Vues Multiples non-calibrées : Applications et Méthodologies

Miguel Carrasco


THÈSE DE DOCTORAT DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité
INFORMATIQUE

Présentée par
M. Miguel CARRASCO

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse

NON-CALIBRATED MULTIPLE VIEWS: APPLICATIONS AND METHODOLOGIES

soutenue le 25 Janvier 2010 devant le jury composé de:

Ryad BENOSMAN	Maître de Conférence-HDR, Université Paris 6	Directeur de thèse
Domingo MERY	Professeur, Pontificia Universidad Católica de Chile	Directeur de thèse
Xavier CLADY	Maître de Conférence, Université Paris 6	Co-directeur de thèse
Ryad CHELLALI	Senior Researcher, Istituto Italiano di Tecnologia - Italie	Rapporteur
Luis RUEDA	Professeur, University of Windsor - Canada	Rapporteur
Bruno GAS	Professeur, Université Paris 6	Examinateur
El Mustapha MOUADDIB	Professeur, Université de Picardie Jules Verne	Examinateur
Álvaro SOTO	Professeur, Pontificia Universidad Católica de Chile	Examinateur
Marcelo GUARINI	Professeur, Pontificia Universidad Católica de Chile	Examinateur
Cristian VIAL	Professeur, Pontificia Universidad Católica de Chile	Examinateur

Dedicated to my wife
Fabiola Iturra

ACKNOWLEDGEMENTS

The work of a PhD student is a great challenge. Studying, doing research, inventing, imagining, and developing are some of the main steps in the process of growth as a doctoral student. After reaching the end of this long road and looking backwards, what is visible and permanent are those persons with whom I shared a friendship, long hours of work, successes and failures. An important ingredient in my doctoral work has been the great trust that I have enjoyed from the Pontificia Universidad Católica de Chile and its constant support of all the activities that I undertook, and from the Colegio Doctoral Franco-Chileno and the National Commission for Scientific and Technological Research (CONICYT), who gave me the great opportunity to carry out my work at the ISIR laboratory of the Université Pierre et Marie Curie - Paris 6.

In Chile I want to express my gratitude to Prof. Domingo Mery. His great trust in my work allowed me to succeed with my research over all these years. He certainly provided me with the main tools for carrying out my research. In France, I want to thank the constant support and dedication of Prof. Xavier Clady, with whom I shared most of my work in France. His permanent help, meetings and comments allowed my research to profit from new concepts and ideas. I also thank Prof. Ryad Benosman, who trusted my work in France, allowing me to do my work under co-tutorship at the ISIR laboratory. I also wish to thank my friend and colleague Dr. Luis Pizarro, with whom I shared so many discussions and long hours of revisions, corrections and improvements that are also part of this thesis and so many other papers that we have worked.

To my doctoral friends both in Chile and in France, with whom I shared my first work, subjects and courses, I thank you for all your friendship, support and team work. In Chile, Roberto, Juan Carlos, Esteban, Luis Felipe, Miguel Ángel, Billy, Alberto and Christian, and in France, my lab-mate at CEA, François, who helped me create the human intention videos, and to my friends at the ISIR laboratory: Charlie, Ali, Sahar, Quoc Dinh, Ilaria, Claire and Nino with whom I shared their great friendship, culture and so many conversations. In particular I thank Ammar Mahdhaoui, whose way of looking at life and constant happiness were fundamental in my adaptation to working in Paris. I thank you all very much.

Above all, I thank my wife, Fabiola. Her unconditional support, spirit, constance and strength were the driving force that allowed me to finish this long work, carried out in Chile as well as in France. Thank you so much, Fabiola, without you I would never have reached this important goal of my professional life.

INTERNATIONAL COOPERATION AGREEMENT

This thesis has been carried out under an international cooperation agreement on jointly supervised PhD. between the Computer Science Department of the Pontificia Universidad Católica de Chile and the Institut des Systèmes Intelligents et de Robotique (ISIR) of the Université Pierre et Marie Curie - Paris 6, thanks to the scholarship of the Colegio Doctoral Franco-Chileno, together with the cooperation of the National Commission for Scientific and Technological Research (CONICYT) and the Council of Chilean University Rectors (CRUCH). The joint work of both laboratories has extended the development of this thesis to multiple papers related to the same area: computer vision.

In Chile the focus of the project was the design of computer vision algorithms oriented at automatic quality control with multiple views. We underscore the design and construction of a completely functional prototype that inspects wine bottle necks. By means of an electromechanical system the prototype turns the bottle and gets multiple images of it at each turn. Then the system determines the quality of the object by means of a tracking algorithm with multiple views.

In France the focus of the project was the detection of the motion of human gestures when performing a grasping task. Its main application is to help people with altered motion, such as parkinson's or other diseases, to assist in the motion of grasping with a robotic arm attached physically to the arm of a person (orthosis). This project falls within the frame of the BRAHMA project, which is the merging of multiple automatic systems (electronic, mechanical and computer vision) for human manipulation and assistance in cases of replacement of an arm, or in assisting motion in case of motor diseases. In this project the main contribution of my research is related to the design of a computer vision system that detects and foresees automatically the kind of movement done by the person, such as approaching or getting away from an object, and pointing at the object that the person wants to grasp.

As a result of merging both projects, we highlight the implementation of an algorithm for the analysis of correspondences with multiple geometric solutions through the analysis of invariant correspondences. All the projects are related by the same principle, which consists of the analysis of multiple views as a way of making inferences. Thanks to the work done at both laboratories, this thesis has been furnished with ideas and knowledge that have led to submitting many papers to international journals and national and international conferences.

PAPERS INCLUDED

This thesis is based in the following papers. A full description of each paper is presented in Section 3 to Section 5. Section 3 shows paper 1 to paper 4. Section 4 shows paper 5. Finally, Section 5 shows paper 6.

Paper 1: M. Carrasco and D. Mery: 'Automated Visual Inspection using trifocal analysis in an uncalibrated sequence of images', Materials Evaluation, vol. 64, pp. 900-906, Sept. (2006)

Paper 2: M. Carrasco and D. Mery: 'Automatic Multiple View Inspection using Geometrical Tracking and Feature Analysis in Aluminum Wheels', Journal of Machine Vision and Applications (**Accepted to be published, Feb. 2010**).

Paper 3: M. Carrasco, L. Pizarro and D. Mery: 'Image Acquisition and Automated Inspection of Wine Bottlenecks by Tracking in Multiple Views', in Proceedings of 8th International Conference on Signal Processing, Computational Geometry and Artificial Vision - ISCGAV 08, Rhodes Island, Greece, Aug. 20-22, (2008). pp. 82-89.

Paper 4: M. Carrasco, L. Pizarrro and D. Mery: 'Visual Inspection of Glass Bottlenecks by Multiple-View Analysis', International Journal of Computer Integrated Manufacturing, (**Second Revision, Feb. 2010.**)

Paper 5: M. Carrasco, D. Mery: 'On solving the point-to-point correspondence problem using multiple geometrical solutions', **Submitted to** Pattern Recognition Letters, Sept. 2009.

Paper 6: M. Carrasco, X. Clady: 'Prediction of user's intentions
based on the eye-hand coordination', **Submitted to** IEEE/ASME Transactions on Mechatronics, Focused Section on Healthcare Mechatronics, Jun. 2009. (*extended version)

OTHER PUBLICATIONS

The following papers have not been included in this thesis. Most part of the work presented in this thesis is based on previous ideas presented on the following papers.

ISI 1: L. Pizarro, D. Mery, R. Delpiano, and M. Carrasco: 'Robust Automated Multiple View Inspection', Journal of Pattern Analysis and Applications, vol. 11(1), pp.22-32. (2008).

ISI 2: D. Mery and M. Carrasco: 'Advances on Automated Multiple View Inspection', Lecture Notes in Computer Science (LNCS), vol. 4319, pp. 513-522, Sept. (2006).

ISI 3: D. Mery and M. Carrasco: 'Automated multiple view inspection based on uncalibrated image sequence', Lecture Notes in Computer Science (LNCS), vol. 3540, pp. 1238-1247, Jun. (2005).

ISI 4: M. Carrasco and D. Mery: 'Segmentation of welding discontinuities using a robust algorithm', Materials Evaluation, vol. 62, pp. 1142-1147, Dec. (2004).

INT 1: M. Carrasco, and X. Clady: 'Utilisation de la coordination main-oeil pour la prediction d'un geste de prehension'. In Proceedings of the 17ème conférence en Reconnaissance des Formes et Intelligence Artificielle (RFIA). Caen, France, Jan 19-22 (2010).

INT 2: M. Carrasco, X. Clady: 'Prediction of user's intention based on the hand motion recognition', in Proceedings of the XXVIII International Conference of the Chilean Computer Society (SCCC), IEEE Computer Society. Santiago, Chile, Nov. 9-14. (2009).

INT 3: M. Carrasco and D. Mery: 'Automatic Multiple Visual Inspection on Non-Calibrated Image Sequence with Intermediate Classifier Block', in Proceedings of Pacific-Rim Symposium on Image and Video Technology (PSIVT), Santiago, Chile, Dec 17-19, (2007), **Best Paper Award on Track Computer Vision and Applications**.

INT 4: D. Mery and M. Carrasco: 'Automated Multiple View Inspection of Metal Castings', IEEE/SPIE, 8th International Conference on Quality Control by Artificial Vision (QCAV), Le Creusot, France, May 23-25, (2007).

CHN 1: M. Carrasco and X. Clady: 'Predicción de la intención humana basado en la coordinación Mano-Ojo', in 1er congreso de estudiantes de Postgrado Ingeniería (CEPIUC), Pontificia Universidad Católica de Chile, Chile, May, (2009).

CHN 2: M. Carrasco and D. Mery: 'Inspección visual automática usando análisis trifocal en una secuencia de imágenes no calibradas', in XII Encuentro Chileno de Computación (ECC). Universidad Austral de Chile, Chile, Nov, (2005).

RÉSUMÉ

La recherche de modèles d'intérêt contenus dans des séquences de vues multiples d'une scène reste l'un des principaux problèmes de la vision par ordinateur actuellement. En dépit des grands progrès observés au cours des 40 dernières années, la vision par ordinateur ne parvient pas encore à répondre adéquatement quant à la manière d'inférer et de détecter des modèles d'intérêt dans des scènes pour lesquelles un ou plusieurs objet(s) sont vus depuis différents points de vue. Afin de surmonter ce probléme, cette thèse propose de nouveaux algorithmes et prototypes capables de caractériser, d'inférer et de détecter des modèles d'intérêt en séquences avec des vues multiples de manière non calibrée, c'est-à-dire sans connaissance à priori de la position du/des objet(s) par rapport ła (aux) caméra(s). Le travail réalisé s'articule autour de trois axes, divisés en six articles qui constituent le corps de la thèse. (1) L'analyse de correspondances point par point à travers de marqueurs explicites et implicites sur les objets. (2) L'estimation de correspondances point par point à travers de multiples relations géométriques indépendantes du/des objet(s) qui composent la scène. (3) La prédiction du flux dynamique du déplacement généré par le mouvement de la caméra autour de l'objet. L'objectif principal de cette thèse est d'appuyer la prise de décision à travers d'une analyse dynamique et/ou géométrique du mouvement du/des objet(s) ou de la (des) caméra(s) pendant que ceux-ci se déplacent. Grâce à cette analyse, il est possible d'accroître l'information sur la scène et l'(les) objet(s) à travers d'un processus inférenciel spécifique pour chaque cas. Il ressort des thématiques exposées qu'il est possible, par exemple, d'assister le processus d'inspection réalisé par un opérateur humain, de déterminer la qualité d'un produit de manière autonome, ou d'exécuter une action spécifique dans un acteur robotique. Bien que ces thématiques présentent des approches différentes, celles-ci ont le même ensemble de pas en ce qui concerne: (1) la détermination de la relation de correspondance de points ou de régions sur plusieurs images, (2) la détermination de la relation géométrique et/ou dynamique existante entre les correspondances estimées précédemment, (3) l'inférence de nouvelles relations sur les points dont la correspondance est inconnue en vue de caractériser le mouvement. Les variations les plus fortes correspondent à la manière dont la correspondance est estimée; au calcul de la dynamique et la géométrie entre les points correspondants; et enfin à la manière dont nous inférons une action particulière suite à un mouvement spécifique. Parmi les principaux résultats, on trouve le développement d'une méthodologie d'inspection non calibrée à vues multiples appliquée à l'analyse de la qualité des jantes de véhicules, le développement d'un prototype fonctionnel appliqué à l'inspection des cols de bouteilles de vin, une méthodologie de correspondance point par point géométrique capable de résoudre le problème de correspondance en deux et trois vues pour tout point d'intérêt, et enfin la reconnaissance de l'intention humaine pour les tâches de *'grasping'* à travers de l'analyse du mouvement des yeux et de la main. À l'avenir, il restera encore à analyser les correspondances dynamiques à travers de caractéristiques invariantes, employer des méthodes d'analyse géométriques en séquences d'images radiologiques, et utiliser des modèles de détection d'intentions pour évaluer la qualité des objets.

Mots clés: vision par ordinateur, géométrie de vues multiples, analyse de correspondances, reconnaissance de gestes, inspection visuelle automatique, tracking.

ABSTRACT

The search for patterns of interest contained in multiple view sequences of the same scene remains one of the major problems in the computer vision community at present. Despite the great advances developed in the last 40 years, computer vision has not yet answered adequately how to infer and detect patterns of interest in scenes in which one or more objects are viewed from multiple points-of-view. To overcome this drawback, this thesis proposes new algorithms and prototypes capable of characterizing, inferring, and detecting patterns of interest in uncalibrated multiple view sequences, i.e. without a priori knowledge about the position of the object(s) in regard to the camera(s). The work is focused on three main themes divided into six papers that constitute the body of the thesis. (1) The point-to-point correspondence analysis by means of explicit and implicit markers on objects. (2) The point-to-point correspondence estimation through multiple geometric relations independent of the object(s). (3) The dynamic flow prediction generated by camera's motion around the object. The main purpose of this thesis is to support the decision making by means of the dynamic and/or geometric analysis of the object(s) movement or the camera(s) movement as these move. Thanks to this analysis, it is possible to increase the information about the scene and the object(s) through an inferential process specific in each case. As a result, it is possible for instance to assist the inspection process performed by a human operator, to determine the product quality automatically, or to execute a specific action in a robotic actuator. Although these issues have different purposes, these resolve the same set of steps concerning to: (1) Determining corresponding points in multiple images. (2) Determining the geometric and/or dynamic relation between the previously estimated correspondences. (3) Inferring new relations at those points whose correspondence is not known in order to characterize the movement. The major changes come in how the correspondence is estimated, in the calculation of the dynamics and geometry between corresponding points, and finally in how we use the inference of motion to identify a new correspondence or to characterize the motion of the points. Among the major results, the development of a new uncalibrated inspection methodology with multiple views applied to the quality analysis in automotive wheel rims, the development of a functional prototype applied to the wine bottleneck inspection, a geometric point-to-point correspondence methodology able to solve the correspondence problem in two and three views for any point of interest, and finally the human intention recognition for grasping tasks through the movement analysis of the eyes, and the hand. As future work, it remains to study the dynamic correspondence by means of invariant features, to use geometric methods in radiologic image sequences, and use intention recognition models in order to evaluate the object-quality.

Keywords: computer vision, multiple view geometry, correspondence problem, motion planing, gesture recognition, automatic visual inspection, tracking.

RESUMEN

La búsqueda de patrones de interés contenidos en secuencias de múltiples vistas de una escena continúa siendo uno de los principales problemas de la visión por computador actual. A pesar de los grandes avances desarrollados en los últimos 40 años, la visión por computador no ha logrado aun responder adecuadamente cómo inferir y detectar patrones de interés en escenas en las que uno o varios objetos son vistos desde distintos puntos de vista. Para superar este problema, esta tesis propone nuevos algoritmos y prototipos capaces de caracterizar, inferir y detectar patrones de interés en secuencias con múltiples vistas en forma no calibrada, es decir, sin conocimiento a priori de la posición de el/los objeto(s) respecto a la(s) cámara(s). El trabajo desarrollado está centrado en tres temáticas divididas en seis artículos que constituyen el cuerpo de la tesis. (1) El análisis de correspondencias punto a punto por medio de marcadores explícitos e implícitos en los objetos. (2) La estimación de correspondencias punto-a-punto a través de múltiples relaciones geométricas independientes de el/los objeto(s) que componen la escena. (3) La predicción del flujo dinámico del movimiento generado por el desplazamiento de la cámara en torno al objeto. El objetivo principal de esta tesis es apoyar la toma de decisión a través de un análisis dinámico y/o geométrico del movimiento de el/los objeto(s) o de la(s) cámaras a medida que estos se desplazan. Gracias a este análisis es posible incrementar la información sobre la escena y de el/los objeto(s) a través de un proceso inferencial específico en cada caso. Como resultado de las temáticas expuestas, es posible por ejemplo asistir el proceso de inspección realizado por un operador humano, determinar la calidad de un producto en forma autónoma, o ejecutar una acción específica en un actuador robótico. Aunque estas temáticas posean enfoques distintos, éstas resuelven el mismo conjunto de pasos concernientes a (1) determinar la relación de correspondencia de puntos o regiones en múltiples imágenes, (2) determinar la relación geométrica y/o dinámica existente entre las correspondencias estimadas anteriormente, (3) inferir nuevas relaciones en aquellos puntos en los cuales no es conocida su correspondencia con el fin de caracterizar el movimiento. Las mayores variaciones provienen en la forma en cómo es estimada la correspondencia; en el cálculo de la dinámica y la geometría entre los puntos correspondientes; y finalmente en cómo inferimos una acción particular como resultado de un movimiento específico. Dentro de los principales resultados, se encuentra el desarrollo de una metodología de inspección no calibrada con múltiples vistas aplicada al análisis de calidad en llantas de vehículos, el desarrollo de un prototipo funcional aplicado a la inspección de cuellos de botellas de vino, una metodología de correspondencia punto a punto geométrica capaz de resolver el problema de correspondencia en dos y tres vistas para cualquier punto de interés y finalmente el reconocimiento de la intención humana para tareas de grasping a través del análisis del movimiento de los ojos y de la mano. Como trabajo futuro resta por analizar las correspondencias dinámicas a través de características invariantes, emplear métodos de análisis geométricos en secuencias de imágenes radiológicas, y utilizar modelos de detección de intenciones para evaluar la calidad de los objetos.

Palabras Claves: visión por computador, geometría de múltiples vistas, análisis de correspondencias, reconocimiento de gestos, inspección visual automática, tracking.

TABLE OF CONTENTS

Chapter 1

Introduction

1. INTRODUCTION

Most living beings of the animal kingdom have diverse and complex perception systems such as hearing, taste, smell, touch and sight. From a general viewpoint, each system consists of specialized receptors that inform the central nervous system of the current state of its environment at all times. Although all the systems are relevant and the lack of some of them would greatly degrade the quality of life, the most highly developed at the neurological level is the visual perception system (Luck, Girelli, McDermott, & Ford, 1997). Depending on the environment of each species, the visual perception has different degrees of specialization. Variations of the range of visible light, the luminosity level, the degree of focusing, among others, are possible evolutions of the visual system of each species (Land & Fernald, 1992). In human beings, the visual perception is responsible for letting us perceive, interpret, store, and recreate the reality of our environment at all times.

In a simplified way, vision can be divided into two stages: (1) acquisition of information, and (2) processing the visual information. First, the acquisition of information is related to how we capture the visual information. The organ responsible for this task is the eye. The eye is a complex multilayered organ that allows capturing the information from the visual environment, varying in terms of sensitivity, resolution and color among the different species (Land & Fernald, 1992). Also, by converting the light changes into electric impulses it reduces the large amount of visual information. This process of converting light into electric impulses at the neuronal level is generated in a light sensitive membrane called the retina (Jessell, Schwartz, & Kandel, 2000; Humayun et al., 1999). Second, after a constant gathering of information from the eyes, we process the visual information in multiple regions of the brain, particularly in the cerebral cortex, the dorsal and ventral stream. In the brain the visual information acquires meaning and in that way we can provide a representation of the sense of sight (Jessell et al., 2000). In general, human vision is understood as the conjunction of various neurological processes interacting with one another to give sense to the visual information. In this way we can locate ourselves in the space that surrounds us, detect the motion of an object at a distance, determine the speed at which an object moves, or do something as usual as holding an object in our hands, among other activities, e.g., (Crawford, Medendorp, & Marotta, 2004; Brouwer & Knill, 2007; Johansson, Westling, Bäckström, & Flanagan, 2001; Mrotek & Soechting, 2007). Both processes, acquisition and processing, must be understood as a synergic process (Hayhoe, Shrivastava, Mruczek, & Pelz, 2003).

The discussion of the visual perception is extensive and profound. Even though in the ancient Greece they already began formulating the first models of the functioning of human vision, its best understanding was brought about at the beginning of the 11th century by the Iraki scientist Ibn al-Haytham in his book *Book of optics* (Sabra, 1989; Tbakhi & Amr, 2007). Various kinds of theories have been developed for centuries, from nativism, which states that we act in an intuitive and innate way, to empirism, which is based on the accumulation of the experience and learning, and finally the *Gestalt* theory which indicates the actions carried out by the brain as a function of memory, previous states, and present vision. Based on this theory, vision is a holistic process that cannot be separated

and analyzed independently (Sternberg, 2003). The sense of sight is certainly one of the main senses of human beings. It allows us to create a projection of our body and environment, generating a representation of the real world to coordinate our actions and in particular the motion of our limbs. Various diseases associated with sight can greatly deteriorate our quality of life. Daily activities such as interacting with our environment, reading, writing, eating, grabbing objects, playing instruments, practicing sports, among others, can be seriously deteriorated due to problems derived from poor eyesight (Wu, Hennis, Nemesure, & Leske, 2008; Nirmalan et al., 2005).

From its beginnings, the study of human vision has been one of the fields that have attracted the greatest interest and led to more research among the scientific community due to the great complexity and importance that vision has to humans. Understanding, deciphering and emulating the various mechanisms that influence the acquisition of information, as well as their representation at the cognitive and neurobiological level still remain as relevant fields of research. In particular, neuroscience is the area that has contributed most to the understanding of vision and of the different factors and mechanisms that take part in the process of the visual perception. Although not yet completely defined and determined, neuroscience has succeeded in determining some of the most important mechanisms that participate in this process, such as the interrelation between the position of our body and limbs in relation to what we see thanks to proprioception, as well as the influence of vision on the grasping motion (Hayhoe, Bensinger, & Ballard, 1998; Donkelaar, Lee, & Drew, 2000; Brouwer & Knill, 2007).

As is natural in the development of science and engineering, emulating the characteristics of human vision has been a permanent challenge, but the results continue to be unfruitful. Let us recall that human vision is composed of a versatile optical system with the added cognitive power of the brain. In this way we can interpret, associate and distinguish visual information rapidly without a priori knowledge of the actions and objects that take part in the scene. It is clearly seen that current technology is far from building systems that reproduce effectively human abilities in terms of visual perception. However, we have seen how in the last 40 years there have been great advances along this line. In fact, there are areas in which technology greatly surpasses the acquisition ability of human vision, e.g., in the range of visible light, focusing distance, thermal vision, night vision, X-rays, among others, but there still remains much work to be done in terms of understanding visual information.

The first contributions to computer vision were aimed at emulating human abilities, a task that has not been solved yet. At present the paradigm is solving problems restricted and defined to solve specific and tedious tasks in a fast, efficient and constant manner, making use of the huge processing capacity of current computers. One of the factors that led to better understanding of this problem was the construction of the first artificial vision or computer vision system that allowed the acquisition of the visual information. The first steps taken in that direction were generated in the 1960s thanks to the development of the space program at the Jet Propulsion Laboratory (JPL) together with the MIT laboratories and Bell Labs in the USA (Rosenfeld, 1969). As a result of that, the development of two products marked the increasing use of digital processing: (1) the invention

of computerized axial tomography (CAT) at EMI Central Research Laboratories, UK; and (2) the invention of CCD technology by the end of the 1970s at Bell Labs. This made it possible to convert the visual information on the objects, either internal or external, into digitized (or electronic) information. This transformation has great advantages because it allows processing the information in computational form, thereby increasing analytical ability and its later support of decision making. Taking advantage of this technology, many branches of science and industry have continued extending the development of these systems, generating a huge variety of applications (Gonzalez & Woods, 2008). Also, thanks to the increasing development and miniaturization of electronic devices as well as the increased computational speed of current systems, it is possible to find autonomous and fast complex artificial systems that are capable of solving problems in real time and with high reliability. For these reasons, commercial applications and products in areas as diverse as industry, the military, biology, physics, medicine, engineering, and safety continue expanding, leading to substantial progress in research on computer vision.

Present day computer vision encompasses many fields of research and is undergoing constant development. The main reasons correspond to the ability to distinguish, detect and reveal patterns of interest from the visual information. Precisely these are the reasons that have the highest complexity, i.e., how to understand the relevant patterns automatically from the visual information. Naturally, technological progress has notably increased the information analysis process, but we are still far from finding a solution to this problem, so research and development are still open. On the other hand, in spite of the large variability of available applications, they are specific to the area in which they are inserted. In this respect there is no single definition that indicates the way in which the problems must be solved through computer vision techniques. Consequently, we will say that computer vision attempts to answer some of the following questions: What do we see in the scene? How do we highlight the relevant information? Where do we look for that information, i.e., provide mechanisms that extract from the visual information features of interest to support decision making. It is desirable to carry out this process in real time, or at least in a period shorter than human processing. In some cases the main requisite is that this process should be performed automatically; however, many of the current systems operate in an assisted way, allowing a human operator to determine the final result of the process.

In general, most of the applications of computer vision are designed to solve specific and well defined problems. An application designed for a problem can hardly be applied directly to another without a change in its design. Some examples of these problems are inspection, recognition, follow-up and modeling of objects, as well as detection, interpretation and human-computer interaction, and finally organization and classification of the visual information. The relevance of these topics is of increasing interest to the computer vision community. This is reflected in the themes presented at the recent International Conference on Computer Vision (ICCV'09) conference, of which we highlight applications in vigilance, human-computer interaction, learning based on the analysis of motion, alignment of objects, 3D recognition of objects, search for objects in video sequences, and multiple camera analysis, among others. Compared to other branches of the science, the computer

vision field is still in its beginnings. As previously mentioned, at present there are very high-level commercial solutions, but in general most of them are under development, and they are limited to solving specific tasks. The challenge is for the applications to be 100% automatic, but there is still much work needed to achieve that goal, and only in limited cases it has been possible to obtain it.

As we have argued, the development of algorithms and commercial systems based on computer vision techniques are still growing and undergoing constant research. This is a strong motivation for the development of this thesis, whose main objective is the development of new algorithms in the computer vision field that will allow the determination of an action as a function of the visual information acquired. In brief, the main contribution of this thesis is the development of algorithms that can infer relevant information on the motion of objects of a scene in multiple views with the purpose of supporting or assisting decision making by either a human, a computer or a robotic agent. Regardless of the intermediate steps to achieve that objective, the results can be: to determine the quality of an object, to determine the geometric relation of an object, or to infer the type of motion of the sequence. In this respect, the common denominator in all applications developed is the analysis in uncalibrated multiple views, i.e., when the position of the objects with respect to the camera is not known. Therefore, the central problem is reduced to describing the dynamics of the motion of the objects or regions as they are displaced with respect to the scene. In general, this procedure is complex because it is subject to multiple variations, mainly geometric or photometric transformations and/or occlusions present in multiple views that limit the formulation of dynamic models stable in time.

At present there is no single method to determine the dynamics of the motion of the objects that compose the scene for any type of motion. In general, there is a set of particular solutions that work better depending on each type of motion. This thesis, in addition to presenting solutions to the specific problems in computer vision, introduces new methods that give rise to the development of new applications both in the industrial field and in human-computer interaction. The possibility of extending the development to other domains using as a basis the methods presented here is open.

From the standpoint of the applications introduced, we highlight those oriented at inspection and quality control, point-to-point correspondence, as well as the development of a new method for inferring human intentions. This wide range of applications has allowed the study and analysis of multiple areas of the computer science, generating a flow of knowledge between the different applications developed. Furthermore, the use of image processing, pattern recognition, and probabilistic methods constitutes an important part in the development of the solutions presented. Below we define the main themes that make up the central body of this thesis:

A. Automatic visual inspection: Developing a geometric tracking system in multiple views with automatic visual inspection applications to automobile wheel rims and wine bottle necks (Chapter 3).

B. Point-to-point correspondence: Geometric correspondence in two and three views through multiple geometric solutions with applications of multiple image types (Chapter 4).

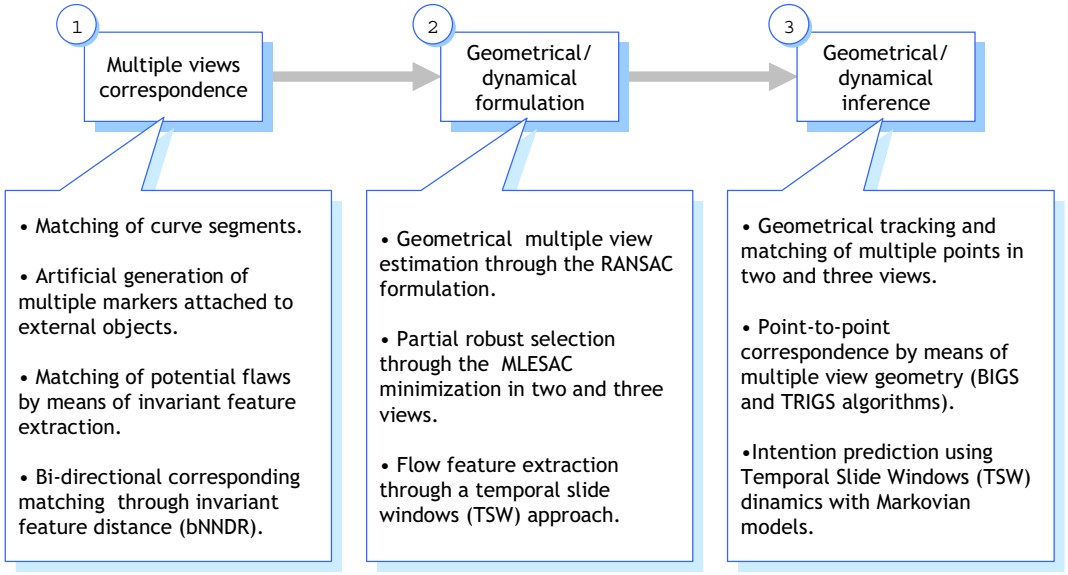


FIGURE 1.1. Main algorithms developed transversely to the development of each of the three themes.

C. Prediction of user's intentions: Inference of the user's intention through a grasping motion by means of a multiple camera system (Chapter 5).

It is evident that these three areas differ in their application domains, but all of them require solving the same steps. (1) Determining corresponding points in multiple images. (2) Determining the geometric and/or dynamic relation between the previously estimated correspondences. (3) Inferring new relations at those points whose correspondence is not known in order to characterize the movement. The steps described above make up the core of all the articles shown, with variations in how the correspondence is estimated, in the calculation of the dynamics and geometry between corresponding points, and finally in how we use the inference of motion to identify a new correspondence or to characterize the motion of the points. An overall review of the implemented algorithms is presented in Fig. 1.1.

A key part in the development of the applications implemented throughout this thesis has been the reuse and adaptation of the different algorithms implemented, even when they have been designed specifically for each domain. This process has allowed extending and improving the solutions proposed initially. For example, the geometric algorithm for inspection in multiple views of Chapter 3 made possible the development of a general point-to-point correspondence algorithm in multiple views in Chapter 4, as well as the extension to the wine bottle inspection system detailed in Chapter 3. In the same way, invariant correspondence method of the human-computer interaction system of Chapter 5 led to the development of a new bidirectional correspondence algorithm detailed in Chapter 4. More examples like these can be found through all the applications presented,

which implies that regardless of the objectives set for each solution, the algorithms that have been developed are transverse and can be adaptable to other domains.

The organization of this document includes the presentation of six articles divided into three central themes. As pointed out earlier, the articles that constitute this thesis develop a similar formulation, corresponding to steps 1, 2 and 3 of Fig 1.1. Thanks to this diversity it has been possible to restate existing solutions, leading to improvements of their initial versions. In what follows we will give a general description of the problem and we will then review in detail each of the stated themes, their hypotheses, objectives, and methodologies, to end with the organization of the document

1.1. Problem description

This section introduces the three main themes that constitute the thesis. As already mentioned, multiple view analysis is the common denominator of all the applications developed. To carry out that analysis the first step is to determine the multiple view correspondences. The correspondence analysis requires solving two steps: First, determining a set of points in an image such that they are identified as the same in other images of the same scene, and second, discarding those correspondences that are false alarms, i.e., eliminating the incorrectly related correspondences. The first step, described in Fig. 1.2, presents the ideal situation. It shows nine correspondent points in three images of the same object that keep their relative positions in each view. In this way each point is reflected in the following images in spite of the existing variations in viewing points, focal distance, and photometric changes. The second step, described in Fig. 1.3, presents the normal situation. It is seen that determining the correspondence of point $\mathbf{r}$ is not a simple procedure because the appearance of correspondences similar to point $\mathbf{r}$ in the following views increases the complexity of the algorithm for determining a unique solution. Solving this problem is extremely important in computer vision because it represents the first step for solving multiple problems such as 3D reconstruction, robotic navigation, tracking in multiple views, estimation of transformation matrices, and homography, among others.

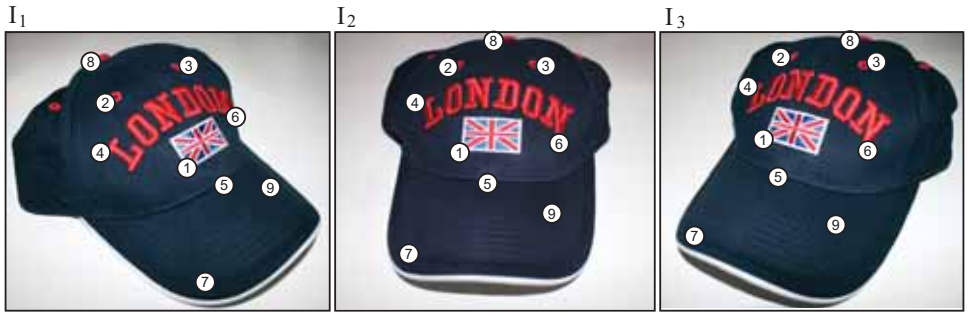


FIGURE 1.2. Corresponding points in the $\mathbf{I}_1$, $\mathbf{I}_2$ and $\mathbf{I}_3$ views; ideal situation.

To humans this problem represents no difficulty because of our enormous ability to relate patterns; however, determining computationally those correspondences is not a simple task. We must consider that the correspondences are subjected to diverse transformations depending on the viewpoints from which they have been captured. This is due to the geometric and photometric transformations caused by the motion of the object as well as by the camera's motion with respect to the object. It is also possible for other points to have a similar texture and color as the point whose correspondence we want to determine. This increases the complexity of the system to discriminate between the set of correct correspondences and the set of incorrect correspondences over all the existing combinations. In some situations it is possible that a corresponding point is not reflected because it is occluded. In those cases there is no photometric method that can determine the correspondence, and there is only a geometric solution.

As we will see in the following chapters, this thesis does not include a single method of analysis of correspondence; each problem has been solved based on the views used and taking into account the methods developed previously. In this sense the applications were implemented by an incremental process. First the correspondences were modeled as an optimization problem that had to determine the best alignment of a curve over another one through their bending and rotation with the purpose of carrying out a geometric tracking of defects in automobile wheel rims (Chapter 3). Then the use was made of correspondence by means of markers in multiple views through a photometric analysis to perform a geometric tracking of defects in wine bottlenecks (Chapter 3). The point-to-point correspondence was then modeled through the combination of multiple geometric solutions which allowed the generalization of the problem of tracking in multiple views (Chapter 4). Finally, the vectorial field of the motion through multiple correspondences in time was determined. These correspondences were estimated by the analysis of invariant features with the aim of determining the user's intention in human-computer interaction problems (Chapter 5). As can be seen, methods that were initially implemented in a restrictive way, finally allowed the development of more general, less restrictive methods regardless of the objects contained in them.

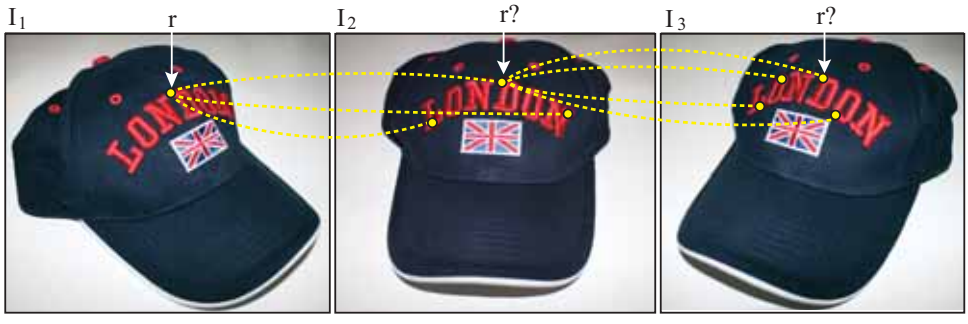


FIGURE 1.3. Which is a corresponding point of point $\mathbf{r}$ in the views $\mathbf{I}_2$ and $\mathbf{I}_3$?

We will now discuss briefly each of the problems of the three main themes that make up the thesis: *A. Tracking in multiple views* (Chapter 3), *B. Point-to-point correspondence* (Chapter 4) and *C. Human-computer interaction* (Chapter 5).

1.1.1. Automatic visual inspection

One of the first problems of computer vision is following regions or objects in time. This problem is known as tracking and its main objective is to determine and predict automatically the position of a region in time, regardless of the kind of motion generated. An effective tracking algorithm should be capable of remaining invariant with respect to photometric and geometric changes generated during the trajectory of the region, and furthermore it must be robust with respect to possible occlusions. Clearly, tracking tries to emulate the ability of human beings to predict the motion of an object due to changes in its position, luminosity, perspective, etc.

Since the 1970s various tracking algorithms have been developed extensively. Currently there is a large variety of algorithms capable of solving the tracking problem regardless of the kind of motion of the object. Restrictive models that consider the motion based on a given distribution, through general models that consider the motion stochastically, have made possible the development of multiple applications in both science and industry (see Yilmaz, Javed, and Shah (2006) for a complete review). The first solutions considered the differences between sequences of images modeling the motion through optical flow (Lucas & Kanade, 1981; Barron, Fleet, & Beauchemin, 1994). Other more general solutions solve this problem by analyzing first the problem of points in correspondence and then formulating the tracking in a geometric way (Scharstein & Szeliski, 2002). In other cases it is necessary to determine a transformation of the edges of a region to detect an object's complete position (Yilmaz, Li, & Shah, 2004). Other algorithms simply follow similar characteristics in sequential images in regions neighboring those where the detection took place (Shafique & Shah, 2003; Veenman, Reinders, & Backer, 2001). Finally, more general applications model the motion of the objects considering that they have a stochastic trajectory. The latter use a probabilistic modeling based on a particle filter or through Kalman's filter (Arulampalam, Maskell, Gordon, & Clapp, 2002).

