, J. W. & A. Overbay (2004) The power of outliers (and why researchers should always check for them). *Practical assessment, research & evaluation,* 9**,** 1-12.

Park, J., H. Kim, M. S. Brown & I. Kweon (2011) High quality depth map upsampling for 3D-TOF cameras. *2011 International Conference on Computer Vision***,** 1623-1630.

Parrish, E. & A. Goksel (1977) Pictorial Pattern Recognition Applied to Fruit Harvesting. *Trans. ASAE,* 20**,** 822-827.

Plebe, A. & G. Grasso (2001) Localization of spherical fruits for robotic harvesting. *Machine Vision and Applications,* 13**,** 70-79.

Ponticelli, R., E. Garcia, P. Santos & M. Armada (2008) A scanning robotic system for humanitarian de-mining activities. *Industrial Robot: An International Journal,* 35**,** 133-142.

Reulke, R. 2006. Combination of distance data with high resolution images. In *ISPRS Commission V Symposium Image Engineering and Vision Metrology*. Dresden, Germany.

Reynolds, M., J. Doboš, L. Peel, T. Weyrich & G. J. Brostow. 2011. Capturing Time-of-Flight data with confidence. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 945-952.

Ringbeck, T. 2007. A 3D time of flight camera for object detection. In *Conference Optical 3-D Measurement Techniques*

ROS. 2007. Robot Operating System.

Rougier, C., E. Auvinet, J. Rousseau, M. Mignotte & J. Meunier. 2011. Fall detection from depth map video sequences. In *Toward useful services for elderly and people with disabilities*, 121-128. Springer.

Sagüés, C., Murillo, A. C., Escudero, F., & Guerrero, J. J. (2006) From lines to epipoles through planes in two views. *Pattern Recognition,* 39**,** 384-393.

Sahu, D. K. & M. Parsai (2012) Different image fusion techniques–a critical review. *International Journal of Modern Engineering Research (IJMER),* 2**,** 4298-4301.

Salinas, C., R. Fernández, H. Montes & M. Armada. 2011. Omnidirectional stereo system for fall detection and in-house intelligent assistance. In *The 2011 IARP Workshopn on The Role of Robotics in Assisted Living*. Korea.

Salinas, C., R. Fernández, H. Montes & M. Armada (2015) A New Approach for Combining Time-of-Flight and RGB Cameras Based on Depth-Dependent Planar Projective Transformations. *Sensors,* 15**,** 24615-24643.

Salinas, C., H. Montes, G. Fernández, P. Gonzalez de Santos & M. Armada (2012) Catadioptric Panoramic Stereovision for Humanoid Robots. *Robotica,* 30**,** 799-811.

Salvi, J., X. Armangué & J. Batlle (2002) A comparative review of camera calibrating methods with accuracy evaluation. *Pattern Recognition,* 35**,** 1617-1635.

Salvi, J., C. Matabosch, D. Fofi & J. Forest (2007) A review of recent range image registration methods with accuracy evaluation. *Image and Vision Computing,* 25**,** 578-596.

Sansoni, G., M. Trebeschi & F. Docchio (2009) State-of-the-art and applications of 3D imaging sensors in industry, cultural heritage, medicine, and criminal investigation. *Sensors,* 9**,** 568-601.

Sarig, Y. (1990) Robotics of Fruit Harvesting. *Journal of Agricultural Engineering Research* 54**,** 265-280.

Scharstein, D. & R. Szeliski (2002) A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision,* 47**,** 7-42.

Schertz, C. E. & G. K. Brown (1968) Basic considerations in mechanizing citrus harvest. *Transactions of the ASAE,* 11**,** 343-346.

Shaw, G. A. (2003) Spectral imaging for remote sensing. *Lincoln Laboratory Journal,* 14**,** 3-28.

Slaughter, D. & R. C. Harrel (1987) Color vision in robotic fruit harvesting. *Transactions of the ASAE,* 30**,** 1144-1148.

Slaughter, D. & R. C. Harrel (1989) Discriminating fruit for robotic harvest using color in natural outdoor scenes. *Transactions of the ASAE,* 32**,** 757-763.