Tracking algorithms are very relevant in all those applications in which it is necessary to predict and determine the movement, position, and direction of an object or a region contained in it, i.e., to determine the object's trajectory. This is exactly the case of industrial control applications that require inspecting objects physically in all directions. Tasks such as rotating, turning and/or transferring objects are procedures that are commonly made by a human operator, a process in which a tracking application can be inserted to support an inspector's decision and in that way increase the quality of the product. An important part of the work done for this thesis is focused on the development of a general tracking system that can be applied as a quality control method. Specifically, a method is presented to evaluate the quality of automobile wheel rims and wine bottles by means of an uncalibrated geometric tracking algorithm. In this case the tracking allows following potential defects in multiple views to evaluate the quality of the inspected object. If a potential defect can be

detected in their relative positions in multiple views, the object is classified as defective, otherwise it is classified as free of defects.

In Chapter 3 we will discuss in detail two applications that deal with the problem of uncalibrated tracking. In particular, we will differentiate the uncalibrated tracking from the calibrated method. The problem is that the calibrated method requires determining with precision the position of 3D points in space with the purpose of calculating the 3D transfer functions to points on the 2D image, a process called calibration (Zhang, 2000; Beardsley, Murray, & Zisserman, 1992). An example of a calibrated automatic visual inspection was solved in (Mery & Filbert, 2002), where the authors introduce the Automatic Multiple View Inspection (AMVI) method, which determines the quality of an object in a calibrated way, using specifically automobile wheel rims as object for inspection. In spite of the advantages and improvements in existing calibration processes, the vibrations inherent to industrial processes induce changes in the position of the cameras, which implies that the transfer functions must be re-estimated periodically. To avoid the problems associated with calibration, we introduce the uncalibrated AMVI methodology.

Among the main contributions in this field we point out the geometric modeling of the tracking problem through the estimation of bifocal and trifocal geometry by means of the uncalibrated AMVI algorithm. Thanks to this method it was possible to determine the quality of an object with multiple views. On the other hand, we present an algorithm for the analysis of the properties of each defect in multiple views, a process that we call Intermediate Classifier Block (ICB), that allows a reduction of the load of the AMVI system to make the inferences. Furthermore, we present a real and functional prototype that allows the inspection of wine bottlenecks. The main advantage of our prototype is that the object to be inspected does not require marks to identify its corresponding points because they are contained in the rotating object. The bottle is turned together with the rotating object, allowing the AMVI algorithm to determine the quality of the bottle. In the four articles that make up Chapter 3 we show the flexibility of the AMVI method to be used in the quality control of completely different products.

1.1.2. Point-to-point correspondence

As mentioned at the beginning of the section, determining the point-to-point correspondence in multiple images is a complex problem due to the various geometric and photometric transformations and/or occlusions that the same point can experience in the different views. Different approaches have been used over the last 40 years to solve this problem. Some of them are, for example, methods based on the analysis of invariant descriptors (Bay, Ess, Tuytelaars, & Gool, 2008; Lowe, 2004; Bosch, Zisserman, & X., 2007), estimation of affine transformations, homographies and estimation of perspective transformations (Caspi & Irani, 2000; Fitzgibbon, 2003), epipolar geometry analysis (Romano, 2002; Vidal, Ma, Soatto, & Sastry, 2006), and methods based on optical flow (Lucas & Kanade, 1981). In general, all these methods differ in the kind of motion of the objects contained in the sequence. If the scene is static and there is no continuous change in the position of the cameras, the problem is reduced mainly to the analysis of the epipolar geometry of two images by stereo

vision (Scharstein & Szeliski, 2002). On the other hand, if the scene is dynamic and the objects have small displacements, differential techniques by optical flow have been shown to be an effective way for determining the correspondences (Barron et al., 1994). Unfortunately, optical flow methods are unable to determine correspondences when displacements are extensive.

In spite of the large number of methods designed to solve this problem, correspondence in images with wide angles of vision has not been solved completely. This last problem is normally found when use is made of two or more independent cameras located in different positions and with wide perspective angles (Caspi, Simakov, & Irani, 2006). On the other hand, this problem can be seen in a monocular way with multiple views when an inter-frames correspondence is used (Carrasco & Mery, 2006). In those cases the position of the objects and the camera vary in time, making the objects appear in very different positions in each view. In the latter it is common to use techniques based on the feature analysis through the extraction of invariant descriptors (Bhat et al., 2006).

Thanks to invariance it is possible to solve and generalize the problem of point-to-point correspondence providing an extension to methods based on stereo vision. This correspondence is made as a function of the points of interest detected previously by some algorithm for the detection of regions of interest (Moreels & Perona, 2007). However, when the point of interest does not correspond to a point detected by current saliency techniques (Kadir, Zisserman, & Brady, 2004; Matas, Chum, Urban, & Pajdla., 2002; Mikolajczyk & Schmid, 2004; Tuytelaars & Mikolajczyk, 2007), how can we determine its corresponding pair in other images? In this case, previous methods do not ensure finding a correct correspondence because they are designed to maximize its performance only in the regions of interest detected by the method and not necessarily in our regions of interest. This problem appears commonly when the images have a low signal-to-noise ratio, thus invariant algorithms will not have a good performance due to the appearance of many incorrect correspondences or false alarms.

Among the main contributions in this matter we point out a new algorithm for determining point-to-point correspondences in multiple views even in those cases in which the displacement angles are wide. Our proposal builds a model based on invariant correspondences followed by the formulation of the point-to-point correspondence. Moreover, since we use a geometric model that is independent of the objects, it is possible to determine the position of corresponding point in those views in which correspondence can be occluded. In relation to the above, this chapter extends the geometric model of correspondence, even allowing point-to-point correspondence only with the geometry in two views. In this way it is possible to carry out a geometric tracking with two views.

1.1.3. Prediction of user's intentions

Human beings are very skillful at reaching and grasping objects with their hands under multiple conditions, even when faced with variable positions, location, structure and orientation of the objects. This natural skill controlled by the brain is called eye-hand coordination. Normally the grasping motion takes place some time before the hand reaches the object. This process is regulated

by the interaction of multiple systems such as the vestibular system and proprioception working together with the control systems of eye, head and hand (Crawford et al., 2004). A large part of this activity occurs in different regions of the brain, especially in the cortical and subcortical regions, with great importance in cognitive processes such as attention and memory. According to Flanagan and Lederman (2001), when we grasp an object the information of what is perceived by our sensors is the result of the preconceived ideas of the shape of the object and an interpretation of what is perceived; this means that our brain uses memory and visual information simultaneously. Researchers in many areas have studied this process for many years, trying to explain the brain mechanisms that control this coordination. However, so far there is no single theory that effectively explains this coordination, and furthermore it is not completely understood (Hayhoe et al., 1998; Brouwer & Knill, 2007).

The main contribution in this area has been limited to the external analysis of human movement. Our proposal is to build a model that uses visual information of the person considering its own visual field. The objective of the model is to detect the movement's flow when performing a grasping task. To that end we need to identify each kind of movement made and indicate the object that the person wants to grasp. Why is it necessary to determine this type of movement? People suffering from neurodegenerative diseases, with motor problems and limitation of movement are greatly hindered for performing grasping tasks. In those cases motion control is altered, causing tremor, slowness, imprecision, etc. Even though visual functions may not be affected, the control system is unable to plan the motion in a normal manner.

This problem has been approached with the development of artificial orthosis designed to assist human movement by means of robotic mechanisms. Examples of these systems are the MIT-manus (Krebs, Hogan, Aisen, & Volpe, 1998) models, pioneers in the development of assistance to people, and the orthosis designs proposed by Kiguchi and Fukuda (2004), Gupta and O'Malley (2006), Sugar et al. (2007), Perry and Rosen (2007), as well as the prediction model proposed by Jarrasse, Paik, Pasqui, and Morel (2008). For the orthoses to be functional and interactive, we propose to characterize human movements by allowing the orthosis to offer assistance only when they are detected, particularly when the grasping motion is detected.

Among the main contributions in this matter we point out an algorithm that infers the user's intention from the perspective of the person, capturing the same scene that the person visualizes. In this way the interpretation of the captured information is the scene and not the body. Our work is designed in contrast with the classical methods of recognition of the body movement and its extremities, because we use the flow of visual information coming from the motion of the cameras toward the object. On the contrary, most methods capture the motion of the body located in front of it. In Chapter 5 we detail the proposed vectorial model to infer the flow of motion in grasping tasks through a Markovian system. In relation to the two previous themes, the vectorial model is a modified version of the geometric analysis in multiple views together with an analysis of invariant correspondences. In this way we only need to extract the properties among the vectors in multiple views.

1.2. Hypothesis

Computer vision applications that use multiple views of the same object are supported by the same principle: whenever corresponding points are determined in multiple views there will be a geometric or dynamic formulation that allows the points of that sequence to be related. This formulation can be described either by equations that determine the three-dimensional geometry of the object by projection on the two-dimensional planes; by the two-dimensional relation between them through epipolar geometry; by estimating the transformation matrices of perspective or homographies; by analysis of invariant features; or by optical flow. Thanks to the multiple view correspondence we can determine the motion and position relations for both the object and the camera.

In general, the existence of a set of points in correspondence in uncalibrated multiple views makes it possible to describe the geometric and dynamic relation that exists between object(s) and camera(s) around the scene. This dynamics is valid if the object is fixed and the cameras are moving, the object is in motion and the cameras are fixed, or combinations of both. Through this model it is possible to infer the movement of the objects or the points of interest, or to describe the complete movement of the scene. If the correspondence is imprecise and corresponding points are concentrated in a limited area, i.e., there is a low dispersion of the position of the points with respect to the scene, then the resultant model will be limited in its ability to infer the correspondence relation at new points. To reduce this error, we use robust selection algorithms that try to determine the best subset of corresponding points that decrease the model's general error. Unfortunately, it is not possible to assure that the model that has been determined is the best. This interrelation constitutes the basis for the formulation of the following general hypothesis:

> *The use of multiple subsets of corresponding points will increase the estimation of the predictive model by using multiple partial models. In this way, to make new inferences on the geometry or dynamics existing between the objects and the cameras it is necessary to use multiple models (each with a corresponding error) which compensate the estimation of the error of the general model. That dynamics is valid in calibrated as well as uncalibrated multiple views. Regardless of the application in which the search process is inserted, the result will support decision making with respect to the action that must be performed.*

With the purpose of increasing the understanding in each domain, in this section we specify each particular hypothesis. Below we will detail three specific hypotheses in relation to each problem in order to detail and specify each developed area. Although the statement of these hypotheses is related to the specific problem that is being solved, it is important to point out that they are all part of the general statement, because the general hypothesis brings together and develops all the applications that make up this thesis, namely to determine the dynamics of the scene with the purpose of supporting decision making.

I. In relation to the system of uncalibrated inspection in multiple views we propose the following hypothesis:

> *The uncalibrated scheme can be applied to automatic visual inspection of products in which there are no clear control points for detection, since they can be generated artificially through external markers that adopt the motion of the object that will be analyzed, or also through modeling of regions of interest of the object such that they allow the determination of a point-to-point correspondence.*

II. In relation to the system of point-to-point correspondence in multiple views with two or three uncalibrated views we propose the following hypothesis:

> *The generation of multiple geometric solutions, some with errors close to the optimum, allows an increase of the performance of the geometric correspondence model. We assume that the intersection of multiple epipolar lines obtained from the best estimations of the epipolar geometry will allow the determination of that correspondence. In the case of three views, we extrapolate the same idea to trifocal geometry.*

III. In relation to the inference of human intentions in gasping tasks we propose the following hypothesis:

> *The grasping motion can be modeled from the flow of visual information captured from the user's perspective. To make that inference we assume that the grasping motion has a unique pattern with respect to other types of human movements, allowing them to be distinguished. In the case of identifying the object, we assume that the multiplicity of information allows distinguishing more precisely the object that the user wants to grasp.*

1.3. Objectives

This section presents the general and specific objectives that motivate to the work developed in this thesis. First, we present the general objective that is related to the transverse topics of all works that make up the body of the thesis. Then, we detail the specific objectives of each of the three main topics divided based on their applications: (A) automatic visual inspection, (B) point-to-point correspondence, and (C) prediction of user's intentions. As we will show below, the chapters that make up this thesis solve the general objective based on each domain and at the same time provide a solution to the general objective of the thesis. This distinction is necessary to provide greater clarity to each topic and to the common objective in all the applications.

1.3.1. General objective

To develop new methodologies and algorithms in the area of computer vision that can be applied to uncalibrated multiple view sequences to support decision making as a means of assistance, allowing the inference of actions, movements and states. Those views can be composed of one or multiple cameras with or without information in correspondence. As a result, the algorithms must be implemented in real prototypes, simulation algorithms, and/or electronic systems. In relation to decision making, it can consist in either supporting the work done by a human inspector, determining the quality of a product automatically, or executing a specific action on a robotic actuator as a function of the result of the algorithm.

A. Automatic visual inspection

   During the last four decades the development of different methods of automatic visual inspection has been a fundamental part of the increased quality and efficiency of most manufactured products at the world level. Recently the introduction of a new methodology called AMVI has allowed the development of new automatic inspection algorithms using multiple views of the same object. Using this potential the general objective in this field is the development of an uncalibrated automatic visual inspection system that can be applied on line with the product to be inspected to objects in which there are no points in correspondence that may be generated artificially.

B. Point-to-point correspondence

   Point-to-point correspondence methods continue to be important problems in the computer vision community because there is still no single way of solving this problem. Thanks to the geometric formulation in multiple views, point-to-point correspondence can be modeled by means of multiple partial geometric solutions. Based on this idea, the general objective is to determine the point-to-point correspondence between two and three images in correspondence by means of a geometric model that must weight the error of each partial solution in such a way that the point-to-point correspondence is determined for any point, even when faced with changes in perspective, photometric, and possible occlusions.

C. Prediction of user's intentions

   Human grasping movements are characterized by their speed and acceleration during the execution of the movement. Furthermore, it has been shown that there is a relation between hand's movements and the sight position in relation to the object. Considering these two properties, the general objective is to detect and characterize different hand's movements when a user is performing a grasping task by merging visual information flow coming from cameras above the head, under the hand, and an eye-tracker. Also, the correlation of information between the views will allow determining the object that a user wants to grasp.

1.3.2. Specific objectives

The specific objectives presented are transverse to the developed applications. As stated in the introduction, all the applications that make up the thesis must solve the same three steps even though they are aimed at specific problems in different areas. Next, we list the specific objectives of the thesis, and then the specific objectives of each application.

- Investigate methods of correspondence in multiple uncalibrated views.
- Investigate and design robust selection correspondence algorithms.
- Investigate and design algorithms that allow getting information on the dynamics and the geometry of the object.
- Implement and build a real prototype for inspection with multiple uncalibrated views.
- Investigate and implement new computer vision techniques oriented at automatic visual inspection.
- Investigate and implement algorithms that allow inferring new relations from the determined motion models.
- Investigate and implement human-computer interaction algorithms that allow making inferences on the actions of human motion.

A. Automatic visual inspection

- Investigate new segmentation algorithms, feature extraction, feature selection, and classification of potential defects based on new image analysis techniques developed recently that have not been used for automatic visual inspection.
- Design an object with markers that will allow an effective estimation of the position of different control points.
- Design an efficient algorithm to detect the changes of the control points of the markers.
- Design a system for inspection with multiple views that detects defects using the object's rotational motion.
- Design an algorithm to determine the correspondence of potential defects.
- Implement an algorithm to filter efficiently false alarms and real defects in multiple views.

B. Point-to-point correspondence

- Review the literature on the correspondence methods through the analysis of invariant features.
- Investigate random solution selection methods.
- Implement an algorithm to determine the best combination of multiple partial solutions.

- Implement a metric to evaluate the partial error of each random solution to re-estimate the partial error of that solution.

C. Prediction of user's intentions

- Review the literature on methods of detection and recognition of human intentions and the eye-hand coordination for grasping tasks.
- Implement an algorithm to identify hand's intentions in the grasping gesture.
- Implement an algorithm to identify the object that a user wants to grasp as a function of the information contained in an object dictionary.
- Implement an algorithm that merges the information from the three channels for the acquisition of visual information, specifically from cameras above the user's head, under the user's wrist, and an eye-tracker device.

1.4. Methodology

This section presents the methodology required to solve the topics proposed in this thesis. For greater clarity, corresponding papers in which each of the proposed methods has been implemented are indicated. In all the algorithms presented we have used a modified version of the Knowledge Discovery Databases (KDD) method, consisting of the following steps: Data acquisition and selection, pre-processing, transformation, data mining, and evaluation (Mitra, Pal, & Mitra, 2002). Our modification is the replacement of the data mining step by multiple views analysis. All the algorithms are developed on MATLAB, C-MEXmulti-camera, Basic Stamp, and Visual C++.

1. **Acquisition and selection:** In relation to the automatic visual inspection system, acquisition and selection require the design and construction of a prototype multi-camera rotation that considers the following aspects: a) a digital camera on a fixed support; b) a lighting system adaptable to the motion of the bottle; c) a bottle rotation system; d) main computer; e) communication driver between the computer and the camera. The camera will be controlled by a computer through a communication driver. The rotating base will be controlled through a mechanical-electric system with a step motor, and with an interval regulator that allows the interspersed capture of images after each rotation, programmed through Basic Stamp (Martin, 2005). In our experiments we will use wine bottles with and without defects provided by Cristaleras Chile S.A. (Results in Paper #3, Paper #4, Paper #5).

   In relation to the system for the detection of user's grasping intentions we consider the following aspects: a) microcamera under the hand; b) camera above the user's head calibrated with the eye-tracker; c) an eye-tracker system that determines the sight position as a function of the second camera; d) an image acquisition system in real time. In our

experiments we will use multiple test objects combined with multiple user's movements (Results in Paper #6).

2. **Preprocessing:** Once the images have been obtained, they will be processed to remove noise, improve contrast, and determine the relevant zones of the study (Gonzalez & Woods, 2008). Furthermore, the segmentation process will be performed using techniques that are generally not used in automatic visual inspection, such as analysis of the defect profile with the Crossing Line Profile algorithm (Mery, 2003a), and robust segmentation threshold selection (Hui-Fuang, 2006). (Some of them in Paper #3, Paper #4, Paper #5, Paper #6).

3. **Transformation:** All the regions relevant to the problem are transformed into the feature space. To that end we must determine the geometric and intensity features. The geometric features determine the spatial and geometric position of the segmented regions, e.g., area, perimeter, roundness, Fourier descriptors, invariant moments, orientation, radius of the major and minor axes of the ellipse, etc. The intensity characteristics are related to the information on the color and intensity of each pixel, e.g., average gray, derivative average, contrast deviation, difference between maximum and minimum intensity level, Hu moments with gray values (Hu, 1962) (Results in Paper #1, Paper #2). A study of invariant features will be made using the following methods: SFSK (Flusser & Suk, 1993), FSKS (Flusser, Suk, & Saic, 1996), CLP (Mery, 2003a) GPSO-PSO-GPD (Mindru, Tuytelaars, Van Gool, & Moons, 2004), SURF (Bay et al., 2008), SIFT (Lowe, 2004), PHOG (Bosch et al., 2007). (Some in Paper #2 Paper #3, Paper #4, Paper #5, Paper #6).

4. **Classification:** In relation to the automatic visual inspection system, the classification requires avoiding the introduction of correlated features, for which we will use a feature selection algorithm through the efficient Branch & Bound (Somol, Pudil, & Kittler, 2004), Sequential Forward Selection (SFS) (Jain, Duin, & Mao, 2000) and Take-L, plus-R (Duda, Hart, , & Stork, 2001; Kudo & Sklansky, 2000) algorithms The separation will be made through Fisher's Discriminant (Stearns, 1976). (Results in Paper #2).
   In relation to the intention detection system, the classification system considers modeling the movement as a variation of local features. Therefore an action described in time as a variation of points in multiple views will be analyzed through Hidden Markov Models (HMM) (Rabiner, 1989). Some techniques proposed by Yamato, Ohya, and Ishii (1992), Starner and Pentland (1995) and Achard, Qu, Mokhber, and Milgram (2007) must be analyzed to determine the best features that define a grasping movement. (Results in Paper #6).

5. **Tracking:** The uncalibrated algorithm requires knowing control points estimated correctly at each view. One of the usual forms will consist in choosing a robust subset of points by means of the RANSAC algorithm (Fischler & Bolles, 1981) or MLESAC (Torr

& Zisserman, 2000) in two and three views. In this way, as the object rotates, the system automatically estimates new positions of valid control points. (Some in Paper #1, Paper #2, Paper #3, Paper #4, Paper #5).

6. **Evaluation:** With respect to the automatic visual inspection system, to evaluate the performance of the uncalibrated AMVI system and the detection-of-intentions system it is necessary to use the cross-validation method proposed in (Mitchell, 1997). In this way, performance will be measured objectively with data that the system does not know at the time of the test. (Results in Paper #1, Paper #2, Paper #3, Paper #4, Paper #5, Paper #6).

1.5. Contributions

This section presents the main contributions of each paper in relation to each topic. The six papers that represent the body of the thesis are divided into three topics. The first topic, Chapter 3, is focused on automatic visual inspection; specifically on tracking in multiple views of automobile wheel rims and wine bottles necks. The second topic, which corresponds to Chapter 4 presents a method of geometric correspondences that is independent of the images used. And finally the third topic, which corresponds to Chapter 5, presents an algorithm for the detection of human intentions that uses multiple sources of visual information, mainly two cameras that observe the scene from the user's perspective in an active form.

1.5.1. Automatic visual inspection

The first topic deals with the automatic inspection of automobile wheel rims in radiographic images and the inspection of wine bottle necks by multiple view analysis. For the inspection of wheel rims we present two papers that contain the first advances of the tracking algorithm with multiple views. For the wine bottle neck inspection we present two papers that show the design of a functional prototype that inspects, through multiple views, different types of bottle necks. By means of an electromechanical system controlled by a computer the bottle is rotated, illuminated internally, and analyzed. This research uses the same previous concepts, and even more so, studies in greater depth the influence of invariant features as a correspondence method; further information in Chapter 3.

Paper 1: This research presents the development of a new flaw detection algorithm in manufactured goods, using an uncalibrated sequence of images. Using the AMVI methodology proposed by Mery and Filbert (2002) we have designed a novel system of automatic calibration based only on the spatial positions of the structures. The proposed approach uses the projection of the epipolar line, generated by the fundamental matrix and the trifocal tensors in a robust manner with the purpose of building a motion model without any a priori knowledge of structure. With respect to the investigation carried out by Mery and

Carrasco (2005), we have extended the analysis from two to three images per sequence through the estimation of trifocal tensors. Furthermore, we have introduced new control points generated artificially through the use of B-Spline curves due to the low number of structures that remain stable in three images of a sequence.

Paper 2: This research introduces the calculation of corresponding points generated artificially through the maximization of the correlation coefficient from two curves. To improve the performance, we designed a false alarm reduction method in two and three views called Intermediate Classifier Block (ICB). The ICB method takes advantage of the classifier ensemble methodology by making use of a feature analysis in multiple views. Using this method, real flaws can be detected with high precision and at the same time most false alarms can be discriminated. The method was tested in a sequence of X-ray images of aluminum wheel rims, but the methodology can be used for other sequences as well by changing the segmentation and control point algorithms, as we showed for a bottle inspection system with multiple views in (Carrasco, Pizarro, & Mery, 2008).

Paper 3: This research presents two main contributions: First, we present a prototype design of an electromechanical device for acquiring image sequences of wine bottle necks using a single camera. Its main novelty is the placement of the illuminating source inside the bottle, which greatly improves the definition of the inspected images. Second, we introduce a new methodology for detecting flaws in the bottle neck based on tracking potential defects along an image sequence. Our inspection system achieves performance rates of 87% true positives and 0% false positives.

Paper 4: This research examines series of two and three images employing multiple view geometry followed by a feature analysis stage to discriminate between real flaws and false alarms. In this way, we classify as real flaws those that present similar features in a set of images taken from different viewpoints. An important ingredient to achieve this goal was the introduction of a novel feature analysis criterion to resolve multiple geometric matches in different views. It can be considered as a bidirectional variant of the *nearest neighbor distance ratio* (NNDR) criterion proposed by Mikolajczyk and Schmid (2005). Our inspection system, tested on image sequences of wine glass bottles with real flaws, obtained a *true positive rate* of 99.1% and a *false positive rate* of 0.9%.

1.5.2. Point-to-point correspondence

As a generalization of the above methodologies, we propose a correspondence method using the same principles of multiple views geometry, extending the error estimation model detailed in Chapter 4.

Paper 5: This paper presents two important contributions. First we present a geometric method that uses multiple partial solutions close to the optimum to determine correspondences in two and three views. Second, for each geometric model we determine the real distance with respect to a corresponding point by means of the MLESAC estimator (Torr & Zisserman, 2000), in this way weighting the error associated with each intermediate solution. The main novelty of our proposal is the geometric methodology for solving the estimation problem of point-to-point correspondences, regardless of the viewpoints of the objects contained in the images and of the point of interest used in each image. It is important to highlight that the point can be occluded in the following views, but its position continues to be valid because our method is based on a geometric model. We will also show that the use of multiple random solutions allows an increase of the correspondence performance.

1.5.3. Prediction of user's intentions

The third topic presents an algorithm to infer user's intentions. The proposed solution required the formulation of equations that characterize hands movements through multiple correspondences in time. Through a predictive model it was possible to determine precisely different types of movements that a user makes when beginning a grasping movement. Furthermore, the algorithm identifies the object that the user wants to grasp through an object dictionary. It is important to point out that this topic influenced the development of the algorithm presented in Chapter 4. Further information on this topic is found in Chapter 5.

Paper 6: The main contribution of this work lies in our clever choice of using human vision combined with an active vision as a means to predict the user's intentions. In our experiments we show that it is possible to detect the hand's intentions using only the objects contained in the scene, and without special markers on the objects' surface. However, the performance of this task varies based on the points of interest detected by the SURF algorithm. Similarly, our method can predict the user's grasping intention, and identify the object placed in the scene by merging two channels of information. Even if the object has been occluded, the system is able to identify it because our approach uses a combination of frames called Temporal Slide Windows (TSW). This approach can allow us to increase the temporal features of the same object in multiple frames. Consequently the paradigm of the frame-by-frame tracking can be effectively replaced by a TSW approach. Although in our experiments the objects analyzed were limited, the results are very promising because we used a limited number of resilient features and the task was conducted by searching a match between both views.

1.6. Document Organization

This section presents the general organization of the thesis. The present work is divided into six chapters. Chapters 1 and 2 include the introduction, theoretical description, and background of the general problem. Chapters 3, 4 and 5 contain the main body of the thesis and the papers that implement the objectives stated earlier. The organization and content of each chapter are described in what follows.

Chapter 2: This chapter introduces the theoretical foundations of the topics of all the work done for the thesis. The objective is to present to the reader the general problem of the methods based on each domain, a discussion of current methods, and indications of how they can be solved.

Chapter 3: The first topic presents four papers related to automatic tracking in multiple views of (1) defects in automobile wheel rims, and (2) defects in wine bottle necks. The first paper is an improvement of a paper presented by the same authors in (Mery & Carrasco, 2005). The second paper introduces the partial elimination algorithm called Intermediate Classifier Block (ICB). This idea allows a greater reduction of false alarms in sequences. The second part extends the multiple views analysis of wine bottle necks. The third paper presents an inspection prototype together with the design of internal illumination using the multiple views methodology of the first and second papers. The fourth paper introduces an improvement of the tracking algorithms through the design of a new bifocal and trifocal correspondence algorithm called bNNDR (bidirectional Nearest Neighbor Distance Ratio) applied to the method of inspection of wine bottle necks.

Chapter 4: The second topic extends the multiple views methodology through multiple solutions. In this chapter, which corresponds to the fifth paper, we extend the methodology to the analysis of correspondence in two and three images with a geometric error reduction formulation.

Chapter 5: The third topic presents a human intention prediction algorithm corresponding to the sixth paper. In this chapter we introduce the general problem of gesture recognition and then we present an algorithm designed to detect grasping intentions, differentiating multiple grasping gestures.

Chapter 6: Finally, the last chapter presents the conclusions related to the implemented algorithms, the work achieved in relation to the specific objectives, and future work on the proposed systems.

Chapter 2

Background

2. BACKGROUND

This chapter introduces the theoretical background that make up the thesis. First we will introduce a brief summary of automatic visual inspection, beginning with manual methods of inspection, then automatic visual inspection, and finally a description of the multiple views methodology. Second, in relation to the topic of human-computer interaction we will introduce the current methods for recognizing gestures and the main paradigms on which they are formulated.

2.1. Visual inspection

In general, most systems are designed specifically for each product depending on the stage of the production process in which quality control is inserted. In spite of the benefits of automation, the most widely used inspection process is still human. The various factors that influence the application of an automatic or manual system are presented. Finally we detail briefly inspection systems with multiple views.

2.1.1. Human inspection

Currently, most inspection and quality control processes in manufacturing industry are done manually, i.e., a human operator carries out the inspection of each object or of a random sample at some stage of the production process (batch inspection) (Newman & Jain, 1995). One of the main reasons for using this kind of inspection is economic. The investment cost to install and develop a specialized machine for inspection tasks is very high compared to the cost of training a human operator. Also, human visual inspection has the great advantage of adapting to unforeseen situations, and is flexible when faced with any change in the objects' position, orientation and shape. This is because human beings have a high cognitive and sensory ability that allows them to carry out complex reasoning and inferences at the time of inspecting the objects (Spencer, 1996). Essentially, human operators use all that ability together with their vision system to detect discontinuities in the objects, even when faced with differences in shape, size, color, depth, brightness, contrast, and/or texture.

Various studies have analyzed the performance of human inspection and its main defects (Drury, 1992; Mital, Govindaraju, & Subramani, 1998; Jacob, Raina, Regunath, Subramanian, & Gramopadhye, 2004; Drury, Saran, & Schultz, 2004). According to them, there is a clear consensus that human inspection does not achieve 100% performance in the detection of defect-free products (error-free). Mital et al. (1998) determined various factors that affect the performance of manual inspections, such as the rhythm and complexity of the task, the time for inspection, fault density, inspection model, luminosity, inspection strategy, training, age, and gender. According to LeBeau (1991), human visual inspection can achieve a maximum of 90% effectiveness; provided it is implemented on a structured inspection system. Other authors have indicated that it has a maximum of 80% effectiveness (Drury et al., 2004). Unfortunately, human inspectors are not always consistent evaluators

because inspection processes require a high rate of constant concentration in time. For that reason some of the largest failures of human inspection are that it is (1) variable, inspection quality is not constant over time because it is dependent on fatigue and monotony caused by the work; (2) irregular, because it depends on the ability, experience and strategy for revision of each inspector; (3) slow, some industries have high production levels and require inspection at a high processing rate, however human inspection can require more time because handling and observation tasks have as limiting factor the speed of human operations; (4) tedious, because the inspection routine can be very repetitive, and that generates a lower concentration level due to the large number of objects that must be revised in a short period; (5) hazardous, because in some environments such as under water inspection, the nuclear industry, and the chemical industry, human inspection can be inviable due to the high risk inherent in those systems; (6) complex, the difference between a product with or without defects can be very subtle, and that is not always easily distinguishable by a human operator; (7) inaccessible, in some cases even access to the object to be inspected can be very complex because of the size of the product.

These factors have made industry gradually replace human inspection by automatic visual inspection (AVI) methods, in that way allowing inspection to be made without any kind of contact with the object to be inspected. As we will see below, the introduction of automatic systems has allowed most of the failures mentioned before to be overcome, even though the implemented solutions have been designed specifically for each object.

2.1.2. Automatic visual inspection

Since the end of the 1970s the first AVI systems started being implemented in the manufacturing industry with the purpose of having a fast and efficient system for inspection and quality control tasks (Jarvis, 1980; Chin & C.A., 1982; Newman & Jain, 1995). The main objective of AVI is to increase productivity ensuring high quality, reliability and consistency standards, i.e., rejecting most of the defective products and accepting all the defect-free products, usually in a shorter time than inspection made by a human operator. Malamas, Petrakis, and Zervakis (2003) and Kumar (2008) have presented extensive reviews of various AVI technologies applied to the manufacturing processes of different products such as electronic components, textiles, glass, mechanical parts, integrated circuits (IC), etc. In general, most of the existing automatic systems are designed specifically for each object, or some part of it, and their main restriction is the position of the object to be inspected. Clearly, a system designed for inspecting a product can hardly be applied to another without requiring a change in the image acquisition process or in the decision making algorithm, depending on the type of inspection that it is desired to make. However, there is consensus that the use of AVI technologies can reduce significantly the cost and time incurred in the inspection process, allowing the replacement of a large number of not well trained inspectors by a single automatic inspection system, or complementarily by the combined work of highly trained operators together with an AVI

system (Mital et al., 1998). It is important to point out that in view of the impossibility of making a general inspection system, AVI systems will continue to be designed to solve visual inspection problems specific for each industry.

Presently, the main disadvantage of AVI systems is that they do not consider information redundancy as a means of increasing performance in the decision making stage, particularly in systems that require inspecting defects on surfaces or in others in which the visual inspection process is sequential. In those cases the segmentation and classification algorithms must be robust for decision making, i.e, accepting or rejecting the product. However, this decision can be more robust if the inspection has additional information on the product; for example, using multiple views with information in correspondence. In this way it would be possible to discard the false alarms that would be found in other views and to detect with higher probability only the real defects instead of concentrating only on designing a single robust segmentation and classification method.

2.1.3. Automatic multiple view inspection

As we have seen, the visual inspection process in manufactured environments is complex and specialized. It requires extracting a significant amount of information from each object based on its physical properties and structural characteristics to then make a decision, of acceptance or rejection, based on some criterion or specification imposed by the manufacturer or by some regulatory agency. This process uses complex reasoning methods, generally human, but also automated. Although in recent years progress in the technologies has allowed an increase in the processing of information and substantial improvement of the optical systems for carrying out the visual inspection process automatically, one of the greatest restrictions continues to be the lack of flexibility to inspect objects in complex positions. One of the current ways of increasing the flexibility of the inspection process is simply to capture more images of the test object from different viewpoints, thereby generating a redundancy of information on the object that is inspected.

Taking into account the advantages and disadvantages of human visual inspection and of current AVI systems, in recent years the Automatic Multiple View Inspection methodology (AMVI) (Mery & Filbert, 2002; Carrasco & Mery, 2006; Pizarro, Mery, Delpiano, & Carrasco, 2008; Carrasco et al., 2008) has been implemented successfully. The main objective of AMVI is to use information redundancy as a means to increase decision making performance. AMVI uses multiple images of the same object captured from different viewpoints to determine the quality of the object by a process called tracking (Yilmaz et al., 2006). In this way if in the defect detection stage a large number of false alarms is generated using a single image –regardless of the segmentation process used–, the use of more views in correspondence produces the opposite effect, i.e., a significant reduction of the number of false alarms. This results in an increase of the final performance of the inspection process.

The main analogy of the AMVI method comes from the process that a human operator would usually carry out. Assuming a sequential inspection process, the AMVI process consists of two stages:

First, an operator identifies all the potential defects in a sequence of images composed of both false alarms and real defects. Second, the operator tracks in all the views only those defects that actually appear in all the views of the sequence. Since the false alarms are normally found in random positions, they do not appear in their relative positions in the other views of the sequence. On the other hand, real defects can always be seen in their relative positions in the image sequence, unless they are occluded. A human inspector may easily distinguish by means of this process the real defects and the false alarms of the set of potential defects. In this way, if there is at least one real defect in the sequence, the object will be classified as *defective*, and on the contrary, if there are no real defects in sequence and only false alarms have been detected, then the object will be classified as *free of defects*.

The AMVI methodology is based on the analysis of multiple view geometry (Hartley & Zisserman, 2000) and has been implemented in calibrated (Mery & Filbert, 2002) and uncalibrated (Mery & Carrasco, 2005; Carrasco & Mery, 2006; Pizarro et al., 2008; Carrasco et al., 2008) sequences. Both methodologies have been used for quality control on X-ray images and on CCD images of different types of wine bottle necks, showing a clear potentiality and flexibility to make it extensible to other types of products. The main advantages and disadvantages of both methodologies are presented below. (1) The Calibrated Methodology allows to know the 3D geometry of an object through a process called calibration (Zhang, 2000). Knowing the 3D geometry, the 2D position of any defect in any view can be determined univocally because it is represented by a transformation function that relates a 3D point with a 2D point of a two-dimensional image. However, one of the great disadvantages of the calibrated systems resides precisely in the stability of the calibration. Generally the calibration is not stable and many times it is necessary to carry out a re-calibration process periodically because of the inherent vibrations that exist in industrial processes, which modify the parameters of the transfer function. (2) The Uncalibrated Methodology allows performing the tracking process without a 3D knowledge of the object because it does not carry out the calibration process. The greatest advantage of the uncalibrated systems is the better adaptation of the inspection process since it is independent of the objects, and it is only necessary to establish correspondences in multiple views to generate the geometric projection model (Hartley & Zisserman, 2000). These correspondences can be implicit (Carrasco & Mery, 2006; Pizarro et al., 2008) or explicit (Carrasco et al., 2008) to the object. However, the greatest disadvantage of the uncalibrated systems is the lack of precision to establish the correspondences in multiple views, which results in a lower performance of the tracking system in two or more views. In spite of the lack of precision, various studies have shown that the use of multiple views generates a real benefit for the inspection process, because it allows confirming and improving the diagnosis compared to systems that use a single image (Gumustekin, 2004; Mery & Carrasco, 2006; Kita, Highnam, & Brady, 2001; Spicer, Bohl, Abramovich, & Barhak, 2006).

The main foundation of AMVI is the fact that only real defects (and not false alarms) can be seen along the sequence of images because they remain stable in their position relative to the object's motion. The same idea has been used by radiologists, who use two or more X-ray views to detect

breast cancer in its early stages (Kita et al., 2001). Other systems, such as that proposed by Spicer et al. (2006) have allowed capturing different parts of the same object and later combining them in a single image through a system called reconfigurable array for machine inspection (RAMVI). In contrast with our work, Spicer et al. (2006) designed a calibration process through a calibration grid with colors to establish the correspondences. On the other hand, Gumustekin (2004) proposed a system of reconstruction of multiple images of an object generating a single 2D image. For that he used a single camera together with multiple mirrors simulating the use of multiple cameras. The same as in the previous system, a calibration process was designed to find the camera's parameters.

At present, AMVI is a useful tool and a powerful alternative to examine complex objects. It has two independent approaches: those based on the calibration of a transfer function $3D \rightarrow 2D$ in the multiple views projection (Mery & Filbert, 2002), and those based on the estimation of the movement of the control points in correspondences of pairs (Mery & Carrasco, 2005) and triplets of views (Carrasco & Mery, 2006; Pizarro et al., 2008; Carrasco et al., 2008) without prior calibration. A brief description of each is given below.

Calibrated method

Mery and Filbert (2002) developed the first calibrated tracking method of defects in sequences of images as a quality control method in aluminum wheel rims. This approach consists of the estimation of the $3D \rightarrow 2D$ model by an off-line process called calibration (Mery, 2003b), which is the process that allows the determination of the parameters of the model to establish the projection matrix of a 3D point of the object on a 2D point of the digital image. Using this model, the multifocal tensors (Hartley & Zisserman, 2000) can be calculated to estimate the correspondence restrictions between the potential defects along the image sequence.

The calibration process must be determined for each image of the sequence by a procedure known as photogrammetric calibration (Zhang, 2000), in which points in the 3D space of the object are known with precision. Unfortunately, the parameters of the model are usually nonlinear, which implies that the optimization problem does not have a closed solution, and therefore it is necessary to have an initial reference to start iterating, and this is very sensitive to the initial value. Also, although the performance of the projection through calibration is ideal in sequences of four views, it is finally impracticable in industrial environments in which there are vibrations and random movements that are not considered in the original transfer function, i.e., the calibration is not stable and the computer vision system must be calibrated periodically in order to avoid this error. This implies a clear increase of the time and costs needed for the inspection (Mery & Carrasco, 2006).

Uncalibrated method

To overcome the problems that exist in the calibrated method, a new automatic visual inspection system was developed initially in (Mery & Carrasco, 2005) using a sequence of uncalibrated images

in two views. The uncalibrated method does not require estimating or performing some calibration process. On the contrary, it allows the estimation of the model of the movement using the correspondences between the images of the sequence; a procedure that can be carried out on-line with the computer vision system. An improvement of this method was made by the same authors in (Carrasco & Mery, 2006) with the inclusion of a third view, because a greater number of images in sequence increases the performance of the AMVI. In this research a new strategy was designed to find control points in three views by means of B-Spline curves (Bartels, Beatty, & Barsky, 1998).

In general, to achieve high precision in the model of the movement it is necessary to determine a large number of correspondences of control points in pairs and triplets of images in sequence. Many times this condition is hard to fulfill, and for that reason in (Carrasco & Mery, 2006) we used the RANSAC algorithm (Fischler & Bolles, 1981). In general, the uncalibrated methodology has turned out to be effective for detecting most real defects in sequence, but there is a large number of false alarms that it has not been possible to eliminate through that process. Also, the difficulty in finding structural points in correspondence limits the generation of an estimation of the model of the movement when the images of the object are not significantly different from the images contained in the sequences, as, for example, in the inspection of bottles, where all the images of the sequence are completely similar. To solve the latter problem we propose the use of markers external to the object that fulfill its rotational motion, allowing the determination of the geometric model in multiple views (Carrasco et al., 2008).

2.2. Human intention recognition

Human activity recognition is currently becoming more and more important. In the last years a wide variety of applications have been developed, such as athletic performance analysis, surveillance, man-machine interfaces, entertainment systems, video conferences. Along this line, Human Computer Interaction (HCI) is one of them, highly extended, especially due to interest in understanding and recognizing human behavior, as well as in designing computer interfaces that are more usable and receptive of the user's needs. This section shows the main approaches to detect human gestures using computer vision methods. Next, we briefly introduce the main paradigms for motion detection.

2.2.1. Approaches on human gestures recognition

For decades, the computer vision community has played an important role in understanding human motion by providing new ways to distinguish several human gestures. Generally, most applications have been designed to recognize some parts of the human body in order to obtain better representations of them, such as arms, lips, hands, legs, face, and body movements or combinations of them. For that reason, much effort has been made to understand human motion in an overall sense, e.g., (Bobick & Davis, 2001; Kim, Kwak, & Ch, 2006; Shechtman & Irani, 2005); interpreting emotions by means of face movements (Cowie et al., 2001; Busso et al., 2004); or using the body language, e.g. (Achard et al., 2007; Yamato et al., 1992). In general, the study of human motion by computer vision methods can be divided in three main approaches: *Passive*, *Active* and *Pointer* paradigms relative to the position of the camera around the user.

Passive approach

In this approach the camera is located in a fixed position, normally in front of the user. This means that the camera's field-of-view (FOV) remains constant. This approach has two main scenarios, whether the subject is captured with one stationary camera or from multiple perspectives in correspondence with multiple cameras (Aggarwal & Cai, 1997). The first scenario uses a single camera located in a stationary position. In order to record the user's actions, internal or external markers are used as features to be tracked such as pixels, lines, blobs, and regions. As a main advantage, this configuration uses the same spatial reference to resolve the matching problem in successive frames (e.g. Fig.2.1). The second scenario employs multiple cameras in correspondence with the user. In this case the main benefit is to increase the FOV of the scene; thus, if the subject disappears from one view, the system is capable of seeing it through another camera. However, this configuration is more complex, since, it is necessary to establish a feature correspondence from multiple viewpoints and coordinate them in the same spatial domain, e.g., (Dockstader & Tekalp, 2001). Both systems are suitable when we are interested in knowing the user's movements.

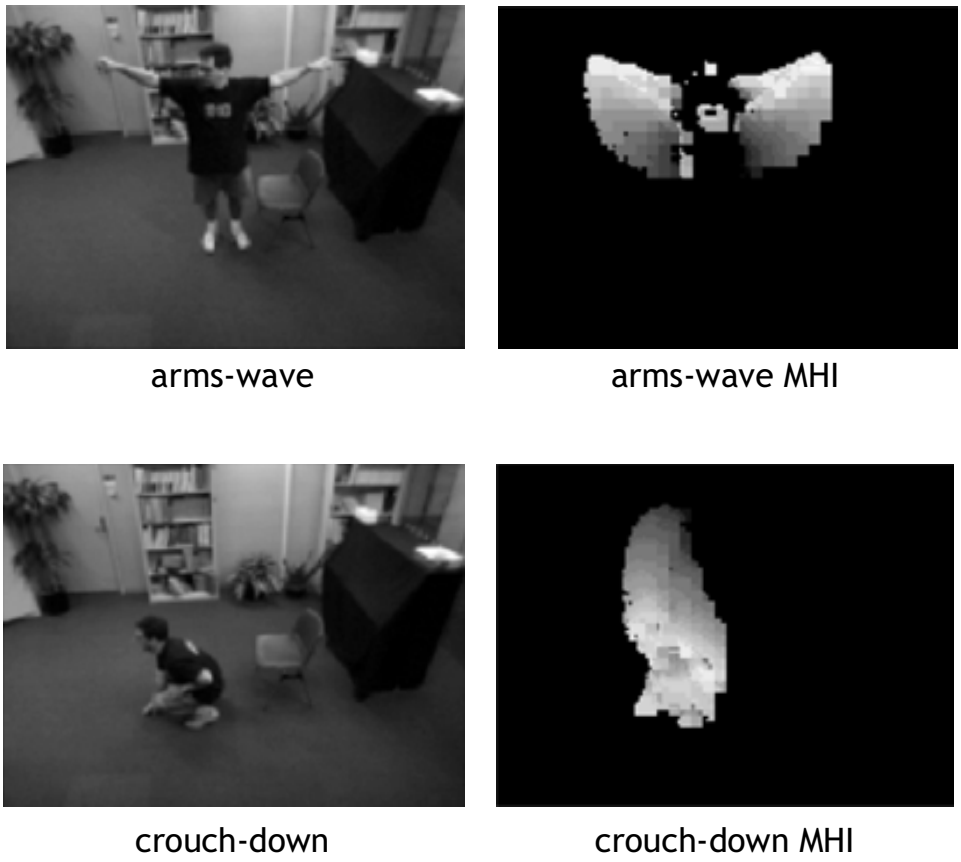


FIGURE 2.1. Passive camera position to gesture recognition proposed by Bobick & Davis (1996).

Active approach

In this approach the camera is not limited to its position and can interact with its environment, achieving a continuous representation of its surrounding. Bajcsy (1988) and Aloimonos (1990) introduced this paradigm to propose new models and control strategies in active perception systems. Active cameras were initially used to provide perception in autonomous robots; furthermore, they are being employed in human operators by adding a new sense of interaction with its world. Normally it uses external cameras and other devices attached to the human body; this last arrangement is called *wearable* because it is worn on the body (e.g. Fig.2.2). At present, wearable active cameras are offering new ways to increase human-computer interactions by allowing the user to gaze at the world and letting him/her move freely. From a computational perspective, this allows us to obtain a better representation of the user's surrounding and thus to infer the user's gestures. Readers may refer to (Mayol, Tordoff, & Murray, 2000; Davison, Mayol, & Murray, 2003; Kurata, Sakata, Kourogi, Kuzuoka, & Billinghurst, 2004; Campos, Mayol, & Murray, 2006) for further details of wearable vision systems.

FIGURE 2.2. Active camera position to gesture recognition proposed by de Campos et al., (2006).

Pointer approach

Passive and active approaches are useful to understand human motion; however, the major drawback of those approaches is that they are not designed to learn about the user's gaze. The *Pointer* paradigm is designed to overcome this disadvantage. It is based on the idea *what I am looking at is what I want*. Currently, the most widely used device to get the user's gaze is the eye-tracker. The eye-tracker allows tracking eyes movement by giving an estimated position of the user's gaze in real-time relative to an image frame, normally after an initial calibration (e.g. Fig.2.3). The system is composed of two head-mounted cameras: i) a camera that looks at the user's gaze. This camera has almost the same user's field-of-view (FOV), so it answers the first part of *what I am looking at*; and ii) a camera that captures eye movement by means of the corneal reflection; thus, it recovers the position of *what I want*.

Initial studies of eye movement were designed to understand the observable surface of the eye when the user was reading (Jacob & Karn, 2003). Today they are widely employed in different areas such as psychology, product design, biology, cognitive-neuroscience, and computer vision. Only in recently they are being used for disabled people with the purpose of increasing the users' interactions with their environment, and thus overcome their motorial difficulties, e.g. (Perini, Soria, Prati, & Cucchiara, 2006). In general, we observe that this technology is opening new opportunities to understand visual perception from a cognitive perspective and to explain the inherent mechanisms that control eye-hand coordination. However, so far there is no clear consensus or a unified theory that can explain this process effectively (see Desmurget, Pelisson, Rossetti, and Prablanc (1998) for detailed discussions). Therefore, it is not possible to use a specific model that explains the procedure underlying eye-hand coordination. Additionally, this technology is not enough to infer the user's intention; required to predict the grasping movement.

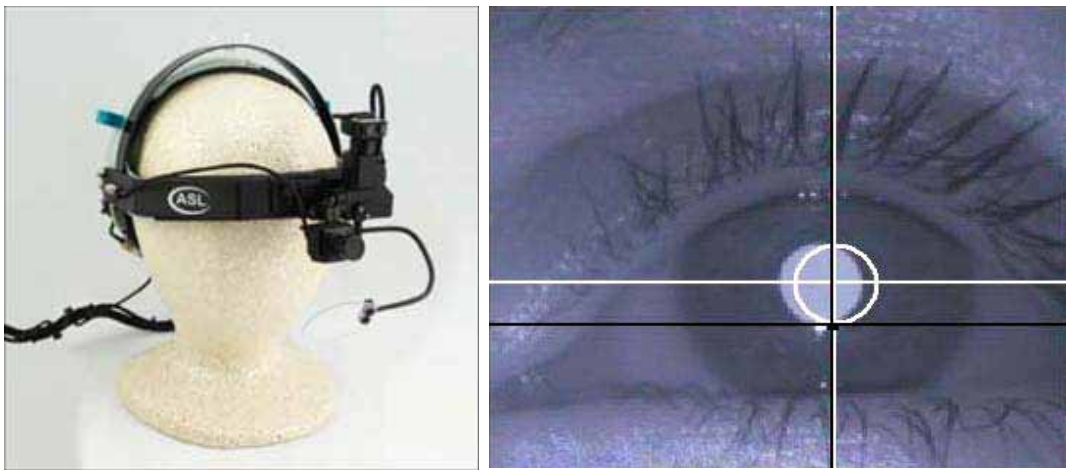


FIGURE 2.3. Left: H6 Optics eye-tracker designed by Applied Science Laboratories (ASL), Right: Pupil detection by the eye tracker.

2.2.2. Gesture recognition

Gesture recognition can be defined as the problem to follow body parts over the space-time in order to interpret the motion behavior as particular gesture. Based on Aggarwal and Cai (1997) definition, the gesture recognition requires to perform three general tasks. First, to identify some human body structure or low-level features such as points, blobs, 2D contours or 3D-volumes; second, to track human movements using low-level features by matching between consecutive frames or using the motion itself; and third, to recognize the human activity by matching the motion descriptor captured in the tracking process against the recognition framework. The last step is considered a higher level task due to the recognition task requires the classification of varying feature data over time (Jain et al., 2000).

The problem of interpreting the human gestures is defined as a learning process. In the training phase, some sequences are used to learn the user's behavior, labeling as a particular human gesture. Later, in the matching phase unknowns test sequences are compared against a model so as to be classified as a particular gesture. Most approaches designed to detect the human gestures are based on template matching or appearance-base models:

Template matching

This approach characterizes the human motion as means of recognition instead of using specific parts of the body. Thus, the action is represented by one robust vector. Polana and Nelson (1994) were pioneered in applying this approach. The main idea is to compute the motion fields between successive frames by dividing each frame into spatial grids, forming a high dimensional feature vector. This feature vector was conformed by optical flow magnitudes and periodicity flow frames. The recognition task was performed using a nearest centroid algorithm by comparing a feature

vector against a reference motion template. Bobick and Davis (1996) designed a similar approach, but they propose to extract features using the motion-energy image (MEI) and the motion-history image (MHI). Motion images are used as a binary representation of temporal difference between successive frames. These motion images are accumulated in time forming the MEI. On the other hand, each MHI uses a temporal decay of each pixel intensity at that position. Thereby, brighter pixels represents the more recently motion. In order to get an invariant representation of the action, invariant features of a set of MHIs and MEIs are extracted . These features are used later to recognize an input action by calculating the Mahalanobis distance between the moment description and known actions. In a similar way, but considering the action template as a space-time 3D volume, Shechtman and Irani (2005) extended the idea of 2D correlation into a 3D space-time volume by defining the action as a spatial-temporal geometric structure. For this, it is computed the correlation of a small video by seeking peaks in the behavioral correlation surface.

Appearance-base models

This approach considers the human motion as a set of local features, where an action is described as a sequence of images. One common technique employed in this approach is the Hidden Markov Models (HMM) (Rabiner, 1989). HMM is a probabilistic technique used to recognize patterns in temporal time series, usually employed in speech recognition. Starting with the work of Yamato et al. (1992), currently, HMMs has been adopted as a tool for recognizing the human motion. Yamato et al. (1992) developed the first human recognition method to recognize six tennis strokes using low-level features as an input to an HMM learning process. Although, low-level features do not provide rich descriptions of the motion, Yamato et al. (1992) shows that these features are enough to identify the human movement. In the same line, Starner and Pentland (1995) proposed a system to interprets the American Sign Language (ASL) using an HMM. They used low-level features such as shape, orientation and trajectory as input to an HMM without describing the hand shape. Recently, a novel approach by Achard et al. (2007) proposes to use semi-global features by estimating micro-movements from 3D spatio-temporal volumes. Finally, they determined invariant 14 moments so as to be used as an input of an HMM framework.

2.2.3. Eye-hand analysis

The previous section highlights the main approaches of the human motion recognition. This problem has been addressed using different approaches, for instance, by analyzing the human motion as itself or using specific parts of the human body. In general, these methods are focused on understanding the human movement from an external perspective. We now consider the problem of analyzing the eye-movement, specifically the eye-hand coordination required for grasping objects using the user's gaze perspective.

For many decades, the eye-hand coordination has been extensively studied from a neurological viewpoint by allowing us to understand this complex process with further details; however, at present researchers have founded that the eye-hand coordination is more complex than they had

thought, involving many system in co-occurrence. Despite this, today, one of the principal concerns is the design of new applications oriented to increment the interaction between the users and its surroundings. Below, we briefly describe the background of eye trackers and the major steps for controlling the eye-hand coordination. For further details of different eye-trackers technologies, readers may refer to Jacob and Karn (2003) for a comprehensive discussion.

Background of eye tracker devices

The study of eye movements has been explored by almost 130 years, starting with Javal's works in 1878. In that period, the earliest methods were too intrusive, including devices that directly manipulated on the cornea. It was at the beginning of the 19th century that Dodge and Cline (1901) designed the first non-invasive eye-tracker focused on capturing the light reflected on the cornea by recording the eye movement photographically. Mostly, those works were designed to record the horizontal and vertical position of the eyes with several constraints, especially avoiding the head movements. It was only in 1950s that Fitts, Jones, and Milton (1950) designed the first application of eye tracking by studying the interaction between the pilots' eyes and its instruments to land an airplane. This study is known now for being the first application in using the concept of *usability engineering*, that is, the study of the interactions with products. Later, in 1960s and 1970s, Mackworth and Thomas (1962) introduced many advances in the design of a head-mounted eye-tracker system less obtrusive and more portable. Also, in the same period the earliest system for capturing reflections of the infrared light from the cornea and the retina was designed. Only in the 1980s several improvements of eye tracking devices on real-time were developed, mainly with the advent of minicomputers and the high-speed in data processing. In this period, much effort was put into the design of computer interfaces to help disable users, e.g. (Perini et al., 2006).

The last 15 years, technological advances in computers have increased our understanding on how our visual processes operates when we are reading (Starr & Rayner, 2001), manipulating, intercepting and reaching objects with hands (Johansson et al., 2001; Mrotek & Soechting, 2007), and studying the brain controls processes that operate in the eye-hand coordination based mainly on the eye movement data (Hayhoe et al., 1998, 2003; Brouwer & Knill, 2007). Today, the challenge is to design eye-trackers more portable, less obstructive, reliable and easier to use. Also, in how to integrate these devices in new applications, mainly oriented to increase human-computers interactions.

Eye-hand coordination

One of the common behaviors of human beings is their ability to control hand movements to reach, grasp and manipulate objects by means of eye-hand coordination. In healthy people these actions are essential for leading a normal life, for instance, doing sports, eating, working, writing, playing instruments, etc. However, any disorder caused by injuries, diseases, mental disorders, or tremor may lead to a deterioration in the quality of life.

The process that controls eye-hand coordination is complex and is the result of many sensor-receptor, control, and cognitive systems working synergically (Brouwer & Knill, 2007; Hayhoe et al., 2003). This process is completely different from human motion as defined by Adams (1981). Human motion is a process strongly influenced by the cognitive process, and as a result, when the movement becomes habitual, the cognitive process is used only to correct or perfect the movement. In contrast, eye-hand coordination requires a sensory signal mechanism that controls the eye and hand movements as a single unit. Such coordination demands three main brain tasks. First, solving a geometric transformation between the internal word, encoded by the retinocentric frame of reference; and the external world, using a body-centered representation by proprioception (Crawford et al., 2004; Engel, Flanders, & Soechting, 2002). Second, developing a plan to reach an object using body-centered coordinates by comparing the gaze signals with the hand coordinates, and by estimating the hand motor difference in relation to the gaze coordinates (Buneo, Jarvis, Batista, & Andersen, 2002). Third, controlling the posture of the hand before reaching an object by seeking the most suitable posture, taking into account the size, shape and orientation of the object (Rizzolatti, Fogassi, & Gallese, 1997).

For many years researchers have been studying this process trying to find the underlying mechanism that controls it; however, at present there is not a single theory that can explain this process effectively and it is still not completely understood, e.g. (Hayhoe et al., 1998; Donkelaar et al., 2000; Brouwer & Knill, 2007). The controversial question is how much information is used to plan a hand movement. More precisely, does human vision rely more on visual information or on memory representations? Over the last decade many researchers have supported the idea that only limited information is acquired from saccades (Hayhoe et al., 1998; McConkie & Currie, 1996; Ballard, Hayhoe, & Pelz, 1995). Humans seem to maximize coordination between eye and hands movements using visual information continuously instead of using a memory representation to plan their movements. The main reason is that memory is too old and uncertain even when nothing has changed; in contrast, visual information is constantly being updated. However, recent studies have shown that people can use both systems simultaneously. Brouwer and Knill (2007) stated that unconscious memory is used all the time to plan hand movements and to point attention to objects. They have shown that the brain can use both sources depending on their relative reliability. If visual information is more reliable than memory representations, the visual source seems to dominate. In contrast, when visual information is degraded, the brain increases the use of memory information to plan hand movements. That explains why people require more time before grasping an object in a low-contrast situation compared to high contrast condition (Brouwer & Knill, 2007).

Another important issue concerning eye-hand coordination is related to gaze fixation. In general, there is clear consensus that the gaze arrives at a specific target quite before the hands can do it (Brouwer & Knill, 2007; Crawford et al., 2004; Hayhoe et al., 2003). Likewise, fixation seems to be stable until the object has been grasped, and later, when the hand arrives, fixations on the object are not needed anymore. Consequently, the number of saccades around the object is reduced

substantially, increasing the visual information on the retina (Mrotek & Soechting, 2007). Therefore, this behavior indicates that fixations have three main features. First, task-dependent, different fixations in time and position are need to carry out different actions based on knowledge and target location (Hayhoe et al., 2003). Second, task-relevant, the sequence of fixations on relevant points allows our brain to estimate the geometric relationship of the world based on the internal coordinates of the body (Crawford et al., 2004; Johansson et al., 2001). Third, memory-dependent, the fixations allow memorizing different spatial positions on objects, and they can later be used for planning the movements (Brouwer & Knill, 2007).

As stated above, the time needed to acquire a gaze fixation has a direct relation with the task context, namely, it depends on the degree of complexity required to manipulate an object with the hands. Therefore, the time is variable and can fluctuate from $100\,\mathrm{ms}$ to $1500\,\mathrm{ms}$, but most distributions take between $100\,\mathrm{ms}$ and $200\,\mathrm{ms}$ (Hayhoe et al., 2003). The long fixations are the result of prolonged action with continuous direction, as stated by Land, Mennie, and Rusted (1999). This important finding is a key factor to understand how our visual system works, and later may lead us to obtain a good idea of what is the user's intention by increasing the probability to detect the object required by the user.

Chapter 3

Automatic Multiple View Inspection

3. AUTOMATIC MULTIPLE VIEW INSPECTION

This chapter presents four papers that introduce Automatic Multiple View Inspection (AMVI). The objective of AMVI is to determine the quality of an object by tracking defects. The first two papers present the AMVI methodology applied to X-ray image sequences of automobile wheel rims. The next two papers present the inspection of wine-bottle necks by means of a functional prototype. The objective of the prototype is to determine the quality of the bottles by multiple image analysis. The inspection process uses the AMVI method, showing the potential of its application as a method of inspection of other objects. Below we detail briefly the main characteristics and achievements related to each paper.

- The first paper presents the AMVI methodology applied to uncalibrated image sequences. To determine the base correspondence between multiple images we designed an algorithm that transforms the edges of corresponding regions on B-Spline curves to then determine matching only in those sections of the curve that maximize their similarity. From that correspondence it is possible to determine the geometric model to carry out the tracking of the potential defects. A potential defect is a region of the image composed of a variation of its gray levels, which would indicate the presence of an irregularity in the object. To evaluate if a potential defect is a false alarm or a real flaw we use the multiple view tracking algorithm. The AMVI method determines the classification of the defect as follows: if a potential defect continues in its relative position with respect to the motion of the object that is being inspected, it is considered as a real flaw and the object is classified as defective, otherwise, if the potential defect does not agree with that motion it is considered a false alarm.

- The second paper presents various improvements of the previous paper on the AMVI methodology. First it presents an intermediate classification system called Intermediate Classifier Block (ICB) that searches in space only those properties in which the potential defects concentrate. This reduces the number of false alarms and real flaws of the set of possible correspondences in two and three views. Second, it introduces an improvement of the correspondence system through the search for the flexion parameters of corresponding curves by means of an optimization process. To carry out this process it is necessary to decrease the noise of the curvature, a process that we carry out through Fourier descriptors. Third, it introduces a modified version of the robust estimation algorithm of the fundamental matrix proposed by (Chen, Wu, Shen, Liu, & Quan, 2000), in which only one subset of correspondences of the combinatorial is used.

- The third paper presents the description of the prototype, the electromechanical system, the internal illumination system, and the diagram of the inspection system. The designed prototype has the advantage that it can be applied to any kind of bottle because it adapts

mechanically to its length. The main novelty is the internal illumination system, which makes the defects easily detectable by means of a CCD camera. This system is quite novel because it allows the bottle to be turned keeping the internal illumination while the camera captures multiple images of the visible section of the bottle. The other big advantage is that the illumination system consists of two zones of markers that are reflected in the bottle when they are backlighted. In this way, when the illumination system is turned, so is the bottle together with the markers. The latter are necessary to determine the base correspondence of the geometric model.

- The fourth paper presents two novelties with respect to the previous one. First, the feature extraction with multiple invariant descriptors is made, evaluating in each case the performance as a function of the re-projection distance. Second, it introduces the Bidirectional Nearest Neighbor Distance Ratio (bNNDR) algorithm applied as a correspondence method of potential defects. The bNNDR algorithm allows the point-to-point correspondence to be analyzed using a bidirectional analysis of the features in two and three views. We show that this algorithm improves by at least 10% the performance of the NNDR algorithm proposed by Mikolajczyk and Schmid (2005).

Paper #1
Automated visual inspection using trifocal analysis in an uncalibrated sequence of images

Miguel Carrasco and Domingo Mery
mlcarras@puc.cl - dmery@ing.puc.cl

Abstract

Automated inspection using multiple views (AMVI) has been recently developed to automatically detect flaws in manufactured objects. The principal idea of this strategy is that, unlike the noise that appears randomly in images, only the flaws remain stable in a sequence of images because they remain in their position relative to the movement of the object being analyzed. AMVI has been successfully applied in sequences of calibrated images for which the 3D→2D transference function for the projection of the views is known precisely. Nonetheless, its application in industrial environments is a complex task because of the instabilities that are inherent to the system. This investigation proposes a new strategy, based on the detection of flaws in a non-calibrated sequence of images. The methodology designed consists of constructing a model and carrying out a trifocal analysis that allows the determination of the real position of a flaw using corresponding control points in the sequence. Experimental results obtained on radioscopic images of die castings illustrate the potential in the detection of defects in non-calibrated images, detecting the totality of the flaws in the sequence

Keywords: Computer vision, multiple view geometry, automated visual inspection, defect detection

1 INTRODUCTION

The quality of manufactured goods is one of the principal objectives of the productive process. In order to evaluate quality, there are a variety of inspection and analysis tools that can be carried out during the manufacturing process, however, all of these depend on security standards set by the manufacturer, or by some regulatory agency. While it is true that some manufacturers tolerate production with flaws, for others safety is a critical issue. Among the latter are high pressure equipment, chemical containers, aluminum wheels, etc. The existence of flaws in these products can cause serious accidents. Generally the inspection is carried out visually by trained personnel due

Miguel Carrasco and Domingo Mery are with the Computer Engineering Department at Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860 (143), Santiago, Chile.

to the fact that human visual inspection is flexible and adaptable to new situations not originally considered. However this process has grave problems such as: deficiency due to the time it takes to inspect an object which depends on fatigues and monotony; and inconsistency because the process depends on the capacity and experience of the inspectors (Newman & Jain, 1995). For this reason the process of analysis is a target for automation which allows an improvement in the quality of inspection, as well as a reduction in costs of production.

In recent years Automated Visual Inspection (AVI) has solved a number of problems in the area of quality control through the establishment of precise and objective control policies (Davis, 2005). Through image processing techniques and pattern recognition protocols, AVI allows the detection of flaws ensuring adherence to two basic conditions in the productive process, namely, efficiency and speed. Nonetheless, the majority of these techniques require the individual processing of each flaw in the image, which implies a subsequent analysis of the individual characteristics of each flaw. Subsequently, through the process of pattern recognition, we determined if the flaw is real or a false alarm.

Recently, a new methodology for automated flaw detection has been developed; Automated Multiple View Inspection, (AMVI) (Mery & Filbert, 2002). Equivalent in form to the flaw detection process carried out by an inspector, AMVI detects flaws using the following two steps. The first step, named *identification*, consists of detecting all the anomalous regions or hypothetical flaws in each image by means of a sequence of movements of the object, without any a priori knowledge of the object's structure. The next step, called *tracking*, consists of a follow-up of the hypothetical flaws found in each image in the sequence during the first step. If the hypothetical flaws continue the length of the image sequence, the hypothetical flaw is tagged as a real flaw, and the object is catalogued as defective. On the other hand if the hypothetical flaws do not show correspondence in the sequence, they are considered to be false alarms. AMVI methodology's founding principle is that only real flaws, (and not false alarms), can be observed throughout the sequence of images because these remain stable relative to the movement of the object. Therefore, with two or more views of the same object, from different points of view, it is possible to improve the performance with regards to real flaw detection.

This original strategy, presented in (Mery & Filbert, 2002), requires a previous calibration of the image sequence acquisition system. In calibration we seek to establish the transference function that projects a 3D point in the object onto a 2D point on the image (Mery, 2003b). Unfortunately, the calibration process is difficult to carry out in industrial environments due to the vibrations and random movements that vary in time and are not considered in the original estimated transference function. An alternative method for the carrying out of the AMVI strategy in non-calibrated sequences was presented in (Mery & Carrasco, 2005) for sequences with two images. Nonetheless, because the robustness of the AMVI methodology increases with the number of images analyzed in the sequence, in the present work we propose a modification to the robust system designed in (Mery & Carrasco, 2005), processing three images instead of two. Additionally, in this work we increase the number of corresponding control points that related the images uses B-Spline curves,

thus significantly improving the estimation of the multi-focal model necessary for carrying out the tracking.

The remainder of this document is divided into the following sections. Section 2 includes background information on AMVI methodology. Section 3, which is dedicated to the proposed method, includes a description of the methodology used to segment hypothetical flaws, estimate the fundamental matrix robustly, generate artificial control points and estimate the trifocal tensors. Section 4 presents the experimental results. Finally, Section 5 presents the conclusions.

2 BACKGROUND

The principal objective of AMVI is to follow only the hypothetical flaws, and not to estimate the structure of the object. Initially the methodology was implemented to automate the inspection of aluminum wheels, using a sequence of images in a calibrated system (Mery & Filbert, 2002). In this case the calibration of the object was generated off-line. A projection model was then generated to track flaws throughout the sequence of images, using the principles of multiple view geometry (Hartley & Zisserman, 2000; Mery, 2003c). The results obtained demonstrated the technical feasibility of detecting the totality of real flaws, together with a high rate of detection of false positives. Nonetheless, in industrial environments, calibration is a complex process due to the vibrations during the acquisition of images of the object, which carries with it a lack of precision for the estimation of the parameters necessary for the multiple view geometric model.

The investigation carried out in (Mery & Carrasco, 2005) puts forth a robust alternative model which requires no calibration of the image acquisition system. One of the principal factors for estimating the movement of the object is that fact that it is a rigid body that has a rotational and/or translational movement with constant velocity and smooth trajectory. The method presented in (Mery & Carrasco, 2005) searches for significant regions of the object to be analyzed that are present throughout the image sequence. Once the correspondence between points in these regions has been established, a two-view model is constructed that serves for establishing correspondence between hypothetical flaws.

The proposed model initially used the following methodology: first, identify the structural points in each image in the sequence; second, find the correspondence, in consecutive images, of the points identified in the first step; third, generate a robust estimate of the fundamental matrix (Hartley & Zisserman, 2000) of possible corresponding points. Upon finishing this process a mathematical model is used to relate motion between pairs of images. The next phase relates hypothetical flaws in both images. Using epipolar geometry (Hartley & Zisserman, 2000), we search for flaws that may agree with the properties of the hypothetical flaws in the previous image. This evaluation is carried out by looking at Euclidean distances over the set of area and intensity properties.

Should the flaw have no correspondence in the next image, the latter is rejected, and is considered as a false positive as it does not fulfill with the epipolar constraint.

There are two relevant factors regarding the process developed in (Mery & Carrasco, 2005). First, the segmentation phase is designed to detect the majority of defects without any a priori knowledge

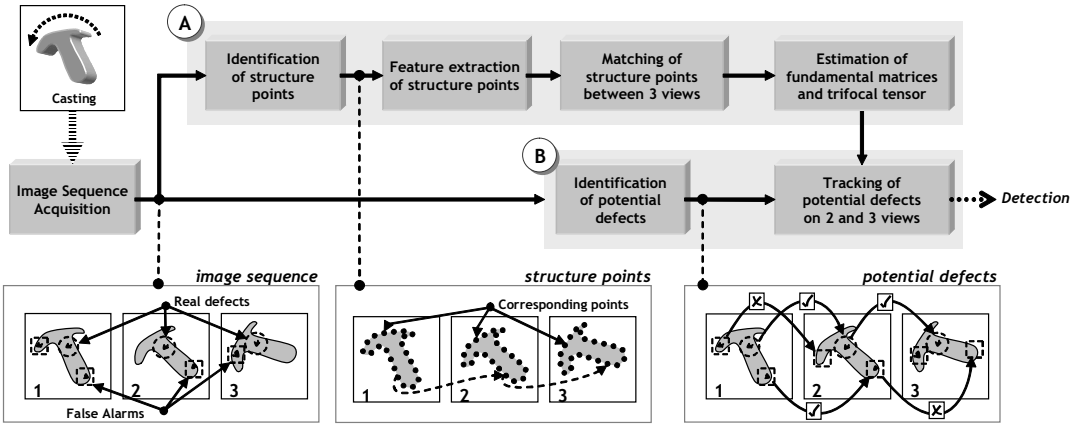


Fig. 1: Block diagram of the uncalibrated automated multiple view inspection: a) estimation of motion model, b) detection of defects acquisition.

of the material and/or position of the object. Within the image processing phases, segmentation occurs in an initial stage, however it has a preponderant role in the entire process because poor performance in this phase can cause poor prediction in the detection of real flaws (Castleman, 1996). Secondly, the system of correspondence must evaluate only those pairs of points related in both views. Nonetheless, there is not always a perfect correspondence due to geometric distortions or other anomalies in the capture process. Therefore, a robust algorithm has been used that rejects positions that contain a projection error, and uses only a set of related pairs, thus minimizing the error between the real and projected position, according to a Euclidean distance metric. This method is known as a RANSAC approximation (Fischler & Bolles, 1981). The term robust refers to the flexibility in the determination of a minimum set of related coordinates, and the rejection of those that do not fulfill with the minimum error allowed between the real and projected position.

3 PROPOSED METHOD

Laboratory results have shown that AMVI performs very well in the detection of flaws in aluminum die castings in calibrated environments (Mery & Filbert, 2002). Nonetheless, in industrial environments calibration is a complex process and of high cost for the manufacturer. This section presents a new AMVI method proposed for the automated detection of flaws using trifocal analysis of non-calibrated images, perfecting thus the method designed by the same authors in (Mery & Carrasco, 2005). The improvement is due to the fact that we have extended the analyses from two to three images, estimating the trifocal tensors robustly, and increasing the points of control artificially in order to establish correspondence between (Fig.1). The steps in this new method are presented below:

1) For each of the three images in the sequence (I, J and K), find the hypothetical flaws using

the crossing line profile segmentation algorithm (Mery, 2003a), which searches in small high contrast areas.

2) For all the structures in the object being analyzed in images I, J, and K, search for relationships between structures and generate artificial control points with B-Spline curves (Bartels, Beatty, & Barsky, 1998).
3) Estimate the fundamental matrix between images I and J using the RANSAC (Fischler & Bolles, 1981) method with the structural and artificial points found in Step 2.
4) For all the hypothetical flaws in I found in Step 1, generate the epipolar projection using the fundamental matrix (Hartley & Zisserman, 2000; Faugeras, Luong, & Papadopoulo, 2001) in image J, and determined which flaws are closest to the epipolar line in search of hypothetical flaws found in Step 1 in image J using the practical bifocal restriction:
   a. If there is more than one hypothetical flaw on the epipolar line, find the best relationship on the basis of the properties of area and intensity by means of the smallest Euclidean distance, and store said relationship.
   b. Should the epipolar line not pass through a hypothetical flaw in image J, this means that there is projected flaw in image J and the hypothetical flaw in image I is rejected.
   c. If there is only one flaw on the epipolar line which meets the Euclidean distance criterion mentioned, then the relationship between the flaws in I and J is stored.
5) Estimate the trifocal tensors between images I, J and K using RANSAC (Fischler & Bolles, 1981) with the structural and artificial points found in Step 2.
6) For the flaw relationships between I and J found in Step 4, find the position of the center of mass of the hypothetical flaws, and re-project these positions using the trifocal tensor:
   a. If there exists a projected flaw that is a minimum distance from the hypothetical flaw in image K from Step 1, assign this position as matching in three views and thus determine that a flaw in the sequence has been detected.
   b. If there is no flaw in image K, related to the projection, then eliminate the hypothetical flaw and catalogue it as a false positive.

An explanation of these steps is presented below

3.1 Identification of hypothetical flaws

The segmentation of hypothetical flaws allows the identification of regions in each image of the sequence which may correspond to real flaws (Fig.2). There are two general characteristics used to identify them: i) a flaw is considered as a connected subset in the image, ii) the differences between the gray levels of the flaw and its neighbors is considerable. Initially, identification takes place through a process with no a priori knowledge using the convolution of the image with a Laplacian-of-Gaussian (LoG) kernel, and then a zero crossing algorithm (Castleman, 1996). The LoG operator intrinsically uses a Gaussian low-pass filter to reduce noise levels in the image. The results of the operator and the zero cross are a binary image that contains real flaws with connected surroundings. Nonetheless, these surroundings are not always closed. This happens when they are

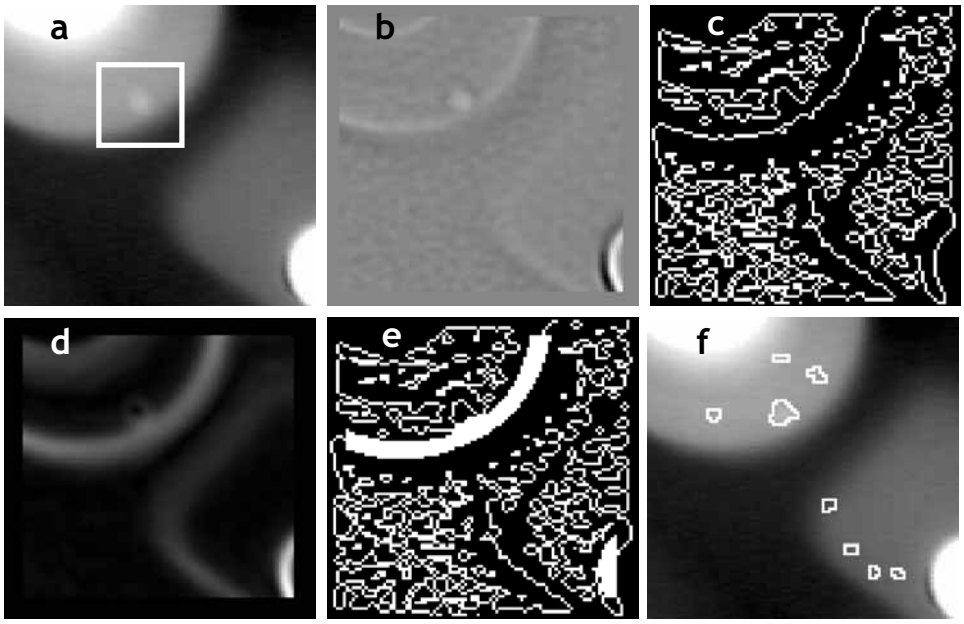


Fig. 2: Flaw detection: a) section of a radioscopic image with a flaw inscribed on the edge of a regular structure, b) application of the Laplacian filter on an image with = 1.25 pixels (kernel = 11×11), c) zero crossing image, d) gradient of the image, e) detection of edges after increasing the edges to the highest levels in the gradient, and f) detection of flaws using the variance of the crossing line profile (Mery, 2003).

close to the edges of a regular structure (Fig.2c). The solution consists of augmenting the borders of the regular structures. This procedure consists of calculating the gradient of the image in order to identify these positions (Fig.2d) and later generate a binary image which employs only the levels with the most energy in the gradient. (Fig.2e). Once each closed region is segmented, characteristics are extracted through the grey tone profile of a straight line that passes through the center of the segmented region. Those that present a high variance profile are identified as hypothetical flaws (see details in (Mery, 2003a)). This hypothetical flaw contains a high number of false positives, it has however the following advantages: i) the same detector is applied to all images, ii) it allows the identification of hypothetical flaws, independently of the position or the structure of the object under study, in other words without a priori knowledge of the design of the structure, iii) the detection of real flaws is very high (higher than 90%).