Song, Y., Glasbey, C. A., Van Der Heijden, G. W. A. M., Polder, G., & Dieleman, J. A. 2011. Combining stereo and time-of-flight images with application to automatic plant phenotyping. In *Lecture Notes in Computer Science* 467-478.

Svoboda, T. & T. Pajdla (2002) Epipolar geometry for central catadioptric cameras. *International Journal of Computer Vision,* 49**,** 23-37.

Tanigaki, K., T. Fujiura, A. Akase & J. Imagawa (2008) Cherry-harvesting robot. *Computer and Electronics in Agriculture,* 63**,** 65-72.

Tippetts, B., D. Lee, K. Lillywhite & J. Archibald (2013) Review of stereo vision algorithms and their suitability for resource-limited systems. *Journal of Real-Time Image Processing***,** 1-21.

Tomasi, C. & R. Manduchi. 1998. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, 839-846.

Torr, P. H. S. & A. Zisserman (2000) MLESAC: a new robust estimator with application to estimating image geometry. *Comput. Vis. Image Underst.,* 78**,** 138-156.

Triggs, B. 1996. Factorization methods for projective structure and motion. In *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on*, 845-851.

Tsai, R. Y. (1987) A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *Robotics and Automation, IEEE Journal of,* 3**,** 323-344.

Van den Bergh, M. & L. Van Gool. 2011. Combining RGB and ToF cameras for real-time 3D hand gesture interaction. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, 66-72. IEEE.

Van Henten, E. J., J. Hemming, B. A. J. van Tuijl, J. G. Kornet, J. Meuleman, J. Bontsema & E. A. van Os (2002) An autonomous robot for harvesting cucumbers in greenhouses. *Autonomous Robots* 13**,** 241-258.

Waske, B. & J. A. Benediktsson (2007) Fusion of Support Vector Machines for Classification of Multisensor Data. *Geoscience and Remote Sensing, IEEE Transactions on,* 45**,** 3858-3866.

Wei, S. & J. R. Cooperstock. 2005. Requirements for Camera Calibration: Must Accuracy Come with a High Price? In *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on*, 356-361.

Weng, J., P. Cohen & M. Herniou (1992) Camera calibration with distortion models and accuracy evaluation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 14**,** 965-980.

Whittaker, D., G. E. Miles, M. O.R. & L. D. Gaultney (1987) Fruit location in a partially occluded image. *Transactions of the ASAE,* 30**,** 591-597.

Wild, D., U. S. Nayak & B. Isaacs (1981) How dangerous are falls in old people at home? *British Medical Journal (Clinical research ed.),* 282**,** 266-268.

Wyawahare, M. V., P. M. Patil & H. K. Abhyankar (2009) Image Registration Techniques : An overview. *International Journal of Signal Processing, Image Processing and Pattern Recognition,* 2**,** 11-28.

Yagi, Y., Y. Nishizawa & M. Yachida (1995) Map-based navigation for a mobile robot with omnidirectional image sensor COPIS. *Robotics and Automation, IEEE Transactions on,* 11**,** 634-648.

Yamazawa, K., Y. Yagi & M. Yachida. 1993. Omnidirectional imaging with hyperboloidal projection. In *Intelligent Robots and Systems '93, IROS '93. Proceedings of the 1993 IEEE/RSJ International Conference on*, 1029-1034 vol.2.

Zambanini, S. & M. K. J. Machajdik. 2010. Computer Vision for an Independent Lifestyle of the Elderly-An Overview of the MuBisA Project. In *AALIANCE Conference – Ambient Assisted Living: Technology and Innovation for Ageing Well*. Malaga, Spain.

Zhang, Z. (2000) A Flexible New Technique for Camera Calibration. *IEEE Trans. Pattern Anal. Mach. Intell.,* 22**,** 1330-1334.

Zhang, Z. & R. S. Blum (2001) A hybrid image registration technique for a digital camera image fusion application. *Information Fusion,* 2**,** 135-149.

Zhu, J., L. Wang, R. Yang & J. Davis. 2008. Fusion of time-of-flight depth and stereo for high accuracy depth maps. In *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*.

Zitová, B. & J. Flusser (2003) Image registration methods: a survey. *Image and Vision Computing,* 21**,** 977-1000.