Once the segmented regions have been determined, the next step is to determine the position of the center of mass for each hypothetical flaw. For each image $\mathbf{m}_i$ will be used to denote the center of mass of the segmented region r_i. In homogenous coordinates, $\mathbf{m}_i$ represents the 2D spatial position of the i-point, $\mathbf{m}_i = [x_i, y_i, 1]$.

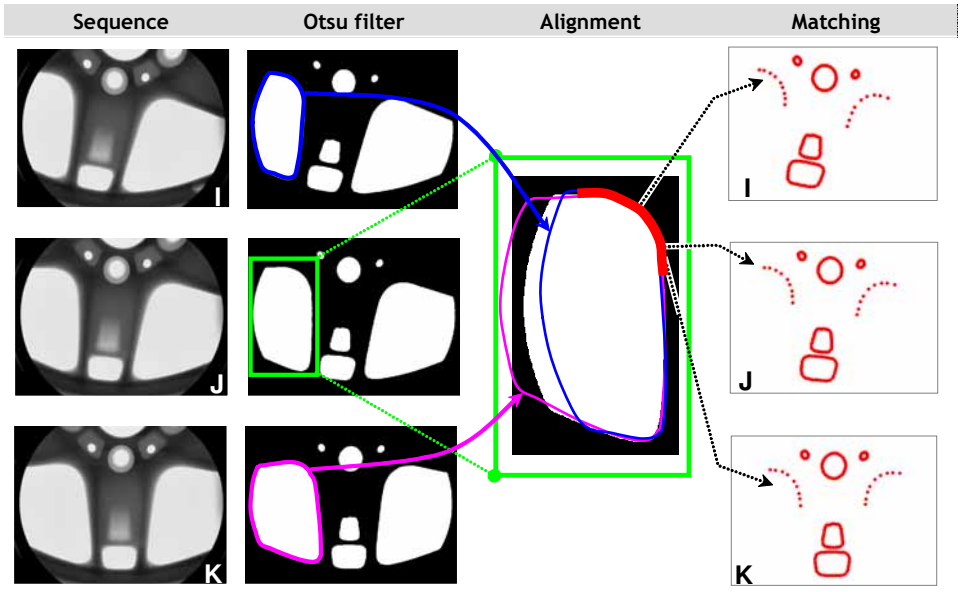


Fig. 3: Non-linear matching algorithm for corresponding points in a sequence of three images: a) sequence of original images, b) application of Otsu's filter to separate the valid tracking structures and generation of border with B-Spline curves, c) alignment of each valid structure with respect to image J, and d) result of alignment with normalized correlation of the pattern of the x-coordinate position

3.2 Identification of control points

The main problem in the correspondence of the control points in the sequence of images used, is the low number of structures that have a valid projective transformation matrix, $\mathbf{H}$. This means that some structures do not have a linear transformation due to the occlusion that results from being near the edge of the image. In these structures, it is possible to see only some regions where there is a correspondence, especially in the interior edges.

Using the rotation information of the sequence analyzed, we designed a practical method for finding a correspondence (Fig.3). The procedure uses the following steps:

1) Construct a binary image using Otsu's method (Haralick & Shapiro, 1992), thus isolating the structures from the image background.
2) Generate a quadratic B-Spline curve using the edges of each segmented structure as control points (Bartels et al., 1998).
3) Align the structures of the first and third view with the second view. Generally, the movement of the piece is known a priori when its inspection is carried out. This information allows the determination of the angle and initial displacement in order to estimate a possible alignment.

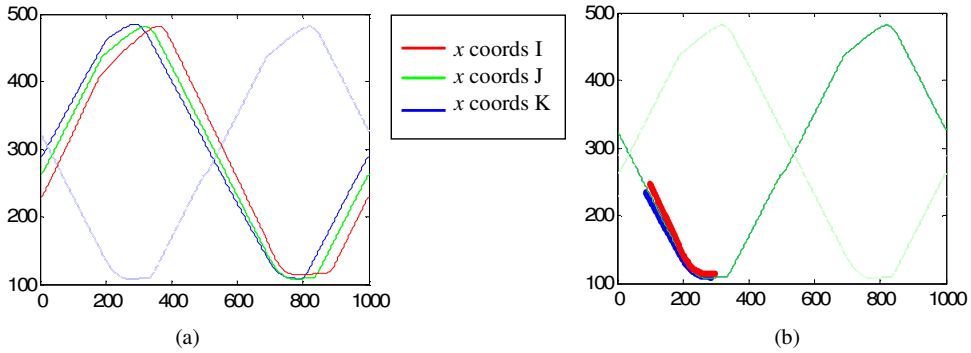


Fig. 4: Determination of x-coordinate position for each the B-Spline curve: a) the x-coordinate position for the first structure of images I, J, K, b) estimation in the position of the pattern of structures I and K over structure J using the pattern correlation over curve J.

4) Determine a common pattern for the rotation of the structures from the first and third view, using the x- coordinates positions from the second view as a base (Jain, 1989).
5) Calculate the position of the pattern relative to the curve of image J, so as to minimize the error by means of a normalized correlation (Fig.4).

3.3 Estimation of the fundamental matrix

The fundamental matrix is vital for the AMVI process as it relates any position in pairs of images. The precision of this matrix allows the correct determination of hypothetical flaws along the length of the epipolar line (Hartley & Zisserman, 2000). Nonetheless, if fundamental matrix is not robust, the epipolar line will be incorrect and the subsequent trifocal tensor process will be failed. In this case, if the point $\mathbf{m}_p$ of the first view corresponds to $\mathbf{m}_q$, in the second view, the following relationship is established,

$$\mathbf{m}_q^\top \mathbf{F}_{pq} \mathbf{m}_p = 0 \tag{1}$$

where $\mathbf{F}_{pq}$ is the fundamental matrix of the projection of points $\mathbf{m}_p$ and $\mathbf{m}_q$ in homogenous coordinates. Once the set of corresponding positions has been generated in each region in both views, we use the robust RANSAC algorithm to estimate the fundamental matrix (Fischler & Bolles, 1981). It should be remembered that there is a probability of error between the position of one region its corresponding pair, nevertheless, RANSAC minimizes this error as it uses the set of pairs of points that generates the best estimate of the fundamental matrix.

The RANSAC algorithm uses seven areas of points to determine the fundamental matrix. For this reason there must be a minimum number of pairs of corresponding points that have been correctly estimated. Should there be an error in the correspondence of control points the fundamental matrix would be incorrectly estimated. Fortunately, it is not necessary for the process that all the points

correspond exactly, this being the principal advantage of the robust algorithm in the estimation of the fundamental matrix.

3.4 Evaluation in two images

Once the robustly generated fundamental matrix has been constructed, it is necessary to calculate the epipolar line for each segmented region of the first view. Using the centers of mass for each hypothetical flaw generated in section 3.1 we generate the epipolar line thus

$$\mathbf{l}_{qi} = \mathbf{F}_{pq}^{\top}\mathbf{m}_{pi} = [l_x, l_y, l_z]_i \tag{2}$$

where $\mathbf{l}_{qi}$ is the epipolar line of flaw i in the second view, and $\mathbf{m}_{pi}$ is the center of mass of flaw i in the first view. The result of $\mathbf{l}_{qi}$ is a line in the xy-plane, and has an equation as follows.

$$\mathbf{l}_{qi} = A_i x + B_i y + C_i \tag{3}$$

where $A_i = l_{xi}$, $B_i = l_{yi}$ and Ci=lzi of flaw i, are the coefficients of the epipolar line. Once the epipolar line of flaw i of the first view has been generated, it is necessary to determine the distance between a corresponding flaw in the second view and the epipolar line. This distance is determined through the practical bifocal restriction (Faugeras et al., 2001). Given that the epipolar constraint is applied to points and not to regions, we consider the center of mass of each hypothetical region to be corresponding points between pairs of images This simplification is subject to error as it supposes that hypothetical regions have their center of mass in a corresponding point in both images. Nonetheless, we use this restriction because the majority of hypothetical flaws are small and the angles of rotation and deformations are small for each image.

For any flaw i in the first view and flaw j in the second view, we define $\mathbf{m}_{pi}$ and $\mathbf{m}_{qj}$ to be the centers of mass of the regions r_{pi} and r_{qj} in each view respectively. If the Euclidean distance between $\mathbf{m}_{qj}$ and the epipolar line of $\mathbf{m}_{pi}$ is less than a given ϵ, this implies that the hypothetical flaw in the second view is related to $\mathbf{m}_{pi}$. If the hypothetical flaw is found in both images, then it is considered to be a flaw in the bifocal correspondence, if this is not the case, the region is discarded.

$$d\left(\mathbf{m}_{pi}, \mathbf{F}, \mathbf{m}_{qj}\right) = \frac{|\mathbf{m}_{qj}^{\top}\mathbf{F}_{pq}\mathbf{m}_{pi}|}{\sqrt{l_x^2 + l_y^2}} < \epsilon \tag{4}$$

Experimentally, a given epipolar can contain more than one hypothetical flaw. In this case, we establish a degree of similarity for each hypothetical flaw, under the assumption that they all fulfill the epipolar constraint. For each flaw the characteristics of area and intensity are compared (Fig.5).

Similarity is established when two flaws (one in image I, and the other in the epipolar line of image J) are at a minimum distance in the normalized space of the properties, using the Euclidean distance as a similarity metric. If there are still false positives after the analysis of the two views, it is possible to use the trifocal tensor to eliminate the remaining false positives.

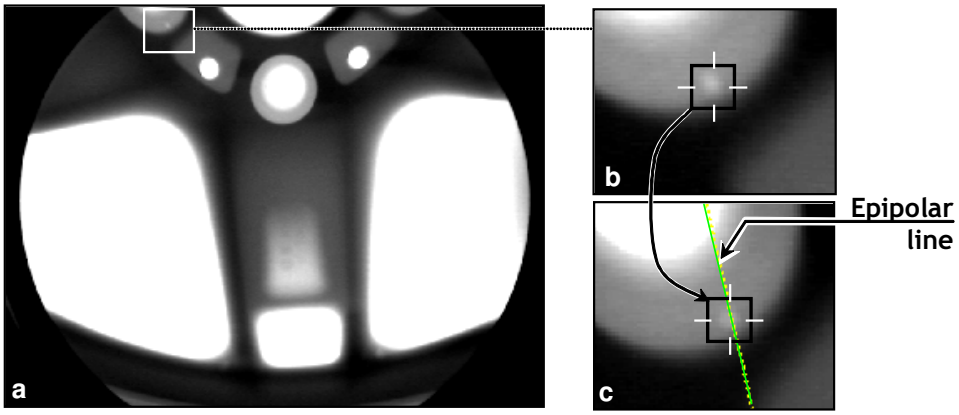


Fig. 5: Epipolar line generated automatically from the fundamental matrix: a) first view, b) identification of a hypothetical flaw, c) intersection of the epipolar line in the second view with one or more corresponding hypothetical flaws.

3.5 Estimation of the trifocal tensors

Trifocal analysis allows the modeling of all the geometric relationships in three views, and is independent of the structure contained in each image (Hartley & Zisserman, 2000). The tensor, a matrix structure similar to the fundamental matrix, only depends on the movement between images and the internal parameters of the cameras. It can be computed directly from the projection matrices of the views. Nonetheless, is can be calculated from the correspondence of the images without any a priori knowledge of the movement or calibration of the object. This characteristic justifies its, because the estimation of the fundamental matrix does not always eliminated the totality of false positives (Mery & Carrasco, 2005). An analysis in three views increases the probability of forming triplets that fulfill with the trifocal condition (Shashua & Werman, 1995).

In order to determine the trifocal tensors, we must have correspondence between the control points of the three views. Nevertheless, this assumption is not met precisely with the non-linear method designed in section 3.2. It is therefore necessary to use the robust algorithm RANSAC (Fischler & Bolles, 1981) to determine the best triplets of points in order to minimize the projection error in the third view.

The initial estimation of the tensors is carried out with Shashua's four trilinearities (Shashua & Werman, 1995). This makes it possible to verify if three corresponding points $\mathbf{m}_p$, $\mathbf{m}_q$ and $\mathbf{m}_s$ in the first, second, and third view respectively, satisfy the trilinearities, in which case they are corresponding points in the three views, and they depend on the projection matrices (Hartley & Zisserman, 2000).

TABLE 1: Performance in the identification of hypothetical flaws in the segmentation phase

Sequence	No. Images	Detected defects/ image	False alarms/image
Left images	70	210/70=3	205/70=2.92
Center images	70	209/70=2.98	205/70=2.92
Right images	70	209/70=2.98	206/70=2.94
All	210	628/210=2.99	616 /210= 2.93

3.6 Evaluation in three views

To determine of the flaw in the third view corresponds to the trifocal relation, we use the center of mass $\mathbf{m}_p$ and $\mathbf{m}_q$ of regions r_p and r_q of the first and second view respectively. For this we use the re-projection of the trifocal tensor in the third view using the positions $\mathbf{m}_p$ and $\mathbf{m}_q$ in the two first views. We use only the centers of mass of the two first views which fulfill with the bifocal relationship from section 3.4.

Let us define $\mathbf{m}_s$ as the center of mass of region r_s from the third view. If the Euclidean distance between the real position of the hypothetical flaw $\mathbf{m}_s$ and that which is estimated with the trifocal tensors, $\hat{\mathbf{m}}_s$ is less than some value , we take the hypothetical flaw to be a real flaw, as it complies with the correspondence in three views. Should the hypothetical flaw in the third view not agree with the projection of the tensor, it is discarded as it does not fulfill with the trifocal condition (Shashua & Werman, 1995).

$$d_r = \|\hat{\mathbf{m}}_r - \mathbf{m}_s\| < \epsilon \tag{5}$$

In general, given that the trifocal condition is analyzed for the sequences that fulfill with the bifocal condition, we reduce the number of false positives generated in two views. Therefore, the number of false positives in three views can only be less than or equal to the number that exists in two views.

4 Experimental Results

This section presents the results of experiments carried out on a sequence of 72 radioscopic images of aluminum wheels (Mery & Filbert, 2002) (Fig.6). There are twelve known real flaws in this sequence. Three of these are impact flaws detected by human visual inspection, ($\emptyset = 2.0 \sim 7.5$ [*mm*]), the remaining nine were generated by a drill which made small orifices ($\emptyset = 2.0 \sim 4.0$ [*mm*]) in positions which would difficult their detection.

We separated analysis into two steps. In the first step, called identification potential defects are automatically identified in each image of the sequence using a single filter and no a priori knowledge of the structure of the test object (Mery, 2003a). The results indicate that it exists 2.99 real flaws in each image, and 2.93 false alarms (Table 1). In the second step, called tracking, an attempt is made to track the identified potential defects in the image sequence. In this last step, we separate the analysis in two phases: first, the detection of pairs of flaws using the estimation of

TABLE 2: Performance in the detection of real flaws with two and three views in sequence

Step	Detected defects in sequence	Real defects in sequence	False alarms in sequence	Detection performance	False alarm rate
Bifocal	190	190	93	100%	32.9%
Trifocal	170	170	19	98.8%	9.9%

TABLE 3: Comparison of the present investigation and the study carried out in Mery & Carrasco (2005).

Technique	Sequence analyzed	Detection performance	False positives
Bifocal AMVI (2005)	12	92.3%	10%
Bifocal AMVI (new)	70	100%	32.9%
Trifocal AMVI (new)	70	98.9%	9.9%

the fundamental matrix in two views, throughout epipolar constraint; second, using the previous results, we re-projected the pairs of hypothetical flaws in the third view using the trifocal tensor estimation. Both last phases are detailed below.

4.1 Performance with two views

The first phase is to evaluate the performance of the algorithm in two views using the bifocal method. This consists of determining corresponding flaws between two images in a sequence through the search for flaws in the epipolar line. The method was applied to 70 pairs of radioscopic images (578×768 pixels) of aluminum wheels generated in (Mery & Filbert, 2002) for which the angle of rotation 5° is known for each sequence in the image. This information is used in order to align the segmented structures (see details in section 3.4).

The results indicate that the model detects 100% of the real flaws that have correspondence (Table 2, bifocal). This validates the assumption of correspondence between the positions of the real flaws and implies that automated detection with a fundamental matrix allows the detection of corresponding flaws contained in the epipolar line, which is in agreement with the results obtained in (Mery & Filbert, 2002) and (Mery & Carrasco, 2005). The study showed a rate of 32.9% of false positives which have correspondence in image pairs. Although this percentage is high, we do not penalize false positives as these can be reduced using a third image.

With respect to the study carried out by the same authors in (Mery & Carrasco, 2005), we have extended our analysis to the entire test image sequence generated in (Mery & Filbert, 2002). The previous study used twelve sequences selected specifically as the matching system was limited to conditions where correspondence was feasible. The present investigation however, includes a non-linear correspondence system based on control points (see details in section 3.2).

The results show an increase in the detection of flaws through the use of the fundamental matrix, using a modified version of the original procedure presented in (Mery & Carrasco, 2005) (Table 3). This is due to the significant increase in the number of control points correlated in pairs and triplets of images through a non-linear correspondence system. The previous investigation used the centers of mass of regions with a variance of less than 7% in the area to be used as corresponding points. The present scheme uses the relation between the edges of corresponding regions. For this reason RANSAC increases the precision of the calculation of control points and minimizes the estimation error of the fundamental matrix.

In general, it is possible to determine precisely the fundamental matrix for each pair of images for the following reasons: i) the majority of the positions of each region agree with the rotation and/or translation of the following image; ii) only those regions that have a variation of less than 4% are considered, thus eliminating possible matching errors: iii) RANSAC uses only the best seven pairs from the total of positions in all the regions.

4.2 Performance with three views

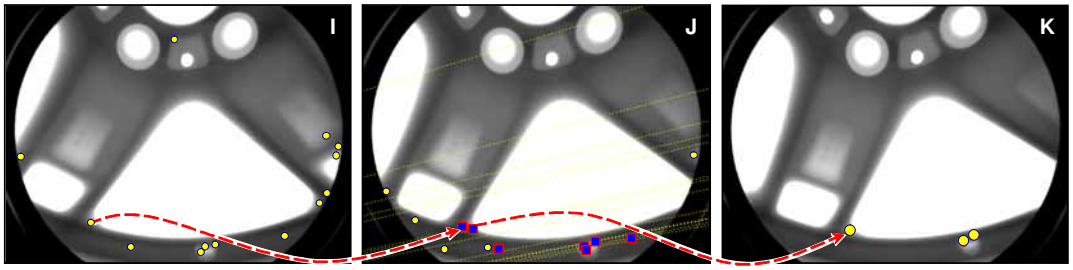
The second phase uses the algorithm proposed in section 3.6. After completing the matching of possible pairs of flaws in both images, we extend the detection of flaws to the third image in the sequence. In the present case the study used 70 triplets of images.

Within the tests performed, the best performance of the trifocal tensor did not achieve perfect results. However, 98.5% of the real flaws were detected (Table 2, trifocal). Moreover, the results indicate a reduction in the quantity of false positives from 32.9% with bifocal AMVI to 9.9% with trifocal AMVI. This agrees with the assumption that in the measure that the number of sequenced images is increased, the number of false positives decreases as in general real flaws are in correspondence.

Our experiments have demonstrated that there is a greater sensitivity in the estimation of the tensor for the re-projection of the flaw in the third view. This phenomenon can be explained by the lower precision in correspondence for the regions in the three views, and by an error in the position of the center of mass for each segmented region. Furthermore, in some cases we were able to show that the estimation of the projection trifocal was correct although in the following image the process of segmentation did not generate flaws that corresponded with previous regions. Thus it was not possible to generate a triplet of flaws in correspondence. Let us remember that our trifocal analysis only allows pairs of flaws contained in the two first images of each sequence in order to estimate the projection in the third view

5 Conclusions

This investigation presents the development of a new flaw detection algorithm in manufactured goods, using a non-calibrated sequence of images. Using new AMVI methodology (Mery & Filbert, 2002) we have designed a novel system of automatic calibration based only on the spatial positions of the structures. The proposed approach uses the projection of the epipolar line, generated by the



(a) (b) (c)

Fig. 6: Generalized flaw estimation process: a) segmentation of hypothetical flaws in the first view, b) projection of the epipolar line in the second view using the robust fundamental matrix, c) projection of the coordinates of image 1 and 2, using trifocal tensors robustly over the third view.

fundamental matrix and the trifocal tensors in a robust manner, with the purpose of building a movement model without any a priori knowledge of structure. We based our investigation on the assumption that hypothetical flaws are real flaws if their positions, in a sequence of images, are in correspondence, because these remain stable in their position relative to the movement of the object.

With respect to the investigation carried out in (Mery & Carrasco, 2005), we have extended the analysis from two to three images per sequence through the estimation of trifocal tensors. Furthermore we have introduced new control points generated artificially through the use of B-Spline curves due to the low quantity of structures that remain stable in three images of a sequence.

Our results indicate that it is possible to generate an automatic model for a sequence of images which represent the movement between the points and the regions contained in these. The possibility of introducing non corresponding control points in triplets of images, is the principal advantage of the RANSAC algorithm for estimating, robustly, the fundamental matrix and the trifocal tensors. In this way we can use as reference points the edges of the structures or areas with no loss of information using a non-linear method. The principal advantage of our model is the automatic estimation of movement. Nonetheless, in some cases we saw some projection errors due to geometric distortions in the acquisition of images. Our future work is to reduce the number of false positives using supervised classification techniques through the use of neuronal networks (Mery, Silva, Caloba, & Rebello, 2003) and to add to the model geometrical distortion correction methods.

ACKNOWLEDGMENTS

This work was supported by FONDECYT – Chile under grant no. 1040210.

REFERENCES

Bartels, R., Beatty, J., & Barsky, B. (1998). *Bezier curves an introduction to splines for use in computer graphics and geometric modelling* (Vol. 10). San Francisco, CA.: Morgan Kaufmann.

Castleman, K. (1996). *Digital image processing*. New Jersey: Prentice-Hall, Englewood Cliffs.

Davis, E. (2005). *Machine vision* (3 ed.). Amsterdam: Morgan Kaufmann Publishers.

Faugeras, O., Luong, Q.-T., & Papadopoulo, T. (2001). *The geometry of multiple images: The laws that govern the formation of multiple images of a scene and some of their applications*. Cambridge MA, London: The MIT Press.

Fischler, M., & Bolles, R. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Haralick, R., & Shapiro, L. (1992). *Computer and robot vision*. New York: Addison-Wesley Publishing Co.

Hartley, R. I., & Zisserman, A. (2000). *Multiple view geometry in computer vision*. Cambridge University Press.

Jain, A. (1989). *Fundamentals of digital image processing*. Prentice Hall, Information and Systems Science Series.

Mery, D. (2003a). Crossing line profile: a new approach to detecting defects in aluminium castings. *Lecture Notes in Computer Science*, *2749*, 725–732.

Mery, D. (2003b). Explicit geometric model of a radioscopic imaging system. *NDT & E International*, *36*(8), 587–599.

Mery, D. (2003c, November). Exploiting multiple view geometry in X-ray testing: Part I, theory. *Materials Evaluation*, *61*(11), 1226–1233.

Mery, D., & Carrasco, M. (2005). Automated multiple view inspection based on uncalibrated image sequence. *LNCS*, *3540*, 1238–1247.

Mery, D., & Filbert, D. (2002, December). Automated flaw detection in aluminum castings based on the tracking of potential defects in a radioscopic image sequence. *IEEE Trans. Robotics and Automation*, *18*(6), 890–901.

Mery, D., Silva, R. da, Caloba, L., & Rebello, J. (2003). Pattern recognition in the automatic inspection of aluminium castings. *Insight*, *45*(7), 475–483.

Newman, T., & Jain, A. (1995). A survey of automated visual inspection. *Computer Vision and Image Understanding*, *61*(2), 231–262.

Shashua, A., & Werman, M. (1995). Trilinearity of three perspective views and its associated tensor. In *5th international conference on computer vision (iccv'95)* (p. 920). Boston MA.

Paper #2
Automatic Multiple View Inspection using Geometrical Tracking and Feature Analysis in Aluminum Wheels

Miguel Carrasco and Domingo Mery
miguel.carrasco@mail.udp.cl - dmery@ing.puc.cl

Abstract

The classic image processing method for flaw detection uses one image of the scene, or multiple images without correspondences between them. To improve this scheme, automated inspection using multiple views has been developed in recent years. This strategy's key idea is to consider as real flaws those regions that can be tracked in a sequence of multiple images because they are located in positions dictated by geometric conditions. In contrast, false alarms (or noise) can be successfully eliminated in this manner, since they do not appear in the predicted places in the following images, and thus cannot be tracked. This paper presents a method to inspect aluminum wheels using images taken from different positions by using a method called *automatic multiple view inspection* (AMVI). Our method can be applied to uncalibrated image sequences, therefore it is not necessary to determine optical and geometric parameters normally present in the calibrated systems. In addition, to improve the performance, we designed a false alarm reduction method in two and three views called Intermediate Classifier Block (ICB). The ICB method takes advantage of the classifier ensemble methodology by making use of feature analysis in multiple views. Using this method, real flaws can be detected with high precision while most false alarms can be discriminated.

Keywords: Automated inspection, tracking, flaw detection, X-ray imaging, nondestructive testing.

1 INTRODUCTION

Over the last 30 years the worldwide manufacturing market has faced heavy competition to produce higher quality products while actively reducing prices. This has led to great advances in the technology required for automating production processes but inspection and quality control problems have yet to be fully resolved. Due to these gaps in the industry, several automatic inspection techniques represent an area of high interest and active research. Traditionally, inspection and quality control

Miguel Carrasco is Escuela de Ingeniería Informática, Facultad de Ingeniería, Universidad Diego Portales, Av. Ejército 441, Santiago, and with the Computer Engineering Department at Pontificia Universidad Católica de Chile. Domingo Mery is with the Computer Engineering Department at Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860 (143), Santiago, Chile.

in manufacturing environments have been carried out by means of an intensive human visual inspection inserted into different phases of the production processes (Newman & Jain, 1995). The economic benefits represent some of the main reasons this kind of inspection is used. The investment cost to install and develop a specialized machine for inspection tasks is very high compared to the cost of training a human operator. Also, human visual inspection has the great advantage of adapting to unforeseen situations and is flexible when faced with any change in the objects' position, orientation or shape. This is because human beings have high cognitive and sensory abilities that allow them to carry out complex reasoning and inferences while inspecting the objects (Spencer, 1996).

Various studies have analyzed the performance of human inspection and its main defects, e.g. (C. Drury, 1992; Mital et al., 1998; Jacob et al., 2004; C. G. Drury et al., 2004)). According to them, there is a clear consensus that human inspection does not achieve 100% performance in the detection of defect-free products (error-free).Mital et al. (1998) determined various factors that affect the performance of manual inspections, such as the rhythm and complexity of the task, the time for inspection, fault density, inspection model, luminosity, inspection strategy, training, age, and gender. Other authors have indicated that human inspection has a maximum of 80% effectiveness (C. G. Drury et al., 2004). Human inspection does have constraints as well as multiple failures, it is (1) variable, inspection quality is not constant over time because it is dependent on fatigue and monotony caused by the work; (2) irregular, because it depends on the ability, experience and strategy for revision of each inspector; (3) slow, some industries have high production levels and require inspection at a high processing rate, however human inspection can require more time because handling and observation tasks have limiting factor, such as the speed of human operations; (4) tedious, because the inspection routine can be very repetitive which generates a lower concentration level due to the large number of objects that must be revised in a short period; (5) hazardous, because in some environments such as under water inspection, the nuclear industry, and the chemical industry, human inspection can be inviable due to the high risk inherent in those systems; (6) complex, the difference between a product with or without defects can be very subtle, and that is not always easily distinguishable by a human operator; (7) inaccessible, in some cases even access to the object to be inspected can be very complex because of the size of the product. All of these factors have lead industry to gradually replace human inspection with automatic visual inspection (AVI) methods which allow contact free inspections to be made of the object.

Since the introduction of AVI methods in the early 1980s (Jarvis, 1980; Chin & Harlow, 1982), several systems for quality inspection have been successfully developed using different image processing techniques. The main objective of AVI is to increase productivity ensuring high quality, reliability and consistency standards, i.e., rejecting most of the defective products and accepting all the defect-free products. AVI inspections normally require less time than inspections performed by human operators. Malamas, Petrakis, and Zervakis (2003) and Kumar (Kumar, 2008) have presented extensive reviews of various AVI technologies applied to the manufacturing processes of different products such as electronic components, textiles, glass, mechanical parts, integrated circuits (IC), etc. Despite their advantages, AVI methods in general also have the following problems. 1) They lack

precision in their performance because of the imbalance between undetected flaws (false negatives) and false alarms (false positives). 2) They are limited by time, the mechanical requirements for placing an object in the desired position can be time consuming. 3) They require high computer cost for determining whether the object is defective or not. 4) They generate high complexity in the configuration and lack of flexibility for analyzing changes in parts design. The issues outlined above show that AVI remains a problem open to the development of new applications.

In many AVI systems the use of one image to carry out quality inspection is sufficient. However, in other cases where the signal-to-noise ratio is low, the identification of real flaws with little contrast implies the appearance of numerous false alarms. It is precisely in these cases where multiple views can improve the inspection performance in the same way a human inspector uses his sight to see multiple parts of an object to evaluate its quality.

In this paper we aim to exploit the redundant information from multiple views that contain corresponding parts of the object. The information captured from different viewpoints can reinforce the diagnosis when a single image is insufficient. In order to discriminate real flaws from false alarms our system tracks every possible flaw. Only real flaws can be successfully tracked along an image sequence. A real flaw entails a spatio-spatiotemporaltemporal relation in different views where it appears while a false alarm corresponds to a random event allowing us to distinguish real flaws from other artifacts. Based on this observation, we propose a three-step methodology for detecting real flaws in uncalibrated image sequences of aluminum wheels: segmentation of potential flaws, computation of corresponding points, and tracking of potential flaws with intermediate classifiers. Similar ideas have been presented in (Spicer, Bohl, Abramovich, & Barhak, 2006; Mery & Filbert, 2002a; Carrasco & Mery, 2006, 2007; Pizarro, Mery, Delpiano, & Carrasco, 2008). The main differences between this contribution and those works lie in the fusion of multiple view geometry and a statistical analysis of each flaw aiming to reduce the number of false alarms while simultaneously improving the true flaws detection in correspondence.

It is important to highlight that our method does not require a calibration process. In general, the calibration process is difficult to carry out in industrial environments due to vibrations and random movements that vary with time. The vibrations of the imaging system induce inaccuracies in the estimated parameters of the multiple view geometric model. Thus, the calibration is not stable and the imaging system must be re-calibrated periodically. In many cases it might be an extremely complicated procedure for real-time applications and manufacturing systems that cannot be stopped temporarily for calibration purposes (Pizarro et al., 2008).

The rest of the paper is organized as follows: Section 2 includes a brief discussion of automatic multiple visual inspection; Section 3 explains our proposed method for uncalibrated image sequences; Section 4 shows the experimental results; and finally, Section 5 presents the conclusions and future work.

2 AUTOMATIC VISUAL INSPECTION

Currently, one of the most widely used flaw detection systems in industry is the X-ray inspection, extensively used by the automotive and aerospace industry, for detecting flaws like: porosity, cracks,

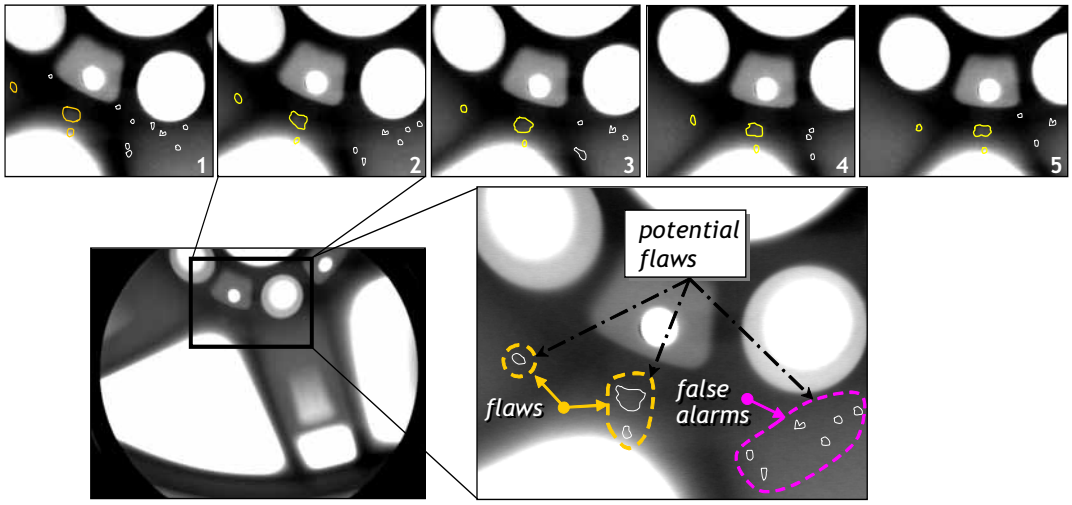


Fig. 1. Top: Example of a radioscopic sequence of five images after the segmentation step. Bottom: Example of false alarms and flaws, called potential flaws at this stage.

corrosion, inclusions, debris, bubbles, and thickness variations, among others (Nguyen, Noble, Mundy, Janning, & Ross, 1998; Filbert, Klatte, Heinrich, & Purshke, 1987; Boerner & Strecker, 1988). It is commonly used because the X-ray attenuation surrounding the flaws is less (or more). The use of X-rays exploits the fact that most material flaws are not visible. However, even in radioscopic images the signal-to-noise ratio (SNR) is low, the flaw signal is slightly greater than the background noise meaning that the identification of real flaws with poor contrast can involve detection of false alarms as well.

Motivated by (human) visual inspections that are able to differentiate between flaws and noise by looking at the objects being tested in motion, a new method of automated inspection was developed using sequences of multiple images (Mery & Filbert, 2002a). The new inspection methodology called *Automated Multiple View Inspection* (AMVI) uses redundant views to perform the inspection task. This novel methodology is opening up new possibilities in the inspection field, mainly by taking into account the useful information in corresponding different views of potential flaws in the test object. The main idea is to consider as false alarms those potential flaws that cannot be tracked in a sequence of multiple images. Therefore, two or more views of the same object taken from different viewpoints can be used to confirm and improve the diagnosis made by analyzing only one image. AMVI has been developed under two schemes: calibrated and uncalibrated. The calibrated scheme uses a 3D calibration object to estimate corresponding points (Mery & Filbert, 2002a). Alternatively, the uncalibrated scheme automatically establishes the correspondences from the information contained in the images through a robust correspondence system (Carrasco & Mery, 2006) (see Fig.2). These steps are equivalent to the work done by an inspector. First, all the possible regions that might contain flaws (or potential flaws) are detected. Second, because of the large number of false alarms that can occur in the identification step, corresponding positions that each

flaw might have in the following views are analyzed, using multiple view tracking (see Fig.1). Both methods share the following two steps: *identification* and *tracking*.

Identification aims at detecting all the anomalous regions or potential flaws in each image of an object's motion sequence, without a priori knowledge of its structure. There are two general features used to identify them: i) a flaw is considered as a connected subset in the image, ii) the differences between the gray levels of the flaw and its neighbors is considerable. Although there are a lot of false alarms detected by this process, the detector has the following advantages: i) the same detector is applied to all the images; ii) it allows for the identification of potential flaws regardless of the position or the structure of the object under study; in other words, without a priori knowledge of the design of the structure; iii) the detection of real flaws is very high (better than 90%)[1]. The process that follows extracts features of each potential flaw after identifying these regions in the previous procedure. This information makes it possible to determine whether a flaw is corresponding in the multiple view analysis, according to the new intermediate classification method.

Tracking aims at "chasing", in subsequent images of a sequence, potential flaws detected in the first step using the positions forced by the geometric restrictions in multiple views (Hartley & Zisserman, 2000). If a potential flaw continues through an image sequence, it is identified as a real flaw and the object is classified as defective. However, if a potential flaw does not have a correspondence in the sequence, it will be considered as a false alarm (details in the segmentation in Fig.1). A similar idea is also used by radiologists that analyze two different X-ray views of the same breast to detect cancer in its early stages. Thus, the number of cancers flagged erroneously as well as missed cancers may be greatly reduced (see for example Kita, Highnam, and Brady (2001), where a novel method that automatically finds correspondences in two different views of the breast is presented).

3 Proposed Method

In this section we provide an explanation of the stages in the uncalibrated AMVI process with intermediate classifiers. The proposed scheme has four major steps (A, B, C and D) detailed in Fig.2. They correspond to the following stages: (A) identification of potential flaws, (B) extraction of control points, (C) tracking, and (D) intermediate classifier block.

A) Segmentation of potential flaws: Numerous investigations have been carried out to segment flaws depending on the product analyzed (Mery, Silva R. R., Calôba, & Rebello, 2003; Mery, 2003; Campbell, Fraley, Murtagh, & Raftery, 1997; Pedreschi, Mery, Mendoza, & Aguilera, 2004). Here we used the segmentation and feature extraction method, described in (Mery, 2003), with the aim of identifying multiple regions which may correspond to real flaws. In particular, the segmentation algorithm used is able to detect most real flaws as well as numerous false alarms. The process consists of the following, each potential flaw extracts a set of measurements (described in Table 1)

1. See (Mery, 2003) for details on the computation of the segmentation algorithm.

TABLE 1
Features extracted from the identification step

Symbol	Feature and Description
A	*Area*: Number of pixels that belong to the region
G	*Mean of the grey*: Mean of the grey values that belong to the region (Mery & Filbert, 2002b)
D	*Mean of the second derivative*: Mean of the second derivative values of the pixels that belong to the boundary of the region (Mery & Filbert, 2002b)
F_1	*Crossing line profiles*: The grey level profiles along straight lines crossing each segmented potential flaw in the middle. The profile that contains the most similar grey levels in the extremes is defined as the best crossing line profile (BCLP). Feature F1 corresponds to the first harmonic of the fast Fourier transformation of BCLP (Mery, 2003)
$K\sigma$	*Contrast*: Standard deviation of the vertical and horizontal profiles without offset (Mery & Filbert, 2002b)
r	*High contrast pixels ratio*: Ratio of number of high contrast pixels to area (Mery, 2006)

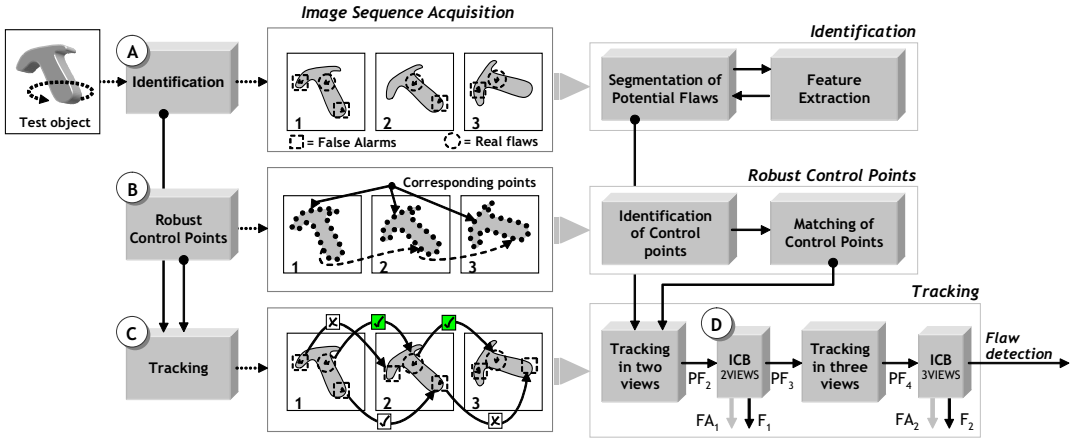


Fig. 2. General block diagram of the Uncalibrated Automatic Multiple View Inspection (AMVI) method with the three phases: Identification, Robust Control Point and Tracking of potential flaws in two and three views with the Intermediate Classifier Block (ICB) method.

and stores them in a normalized feature vector. For instance, let $m_a^i = [x_a^i, y_a^i, 1]$ be the centre of mass stored in homogenous coordinates of the segmented region i in the a-th view, and let $\mathbf{v}_a^i$ be the feature vector of the region i in the a-th view. As a result, numerous potential flaws appear as observed in the segmented image (Fig.1).

B) Robust control points:

As stated before, our final goal is tracking real flaws in an image sequence. For this purpose, accurate corresponding points between every pair of views are required. In general, the estimation of control points can be solved by various mechanisms that use the intrinsic information of the

structures after a process of segmentation, edges extraction, normalization, and smoothing (Liu & Srinath, 1990; Sebastian, Klein, & Kimia, 2003). In general, there are two curve alignment categories: methods based on rigid transformations (Umeyama, 1993) and methods based on non-rigid deformation (Cohen, Ayache, & Sulger, 1992). First, methods based on rigid transformations determine the control points by estimating the rotation, lineal displacement, and scaling parameters (Liu & Srinath, 1990). However, due to the rigidity assumption they are sensitive to occlusions, deformations, articulations, perspective projections, and other variations of the edges (Sebastian et al., 2003). Second, methods based on non-rigid deformations try to match one curve over the other. The goal is minimizing a function of elasticity through the transformation of the curve flexion, orientation or stretching. Generally, this transformation is not invariant under rotation and scaling (Gdalyahu & Weinshall, 1999), it is very sensitive to noise because it is defined in terms of the curvature, and it requires the evaluation of second order derivatives (Sebastian et al., 2003).

Our investigation proposes a simple and effective curve alignment method by minimizing the Pearson's correlation coefficient using an isometric transformation between two curves. We use this scheme because in the analysis of manufactured products the object being analyzed is usually not deformable. This premise justifies the use of a rigid transformation method with which, given a rotation and a lineal displacement, it is possible to estimate a correspondence between the object's control points. However, due to the object's rotation, some regions can remain occluded, and therefore the proposed system must consider that only some regions retain this transformation. The proposed robust system of control points consists of two stages that are detailed below: matching of regions, and matching of control points.

B.1) Matching of regions: This consists of establishing correspondences between regions of each view and not between control points. The designed process is composed by four stages: First, segmentation of those regions in which the intensity of the object is distinguishable from the background by using Otsu's method (Haralick & Shapiro, 1992) (Fig.3a). Second, extraction of a set of features for each segmented region. This consists of extracting the moments of Flusser-and-Suk (Sonka, Hlavac, & Boyle, 1999) of each region in three views. Third, determination of a region-correspondence using the features extracted before by relating those regions with greater similarity. The similarity relation is fulfilled when two or three regions have little variation in their normalized features according to the Euclidean distance metric (Fig.3b). Fourth, smoothing the edges of each region in correspondence, in order to decrease the noise of each curvature. For that we calculate the perimeter of each segmented region and generate a list in a parametric form as $Z_s = [x_s, y_s]$, where $s = 0, \ldots, L-1$ is the index of the list of pixels ordered in a turning direction, and L is the number of pixels of the region's perimeter. Using this parametric form, we generate the Fourier descriptors, transforming the Z_s coordinates into a complex value $u_s = x_s + j \cdot y_s$. This signal with period L is transformed into the Fourier domain by means of a discrete Fourier transform (DFT):

$$F_n = \sum_{s=1}^{L-1} u_s \cdot e^{-j \cdot \frac{2\pi \cdot s \cdot n}{L}}$$

The modulus of the complex Fourier coefficients describes the energy of each descriptor. Therefore, if we choose the highest energy coefficients (above 98%) and return to real space with the inverse discrete Fourier transform (IDFT) we get a smoother curve with less noise. This transformation produces the same number of points as the original curve. Likewise, the spacing between the original points remains constant. However, when applying the elimination of some Fourier coefficients, the original curve is transformed into a new curve $C_s = [x'_s, y'_s]$, where, $C_s \neq Z_s$.

B.2) Matching of control points: The estimation of control points is a process in which the correspondence of pairs-points on the border of a region is established (Fig.3c). Using Fourier procedure as described above, we define a curve C_1 corresponding to a region in the first view, and a curve C_2 corresponding to C_1 in the second view. Both curves do not have the same length because they correspond to the perimeter of corresponding regions. However, these regions have an isometric transformation, and in cases of occlusion the curves will have different sizes. For both curves, to keep the same distance and to be aligned, it is necessary to select a section of equal length from each list. Let P, a section of curve C, be such that $P = C(\delta)$, where $\delta = [s_i, \cdots, s_j]$, for $i, j \in [1, \cdots, n]$. In this way there is a section P_1 in the first view that has the same length as section P_2 in the second view. These sections of the curve do not necessarily have a correspondence, and for that we define a shift operator $\Theta(P, \lambda)$ that displaces the list P by λ positions in a turning

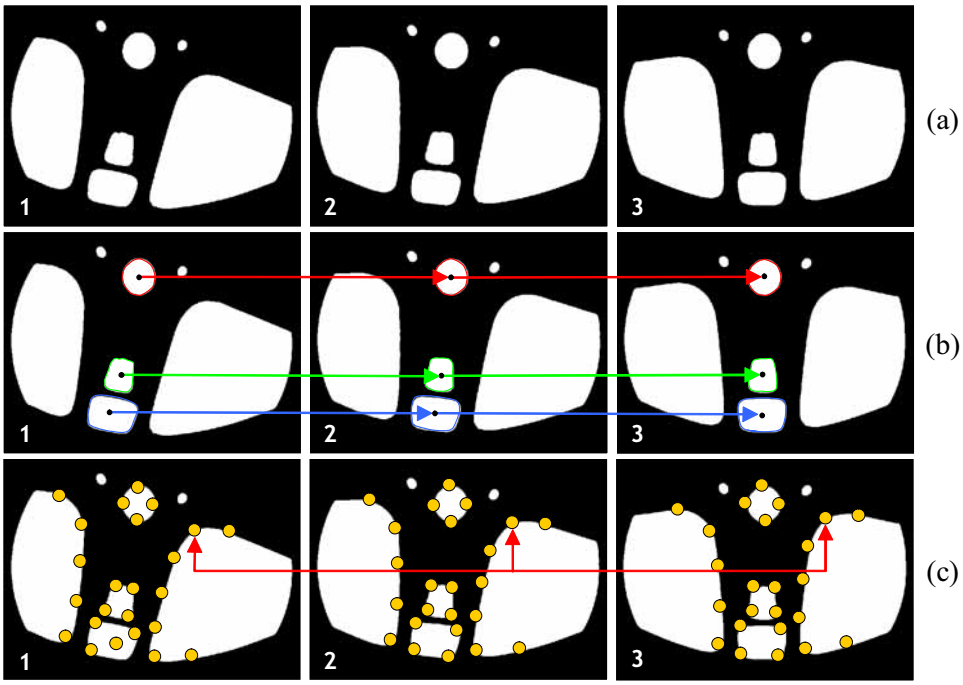


Fig. 3. Matching of regions: (a) Segmentation of regions in a sequence by Otsu's method. (b) Correspondence between regions according to a similarity criterion between the extracted features. (c) Matching of control points on the border (i.e., the curve) of a region.

direction. Operator Θ uses the function "mod" (modulus after division) to determine the λ relative positions that list C, of length P, must turn.

Using the above definitions, we implemented an alignment function $\mu(\Omega)$ as the maximization of the Pearson's correlation coefficient $\rho(\alpha, \beta)$ (Dunn & Clark, 1974) between the isometric transformation of a section of P_1, with the shift of section P_2 with a jump λ, composed by four parameters $\Omega = \{\theta, \Delta s_x, \Delta s_y, \lambda\}$

$$\mu(\Omega) = |1 - \rho\left([R, t][P_1], \Theta(P_2, \lambda)\right)| \rightarrow \min \tag{1}$$

where,

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad t = \begin{bmatrix} \Delta s_x \\ \Delta s_y \end{bmatrix} \tag{2}$$

The minimization of $\mu(\Omega)$ must find Ω parameters to estimate an alignment between sections P_1 and P_2. The main advantage of this function is that it does not require a perfect alignment because the correlation coefficient takes a maximum if the displacement is linear. Another advantage is that curves P_1 and P_2 are open, the alignment determines only sections that are corresponding, allowing control points to be obtained for curves that have partial occlusion in corresponding regions. Also, the use of the parameter λ allows finding a position relation for curve C_2 with P_1, and in this way, while curve P_2 adjusts its shift, curve P_1 adjusts its lineal displacement and rotation angle to become aligned.

C) Tracking of potential flaws: In the previous steps we have segmented all potential flaws along an image sequence and we have established corresponding points in a sequence. We now turn to the problem of separating real flaws from false alarms. The essential point is that only real flaws can be tracked along an image sequence. A real flaw entails a spatiotemporal relation in different views where it appears, while a false alarm corresponds to a random event.

C.1) Two views:

If a potential flaw $\mathbf{m}_a^i$ in view a is actually a real i-flaw it must have a corresponding point $\mathbf{m}_b^j$ in another consecutive view b where a potential flaw j was also segmented. According to the *principle of multiple view geometry* (Hartley & Zisserman, 2000), points $\mathbf{m}_a^i$ and $\mathbf{m}_b^j$ are in correspondence if matrix $\mathbf{F}_{a,b}$ exists such that

$$\mathbf{m}_b^{j\top} \cdot \mathbf{F}_{a,b} \cdot \mathbf{m}_a^i = \mathbf{m}_b^{j\top} \cdot \begin{bmatrix} f_1 & f_2 & f_3 \\ f_4 & f_5 & f_6 \\ f_7 & f_8 & f_9 \end{bmatrix} \cdot \mathbf{m}_a^i = 0 \tag{3}$$

where $\mathbf{F}_{a,b}$ is known as the fundamental matrix of the projection of points $\mathbf{m}_a^i$ and $\mathbf{m}_b^j$, in homogenous coordinates as $[x_a^i, y_a^i, 1]^\top$ and $[x_b^j, y_b^j, 1]^\top$, respectively. Once the set of corresponding positions has been generated in both views, we use the method proposed by Chen, Wu, Shen, Liu, and Quan (2000) to make an initial estimation of the fundamental matrix. For the sake of completeness, we briefly describe the Chen's method below.

The method proposed by Chen et al. is based on choosing a subset of candidate points by exploiting the fact that the rank of the fundamental matrix is two and the value of the epipolar restriction is zero ($\mathbf{m}_b^{j\top} \cdot \mathbf{F}_{a,b} \cdot \mathbf{m}_a^i = 0$). Based on this, it is possible to define the fundamental matrix without losing generalization by means of a lineal combination of its values, where it is only necessary to estimate four parameters $\Gamma = \{\alpha, \beta, \alpha', \beta'\}$ according to

$$\begin{aligned} f_3 &= -\alpha' f_1 - \beta' f_2 \\ f_6 &= -\alpha' f_4 - \beta' f_5 \\ f_7 &= -\alpha f_1 - \beta f_4 \\ f_8 &= -\alpha f_2 - \beta f_5 \\ f_9 &= -\alpha'\alpha f_1 - \alpha'\beta f_4 + \beta'\alpha f_2 + \beta'\beta f_5 \end{aligned} \tag{4}$$

Using the above parameters and the point correspondences detected in the control point step, we define a new problem as $\mathbf{A} \cdot \mathbf{f} = 0$, where $\mathbf{f} = [f_1, f_2, f_4, f_5]^\top$. Since the values of $\mathbf{f}$ are not null, only $\det(\mathbf{A}) = 0$ is possible. Likewise, to find a solution of matrix $\mathbf{A}$, the Γ parameters are defined randomly. Thus, it is necessary to build a matrix $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix} (x_1 - \alpha)(x_1' - \alpha') & \dots & (x_n - \alpha)(x_n' - \alpha') \\ (x_1 - \alpha)(y_1' - \beta') & \dots & (x_n - \alpha)(y_n' - \beta') \\ (y_1 - \beta)(x_1' - \alpha') & \dots & (y_n - \beta)(x_n' - \alpha') \\ (y_1 - \beta)(y_1' - \beta') & \dots & (y_n - \beta)(y_n' - \beta') \end{bmatrix}^\top_{4 \times N} \tag{5}$$

Since the matrix $\mathbf{A}$ is composed by N stereo correspondences, Chen proposed that using only a combination of four restrictions will be enough to seek a fundamental matrix. For this reason, we define a row vector $\mathbf{B}_i$ of the matrix $\mathbf{A}$ as

$$\mathbf{B}_i(\alpha, \beta, \alpha', \beta') = \begin{bmatrix} (x_i - \alpha)(x_i' - \alpha') \\ (x_i - \alpha)(y_i' - \beta') \\ (y_i - \beta)(x_i' - \alpha') \\ (y_i - \beta)(y_i' - \beta') \end{bmatrix}^\top . \tag{6}$$

Then, choosing randomly four restrictions (i, j, k, l) between N rows, the problem that follows must fulfill the following condition

$$\mathbf{B}_{ijkl} = \det\left(\begin{bmatrix} \mathbf{B}_i & \mathbf{B}_j & \mathbf{B}_k & \mathbf{B}_l \end{bmatrix}^\top \right) = 0 \tag{7}$$

where rows (i, j, k, l) belong to the combinatorial subset of $R = C_4^N$. As we stated before, to estimate the Γ parameters it is necessary to solve the optimization problem by looking for four $\mathbf{B}_i$ restrictions such that (7) is a minimum according to

$$(\alpha, \beta, \alpha', \beta') = \text{argmin}_{\alpha,\beta,\alpha',\beta'} \sum_{1}^{N} |\mathbf{B}_{ijkl}| \tag{8}$$

Next, replacing the Γ parameters in (5) allows us to compute the vector $\mathbf{f}$ by decomposing into singular values $[\mathbf{S}, \mathbf{V}, \mathbf{D}] = \text{svd}(\mathbf{A})$.

Thus, the solution of (5) corresponds to the last column vector $\mathbf{D}$. Finally, the Γ parameters obtained in (8) are replaced in (4) defining an initial solution of $\mathbf{F}_{a,b}$. In our research, we combine this procedure with the RANSAC algorithm, thus the solution with the largest number of inliers is used to compute the fundamental matrix.

The next procedure is to estimate the epipolar line based on previous results. More formally, let $\mathbf{l}_a^i$ be the epipolar line defined as the product between the fundamental matrix $\mathbf{F}_{a,b}$ and the point $\mathbf{m}_a^i$ as

$$\mathbf{l}_b^i = [l_{b,x}^i, l_{b,y}^i, l_{b,z}^i] = \mathbf{F}_{a,b}^\top \cdot \mathbf{m}_a^i \,, \tag{9}$$

where $\mathbf{l}_b^i$ is the epipolar line of flaw i in view b, $\mathbf{m}_a^i$ is the centre of mass of flaw i in view a and $[l_{b,x}^i, l_{b,y}^i, l_{b,z}^i]$ are the coefficients of the epipolar line. Once the epipolar line $\mathbf{l}_b^i$ has been generated, it is necessary to determine the distance between corresponding flaw in view b. This distance is determined through the *practical bifocal constraint* (Hartley & Zisserman, 2000). Given that the epipolar constraint is applied to points and not to regions, we consider the centre of mass of each potential flaw to be a corresponding point between pairs of images. Using the epipolar line $\mathbf{l}_a^i$, we identify the correspondence associated with the potential flaw i as the potential flaw j in view b that satisfies the constraints

$$\frac{|\mathbf{m}_b^{j\top} \mathbf{F}_{a,b} \mathbf{m}_a^i|}{\sqrt{(l_{b,x}^i)^2 + (l_{b,y}^i)^2}} < \varepsilon_1 \tag{10}$$

for small $\varepsilon_1 > 0$. If this constraint is fulfilled, a potential flaw is thus found in two views. In this case it could be considered a real flaw with a bifocal correspondence. Otherwise, it is regarded as a false alarm. An example is shown in Fig. 4. The same procedure is applied to every potential flaw in view a that is to be found in view b. It is important to recall that the precision of the fundamental matrix allows the correct determination of potential flaws along the length of the epipolar line. However, if the fundamental matrix is not robust, the epipolar line will be incorrect and the subsequent trifocal tensor process will fail

C.2) Three views: Trifocal analysis allows modeling all the geometric relationships in three views, and is independent of the structure contained in each image (Hartley & Zisserman, 2000). The tensor, a matrix structure similar to the fundamental matrix, only depends on the movement between images and the internal parameters of the cameras. Its main advantage is that it can be calculated from the correspondences of the images without any a priori knowledge of the movement or calibration of the object. This characteristic justifies it because the estimation of the fundamental matrix does not always eliminate all false positives.

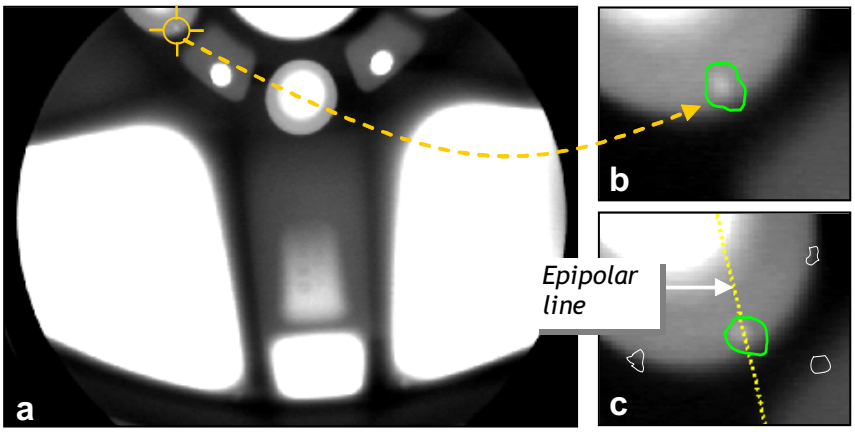


Fig. 4. Epipolar line generated from the fundamental matrix: a) First view. b) Segmentation of a potential flaw in the first view. c) Intersection of the epipolar line in the second view with one or more corresponding potential flaws.

Based on previous results, to confirm that a bifocal correspondence indeed represents a real flaw, we try to discover a new correspondence using a third view with the help of trifocal tensors. Let $\mathbf{T} = (T_t^{rs})$ be a $3 \times 3 \times 3$ matrix representing the trifocal tensor that encodes the relative motion among views a, b, c.[2] Then, we can estimate the hypothetical position of a flaw k in a third view c using the correspondences $\mathbf{m}_a^i, \mathbf{m}_b^j$ and the tensor $\mathbf{T}$ as[3]

$$\widehat{\mathbf{m}}_c^k = \frac{1}{\mathbf{m}_a^{i\top}(T_\cdot^{13} - x_b^j T_\cdot^{33})} \begin{bmatrix} \mathbf{m}_a^{i\top}(T_\cdot^{11} - x_b^j T_\cdot^{31}) \\ \mathbf{m}_a^{i\top}(T_\cdot^{12} - x_b^j T_\cdot^{32}) \\ \mathbf{m}_a^{i\top}(T_\cdot^{13} - x_b^j T_\cdot^{33}) \end{bmatrix} . \tag{11}$$

We compare the estimated position with all potential flaws of view c, regarding the potential flaw k as a real flaw if

$$\|\widehat{\mathbf{m}}_c^k - \mathbf{m}_c^k\| < \varepsilon_1 \,, \tag{12}$$

for small $\varepsilon_1 > 0$. If this constraint is fulfilled, we take the potential flaw to be a real flaw, since it complies with the correspondence in three views. Should the potential flaw in the third view not coincide with the projection of the tensor, it is discarded, as it does not fulfill the trifocal condition. In general, given that the trifocal condition is analyzed for the sequences that fulfill the bifocal condition, we reduce the number of false positives generated in two views.

D) Intermediate Classifier Block method:

The goal of the ICB method is to eliminate those correspondences between potential flaws

2. See (Hartley & Zisserman, 2000) for details on the computation of the trifocal tensors.
3. The estimated projection in the third view can be improved applying the point-line-point method proposed in (Hartley & Zisserman, 2000, pp.373).

that have a low probability of being true positives. The ICB method uses the classifier ensemble methodology (Polikar, 2006), in which multiple linear classifiers do the classification and then, through the majority of votes technique, a final decision is made. According to the multiple view hypothesis, the key idea is to consider as false alarms those potential flaws that cannot be tracked in a sequence of multiple images. Nonetheless, there are false alarms that fulfill the above condition and must be eliminated in the multiple views analysis using a partial elimination classification system. The ICB method has as input the distribution of two classes: flaws (F) and false alarms (FA). According to this distribution, the classifier must determine the region of space where there are actually flaws only starting from point θ_F, and false alarms from θ_{FA} (Fig.5a). Once these regions are extracted only flaws or false alarms (which the classifier cannot verify with high probability the class to which they belong) are assigned to a new class called Potential Flaw (PF) (Fig.5b). As a result, the ICB method generates the separation of three classes (F, FA and PF). Accordingly, flaws or false alarms contained in the PF region are used as new potential flaws in the following step of the multiple views analysis. This reduction avoids the analysis of the trajectories of all flaws in correspondence; consequently, improving the performance. The simplest form of the previous classifier is reflected in the linear separation of the F, FA and PF regions, using the V_1 and V_2 features (Fig.5b). The methodology used by the ICB is composed by a series of stages detailed below (see Fig.6).

D.1) Assessment method of the ICB classifier: Our problem falls within the framework of supervised classification problems, since the class which each potential flaw belongs to is known. Using this information, the classification model is designed by means of the cross-validation method. To compare the results of the various configurations of the classifier we use the ROC curve (Egan, 1975). The main advantage of the ROC curve is that it allows the comparison to be independent of the sample. For this research, the classes are a set of registers with flaws and false alarms, we

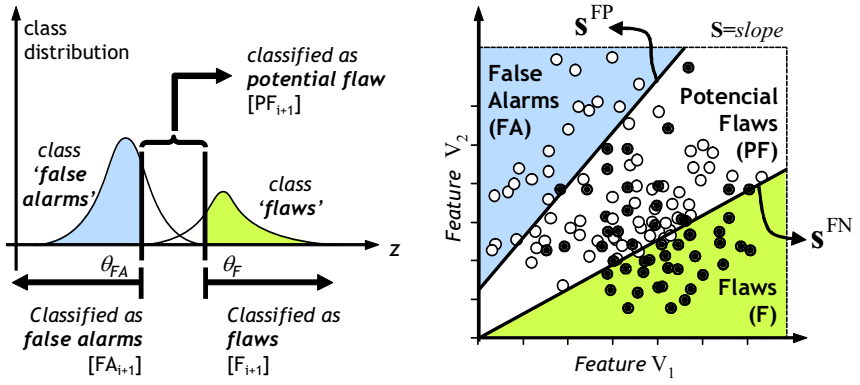


Fig. 5. (a) Distribution of classes of potentials flaws between views. (b) Distribution of three classes in two dimensions with a linear separation between F, FA and PF regions.

determine the sensitivity (S_n) and 1-specificity $(1-S_p)$ as $S_n = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ and $S_p = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$, where TP is the number of true positives (classified correctly as flaws), TN is the number of true negatives (classified correctly as false alarms), FP is the number of false positives (classified incorrectly as flaws), and FN is the number of false negatives (classified incorrectly as false alarms). The objective is for the sensitivity to be maximum (100%) and simultaneously the 1-specificity to be minimum (0%), this way the classifier guarantees an ideal classification for two classes. In practice this is difficult to achieve because it depends on the classifier's internal parameters which can vary with respect to the noise existing in the data.

D.2) Selection of features: The features selected by the ICB classifier are determined automatically using the information contained in each potential flaw, each of which has an associated feature vector $\mathbf{v}$. Each feature vector is composed of twelve measures extracted previously in the segmentation step (see details in Table 1). Here we used the Take-L-Plus-R feature selection algorithm (Duda, Hart, & Stork, 2001) to determine a combination of features that separate the classification space. The objective of this algorithm is to determine the best features that allow greater separation between the classes space[4].

In the multiple views analysis above, we combined multiple flaws only by means of a geometric analysis. In this next step we carry out a fusion of its features to seek a separation of its classes. For instance, let $\mathbf{v}_a^i$ be a feature vector of flaw i in view a and let $\mathbf{v}_b^j$ be a feature vector of flaw j in view b. If these two regions are corresponding, in theory the distance of their features should be short, since both regions are the same in two views, otherwise if there are false alarms the distance between them should be longer. With this simple criterion we define a unique vector of features $\mathbf{v}_{ab}^{ij}$ that relates two regions for the bifocal case and $\mathbf{v}_{abc}^{ijk}$ which relates three regions in the trifocal

4. As a criterion function, we used the Fisher discriminant (Stearns, 1976).

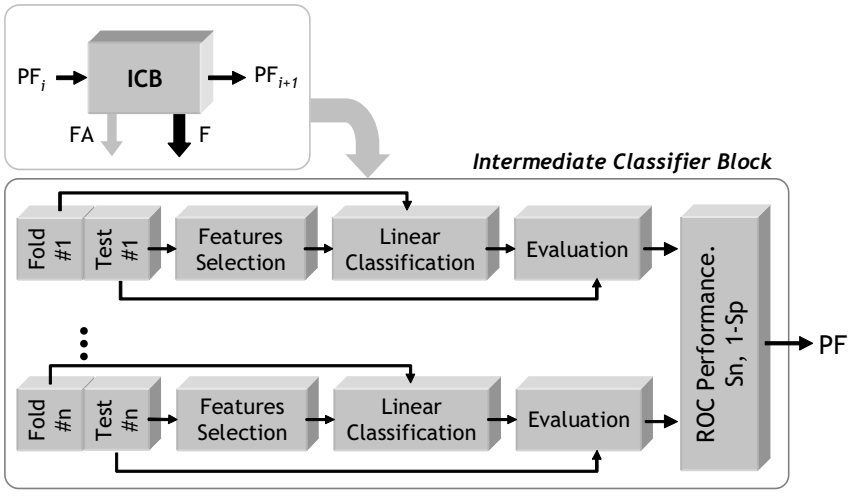


Fig. 6. General model of the internal process of the ICB linear classifier with cross validation selection and automatic feature selection by means of Fisher discriminant.

case as

$$\begin{aligned} \mathbf{v}_{ab}^{ij} &= \left[(\mathbf{v}_a^i - \mathbf{v}_b^j)\right] \\ \mathbf{v}_{abc}^{ijk} &= \left[(\mathbf{v}_a^i - \mathbf{v}_b^j)(\mathbf{v}_a^i - \mathbf{v}_c^k)(\mathbf{v}_b^j - \mathbf{v}_c^k)\right] \end{aligned} \tag{13}$$

For the next analysis, and for simplicity, we define matrix $\mathbf{v}$ as $\mathbf{v} \equiv \mathbf{v}_{ab}^{ij} \equiv \mathbf{v}_{abc}^{ijk}$, because the feature vector contains the relations between two and three regions and both elimination processes are independent (see ICB process in Fig.2).

D.3) Linear classification: We use a linear discriminative analysis (LDA) classification (Duda et al., 2001) that allows finding the hyper-planes that best separate the solution space. For that, the classification process must fit the following linear equation

$$\mathbf{w}^{\mathsf{T}} \cdot \mathbf{v} + w_0 > 0\,, \tag{14}$$

where $\mathbf{w} = \Sigma_{\mathrm{w}}^{-1} \cdot (\bar{v}_1 - \bar{v}_2)$ are the hyper-plane parameters, Σ_{w} is the interclass covariance matrix, and $\mathbf{v}$ corresponds to the feature vector chosen earlier. Finally, the value w_0 for two classes is determined according to the mean of features $\bar{v}_1$ and $\bar{v}_2$ and the probabilities of each class p_{e1} and p_{e2} according to

$$w_0 = -\frac{1}{2} \cdot (\bar{v}_1 + \bar{v}_2) \cdot \Sigma_{\mathrm{W}}^{-1} \cdot (\bar{v}_1 - \bar{v}_2) - \log\left(\frac{p_{e1}}{p_{e2}}\right) \tag{15}$$

Once an initial solution is obtained for the vector $\mathbf{w}$, the optimization problem tries to fit the hyper planes so that (16) is the maximum, this ensures that we are obtaining a high performance of (S_n) and $(1 - S_p)$ for each sub-selection of its features.

$$\{\mathbf{w}, w_0\} = \arg\max\{S_n(\mathbf{w}, w_0)\} \;\; \textit{s.t.} \;\; S_p(\mathbf{w}, w_0) = 1 \tag{16}$$

This problem has been solved by the Nelder-Mead Simplex method (Lagarias, Reeds, Wright, & Wright, 1998). Then the information from the selected straight lines and features is used to evaluate the performance of the classifier on the test data.

D.4) Joint classification :

The linear discrimination analysis model has been used together with the cross-validation technique. This way the optimization process generated by each combination of features generates, as a result, a set of straight lines specific for each combination. Finally, this model is used for the testing data by the classifier, therefore a set of weak classifiers makes it possible to generate a robust classification. For example, let us assume that we have used a set C consisting of three features in the training phase $C = \{V_1, V_2, V_3\}$. The separation between them generates a three-dimensional volume bounded by the cuts of the two-dimensional separations, containing only potential flaws (PF) (Fig.7). This three-dimensional volume generated from the combination of the two-dimensional features $[\mathrm{V}_1, \mathrm{V}_2]$, $[\mathrm{V}_1, \mathrm{V}_3]$ and $[\mathrm{V}_2, \mathrm{V}_3]$ contains potential flaws. Conversely, the space outside the three-dimensional volume could be flaws or false alarms, depending on the

position in which the hyper-planes are projected.

Our final classification method is based on the use of multiple linear separation models. The objective of the linear separation is to find a dividing line for two classes, but we use the same LDA algorithm with two purposes: First, to find the best separation line that minimizes the FPs subject to $S_n = 1$ defined as $\mathbf{s}^{\mathrm{FP}}$. Second, to find the best separation line that minimizes the FNs subject to a $S_p = 0$, defined as $\mathbf{s}^{\mathrm{FN}}$ (see slopes in Fig.5b). These two separation lines generate a two-dimensional separation space, and the total set of combinations of features generates a hyper-plane.

Firstly, we calculate the separation line set in order to evaluate the joint classifier in the testing data for each combination pair. Thus, for the $\mathrm{V_m}, \mathrm{V_n}$ combination we generate a linear separation between the straight lines $\mathbf{s}^{\mathrm{FP}}_{m,n}$ and $\mathbf{s}^{\mathrm{FN}}_{n,n}$. More formally, let $PF_{m,n}$, $FA_{m,n}$ and $F_{m,n}$ be the space generated by the linear intersection between the features m and n, defined as

$$\begin{aligned} PF_{m,n} &= \begin{bmatrix} \mathbf{v} & 1 \end{bmatrix} \cdot \mathbf{s}^{\mathrm{FP}}_{m,n} < 0 \wedge \begin{bmatrix} \mathbf{v} & 1 \end{bmatrix} \cdot \mathbf{s}^{\mathrm{FN}}_{m,n} > 0 = \{0, 1\} \\ FA_{m,n} &= \begin{bmatrix} \mathbf{v} & 1 \end{bmatrix} \cdot \mathbf{s}^{\mathrm{FP}}_{m,n} < 0 \wedge \begin{bmatrix} \mathbf{v} & 1 \end{bmatrix} \cdot \mathbf{s}^{\mathrm{FN}}_{m,n} < 0 = \{0, 1\} \\ F_{m,n} &= \begin{bmatrix} \mathbf{v} & 1 \end{bmatrix} \cdot \mathbf{s}^{\mathrm{FP}}_{m,n} > 0 \wedge \begin{bmatrix} \mathbf{v} & 1 \end{bmatrix} \cdot \mathbf{s}^{\mathrm{FN}}_{m,n} > 0 = \{0, 1\} \end{aligned} \tag{17}$$

The next step is to verify the classification for each feature vector $\mathbf{v}$ with unknown values. To that end, let $N = C_2^P$ be the number of possible combinations of a feature vector P. For every vector of length P we generate N combinations of two features, therefore the classification result requires a set of N results, and then by simple majority vote, the final classification is evaluated. Let us assume that $\mathbf{v_1}$ is the first feature vector and we want to find its classification. Its result for each class $\{PF, FA, F\}$ is defined by

$$\mathbf{M}(\mathbf{v_1}) = \begin{bmatrix} PF_{1,2} & FA_{1,2} & F_{1,2} \\ \vdots & \vdots & \vdots \\ PF_{m,n} & FA_{m,n} & F_{m,n} \\ \vdots & \vdots & \vdots \\ PF_{N-1,N} & FA_{N,N-1} & F_{N,N-1} \end{bmatrix}^{\top}_{N \times 3} \tag{18}$$

where the matrix $\mathbf{M}$ is the binary outcome of multiple partial classifications for the feature vector $\mathbf{v_1}$. Finally, the classification of the matrix $\mathbf{M}(\mathbf{v_1})$ is defined as

$$p(class|\mathbf{v}_1) = \max \frac{\sum_1^N \mathbf{M}(\mathbf{v_1})}{\sum_1^N \sum_1^3 \mathbf{M}(\mathbf{v_1})} \tag{19}$$

This process is carried out for each of the testing vectors. In our analysis, we consider the combination of two to seven features. This is because more than seven features turn the performance of the ICB down to zero, and therefore it is not possible to filter more false alarms. However this number can vary as a result of the linear classification inserted in each ICB.

TABLE 2
Performance of the Uncalibrated Tracking

Step	Flaws in sequence	False Alarms in sequence	Rate of Real Flaws	Rate of False Alarms
2-Views Track	190	198	100%	51.0%
ICB-2	151	94	100%	24.2%
3-Views Track	137	45	100%	11.6%
ICB-3	18	17	100%	4.4%

4 EXPERIMENTAL RESULTS

This section presents the results of experiments carried out on a sequence of 72 radioscopic images of aluminum wheels (see some of them in Fig.8). The dimensions of the wheel are 470 [*mm*] diameter and 200 [*mm*] height. The image size is 572×768 pixels with a dynamic range of 8 bits. There are twelve known real flaws in this sequence. Three of these flaws were detected by human visual inspection ($\emptyset = 2.0 \sim 7.5$ [*mm*]) however the remaining nine flaws (small holes generated by a drill ($\emptyset = 2.0 \sim 4.0$ [*mm*]) in positions making their detection difficult) were not detected. A pattern of 1 [*mm*] in the middle of the wheel is projected in the X-ray projection coordinate as a pattern of 2.96 pixels in the image, i.e., the flaws are very small. In addition, since the signal-to-noise ratio in our radioscopic images is low, the flaw signal is slightly greater than the background

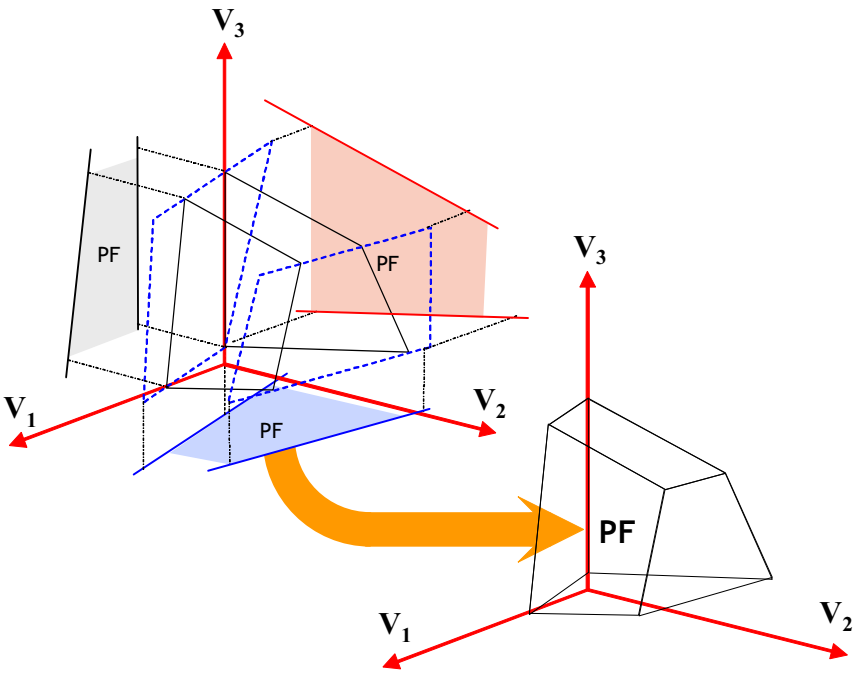


Fig. 7. Three dimensional representation of the ICB classification method.

noise, as illustrated in Fig.8. In our experiments, the mean gray level of the flaw signal ranged from 2.4 to 28.8 gray values with a standard deviation of 6.1. Analyzing a homogeneous background in different areas of interest, we obtain noise within ± 13 gray values with a standard deviation of 2.5. Due to the reasons stated above, the segmentation of real flaws with poor contrast can also involve detection of false alarms.

We separated the analysis into three steps. (1. Identification) Potential flaws are automatically identified in each image of the sequence using a single filter without a priori knowledge of the object structure (Fig.8b). The result of the identification generates a data base that contains 424 registers with twelve features of the total potential flaws detected in the sequence. From them, 214 are real flaws, which correspond to the twelve real flaws mentioned above, and 210 registers are false alarms that must be reduced. (2. Tracking) In this step we separate the analysis into two phases: a) the detection of pairs of flaws using the estimation of the fundamental matrix in two views, through the epipolar constraint; b) using the previous results, we re-projected the pairs of potential flaws in the third view using the trifocal tensor estimation. (3. ICB method) Classifiers are inserted into two and three views to filter potential flaws between the views, according to the general model proposed in Fig.2. All the phases are detailed below.

4.1 Performance with two views

The first phase is to assess the performance of the algorithm in two views using the bifocal method. This consists of determining corresponding flaws between two images in a sequence through the search for flaws along the epipolar line (Fig.8c). The results indicate that the model detects 100% of the real flaws that are corresponding in two views (Table 2, 2-Views Track). This validates the assumption of correspondence between the position of real flaws and implies that automatic detection with the fundamental matrix allows the detection of corresponding flaws that are contained on the epipolar line. There is, however, a large number of false alarms in the sequence (198/388=51%), which must be reduced using a third view.

4.2 Performance with three views

After completing the matching of possible pairs of flaws in both images, we extend the detection of flaws to the third image in the sequence (Fig.8d). In this case the performance remains at 100% of real flaws detected in the sequence, however, it has not been possible to eliminate all false alarms (Table 2, Rate of False Alarms). Furthermore, the ICB method in two and three views has allowed the detection of a large part of the real flaws (F) and false alarms (FA) with high probability, allowing them to be separated from the multiple views analysis. Thus, in two views the reduction of false alarms with ICB-2 reaches 24.2%, and with ICB-3 it reaches 4.4%. These results indicate that the proposed method has generated a sustained reduction of potential flaws (PF) in the sequence.

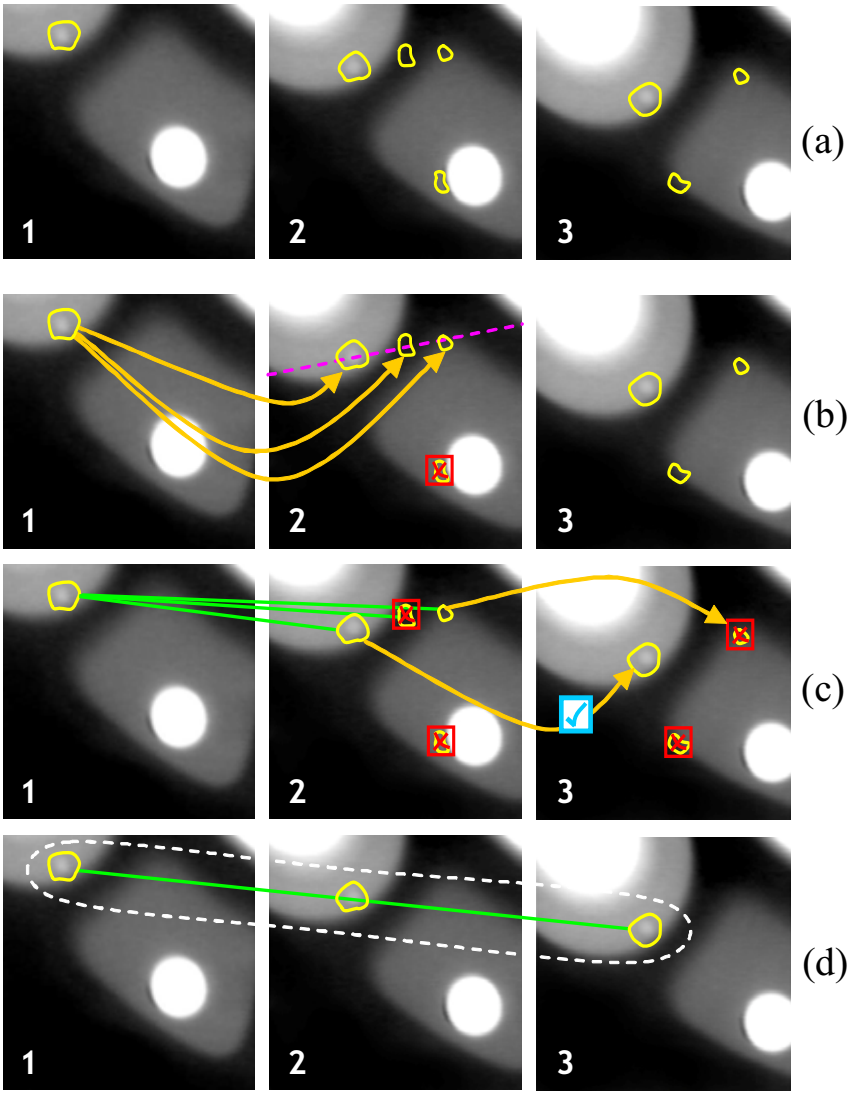


Fig. 8. Generalized flaw tracking process in one sequence of three views: a) Identification of potential flaws. b) Searching for a match two views with the fundamental matrix. c) Searching for a match in three views with the trifocal tensor. d) Final detection, the false alarms are eliminated without discriminating real flaws.

4.3 Performance of ICB

The results indicate a clear correlation between the performance of the ICB method for two views and the number of features chosen. By using the five best feature combinations, the performance of the classification is ideal, but there is a clear decrease in the number of false alarms extracted by the ICB method (Table 3. 2 Views). For example, in the case of combining the five best features,

only 17.3% of the total potential flaws in correspondence are extracted, leaving 82.7% which is then transferred to the following matching block in three views. In the case of three views, the number of correspondences is drastically reduced because the correspondence of a false alarm in three images has a lower probability of occurrence. The highest performance results occur when the six best features are combined by means of the Take-L-Plus-R selection process included in each ICB (Fig.6).

With respect to real-time capabilities, we tested our method in two ways, with and without ICB. First, with ICB, the required computation time to process a sequence of three images, was in average 38.3 [s], with a remainder of 4.4% of false alarms. Second, without ICB, using the same sequence, the time was in average of 31.1 [s][5], with a remainder of 32.3% of false alarms. These results are very promising because the ICB method can filter the majority of false alarms in sequence, although it did require more time. We want to clarify that the ICB method follows a classification ensemble methodology to filter false alarms; without ICB each potential flaw is tracked in two and three views, thus the number of tests in multiple views is increased.

4.4 Comparison with other methods

Finally, we present a summary of the performance obtained with the calibrated and uncalibrated AMVI method (Table 4). For comparison purposes, we show the performances carried out with the same sequence of X-ray images designed in (Mery & Filbert, 2002a). The given performances correspond to the 'true positives' and 'false positives'. The true positives are the percentages of flaws correctly detected in a sequence. The false positives (or false alarms) correspond to the percentage of 'non-flaws' that have been classified incorrectly as flaws. Current results indicate that it is possible to obtain 100% of the real flaws in a sequence detected correctly. These results have been generated in spite of the optical and geometric perturbations and the low SNR level that corresponds to X-ray images. However, false alarms remain in the sequence and reducing them has not been possible. Despite the false alarms, our method has achieved better performance than the system proposed by Pizarro et al. (2008), mainly because it is not necessary to carry out matching of the potential flaws, only a tracking analysis. According to the results generated in two and three views in (Carrasco & Mery, 2006), the ICB technique has allowed a reduction of 8.7% in the correspondence number in two views, and of 5.5% in the case of three views with a 4.4% remainder, which has been impossible to eliminate so far by geometric analysis.

5 Conclusions

Automated visual inspection remains an open question. Many research directions have been exploited, some very different principles have been adopted and a wide variety of algorithms have appeared in literature on automated visual inspection. Although there are several approaches in the last 30 years that have been developed, automated visual inspection systems still suffer from i) detection accuracy, because there is a fundamental trade off between false alarms and missed

5. The method was programmed in Matlab 7.0 under Windows XP SP2 on a Pentium Centrino Duo/2 GHz

TABLE 3
Sensitivity and 1-Specificity performance of the ICB classifier, and percentage of flaws reduction of ICB classifier

Views	Features	2	3	4	5	6	7
2 Views	S_n	92.7%	95.6%	98.4%	100%	100%	100%
	$1 - S_p$	2.4 %	0%	0%	0%	0%	0%
	ICB reduction	52.3%	49.5%	43.6%	17.3%	12.8%	10.3%
3 Views	S_n	95.8%	98%	96%	99.1%	99.1%	100%
	$1 - S_p$	25 %	16.7%	26.7%	22.2%	0%	6.7%
	ICB reduction	81.6%	61.17%	79.7%	89.4%	75.3%	75.7%

TABLE 4
Comparison between different calibrated and uncalibrated tracking techniques

Method	Images tracked	Year and reference	Analyzed Images	True Positives	False Positives
Calibrated	3	2002 (Mery & Filbert, 2002a)	70	100%	25%
	4	2002 (Mery & Filbert, 2002a)	70	100%	0%
	5	2002 (Mery & Filbert, 2002a)	70	83%	0%
Uncalibrated	2	2005 (Mery & Carrasco, 2005)	24	92.3%	10%
	2	2006 (Carrasco & Mery, 2006)	70	100%	32.9%
	3	2006 (Carrasco & Mery, 2006)	70	98.8%	9.9%
	2	2008 (Pizarro et al., 2008)	70	86.7%	14%
	2	2008 ^{new}	70	100%	24.2%
	3	2008 ^{new}	70	100%	4.4%

detections; and ii) strong bottleneck derived from mechanical speed (required to place the test object in the desired positions) and iii) high computational cost (to determine whether the test object is defective or not). In this sense, Automated Multiple View Inspection offers a robust alternative method that uses redundant views to perform the inspection task. In this paper we have developed a new flaw detection algorithm using an uncalibrated sequence of images. Using the new uncalibrated AMVI methodology, we have designed a novel system based only on the spatial positions of the structures. The proposed approach uses the projection of the epipolar line, generated by the fundamental matrix and the trifocal tensors in a robust manner, with the purpose of building a motion model without a priori knowledge of the object structure. The key idea of our strategy is to consider as false alarms those potential flaws that cannot be tracked in a sequence of multiple images. In this research we have introduced the calculation of corresponding points generated artificially through the maximization of the correlation coefficient from two curves and the intermediate classifier block (ICB) method in order to filter false alarms. The method was tested in a sequence of X-ray images of aluminum wheels but the methodology can be applied to other sequences as well by changing the segmentation and control point algorithms as we demonstrated in bottle inspection system with multiple views in (Carrasco, Pizarro, & Mery, 2008).

Our results indicate that it is possible to generate an automatic model for a sequence of images

which represent the movement between the points and regions they contain. This way we can use as reference points the edges of the structures or areas, without loss of information, using a nonlinear method. The main advantage of our model is the automatic estimation of movement thus avoiding the calibration process. Our future aim is to reduce the number of false alarms by means of a method of final verification of the flaws in correspondence, and an analysis of the ICB classification method with other ensemble classification and probabilistic techniques. Another possibility is to change the Fisher Discriminant into another Linear Dimensionality Reduction (LDR) technique inside of each ICB. That way we can maximize the Chernoff distance based in (Rueda & Herrera, 2008). The essential idea is to increase the distance defining linear class separability. This will allow the separation of more false alarms and flaws in each ICB. Thus, with less potential flaws in a sequence, the process will be faster.

ACKNOWLEDGEMENTS

The author acknowledges the financial support from Escuela de Ingeniería, Pontificia Universidad Católica de Chile and from FONDECYT – Chile (grant no. 1040210).

REFERENCES

Boerner, H., & Strecker, H. (1988). Automated x-ray inspection of aluminum casting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*(1), 79–91.

Campbell, J. G., Fraley, C., Murtagh, F., & Raftery, A. E. (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*, *18*(14), 1539-1548.

Carrasco, M., & Mery, D. (2006). Automated visual inspection using trifocal analysis in an uncalibrated sequence of images. *Materials Evaluation*, *64*(9), 900–906.

Carrasco, M., & Mery, D. (2007, Dec. 17-19). Automatic multiple visual inspection on non-calibrated image sequence with intermediate classifier block. In *Pacific-rim symposium on image and video technology (psivt'07)* (pp. 371–384). Springer-Verlag.

Carrasco, M., Pizarro, L., & Mery, D. (2008, Aug. 20–22). Image acquisition and automated inspection of wine bottlenecks by tracking in multiple views. In *Proc. of 8th international conference on signal processing, computational geometry and artificial vision –iscgav'08* (pp. 84–89). Rhodes Island, Greece.

Chen, Z., Wu, C., Shen, P., Liu, Y., & Quan, L. (2000). A robust algorithhm to estimate the fundamental matrix. *Patter Recognition Letters*, *21*, 851–861.

Chin, R. T., & Harlow, C. A. (1982). Automated visual inspection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *4*(6), 557-573.

Cohen, I., Ayache, N., & Sulger, P. (1992). Tracking points on deformable objects using curvature information. In *Proc. of the 2nd european conference in computer vision* (p. 458-466).

Drury, C. (1992). Inspection performance. In (pp. 2282–2314). New York, NY, USA: John Wiley and Sons.

Drury, C. G., Saran, M., & Schultz, J. (2004, Jan). *Effect of fatigue / vigilance/ environment on inspectors performing fluorescent penetrant and/or magnetic particle inspection* (Interim

Report No. 03-G-012). Federal Aviation Administration William J. Hughes Technical Center: University at Buffalo.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (Second ed.). New York: John Wiley & Sons, Inc.

Dunn, O. J., & Clark, V. A. (1974). *Applied statistics: analysis of variance and regression*. Wiley.

Egan, J. (1975). *Signal detection theory and roc analysis*. New York: Academic Press.

Filbert, D., Klatte, R., Heinrich, W., & Purshke, M. (1987). Computer aided inspection of castings. In *Ieee-ias annual meeting* (pp. 1087–1095). Atlanta, USA.

Gdalyahu, Y., & Weinshall, D. (1999). Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(12), 1312–1328.

Haralick, R. M., & Shapiro, L. G. (1992). *Computer and robot vision*. New York: Addison-Wesley Publishing Co.

Hartley, R., & Zisserman, A. (2000). *Multiple view geometry in computer vision* (First ed.). Cambridge, UK: Cambridge University Press.

Jacob, R., Raina, S., Regunath, S., Subramanian, R., & Gramopadhye, A. K. (2004). Improving inspector's performance and reducing errors - general aviation inspection training systems (gaits). In *In proceedings of the human factors and ergonomics society annual meeting proceedings*. Human Factors and Ergonomics Society.

Jarvis, J. F. (1980). Visual inspection automation. *Computer*, *13*(5), 32–38.

Kita, Y., Highnam, R., & Brady, M. (2001). Correspondence between different view breast x-rays using curved epipolar lines. *Computer, Vision and Understanding*, *83*(1), 38–56.

Kumar, A. (2008, Jan). Computer-vision-based fabric defect detection: A survey. *IEEE Transactions on Industrial Electronics*, *55*(1), 348–363.

Lagarias, J., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, *9*(1), 112–147.

Liu, H., & Srinath, M. (1990). Partial shape classification using contour matching in distance transformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(11), 1072–1079.

Malamas, E., Petrakis, E. G., & Zervakis, M. (2003). A survey on industrial vision systems, applications and tools. *Image and Vision Computing*, *21*(2), 171–188.

Mery, D. (2003). Crossing line profile: a new approach to detecting defects in aluminium castings. *Lecture Notes in Computer Science*, *2749*, 725–732.

Mery, D. (2006). High contrast pixels: a new feature for defect detection in x-ray testing. *Insight*, *46*(12), 751–753.

Mery, D., & Carrasco, M. (2005). Automated multiple view inspection based on uncalibrated image sequence. *Lecture Notes in Computer Science*, *3540*, 1238–1247.

Mery, D., & Filbert, D. (2002a). Automated flaw detection in aluminum castings based on the

tracking of potential defects in a radioscopic image sequence. *IEEE Trans. Robotics and Automation*, *18*(6), 890-901.

Mery, D., & Filbert, D. (2002b). Classification of potential defects in automated inspection of aluminium castings using statistical pattern recognition. In *Proc. of 8th european conference on non-destructive testing (ecndt 2002)*. Barcelona, Spain.

Mery, D., Silva R. R. da, Calôba, L. P., & Rebello, J. M. A. (2003). Patter recognition in the automatic inspection of aluminium castings. *Insight*, *45*(7), 475-483.

Mital, A., Govindaraju, M., & Subramani, B. (1998). A comparison between manual and hybrid methods in parts inspections. *Integrated Manufacturing Systems*, *9*(6), 344–349.

Newman, T. S., & Jain, A. K. (1995). A survey of automated visual inspection. *Computer Vision and Image Understanding*, *61*(2), 231–262.

Nguyen, V.-D., Noble, A., Mundy, J., Janning, J., & Ross, J. (1998). Exhaustive detection of manufacturing flaws as abnormalities. In *Proc. of ieee computer society conference on computer vision and pattern recognition* (pp. 945–952).

Pedreschi, F., Mery, D., Mendoza, F., & Aguilera, J. (2004). Classification of potato chips using pattern recognition. *Journal of Food Science*, *69*(6), E264–E270.

Pizarro, L., Mery, D., Delpiano, R., & Carrasco, M. (2008). Robust automated multiple view inspection. *Pattern Analysis and Applications*, *11*(1), 21–32.

Polikar, R. (2006). Ensemble systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45.

Rueda, L., & Herrera, M. (2008). Linear dimensionality reduction by maximizing the chernoff distance in the transformed space. *Pattern Recognition*, *41*(10), 3138–3152.

Sebastian, T., Klein, P., & Kimia, B. (2003). On aligning curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(1).

Sonka, M., Hlavac, V., & Boyle, R. (Eds.). (1999). *Image processing, analysis and machine vision* (2nd ed.). Pacific Grove, CA: PWS Publishing.

Spencer, F. (1996, Sep). *Visual inspection research project report on benchmark inspections* (Technical Report No. DOT/FAA/AR-96/65). Office of Aviation Research Washington, D.C. 20591: U.S. Department of Transportation, Federal Aviation Administration,Washington, DC.

Spicer, P., Bohl, K., Abramovich, G., & Barhak, J. (2006, Feb. 25-28). Robust calibration of a reconfigurable camera array for machine vision inspection (RAMVI): Using rule-based colour recognition. In *Proc. of the 1st international conference on computer vision theory and applications ultrasonics symposium (visapp)* (pp. 131–138). Setúbal, Portugal.

Stearns, S. (1976). On selecting features for patterns classifiers. In *Proc. of iapr international conference on pattern recognition* (pp. 71–75).

Umeyama, S. (1993, Feb). Parameterized point pattern matching and its application to recognition of object families. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(2), 136–144.

Paper # 3
Image Acquisition and Automated Inspection of Wine Bottlenecks by Tracking in Multiple Views

Miguel Carrasco, Luis Pizarro and Domingo Mery
mlcarras@puc.cl, pizarro@mia.uni-saarland.de, dmery@ing.puc.cl

Abstract

In this paper we propose a prototype for image sequence acquisition of glass wine bottlenecks, whose main novelty is the use of an inner lighting source for better capturing of potential defects. A novel approach for automatic inspection of the bottlenecks based on tracking potential flaws along the acquired sequence is also presented. Our inspection system achieves performance rates of 87% true positives and 0% false positives.

Index Terms

Image acquisition; Automated inspection; Tracking; Multiple views; Wine bottle inspection; Defect detection; Control quality.

1 Introduction

The bottle inspection systems appeared in the literature can be classified in two categories: Approaches that make use of a single view/camera for detecting flaws, e.g. (Canivet, Zhang, & Jourlin, 1994; Mery & Medina, 2004; Y.-N. Wang, H.-J., & Duan, 2005; Yan & Cui, 2006; Duan, Wanga, Liua, & Li, 2007; Yepeng, Yuezhen, & Zhiyong, 2007); and frameworks that exploit the utilization of multiple views/cameras to reinforce the detection process, e.g. (Firmin, Hamad, Postaire, & Zhang, 1997; Hamad, Betrouni, Biela, & Postaire, 1998; Ma, Su, & Ni, 2002; Shafait, Imran, & Klette-Matzat, 2004; Katayama, Ishikura, Kodoma, Fukuchi, & Fujiwara, 2008). Our proposed inspection device employs a single camera for image acquisition. However, we emulate the use of multiple cameras by recording an image sequence of a wine glass bottle in successive rotations along its principal axis.

Miguel Carrasco and Domingo Mery are with the Computer Engineering Department at Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860 (143), Santiago, Chile.
Luis Pizarro is with the Mathematical Image Analysis Group at Saarland University, Building E1.1, 66041 Saarbrücke, Germany

In every image acquisition device the lighting conditions play a major role in the quality of the acquired images and therefore in the inspection task. Since natural lighting conditions are dynamic and change all the time it is not feasible to implement algorithms that are robust to illumination changes without burning important computational time (Kopparapu, 2006). Therefore, the use of artificial lighting is a requisite for reaching good and uniform illumination for real-time inspection systems. There exist several studies, e.g., (Yi, Haralick, & Shapiro, 1995), concerning the placement of external light sources around the object under examination. However, we do not know any work reporting on light sources placed inside a glass bottle. We proposed the design of an electro-mechanical device for image acquisition and inspection of glass wine bottles using an internal illuminating system. This allows us to obtain high-quality images for capturing very small defects, and to avoid the intrinsic reflections produced by external light sources.

Numerous methods for automated glass bottle inspection attempt to identify defects in the lips, body and bottom of the bottles (Canivet et al., 1994; Firmin et al., 1997; Hamad et al., 1998; Ma et al., 2002; Shafait et al., 2004; Y.-N. Wang et al., 2005; Yan & Cui, 2006; Duan et al., 2007; Yepeng et al., 2007; Katayama et al., 2008). However, there are only few works that deal with the problem of detecting flaws in the bottleneck (Ma et al., 2002; Mery & Medina, 2004). The neck is the bottle's part where most of the defects appear during fabrication, due to its narrow and hence difficult to manipulate structure. In this paper we focus our research on the inspection of necks in empty wine glass bottles.

For the inspection of the bottlenecks we propose a novel methodology that performs tracking of potential flaws along the acquired image sequence. The key observation is that only real flaws can be successfully traced, since they do induce spatiotemporal relations between the views where they appear. Conversely, potential defects that cannot be tracked correspond to false alarms. We effectively track and thus identify real flaws by means of geometry of multiple views (Hartley & Zisserman, 2000). In particular, we employ bifocal and trifocal analysis. Although several published methods work with multiple views, this is, to the best of our knowledge, the first work on multiple view tracking for inspection of wine glass bottles.

Our paper is organized as follows: In Section 2 we present our prototype for image acquisition and inspection of wine bottlenecks. Section 3 describes our proposed algorithm for tracking real flaws along multiple views. Section 4 shows the performance achieved by our inspection system in comparison with other methods proposed in the literature. Finally, we summarize our contributions and succinctly describe some ongoing and future work in Section 5.

2 Electro-Mechanical System for Image Sequence Acquisition

In this section we describe an electro-mechanical device we have designed for the automatic acquisition of an image sequence of the bottleneck of an empty glass bottle under inspection. Two are the main components of our mechanism: An internal illumination system and a rotor that rotates the bottle during acquisition. The image sequence is recorded by a standard CCD camera. The device we have built is schematically shown in the Fig. 1.

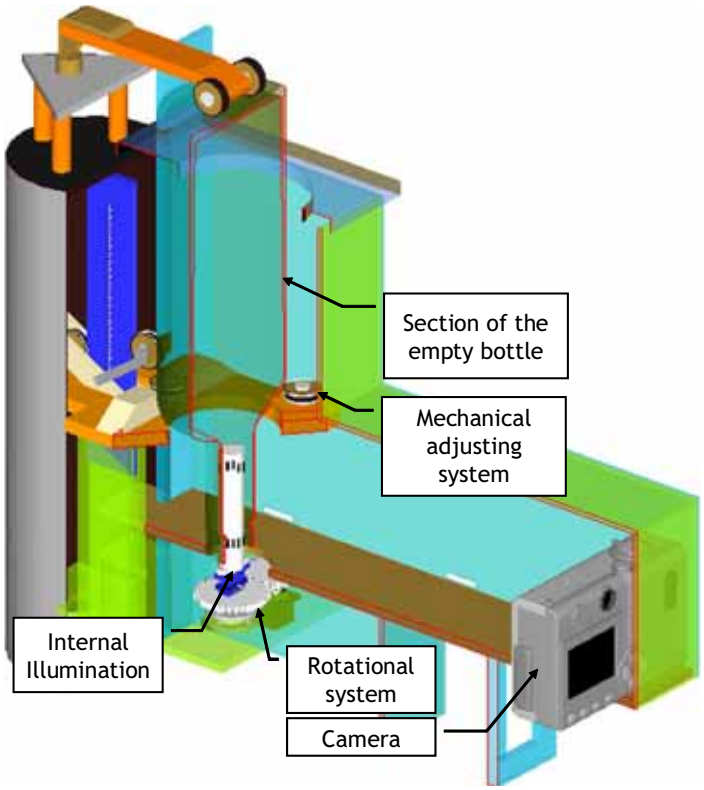
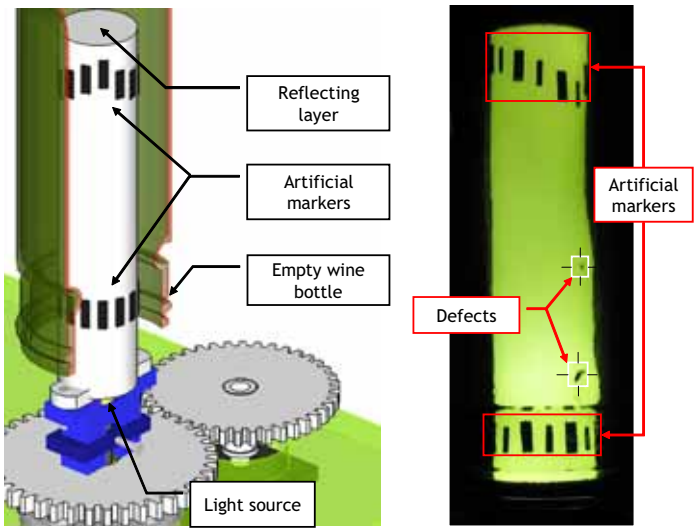


Fig. 1. Proposed electro-mechanical prototype for image acquisition.

As observed in the Fig. 1, an illuminating tube has been placed inside the bottle. Four LEDs (T1 3.5v-20mA) emitting white light uniformly are located at the bottom of the tube. To improve light uniformity a reflecting layer has been fixed at the other extreme of the tube. To the best of our knowledge, there is no inspection system for glass bottles proposed in the literature that places the illumination system inside the bottle. This greatly improves the definition of the acquired images, increasing therefore the probability of capturing the smallest defects around the bottleneck. Another important characteristic of the illuminating tube is the set of artificial markers situated on both extremes, as displayed in the Fig. 2. They will later allow us to know the relative position of a defect along the image sequence.

The rotational system shown in the Fig. 1 permits rotating the bottle and the light source at the same time. An image sequence of the bottleneck is thus composed by views taken at successive rotations by a configurable spin angle α, controlled by a step motor. The images are captured by a CCD sensor with high resolution. In our experimental prototype we use a CANON S3 IS camera with a resolution size of 2592 $\times$ 1944 pixels and a dynamic range of 24 bits. The camera is placed around 20 cm from the bottleneck. No additional light sources are utilized. The device also includes a mechanical adjusting system to adapt the inspection to different bottleneck lengths. An adjustable



(a) (b)

Fig. 2. (a) Details of the illuminating tube; (b) Example image captured by the CCD camera.

arm holds the bottle from its body, and a press mechanism pushes the bottle against the rotor to keep its vertical position.

The electro-mechanical system is commanded by a Basic Stamp micro-controller PIC16C57 connected to a standard personal computer via a RS232 communication port. The micro-controller is programmed in Pbasic. For a specified spin angle α (degrees) the micro-controller synchronizes the step motor with the illumination system. The camera's acquisition method is triggered by a camera control system via Matlab. The image sequence consists thus of $\lfloor 360/\alpha \rfloor$ different views, where $\lfloor \cdot \rfloor$ is the floor function.

For the sake of simplicity, we have built our prototype considering an upside down bottle[1]. However, it is also possible to assemble the system with a right side up bottle. Our bottle inspection apparatus employs a single camera only. However, by recording an image sequence of the bottle under examination we are able to emulate a system using multiple cameras (Firmin et al., 1997; Hamad et al., 1998; Ma et al., 2002; Shafait et al., 2004; Katayama et al., 2008). It is important to mention that no camera calibration procedure is considered at all. Therefore, we actually obtain *uncalibrated* image sequences, which we will utilize in the next section for tracking and detecting real defects in bottlenecks using geometry of multiple views (Hartley & Zisserman, 2000).

3 Detection of Real Flaws by Tracking in Multiple Views

In the previous section we described an electro-mechanical device specially designed for capturing image sequences of bottlenecks using a single camera. In some applications, a unique image might

1. Similar to the inspection of wineglass in (J. Wang & Asundi, 2000).

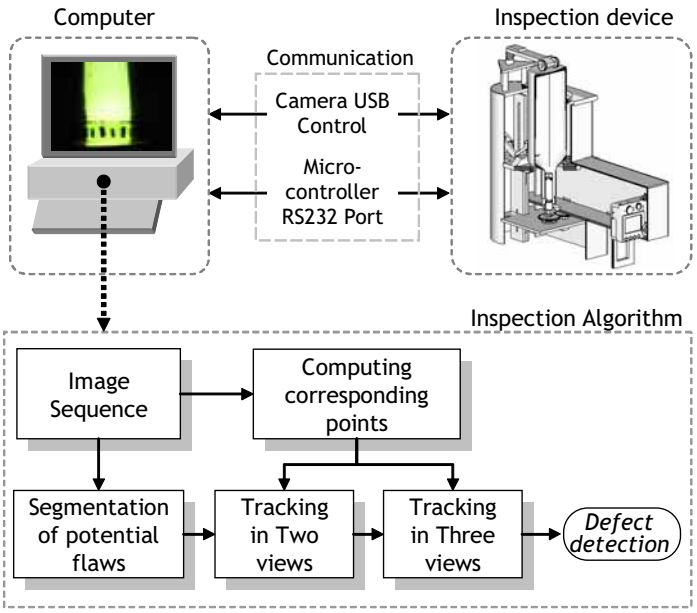


Fig. 3. Proposed bottle inspection system.

be enough for inspecting certain objects or materials. However, the use of multiple views can reinforce the diagnosis made with a single image. That is the case for example for low signal-to-noise ratio imaging systems, where the identification of real defects with poor contrast entails the appearance of numerous false alarms as well. Here, we aim at exploiting the redundant information present in the multiples views of the bottle under inspection in order to discriminate real defects from false alarms. In fact, only real flaws can successfully be tracked along an image sequence. This is the main idea that will allow us to distinguish real flaws from other artifacts. Based on such observation, we propose a three-steps methodology for detecting real flaws in the bottleneck of a glass bottle: segmentation of potential flaws, computation of corresponding points, and tracking potential flaws. Similar ideas has been treated in (Mery & Filbert, 2002; Mery & Medina, 2004; Carrasco & Mery, 2006; Pizarro, Mery, Delpiano, & Carrasco, 2008). Apart from dealing with another application and proposing a system for image acquisition, the main differences between this contribution and these works lie in our clever choice of corresponding points and in the utilization of several filters for defect segmentation. The Fig. 3 shows a general overview of our proposed methodology for image acquisition and defect detection.

3.1 Segmentation of potential flaws

The segmentation of potential defects in every image of the sequence is outlined in the Fig. 4. First, the original image is filtered with a Gaussian filter in order to reduce the amount of noise intrinsic to any CCD image acquisition process (I_1). Second, a bottom-hat filter is applied to isolate potential defects from the background (I_2). Third, potential defects are segmented in every image

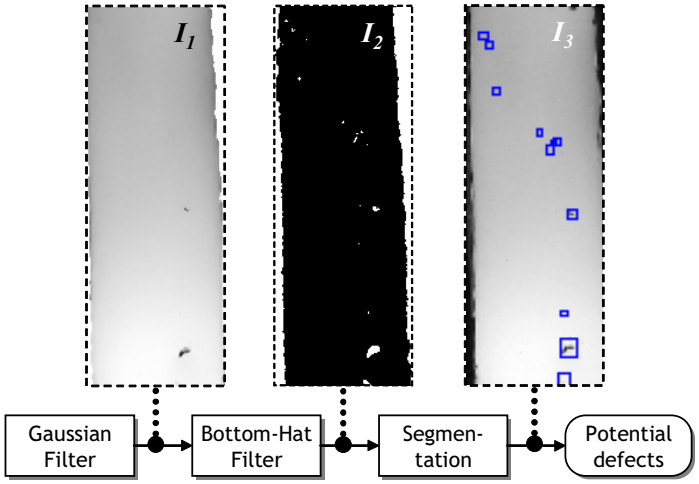


Fig. 4. Segmentation of potential defects.

using the Valley Emphasis method (Hui-Fuang, 2006). For each potential defect, the mass centre is taken and stored in homogeneous coordinate. As a result, numerous potential defects appear as observed in the segmented image (I_3). Nevertheless, only few of them correspond to real flaws.

3.2 Computation of corresponding points

As stated before, our final goal is tracking real defects in an image sequence. For this purpose, accurate corresponding points between every pair of views are required. We solve this problem by placing equidistant artificial markers on both extremes of the illuminating source, as shown in the Fig. 2. The lower markers are positioned at the same vertical level, while the upper ones follow a sinusoidal wave. Using these markers we can compute a set of corresponding points between each pair of consecutive or non-consecutive views. This is schematically outlined in the Fig. 5. The upper and lower markers of each view are connected through vertical lines between their mass centre, which were also extracted in the segmentation step. Since the length of a vertical line connecting two particular markers remains constant along the image sequence, we know the relative position of these markers in different views. Therefore, the set of corresponding points between two views a and b is conformed by the relative positions of their markers. Such correspondences are later used to estimate the *fundamental matrix* $\mathbf{F}_{a,b}$ that relates any pair of points in the views a and b.[2]

3.3 Tracking of potential flaws

In the previous step we have segmented all potential defects along the image sequence. We now turn to the problem of separating real flaws from false alarms. The key observation is the fact that

2. In estimating $\mathbf{F}_{a,b}$ we combine the algorithm of Hartley (Hartley & Zisserman, 2000) with the biepipolar restriction presented in (Chen, Wu, Shen, Liu, & Quan, 2000).

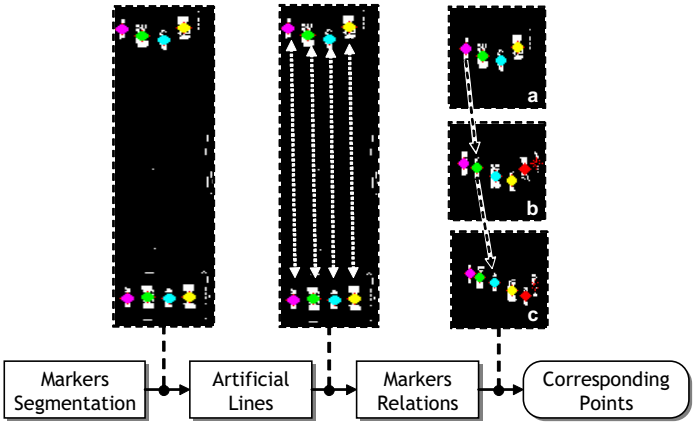


Fig. 5. Computation of corresponding points.

only real flaws could be tracked along the image sequence. A real flaw entails a spatiotemporal relation in the different views where it appears, while a false alarm corresponds to a random event.

In the previous segmentation step each identified potential defect is represented by its mass centre. For instance, the mass centre of the $j\text{-}th$ potential defect in the $a\text{-}th$ view is stored in homogenous coordinates as $\mathbf{m}_a^i = [x_a^i, y_a^i, 1]^\top$. If this potential defect is actually a real flaw it must have a corresponding point $\mathbf{m}_b^j$ in another consecutive or non-consecutive view b where a potential defect j was also segmented. According to the *principle of multiple view geometry* (Hartley & Zisserman, 2000), the points $\mathbf{m}_a^i$ and $\mathbf{m}_b^j$ are in correspondence if they are related by the fundamental matrix $\mathbf{F}_{a,b}$ such that

$$\mathbf{m}_b^{j\top}\,\mathbf{F}_{a,b}\,\mathbf{m}_a^i = 0\,.$$

This relation is known as *epipolar constrained*. It indicates that the point $\mathbf{m}_b^j$ can only lie on the epipolar line of the point $\mathbf{m}_a^i$ defined as $\mathbf{l}_a^i = \mathbf{F}_{i,j}\,\mathbf{m}_a^i = [l_{a,x}^i, l_{a,y}^i, l_{a,z}^i]$. Then, knowing $\mathbf{l}_a^i$ we identify the correspondence associated with the potential defect i as the potential defect j in the view b that satisfies the constraint

$$\frac{|\mathbf{m}_b^{j\top}\,\mathbf{F}_{a,b}\,\mathbf{m}_a^i|}{\sqrt{(l_{a,x}^i)^2 + (l_{a,y}^i)^2}} < \varepsilon_1\,,$$

for small ε_1. If this constraint is fulfilled a potential defect is thus found in the two views. In this case it could be considered as a real flaw with a bifocal correspondence. Otherwise, it is regarded as a false alarm. An example is shown in the Fig. 6a. The same procedure is applied to every potential defect in the view a that is to be found in the view b.

To confirm that a bifocal correspondence represents indeed a real flaw, we try to discover a new correspondence in a third view with the help of trifocal tensors. Let $\mathbf{T} = (T_t^{rs})$ be a $3\times3\times3$ matrix representing the trifocal tensor that encodes the relative motion among the views a, b, c.[3] Then, we

3. See (Hartley & Zisserman, 2000) for details on the computation of the trifocal tensors.

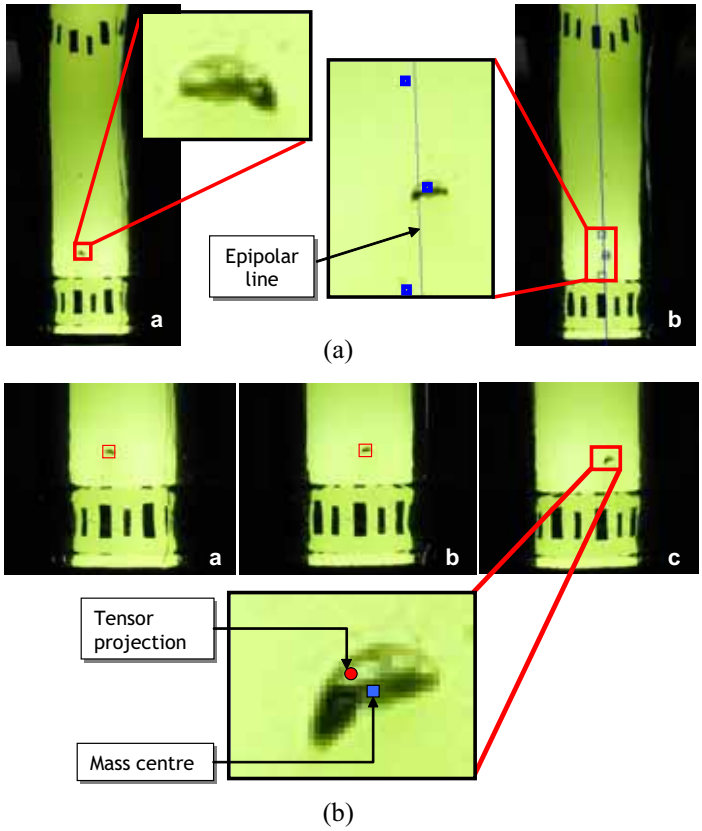


Fig. 6. (a) Example of bifocal correspondences; (b) Example of trifocal correspondences.

can estimate the hypothetical position of a defect k in a third view c using the correspondences $\mathbf{m}_a^i, \mathbf{m}_b^j$ and the tensor $\mathbf{T}$ as[4]

$$\widehat{\mathbf{m}}_c^k = \frac{1}{\mathbf{m}_a^{i\top}(T_{\cdot}^{13} - x_b^j T_{\cdot}^{33})} \begin{bmatrix} \mathbf{m}_a^{i\top}(T_{\cdot}^{11} - x_b^j T_{\cdot}^{31}) \\ \mathbf{m}_a^{i\top}(T_{\cdot}^{12} - x_b^j T_{\cdot}^{32}) \\ \mathbf{m}_a^{i\top}(T_{\cdot}^{13} - x_b^j T_{\cdot}^{33}) \end{bmatrix}.$$

We compare the estimated position with all potential flaws of the view c, regarding the potential defect k as a real flaw if the constraint

$$\|\widehat{\mathbf{m}}_c^k - \mathbf{m}_c^k\| < \varepsilon_1$$

is fulfilled. In this case a potential defect is thus found in the three views, i.e., a real flaw with a trifocal correspondence has been detected. Potential defects that do not find correspondence in three views are finally discarded and considered false alarms. An example is shown in the Fig. 6b.

4. The estimated projection in the third view can be improved applying the point-line-point method proposed in (Hartley & Zisserman, 2000, pp.373).

TABLE 1
Comparison with other inspection systems

Inspected parts	Views	Tracking	TPR	FPR
neck (Mery & Medina, 2004)	Single	No	85%	4%
lips, body, bottom (Duan et al., 2007)	Single	No	97%	>1%
lips (Y.-N. Wang et al., 2005)	Single	No	98%	0%
body (Firmin et al., 1997; Hamad et al., 1998)	Multiple	No	80%-85%	2%
body, bottom (Shafait et al., 2004)	Multiple	No	100%	>1%
lips, neck (Ma et al., 2002)	Multiple	No	98%	2%
neck (our method)	**Multiple**	**Yes**	**87%**	**0%**

4 COMPARATIVE RESULTS

We now evaluate the performance of our proposed methodology for inspecting bottlenecks of empty glass wine bottles. In our experiments we used 13 color image sequences of a dozen bottles with real flaws. The area of the smallest defect was around 15 pixels. Each sequence consists of 24 images ($\alpha = 15$ degrees). From the recorded images we extract sub-images of the bottlenecks of 1000×250 pixels. The inspection was performed considering trifocal correspondences in consecutive images, where the number of real flaws fluctuates between 0 and 8, with an average of 3.3 flaws/image. We therefore expect our method to identify $13 \times \binom{24}{3} \times 3.3$ flaws approximately. The performance is assessed considering two indicators: the true positive rate (TPR) and false positive rate (FPR), defined respectively as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{RD}}, \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{RD}},$$

where TP is the number of true positives (defects correctly classified), FP is the number of false positives (regular structures classified as defects, i.e., false alarms), and RD is the number of existing real defects. Ideally, $\mathrm{TPR} = 100\%$ and $\mathrm{FPR} = 0\%$, i.e., all defects are detected without flagging false alarms.

Our tests showed good performance with $\mathrm{TPR} = 87\%$ and $\mathrm{FPR} = 0\%$. In Table 1 we juxtapose our results with other inspection systems proposed in the literature, indicating the use of single or multiple images. It is important to mention that this is just a quantitative comparison, since the listed methods were tested on different images (or image sequences) and type of bottles, and they inspect one or several bottle parts (lips, mouth, bottleneck, body, bottom). Nevertheless, to the best of our knowledge, our methodology for glass bottle inspection is the first one that performs tracking of potential flaws in multiple views.

Concerning the real-time capabilities of our methodology, the computational time required to process trifocal correspondences was 2.8 sec in average, using a Matlab 7.0's implementation running on a Pentium Centrino 2.0 GHz under Windows XP SP2. 34% of computational time is spent reading the images, 35% in the segmentation, 11% in the trifocal analysis, and 20% in other Matlab's internal operations.

5 CONCLUSIONS

Our principal contribution was twofold: First, we present a prototyping design of an electro-mechanical device for acquiring image sequences of wine bottlenecks using a single camera. Its main novelty is the placement of the illuminating source inside the bottle, which greatly improves the definition of the inspected images. Second, we introduce a new methodology for detecting flaws in the bottleneck based on tracking potential defects along an image sequence. Our inspection system achieves performance rates of 87% true positives and 0% false positives. Although these good results, our implementation is not yet competitive in terms of computational time. In this sense, several improvements are matter of ongoing work. For instance: the use of a fast industrial camera instead of a slow commercial camera together with a faster bottle rotating mechanism; the transfer of our Matlab implementations to a highly efficient low-level programming language; and the exploitation of multi-grid implementation strategies. Additional future work includes the adaptation of our electro-mechanical device for inspecting not only bottlenecks, but other bottle parts as well.

Acknowledgements: We gratefully acknowledge partial funding by CONICYT – Colegio Doctoral Franco–Chileno, grant no. 21050185.

References:

Canivet, M., Zhang, R. D., & Jourlin, M. (1994). *Finish inspection by vision for glass production* (No. 1). SPIE.

Carrasco, M., & Mery, D. (2006). Automated visual inspection using trifocal analysis in an uncalibrated sequence of images. *Mater Eval*, *64*, 900-906.

Chen, Z., Wu, C., Shen, P., Liu, Y., & Quan, L. (2000). A robust algorithhm to estimate the fundamental matrix. *Pattern Recogn Lett*, *21*, 851–861.

Duan, F., Wanga, Y.-N., Liua, H.-J., & Li, Y.-G. (2007). A machine vision inspector for beer bottle. *Eng Appl Artif Intell*, *20*(7), 1013–1021.

Firmin, C., Hamad, D., Postaire, J., & Zhang, R. (1997). Gaussian neural networks for bottles inspection: a learning procedure. *Int J Neural Syst*, *8*(1), 41-46.

Hamad, D., Betrouni, M., Biela, P., & Postaire, J. (1998). Neural networks inspection system for glass bottles production: A comparative study. *Int J Pattern Recogn Artif Intell*, *12*(4), 505-516.

Hartley, R., & Zisserman, A. (2000). *Multiple view geometry in computer vision*. Cambridge, UK: Cambridge University Press.

Hui-Fuang, N. (2006). Automatic thresholding for defect detection. *Pattern Recogn Lett*, *27*(14), 1644-1649.

Katayama, K., Ishikura, T., Kodoma, Y., Fukuchi, H., & Fujiwara, A. (2008). *Optical inspection of glass bottles using multiples cameras.* US. Patent. 7.329.855 B2.

Kopparapu, S. K. (2006). Lighting design for machine vision application. *Image Vis Comput*, *24*(7), 720-726.

Ma, H.-M., Su, G.-D., & Ni, J.-Y. W. Z. (2002). *A glass bottle defect detection system without touching*.

Mery, D., & Filbert, D. (2002). Automated flaw detection in aluminum castings based on the tracking of potential defects in a radioscopic image sequence. *IEEE Trans Robot Autom*, *18*(6), 890-901.

Mery, D., & Medina, O. (2004). Automated visual inspection of glass bottles using adapted median filtering. *LNCS*, *3212*, 818-825.

Pizarro, L., Mery, D., Delpiano, R., & Carrasco, M. (2008). Robust automated multiple view inspection. *Pattern Anal Appl*, *11*(1), 21–32.

Shafait, F., Imran, S., & Klette-Matzat, S. (2004). Fault detection and localization in empty water bottles through machine vision. In *Emerging technology conference* (pp. 30–34).

Wang, J., & Asundi, A. (2000). A computer vision system for wineglass defect inspection via gabor-filter-based texture features. *Inform Sci*, *127*, 157–171.

Wang, Y.-N., H.-J., L., & Duan, F. (2005). A bottle finish inspect method based on fuzzy support vector machines and wavelet transform. In *International conference on machine learning and cybernetics* (Vol. 8, pp. 4588–4592).

Yan, T.-S., & Cui, D.-W. (2006). The method of intelligent inspection of product quality based on computer vision. In *7th int. conf. on computer-aided industrial design and conceptual design (caidcd '06)* (pp. 1–6). Hangzhou.

Yepeng, Z., Yuezhen, T., & Zhiyong, F. (2007). Application of digital image process technology to the mouth of beer bottle defect inspection. In *8th int. conf. on electronic measurement and instruments* (Vol. 2, pp. 905–908).

Yi, S., Haralick, R., & Shapiro, L. (1995). Optimal sensor and light source positioning for machine vision. *Comput. Vis. Image Underst.*, *61*(1), 122–137.

Paper# 4
Visual Inspection of Glass Bottlenecks by Multiple-View Analysis

Miguel Carrasco, Luis Pizarro and Domingo Mery
mlcarras@puc.cl, pizarro@mia.uni-saarland.de, dmery@ing.puc.cl

Abstract

Automated multiple view inspection (AMVI) was developed by (Mery & Filbert, 2002) to automatically detect flaws in calibrated radioscopic images of aluminum die castings. This technique involves an initial step that extracts numerous segmented regions from a set of views of the object under inspection. These regions are subsequently classified either as real flaws or as false alarms. The main idea exploited in the classification process considers, on the one hand, that image noise and false alarms occur as random events in the different views. Real flaws, on the other hand, induce geometric and featural relations in the views where they appear. Therefore, by analysing such relations it is possible to successfully localize real flaws and to discard a large number of false alarms. In this paper, we propose an adaptation of the AMVI technique for the automatic detection of flaws in uncalibrated images of glass bottlenecks. The narrow structure of bottlenecks poses a very challenging problem for their automated visual inspection, which has received little attention in the literature. This is a highly relevant issue in the fabrication of glass bottles e.g. for the wine and beer industry. Our inspection approach utilizes geometry of multiple views and a rich set of feature descriptors to discriminate real flaws from false alarms. We also contribute to the design of a prototype for an electro-mechanical device for image acquisition that considers illumination sources inside the bottle. It allows us to capture very clean images around the bottleneck and to detect very small flaws. Our inspection system achieves a *true positive rate* of 99.1% and a *false positive rate* of 0.9%.

Index Terms

Automated visual inspection; Flaw detection; Multiple views; Uncalibrated images; Glass bottlenecks.

1 INTRODUCTION

Visual inspection is defined as a quality control task that determines if a product deviates from a given set of specifications using visual data (Malamas, Petrakis, & Zervakis, 2003; Kumar, 2008). Inspection usually involves measurements of specific part features such as assembly integrity, surface

Miguel Carrasco and Domingo Mery are with the Computer Engineering Department at Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860 (143), Santiago, Chile.
Luis Pizarro is with the Mathematical Image Analysis Group at Saarland University, Building E1.1, 66041 Saarbrücke, Germany

finish and geometric dimensions. If the measurements lie within a determined tolerance, the inspection process considers the product as accepted for use. In industrial environments, inspection is performed by human inspectors and/or automated visual inspection (AVI) systems. Human inspectors are not always consistent and effective evaluators because the inspection tasks are monotonous and exhausting (Mital, Govindaraju, & Subramani, 1998). Typically, there is one rejected in hundreds of accepted products. According to (Drury, Saran, & Schultz, 2004), human inspection is at best 80% effective. Achieving 100% effectivity would require a high level of redundancy. That is, several human controls need to be made, which naturally increases costs and slows down the inspection task itself and subsequent manufacturing processes (Jacob, Raina, Regunath, Subramanian, & Gramopadhye, 2004). Human visual inspection has been estimated to account for at least 10% of the total fabrication costs. In some applications, it suffices to select a reduced but representative set of products to inspect, from which statistical inference is applied to estimate the amount of defective products in the total production. Nevertheless, there exist applications requiring every product to be inspected thoroughly, i.e. an inspection process 100% effective needs to be ensured.

One of the applications where every product needs to be flawless is the fabrication of glass bottles for the wine and beer industry. Defects in glass bottles can arise from an incompletely reacted batch, from batch contaminants which fail to melt completely, from interactions of the melted material with glass-contact refractories and superstructure refractories, and by devitrification. If conditions are abnormal many flaws can be produced and even just one flaw of only 1-2 mg in every 100 mg article can be enough to give 100% rejection rates (Parker, 2000). Most of the methods proposed in the literature for automated glass bottle inspection attempt to identify flaws in the lips, body and bottom of the bottles (Canivet, Zhang, & Jourlin, 1994; Firmin, Hamad, Postaire, & Zhang, 1997; Hamad, Betrouni, Biela, & Postaire, 1998; Ma, Su, & Ni, 2002; Shafait, Imran, & Klette-Matzat, 2004; Y.-N. Wang, H.-J., & Duan, 2005; Yan & Cui, 2006; Duan, Wanga, Liua, & Li, 2007; Yepeng, Yuezhen, & Zhiyong, 2007; Katayama, Ishikura, Kodoma, Fukuchi, & Fujiwara, 2008). Only few of them deal with the problem of detecting flaws in the bottleneck (Ma et al., 2002; Mery & Medina, 2004). Nevertheless, this bottle part makes the inspection task very challenging due to its narrow and difficult to manipulate structure. The smallest undetected flaw around this region can cause a grave danger for consumers. For example, the introduction of the cork in a wine bottle could detach glass particles belonging to a defective region. Encouraged by this difficulty, we focus our research on the inspection of necks in empty glass bottles.

The selection of the lighting for an inspection system is a crucial problem. Since natural lighting conditions are dynamic and change all the time it is not feasible to implement algorithms that are robust to illumination changes without burning important computational time (Kopparapu, 2006). Therefore, the use of artificial lighting is a requisite for reaching good and uniform illumination for real-time inspection systems. There exist several studies, e.g. (Vazquez, 2007; Marchand, 2007), concerning the placement of external light sources around the object under examination. However, we do not know any work reporting on light sources placed inside a glass bottle. We proposed the design of an electro-mechanical device for image acquisition and inspection of glass bottlenecks

using an internal illumination system. This allows us to obtain high-quality images for capturing very small flaws, avoiding in this way intrinsic reflections due to external light sources.

We can classify bottle inspection systems in two categories: Approaches that make use of a single view/camera for detecting flaws, e.g. (Canivet et al., 1994; Mery & Medina, 2004; Y.-N. Wang et al., 2005; Yan & Cui, 2006; Duan et al., 2007; Yepeng et al., 2007); and frameworks that exploit the utilization of multiple views/cameras to reinforce the detection process, e.g. (Firmin et al., 1997; Hamad et al., 1998; Ma et al., 2002; Shafait et al., 2004; Katayama et al., 2008). Our inspection device employs a single camera for image acquisition. However, it captures multiple views of the bottleneck, which are taken at successive rotations of the bottle along its principal axis, i.e. we record an image sequence of the bottleneck.

Concerning the problem of detecting flaws itself, we propose a novel methodology that performs tracking of potential flaws along the acquired image sequence. The key observation is that only real flaws can be successfully tracked, since they do induce spatial relations between the views where they appear. Conversely, potential flaws that cannot be tracked will be considered as false alarms. This idea was originally proposed in (Mery & Filbert, 2002) for flaw detection in calibrated X-ray images. Here we extend that approach to the analysis of uncalibrated image sequences. Moreover, we combine the use of geometry of multiple views with several feature descriptors to achieve a precise distinction between real flaws and false alarms. This paper also extends previous ideas presented in (Carrasco & Mery, 2006; Pizarro, Mery, Delpiano, & Carrasco, 2008; Carrasco, Pizarro, & Mery, 2008).

Our paper is organized as follows. In Section 2 we present our prototype for image acquisition of glass bottlenecks. In Section 3 we describe the inspection algorithm for tracking real flaws in multiple views. In Section 4 we report on the performance achieved by our inspection system in comparison with other methods proposed in the literature. Finally, we summarize our contributions and succinctly describe some ongoing and future work in Section 5.

2 Electro-Mechanical System for Image Sequence Acquisition

In this section, we describe our electro-mechanical device for image sequence acquisition. Fig. 1a displays the prototype holding an empty glass bottle. We make a close-up around the bottleneck (Fig. 1b) to show an illuminating tube we have placed inside the bottle. Four LEDs (T1 3.5v-20mA) emitting white light uniformly are located at the bottom of the tube. To improve light uniformity a reflecting layer has been fixed at the other extreme of the tube. In this way, we greatly improve the definition of the acquired images, which increasing the probability of capturing the smallest flaws around the bottleneck. To the best of our knowledge, there is no inspection system for glass bottles in the literature that uses an internal illumination system. No additional light sources are utilized. Another important characteristic of the illuminating tube is the set of control markers situated on both extremes. As we will see later in Section 3.2, they allow us to compute accurate corresponding points between different views, which is a fundamental step for detecting flaws in multiples views. Fig. 1c shows an image, taken by a standard CCD camera, that contains two highlighted flaws.

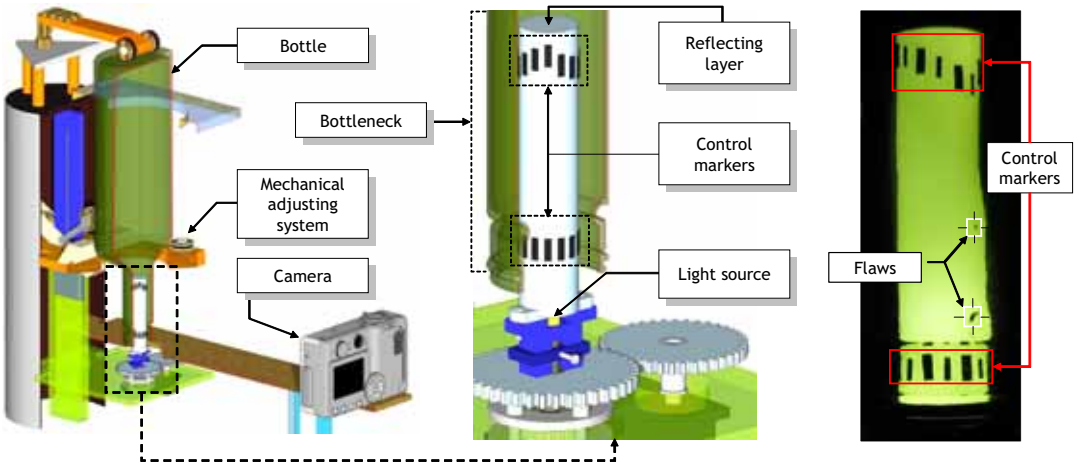


Fig. 1. (a) Proposed electro-mechanical prototype for image acquisition; (b) Details of the illuminating tube; (c) Example of an image captured by the CCD camera.

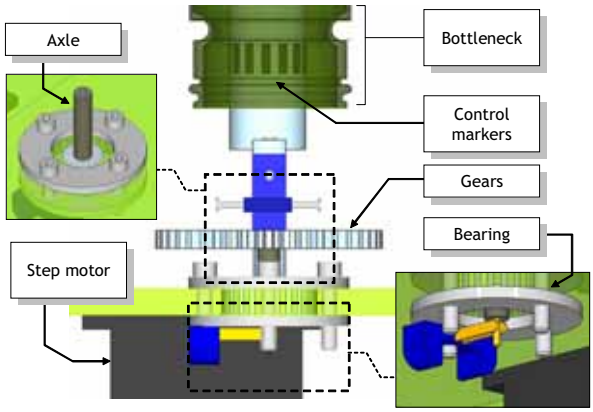


Fig. 2. Mechanical system that simultaneously rotates the bottle and the illuminating tube.

Our premise is that we can detect flaws more efficiently by looking at several images of the object under inspection. Therefore, we need our image acquisition system to be able to capture – with a unique camera – a number of images taken from different viewpoints. To this end, we provide our prototype with an electro-mechanical system (Fig. 2) that simultaneously rotates the bottle and the illuminating tube by configurable spin angle that is controlled by a step motor. An image sequence of the bottleneck is then taken at successive rotations of the glass bottle. In our experimental prototype we use a Canon S3 IS camera with resolution 2592×1944 pixels and a dynamic range of 24 bits. The camera is placed around 20 cm from the bottleneck. The device also includes a mechanical adjusting system to adapt the inspection to different bottleneck lengths. An

adjustable arm holds the bottle from its body, and a press mechanism pushes the bottle against the axle to keep its vertical position.

The electro-mechanical system is commanded by a Basic Stamp micro-controller PIC16C57 connected to a standard personal computer via a RS232 communication port. The micro-controller is programmed in Pbasic (Martin, 2005). For a specified spin angle α (in degrees) the micro-controller synchronizes the step motor with the illumination system. The camera's acquisition process is triggered by a control system via Matlab. The image sequence consists thus of $\lfloor 360/\alpha \rfloor$ different views, where $\lfloor \cdot \rfloor$ is the floor function.

We have built our prototype considering an upside down bottle similar to (J. Wang & Asundi, 2000), but it is also possible to assemble the system with a right side up bottle. Note that in contrast to several inspection systems that make use of multiple cameras (Firmin et al., 1997; Hamad et al., 1998; Ma et al., 2002; Shafait et al., 2004; Katayama et al., 2008) the proposed inspection mechanism employs a single camera only. This simplifies the scene's geometry and suffices to build a robust flaw detection system by analyzing the acquired image sequence. It is important to mention that no camera calibration procedure is considered at all[1]. In the next section we focus on the flaw detection process in uncalibrated image sequences.

3 Inspection System for Flaw Detection in Uncalibrated Images

In some applications a unique image might be enough for inspecting certain objects or materials. However, the use of multiple views can reinforce the diagnosis made with a single image. That is the case for example for low signal-to-noise ratio imaging systems, where the identification of real flaws with poor contrast entails the appearance of numerous false alarms as well. Here, we aim at exploiting the redundant information contained in the bottleneck's multiples views to accurately detect real flaws and to discard false alarms. As it was mentioned before, only real flaws induce geometric and featural relations in the different object's view, while noise and false alarms appear as random events. Geometric characteristics in uncalibrated images are established by both bifocal and trifocal analysis of multiple views (Hartley & Zisserman, 2000), whereas the featural characteristics are extracted by several feature descriptors proposed in the literature. Potential flaws that match this set of characteristics in multiple views are thus regarded as real flaws. This can also be seen as a tracking process. Indeed, we analyze the image sequence taken at successive rotations of the bottle, and classify as real flaws those that can be tracked along it.

Considering the above aspects, we propose a three-step methodology to flaw detection in bottlenecks: i) Segmentation and feature extraction of potential flaws, ii) Computation of corresponding points between views, and iii) Tracking flaws in multiple views. Fig. 3 shows a general overview of the proposed approach for image acquisition and inspection of glass bottlenecks. In the following sections we explain each step in further detail.

1. We refer to (Mery & Carrasco, 2006) for a further discussion about the calibration problem in industrial environments.

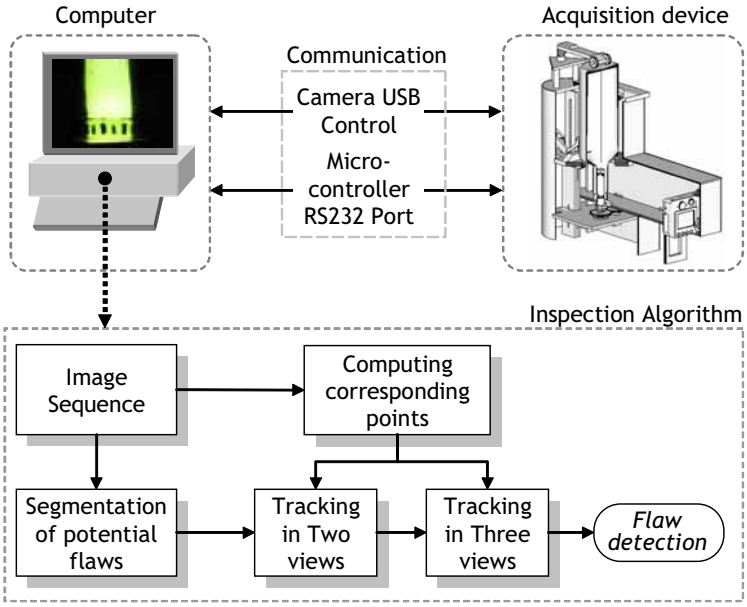


Fig. 3. Proposed system for image acquisition and inspection of glass bottlenecks.

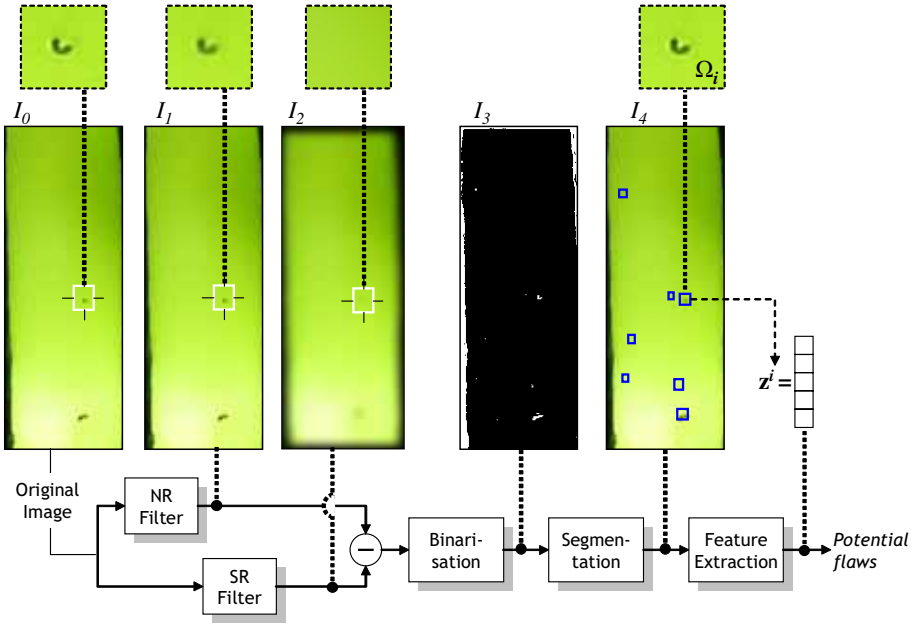


Fig. 4. Segmentation and feature extraction of potential flaws.

3.1 Segmentation and Feature Extraction of Potential Flaws

The segmentation of potential flaws for every image of the sequence is outlined in Fig. 4. Each image I_0 of the sequence is preprocessed by two independent filters. A *noise removal* (NR) filter is used to

reduce the amount of noise intrinsic to any CCD sensor. For this purpose we employ a Gaussian filter with fixed kernel size (3×3), obtaining a filtered image I_1. A *structure removal* (SR) Gaussian filter with kernel size $(n \times n)$, $n \gg 3$[2], is used to blur any structure present in the image. In this way we obtain a uniform background image I_2. Subsequently, the absolute difference $|I_1 - I_2|$ is binarized using the *valley-emphasis* method (Hui-Fuang, 2006), obtaining an image mask I_3 with potential flaws. Finally, for each potential flaw segmented in a region Ω_i a set of features are computed and stored in a feature vector $\mathbf{z}^i$, as shown in the image I_4. Table 1 describes the numerous feature descriptors for greyscale and color images we employ in this study. In Section 3.3 we detail how potential flaws are tracked in two and three different views by matching their feature vectors. The tracking task itself can be highly demanding due to the possibly large number of potential flaws generated by the segmentation, though only a few might correspond to real flaws. To maximize the probability of capturing all real flaws the segmentation process is tuned by the kernel size parameter n. Later on in the experimental section we discuss the effect of varying n, as well as we show the performance of each feature descriptor in the flaw tracking process.

3.2 Computation of Corresponding Points between Views

The geometric relations between potential flaws in different views can be described by sound mathematical concepts from *multiple view geometry* (Hartley & Zisserman, 2000). In particular, we can relate points in two and three views via the *so-called* bifocal and trifocal analysis, respectively. In the next section we develop such analyses. Here we want to point out that in order to compute these geometric relations it is necessary to have accurate corresponding points between the views. Such points will later allow us to match potential flaws along the bottleneck's image sequence.

We can easily find corresponding points by placing equidistant control markers on both extremes of the illumination source (Fig. 1b). The mass centre of these markers is known to us since they were also extracted in the previous segmentation step. The lower control markers are positioned at the same vertical level, while the upper ones follow a sinusoidal wave. Using these markers we can compute a set of corresponding points between each pair of consecutive or non-consecutive views. This is schematically outlined in Fig. 5. The upper and lower control markers of each view are connected through vertical lines between their mass centre. Since the length of a vertical line connecting two particular markers remains constant along the image sequence, we know the relative position of these markers in different views. Therefore, the set of corresponding points between two or three views is conformed by the relative positions of their markers. We employ such correspondences in the following Section 3.3 to establish geometric relations between potential flaws in two and three views.

3.3 Tracking Flaws in Multiple Views

After having segmented and extracted features for all potential flaws in an image sequence, and having computed a set of corresponding points between every pair of images, we now turn to the

2. We efficiently compute Gaussian filtering with large spatial supports using the Fast Fourier Transform (FFT).

TABLE 1
Different feature descriptors we employ to characterize the segmented potential flaws.

Descriptor	Notes
HU	(Hu, 1962) proposed a descriptor composed by seven moment invariant under scale, rotation, translation and skew transformations.
Co-occurrence	(Haralick, Shanmugam, & Dinstein, 1973) introduced a descriptor based on the computation of the co-occurrence matrix. We compute it for the contrast, homogeneity and energy of each color channel (Castleman, 1996).
FSK (A)	(Flusser & Suk, 1993) proposed a set of moments invariant under affine geometric transformations.
FSKS (B)	(Flusser, Suk, & Saic, 1996) designed a set of moments invariant under motion blur.
FSKS (B+S)	(Flusser et al., 1996) extended the FSKS (B) moments by introducing 4 new moments invariant under scale and rotational transformation.
CLP	(Mery, 2003) proposed the *crossing line profiles* (CLP) descriptor whose features correspond to the first five harmonics of the fast Fourier transform.
PSO*	(Mindru, Tuytelaars, Van Gool, & Moons, 2004) introduced a descriptor invariant under photometric transformations of scale and translation (robust to illuminations changes).
GPSO	(Mindru et al., 2004) extended the PSO* descriptor by including invariance under affine geometric deformations.
GPD	(Mindru et al., 2004) proposed another descriptor invariant under diagonal photometric and geometric transformations.
SIFT	(Lowe, 2004) introduced a descriptor invariant under scale and rotation transformations and partially invariant to illumination changes.
PHOG	(Bosch, Zisserman, & X., 2007) proposed a descriptor based on a vectorial representation of the spatial distribution of edges.
SURF	(Bay, Ess, Tuytelaars, & Gool, 2008) introduced a descriptor invariant under scale and rotation transformations.

problem of separating real flaws from false alarms. We achieve this goal by performing tracking of the potential flaws in two and in three views. We know that only real flaws can be tracked along multiple views, since they do induce geometric and featural relations in the different views where they appear, while false alarms correspond to random events. The tracking processes in two and three views are detailed below.

3.3.1 Tracking in two views

In the following we consider that the mass centre of the potential flaw i in the view a is stored in homogenous coordinates as $\mathbf{m}_a^i = [x_a^i, y_a^i, 1]^\top$. If this potential flaw is actually a real flaw it must have a corresponding point $\mathbf{m}_b^j$ in another (non-)consecutive view b where a potential flaw j was also segmented. According to the *principle of multiple view geometry* (Hartley & Zisserman, 2000), the points $\mathbf{m}_a^i$ and $\mathbf{m}_b^j$ correspond each other if they are related by the *fundamental matrix* $\mathbf{F}_{a,b}$

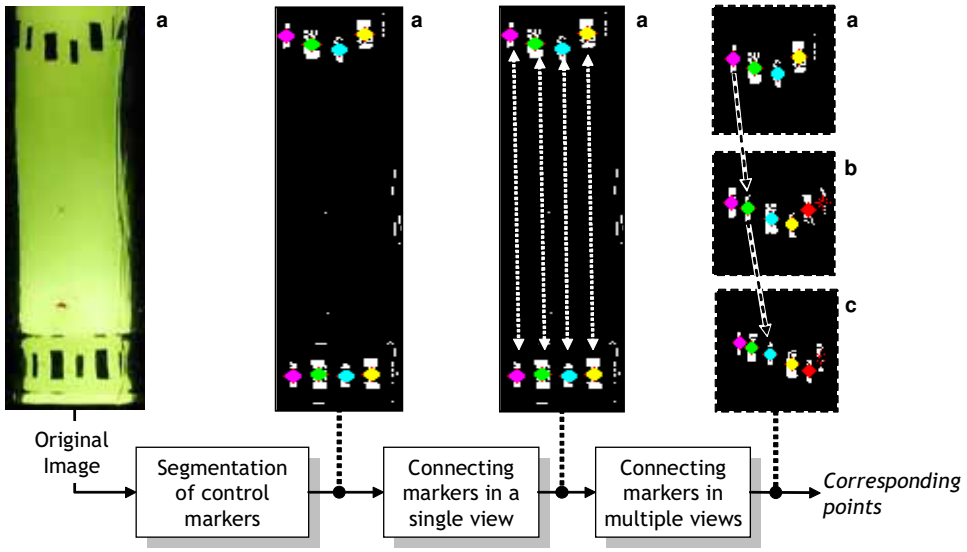


Fig. 5. Computation of corresponding points in two and three views $a, b,$ and c.

such that

$$\mathbf{m}_b^{j\top}\,\mathbf{F}_{a,b}\,\mathbf{m}_a^i = 0\,. \tag{1}$$

Note that the matrix $\mathbf{F}_{a,b}$ is known to us provided that we have estimated it by the algorithm of (Chen, Wu, Shen, Liu, & Quan, 2000) using the correspondences found in Section 3.2. The relation (1) is known as *epipolar constraint*. It indicates that a corresponding point $\mathbf{m}_b^j$ can only lie on the epipolar line of the point $\mathbf{m}_a^i$ defined as $\mathbf{l}_a^i = \mathbf{F}_{i,j}\,\mathbf{m}_a^i = [l_{a,x}^i, l_{a,y}^i, l_{a,z}^i]$. It can happen that there are several points lying on the epipolar line as shown in Fig. 6a. In such a case, we identify the correspondence associated with $\mathbf{m}_a^i$ as the point $\mathbf{m}_b^j$ that satisfies the following condition

$$\frac{|\mathbf{m}_b^{j\top}\,\mathbf{F}_{a,b}\,\mathbf{m}_a^i|}{\sqrt{(l_{a,x}^i)^2 + (l_{a,y}^i)^2}} < \varepsilon\,, \tag{2}$$

for small $\varepsilon > 0$. That is, we choose the point $\mathbf{m}_b^j$ with the smallest (perpendicular) distance to the epipolar line $\mathbf{l}_a^i$. In this case, a geometric match has been found, which could be regarded as a real flaw with a bifocal relationship. However, it is important to emphasize that the condition (2) is not enough to ensure correct matches in two views. Fig. 7 shows the different types of matches that can result:

i) *one-to-one*: Fig. 7a displays the case of only one possible geometric match for every potential flaw, although it does not necessarily mean that every match is correct.

ii) *one-to-many*: Fig. 7b exhibits a possible wrong match and a match that could not be established. This last case occurs when the condition (2) is not fulfilled.

iii) *many-to-one*: Fig. 7c shows multiple matches to the same flaw in the second view.

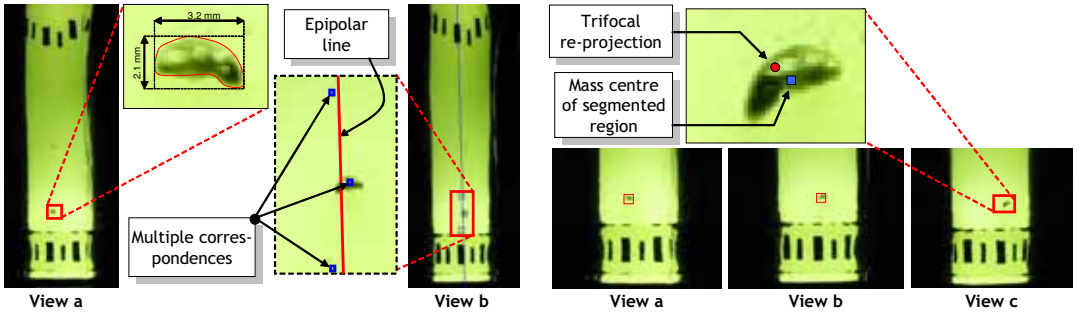


Fig. 6. Examples of (a) bifocal and (b) trifocal correspondences.

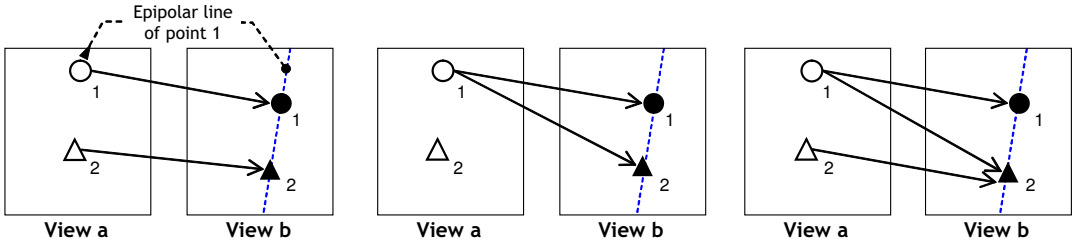


Fig. 7. Possible geometric matches in two views. (a) Ideal *one-to-one* match, (b) *one-to-many* matches that can be resolved by the NNDR criterion, and (c) *many-to-one* matches that can be resolved by the proposed bNNDR criterion.

This tell us that purely geometric relations can lead to wrong matches. Such consequences are less favorable when there is a large number of false alarms, and even worse when we look for correspondences in three views. As a remedy, we analyze not only the geometric characteristics of the potential flaws but also their featural characteristics. In this way, we are able to filter out most of the wrong matches outlined in Fig. 7.

We need to introduce a short notation to describe the feature analysis that we carry out. Let $\mathbf{z}_a^i$ and $\mathbf{z}_b^j$ be the feature vectors (see Section 3.1 and Fig. 4) of the potential flaw i in the first view a and the potential flaw j in the second view b, respectively. We then compute the Euclidean distance between both feature vectors as $d_{a,b}^{i,j} = \|\mathbf{z}_a^i - \mathbf{z}_b^j\|$, and define the array $\mathbf{X}_{a,b}$ containing all possible geometric matches in both views and the vector $\mathbf{D}_{a,b}$ with corresponding featural distances:

Algorithm 1 : *Nearest neighbor distance ratio* (NNDR) criterion for two views a and b.

Input: $\mathbf{X}_{a,b}, \mathbf{D}_{a,b}$, and default parameter $\sigma = 0.7$
Output: $\mathbf{X'}_{a,b}$

1: $p \leftarrow length(\mathbf{X}_{a,b})$
2: **for all** $t \in [1, \ldots, p]$ **do**
3: $\quad \mathbf{X'}_{a,b}(t) \leftarrow$ NULL
4: $\quad \mathbf{Q} \leftarrow$ set of indices q such that $\mathbf{X}_{a,b}(q) = \{t, j\}$, for all j in view b
5: $\quad \mathbf{J}, \mathbf{V} \leftarrow sort\big(\mathbf{D}(\mathbf{Q}),$'increasing order'$\big)$ {$\mathbf{J}$ gets the indices, $\mathbf{V}$ the distances}
6: $\quad$ **if** $\mathbf{V}(1) < \mathbf{V}(2) \cdot \sigma$ **then**
7: $\quad\quad \mathbf{X'}_{a,b}(\mathbf{J}(1)) \leftarrow \{t, \mathbf{J}(1)\}$
8: $\quad\quad \mathbf{X'}_{a,b}(\mathbf{J}(2)) \leftarrow$ NULL
9: $\quad$ **else**
10: $\quad\quad \mathbf{X'}_{a,b}(\mathbf{J}(1)) \leftarrow \{t, \mathbf{J}(1)\}$
11: $\quad\quad \mathbf{X'}_{a,b}(\mathbf{J}(2)) \leftarrow \{t, \mathbf{J}(2)\}$
12: $\quad$ **end if**
13: **end for**

$index$	$\mathbf{X}_{a,b}(index)$	$\mathbf{D}_{a,b}(index)$
1	$\{1,1\}$	$d_{a,b}^{1,1}$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\{i,j\}$	$d_{a,b}^{i,j}$
$\vdots$	$\vdots$	$\vdots$
p	$\{n,m\}$	$d_{a,b}^{n,m}$

Now, given $\mathbf{X}_{a,b} \in \mathbb{Z}^{p\times 2}$ and $\mathbf{D}_{a,b} \in \mathbb{R}^{p\times 1}$ we execute the *nearest neighbor distance ratio* (NNDR) criterion[3] (Mikolajczyk & Schmid, 2005), described in Algorithm 1, to obtain the set of matches $\mathbf{X'}_{a,b} \in \mathbb{Z}^{p\times 2}$ that minimize the featural distances among all possible matches. As it can be noted in Algorithm 1, line 4, this criterion is useful to resolve matches of type *one-to-many*, as depicted in Fig. 7b. However, it fails to correct the wrong *many-to-one* matches shown in Fig. 7c.

To overcome this problem, we introduce a novel and simple modification to the NNDR criterion that is able to resolve the general case of having *many-to-many* matches. We call it *bidirectional* NNDR (bNNDR) criterion because it checks both the possible matches going from the first view a to the second view b (as in Algorithm 1, line 4) and those going from the second to the first view. By looking at both directions we adjust the distance feature vector $\mathbf{D}_{a,b}$ with a weighting vector $\mathbf{W}_{a,b} \in \mathbb{R}^{p\times 1}$. Algorithm 2 shows the computation of $\mathbf{W}_{a,b}$. The weights are set to 1 for *one-to-one* matches and for the matches with minimal *backward* distance, while matches with larger distances get weights larger than 1. In this way, the distance vector $\mathbf{D}_{a,b}$ is cross-correlated with information from all multiple matches. The proposed bNNDR criterion, fully described in Algorithm 3, outperforms

3. There exist alternative criteria in the literature, but we chose the NNDR because it is computationally inexpensive. See (Sidibe, Montesinos, & Janaqi, 2007).

Algorithm 2 : *W-weights* for two views a and b.

Input: $\mathbf{X}_{a,b}, \mathbf{D}_{a,b}$
Output: $\mathbf{W}_{a,b}$

$p \leftarrow length(\mathbf{X}_{a,b})$
for all $t \in [1, \ldots, p]$ **do**
  $\mathbf{W}_{a,b}(t) \leftarrow$ NULL
  $\mathbf{J} \leftarrow$ set of indices q such that $\mathbf{X}_{a,b}(q) = \{t, j\}$, for all j in view b
  $\mathbf{I} \leftarrow$ set of indices q such that $\mathbf{X}_{a,b}(q) = \{i, t\}$, for all i in view a
  if $cardinality\big(\mathbf{I}\big) == 1$ **then**
    $\mathbf{W}_{a,b}(\mathbf{J}) \leftarrow 1$
  else
    $\mathbf{W}_{a,b}(\mathbf{I}) \leftarrow \dfrac{\mathbf{D}(\mathbf{I})}{min\big(\mathbf{D}(\mathbf{I})\big)}$
  end if
end for

Algorithm 3 : *Bidirectional* NNDR (bNNDR) criterion for two views a and b.

Input: $\mathbf{X}_{a,b}, \mathbf{D}_{a,b}$
Output: $\mathbf{X'}_{a,b}$

$p \leftarrow length(\mathbf{X}_{a,b})$
$\mathbf{W}_{a,b} \leftarrow$ *W-weights*$(\mathbf{X}_{a,b}, \mathbf{D}_{a,b})$ {by Algorithm 2}
for all $t \in [1, \ldots, p]$ **do**
  $\mathbf{D'}_{a,b}(t) \leftarrow \mathbf{W}_{a,b}(t) \cdot \mathbf{D}_{a,b}(t)$
end for
$\mathbf{X'}_{a,b} \leftarrow$ *NNDR*$(\mathbf{X}_{a,b}, \mathbf{D'}_{a,b})$ {by Algorithm 1}

the NDDR criterion in identifying the correct matches from sets of *many-to-many* possible ones, which will be demonstrated in the experimental section.

3.3.2 Tracking in three views

If the segmentation stage (Section 3.1) outputs a large number of potential flaws of small size, which might have very similar feature vectors, it is likely that the precedent two-view tracking process classifies matches that correspond to false alarms as real flaws. In order to avoid such a result, we now elaborate on a three-view tracking mechanism to discard wrong matches and to confirm those that indeed represent real flaws.

In terms of geometry of multiple views, given a potential flaw with corresponding coordinates $\mathbf{m}_a^i$ and $\mathbf{m}_b^j$ in the views a and b, one can estimate the hypothetical position of a third correspondence $\mathbf{m}_c^k$ in a view c by

$$\widehat{\mathbf{m}}_c^k = \frac{1}{\mathbf{m}_a^{i\top}(T_\cdot^{13} - x_b^j\, T_\cdot^{33})} \begin{bmatrix} \mathbf{m}_a^{i\top}(T_\cdot^{11} - x_b^j\, T_\cdot^{31}) \\ \mathbf{m}_a^{i\top}(T_\cdot^{12} - x_b^j\, T_\cdot^{32}) \\ \mathbf{m}_a^{i\top}(T_\cdot^{13} - x_b^j\, T_\cdot^{33}) \end{bmatrix},$$

where the $3 \times 3 \times 3$ trifocal tensor[4] $\mathbf{T} = (T_t^{rs})$ encodes the relative motion among the views $a, b,$

4. Computed using the *point-line-point* method (Hartley & Zisserman, 2000).

Algorithm 4 : *Bidirectional* NNDR (bNNDR) criterion for three views a, b, and c.

Input: $\mathbf{X}_{a,b,c}$
Output: $\mathbf{X'}_{a,b,c}$
1: $p \leftarrow length(\mathbf{X}_{a,b,c})$
2: **for all** view-pair $(x, y) \in \{(a, b), (a, c), (b, c)\}$ **do**
3: $\quad \mathbf{X}_{x,y} \leftarrow$ *bifocal matches*$(x, y, \mathbf{X}_{a,b,c})$
4: $\quad \mathbf{D}_{x,y}, \leftarrow$ *feature distances*$(\mathbf{X}_{x,y})$
5: $\quad \mathbf{W}_{x,y} \leftarrow$ *W-weights*$(\mathbf{X}_{x,y}, \mathbf{D}_{x,y})$
$\quad$ {Note: $length(\mathbf{X}_{x,y}) = length(\mathbf{D}_{x,y}) = length(\mathbf{W}_{x,y}) = p$}
6: **end for**
7: **for all** $t \in [1, \ldots, p]$ **do**
8: $\quad \mathbf{D'}_{a,b,c}(t) \leftarrow \mathbf{W}_{a,b}(t) \cdot \mathbf{D}_{a,b}(t) \cdot \mathbf{W}_{a,c}(t) \cdot \mathbf{D}_{a,c}(t) \cdot \mathbf{W}_{b,c}(t) \cdot \mathbf{D}_{b,c}(t)$
9: **end for**
10: $\mathbf{X'}_{a,b,c} \leftarrow$ *NNDR*$(\mathbf{X}_{a,b,c}, \mathbf{D'}_{a,b,c})$

and c. We compare the projected position $\widehat{\mathbf{m}}_c^k$ to all segmented potential flaws $\mathbf{m}_c^k$ in the third view c, and choose as corresponding the one that satisfies

$$\|\widehat{\mathbf{m}}_c^k - \mathbf{m}_c^k\| < \varepsilon\,. \tag{3}$$

If condition (3) is fulfilled, as in the example of Fig. 6b, a geometric correspondence in three views has been found and it remains to check whether their respective feature vectors are also close to each other. If no segmented region in the view c satisfies (3) we then regard the correspondences $\mathbf{m}_a^i$ and $\mathbf{m}_b^j$ as false alarms.

As it happened in the two-view tracking process, it can also occur that several potential flaws in the view c fulfill the condition (3). However, at most one of these multiple geometric matches is correct. Let $\mathbf{X}_{a,b,c} \in \mathbb{Z}^{p\times 3}$ be an array with all triplets $\{i, j, k\}$ that fulfill the condition (3). We filter out the wrong matches by analyzing their featural characteristics with the adapted bNNDR criterion for three views described in Algorithm 4. The output array $\mathbf{X'}_{a,b,c}$ contains the correct triplets whose feature vectors match in the three views. In the following section we show that this algorithm allows us to distinguish real flaws from false alarms with high accuracy. Although our methodology can be easily extended to process n views, it suffices to take up to three for the application considered in this paper.

4 Experimental Results

We now evaluate the performance of the proposed methodology for inspecting bottlenecks of empty wine bottles. In our experiments we use 120 color image sequences. Each sequence consists of 3 views with a rotation angle $\alpha = 15$ degrees between them. From the recorded images we extract sub-images of the bottlenecks of 1000×250 pixels. The number of real flaws per image fluctuates between 0 and 4 with an average of 2.8, and the number of false alarms per image fluctuates between 0 and 10 with an average of 9.5 per image. The area of the smallest flaw is around 9 pixels equivalent to 0.16 mm^2 on the bottle's surface. The performance is assessed considering two standard indicators (Olson, 2008): $r = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ (recall) and $p = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$ (precision). TP is the

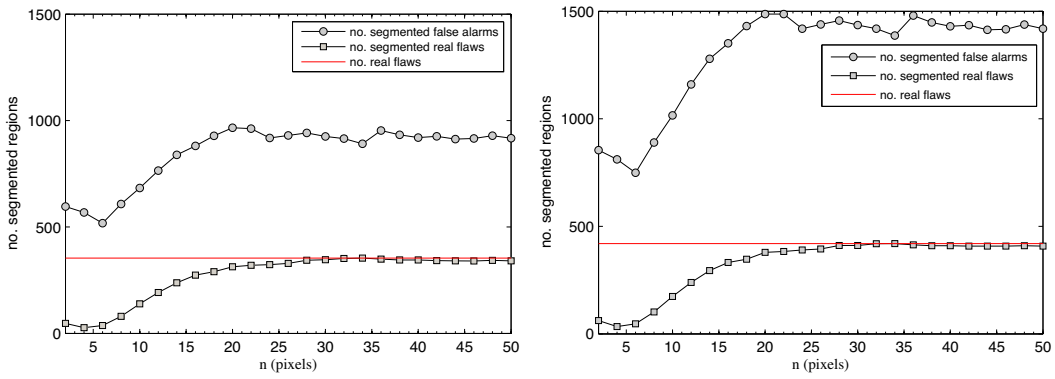


Fig. 8. Influence of the kernel size $(n \times n)$ of the *structure removal* (SR) filter in the segmentation process. The larger n the more real flaws can be successfully segmented (a) in two views and (b) in three views. The solid line indicates the total number of real flaws present in the image sequences.

number of *true positives* or flaws correctly classified as such by the inspection system. FN is the number of *false negatives* or existing real flaws not detected. FP is the number of *false positives* or flawless regions that are incorrectly classified as defective. These two indicators can be cast in a unique measure F-score $= \frac{2 \cdot p \cdot r}{p+r}$ (Olson, 2008). Ideally, one can expect $r = 100\%, p = 100\%$, and F-score $= 1$. In the following subsections we evaluate several aspects of the proposed framework for detecting flaws in uncalibrated images of glass bottlenecks.

4.1 Evaluation of the Segmentation Process

The segmentation process described in Section 3.1 outputs a set of regions with potential flaws. This process has to be able to segment all real flaws present in the bottlenecks. Fig. 8 shows how good this task is performed as a function of the kernel size of the *structure removal* (SR) filter utilized to obtain a background image. The more uniform this image, the more real flaws are successfully extracted by the segmentation. However, the number of false alarms, i.e. regions that do not represent real flaws, also augment with increasing n. In the subsequent experiments we have set $n = 34$.

4.2 Evaluation of the Tracking Processes

We now test the performance of our inspection methodology for detecting reals flaws and discarding false alarms. In the following, we denote as 2V (two views) and 3V (three views) the inspection results obtained using only the geometric conditions (2) and (3), respectively. Similarly, 2V+*criterion* and 3V+*criterion* represent the inspection results obtained by further utilization of one of the feature analysis criteria NNDR or bNNDR. Fig. 9 and Fig. 10 exhibit the F-scores of the tracking processes in two and three views using all feature descriptors described in Table 1,

considering both the NNDR criterion (Mikolajczyk & Schmid, 2005) and the proposed *bidirectional* NNDR (bNNDR) criterion. The results are displayed as a function of the parameter ε that denotes the maximal distance (in pixels) at which one looks for geometric matches in two and three views, cf. conditions (2) and (3). Note that in all plots the three-view inspection process largely outperforms its two-view counterpart. It is also notorious the improvement introduced by our bNNDR criterion in contrast to the NNDR criterion. We recall that these criteria are useful to resolve misleading *many-to-many* geometric matches (Fig. 7).

The non-calibrated nature of the images we work with can produce imprecisions in the estimation of the fundamental matrices and trifocal tensors. Therefore it is sometimes necessary to use a larger parameter ε, as it is drawn in Fig. 9 and Fig. 10. Although this can lead to the appearance of more misleading geometric matches, this problem is effectively overcome thanks to the utilization of the feature analysis criteria. Fig. 11 shows the performance of the tracking algorithms when only geometric matches are sought in contrast to the performance obtained by additionally employing the proposed bNNDR criterion for feature analysis. It is clear that the latter case provides better results, specially when only two views are considered. Fig. 12 displays the performance using optimal ε-values for the geometric matches in combination with the bNNDR criterion with each feature descriptor. Fig. 13 shows the relative performance improvement of the bNNDR criterion compared to the NNDR criterion.

In Table 2 we juxtapose our inspection results with those from other inspection systems proposed in the literature, indicating the usage of single or multiple images. We compare the different methods in terms of the *true positive rate* TPR = TP / (TP + FN) and the *false positive rate* FPR = FP / (FP + TN), where TN is the number of segmented regions correctly classified as false alarms. An ideal inspection system would have 100% TPR and 0% FPR. It is important to mention that Table 2 corresponds just a quantitative comparison, since the other methods proposed in the literature were tested on different images, types of bottles, and they inspect one or several bottle parts (lips, mouth, bottleneck, body, bottom). Nevertheless, our results correspond to the most accurate reported in the literature regarding the inspection of glass bottlenecks. We want to emphasize the relevance of the combined use of geometry of multiple views and feature analysis of the potential flaws to distinguish between real flaws and false alarms effectively.

Concerning the computational aspects of our inspection system, the total time required to process three views was 1.3 sec in average, running Matlab (7.0) code on a Pentium Centrino 2.0 GHz with Windows XP SP2. The time was spent as follows: Reading images from hard disk (34%), segmenting potential flaws (35%), geometric and featural analysis in three views (11%), and other Matlab's internal operations (20%). This indicates that there is a lot of room for improvements in each of these tasks.

5 CONCLUSIONS

In this paper we have developed two independent contributions. First, we presented the prototype of an image acquisition system for capturing image sequences of glass bottlenecks using a single

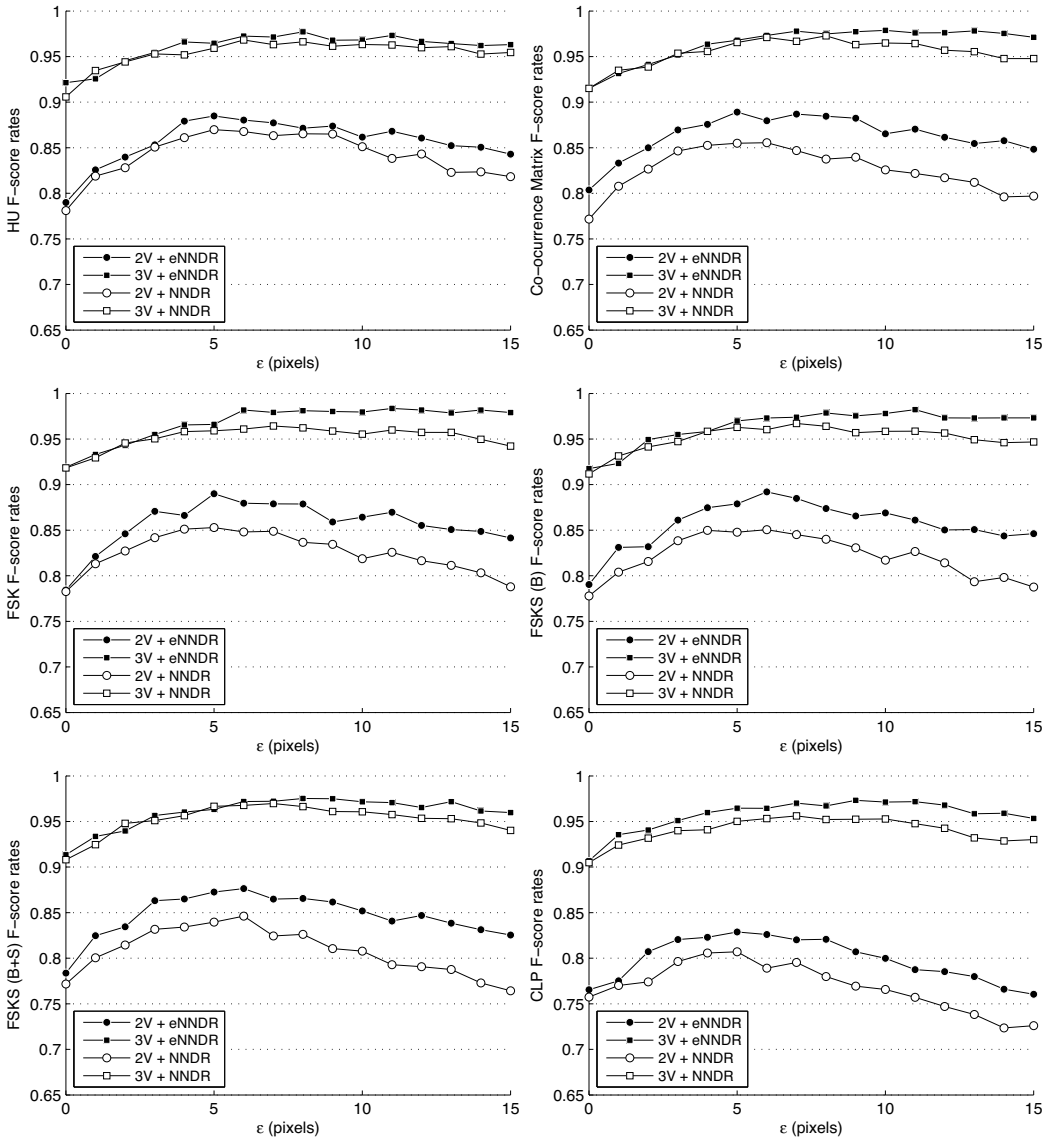


Fig. 9. Performance comparison of the proposed two- and three-view inspection processes. We test the influence of the parameter ε for obtaining the geometric matches, as well as the NNDR and bNNDR criteria for feature analysis using the feature descriptors (from left to right, top to bottom): HU, Co-occurrence, FSK, FSK(B), FSK(B+S), CLP (see Table 1).

camera, where no camera calibration process is considered. The main novelty of this prototype is the placement of the illumination source inside the bottle, which greatly improves the quality of the acquired images, avoiding the intrinsic reflections produced by external light sources. Second, we introduced a novel methodology for inspecting glass bottlenecks using uncalibrated images.

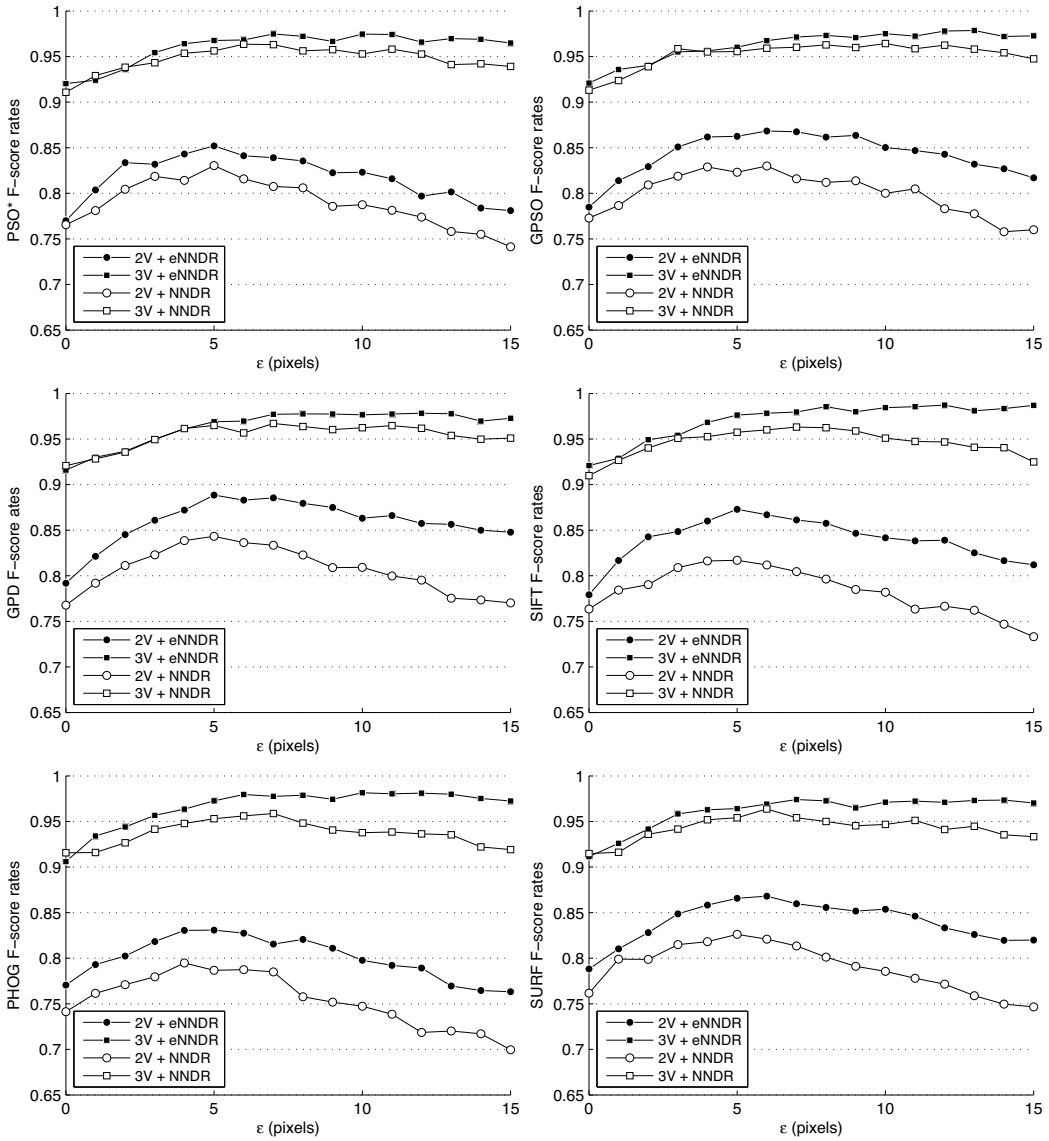


Fig. 10. Performance comparison of the proposed two- and three-view inspection processes. We test the influence of the parameter ε for obtaining the geometric matches, as well as the NNDR and bNNDR criteria for feature analysis using the feature descriptors (from left to right, top to bottom): PSO*, GPSO, GPD, SIFT, PHOG, SURF (see Table 1).

Our inspection system examines series of two and three images employing geometry of multiple views followed by a feature analysis stage to discriminate between real flaws and false alarms. In this way, we classify as real flaws those that present similar characteristics in a set of images taken from different viewpoints. An important ingredient to achieve this goal was the introduction

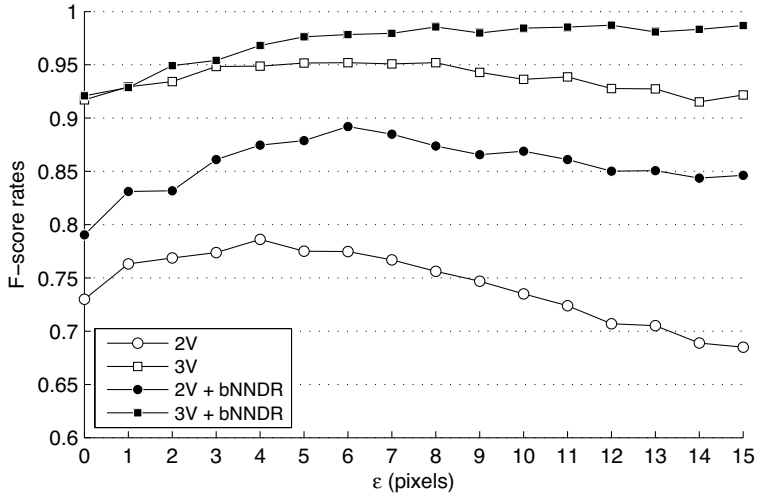


Fig. 11. Performance comparison of the inspection using only geometric constraints (2V and 3V) and including the proposed bNNDR criterion for feature analysis. In the latter case, we use the feature descriptors FSKS(B) and SIFT in two and three views, respectively.

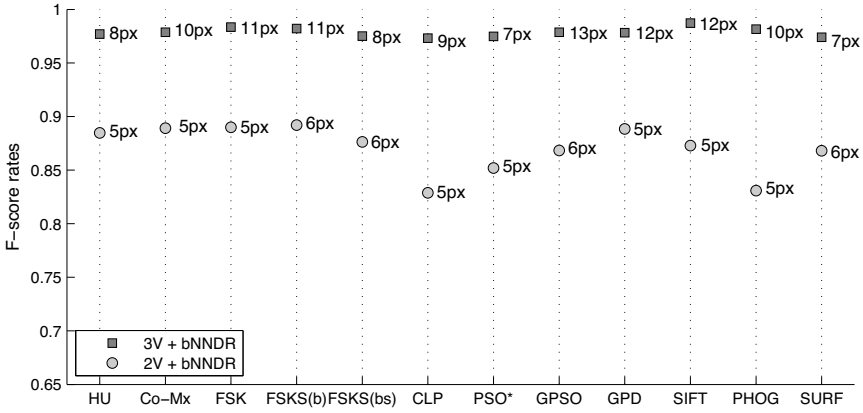


Fig. 12. Best inspection performance running the bNNDR criterion for each feature descriptor. Optimal ε-values has been chosen.

of a novel feature analysis criterion to resolve multiple geometric matches in different views. It can be considered as a bidirectional variant of the *nearest neighbor distance ratio* (NNDR) criterion proposed by (Mikolajczyk & Schmid, 2005). Our inspection system, tested on image sequences of wine glass bottles with real flaws, obtained a *true positive rate* of 99.1% and a *false positive rate* of 0.9%.

An important characteristic of the proposed methodology for flaw detection is that no camera calibration is considered at all. This makes our method suitable for applications where camera calibration is difficult or expensive to carry out. Moreover, our approach is generic in the sense that

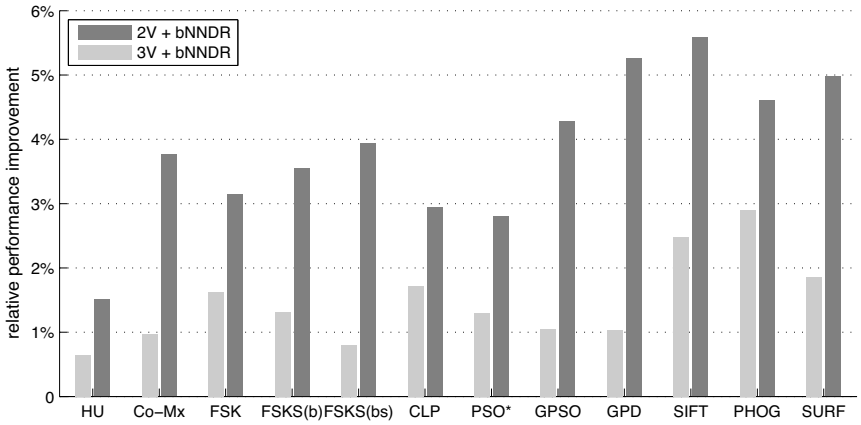


Fig. 13. Mean relative performance improvement achieved by the proposed bNNDR criterion with respect to the NNDR criterion for each feature descriptor.

TABLE 2
Quantitative comparison of inspection systems for flaw detection in glass bottles. Our proposed methodology achieves the highest performance detecting flaws in the bottlenecks.

Inspected bottle part	Views	Tracking	TPR	FPR
neck (Mery & Medina, 2004)	Single	No	85%	4%
lips, body, bottom (Duan et al., 2007)	Single	No	97%	>1%
lips (Y.-N. Wang et al., 2005)	Single	No	98%	0%
body (Firmin et al., 1997; Hamad et al., 1998)	Multiple	No	85%	2%
lips, **neck** (Ma et al., 2002)	Multiple	No	98%	2%
body, bottom (Shafait et al., 2004)	Multiple	No	100%	>1%
neck (our method)	2V	Yes	99.9%	46.4%
	2V + bNNDR	Yes	99.9%	30.2%
	3V	Yes	99.1%	2.5%
	3V + bNNDR	Yes	**99.1%**	**0.9%**

it can be used for the visual inspection of other manufactured objects as well.

ACKNOWLEDGEMENTS

M.C. was partially supported by the *National Commission for Scientific and Technological Research* (CONICYT), and L.P. was partially supported by the *German Academic Exchange Service* (DAAD) under grant A/05/21715.

REFERENCES

Bay, H., Ess, A., Tuytelaars, T., & Gool, L. (2008). Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, *110*(3), 346–359.

Bosch, A., Zisserman, A., & X., M. (2007, July 9–11). Representing shape with a spatial pyramid kernel. In ACM (Ed.), *Proceedings of the 6th acm international conference on image and video retrieval (civr)* (pp. 401 – 408). Amsterdam, The Netherlands: ACM.

Canivet, M., Zhang, R. D., & Jourlin, M. (1994). Finish inspection by vision for glass production. In B. M. Dawson, S. S. Wilson, & F. Y. Wu (Eds.), *Machine vision applications in industrial inspection ii* (Vol. 2183, pp. 164–169). Proceedings of SPIE.

Carrasco, M., & Mery, D. (2006). Automated visual inspection using trifocal analysis in an uncalibrated sequence of images. *Materials Evaluation*, *64*(9), 900–906.

Carrasco, M., Pizarro, L., & Mery, D. (2008). Image acquisition and automated inspection of wine bottlenecks by tracking in multiple views. In *Proc. of 8th int. conf. on signal processing, computational geometry and artificial vision (iscgav'08)* (pp. 82–89). Rhodes Island, Greece: WSEAS Press.

Castleman, K. (1996). *Digital image processing*. New Jersey: Prentice-Hall.

Chen, Z., Wu, C., Shen, P., Liu, Y., & Quan, L. (2000). A robust algorithhm to estimate the fundamental matrix. *Pattern Recogn Letters*, *21*, 851–861.

Drury, C. G., Saran, M., & Schultz, J. (2004, Jan). *Effect of fatigue / vigilance/ environment on inspectors performing fluorescent penetrant and/or magnetic particle inspection* (Interim Report No. 03-G-012). Federal Aviation Administration William J. Hughes Technical Center: University at Buffalo.

Duan, F., Wanga, Y.-N., Liua, H.-J., & Li, Y.-G. (2007). A machine vision inspector for beer bottle. *Eng Appl Artif Intell*, *20*(7), 1013–1021.

Firmin, C., Hamad, D., Postaire, J., & Zhang, R. (1997). Gaussian neural networks for bottles inspection: a learning procedure. *Int J Neural Syst*, *8*(1), 41-46.

Flusser, J., & Suk, T. (1993). Pattern recognition by affine moment invariants. *Pattern Recogn*, *26*(1).

Flusser, J., Suk, T., & Saic, S. (1996). Recognition of images degraded by linear motion blur without restoration. *Computing. Supplement*, *11*, 37–51.

Hamad, D., Betrouni, M., Biela, P., & Postaire, J. (1998). Neural networks inspection system for glass bottles production: A comparative study. *Int J Pattern Recogn Artif Intell*, *12*(4), 505-516.

Haralick, R., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. on Systems, Man, and Cybernetics*, *SMC-3*(6), 610-621.

Hartley, R., & Zisserman, A. (2000). *Multiple view geometry in computer vision*. Cambridge, UK: Cambridge University Press.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Trans Info Theory*, *IT*(8), 179-187.

Hui-Fuang, N. (2006). Automatic thresholding for defect detection. *Pattern Recogn Letters*, *27*(14), 1644-1649.

Jacob, R., Raina, S., Regunath, S., Subramanian, R., & Gramopadhye, A. K. (2004). Improving inspector's performance and reducing errors - general aviation inspection training systems (gaits). In *In proceedings of the human factors and ergonomics society annual meeting proceedings*. Human Factors and Ergonomics Society.

Katayama, K., Ishikura, T., Kodoma, Y., Fukuchi, H., & Fujiwara, A. (2008, Feb). Optical inspection of glass bottles using multiples cameras. *USA Pantent*. http://www.patentstorm.us/patents/7329855/claims.html. (No. 7.329.855 B2)

Kopparapu, S. K. (2006). Lighting design for machine vision application. *Image Vis Comput*, *24*(7), 720-726.

Kumar, A. (2008, Jan.). Computer-vision-based fabric defect detection: A survey. *IEEE Transactions on Industrial Electronics*, *55*(1), 348–363.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int J Comput Vis*, *60*(2), 91-110.

Ma, H.-M., Su, G.-D., & Ni, J.-Y. W. Z. (2002). A glass bottle defect detection system without touching. In *Int. conf. on machine learning and cybernetics* (Vol. 2, p. 628-632). IEEE.

Malamas, E., Petrakis, E. G., & Zervakis, M. (2003). A survey on industrial vision systems, applications and tools. *Image and Vision Computing*, *21*(2), 171–188.

Marchand, E. (2007, April). Control camera and light source positions using image gradient information. In *Ieee int. conf. on robotics and automation (icra)*. Rome, Italy: IEEE.

Martin, J. (2005). Basic stamp syntax and reference manual. *Parallax USA*. (accessed in 2009, www.parallax.com/dl/docs/prod/stamps/web-BSM-v2.2.pdf)

Mery, D. (2003). Crossing line profile: a new approach to detecting defects in aluminium castings. *LNCS*, *2749*, 725–732.

Mery, D., & Carrasco, M. (2006). Advances on automated multiple view inspection. *LNCS*, *4319*, 513–522.

Mery, D., & Filbert, D. (2002). Automated flaw detection in aluminum castings based on the tracking of potential defects in a radioscopic image sequence. *IEEE Trans. Robotics and Automation*, *18*(6), 890-901.

Mery, D., & Medina, O. (2004). Automated visual inspection of glass bottles using adapted median filtering. *LNCS*, *3212*, 818-825.

Mikolajczyk, K., & Schmid, C. (2005, Oct.). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(10), 1615–1630.

Mindru, F., Tuytelaars, T., Van Gool, L., & Moons, T. (2004). Moment invariants for recognition under changing viewpoint and illumination. *Comput Vis Image Underst*, *94*(1-3), 3–27.

Mital, A., Govindaraju, M., & Subramani, B. (1998). A comparison between manual and hybrid methods in parts inspections. *Integrated Manufacturing Systems*, *9*(6), 344–349.

Olson, D., David L.; Delen. (2008). *Advanced data mining techniques*. Springer.

Parker, J. (2000, 9-10 Oct). Defect in glass and their origin. In *First balkan conference on glass science and technology*. Vollos, Greace: University of Thessaly.

Pizarro, L., Mery, D., Delpiano, R., & Carrasco, M. (2008). Robust automated multiple view inspection. *Pattern Analysis and Applications*, *11*(1), 21–32.

Shafait, F., Imran, S., & Klette-Matzat, S. (2004). Fault detection and localization in empty water bottles through machine vision. In *Emerging technology conference (e-tech)* (pp. 30–34). IEEE.

Sidibe, D., Montesinos, P., & Janaqi, S. (2007, March). Fast and robust image matching using contextual information and relaxation. In *2nd international conference on computer vision theory and applications, visapp*. Barcelona, Spain: Springer.

Vazquez, P.-P. (2007, June). Automatic light source placement for maximum visual information recovery. *Computer Graphics Forum*, *26*(2), 143–156.

Wang, J., & Asundi, A. (2000). A computer vision system for wineglass defect inspection via gabor-filter-based texture features. *Inform Sci*, *127*, 157–171.

Wang, Y.-N., H.-J., L., & Duan, F. (2005). A bottle finish inspect method based on fuzzy support vector machines and wavelet transform. In *International conference on machine learning and cybernetics* (Vol. 8, pp. 4588–4592). IEEE.

Yan, T.-S., & Cui, D.-W. (2006). The method of intelligent inspection of product quality based on computer vision. In *Proc. of the 7th int. conf. on computer-aided industrial design and conceptual design (caidcd '06)* (pp. 1–6). Hangzhou: IEEE.

Yepeng, Z., Yuezhen, T., & Zhiyong, F. (2007). Application of digital image process technology to the mouth of beer bottle defect inspection. In *Proc. of the 8th int. conf. on electronic measurement and instruments* (Vol. 2, pp. 905–908). IEEE.

Chapter 4

Point-to-point correspondence

4. POINT-TO-POINT CORRESPONDENCE

This chapter presents a paper on point-to-point correspondence in uncalibrated image sequences. This problem has a fundamental role in the field of computer vision because through the correspondence it is possible to determine the mathematical relation that exists between two or more images. This relation allows, for example, determining the object's geometry, making a 3D reconstruction, determining the transformation function of an image into another, detecting patterns or regions of permanent interest in time, etc. Our proposal consists of extending the geometric model used in the previous chapters.

According to the literature, current methods of correspondence and extraction of points of interest are robust toward diverse geometric and photometric transformations in corresponding images. However, these methods maximize their performance in some zones of the image according to a particular metric. The problem is that these techniques do not necessarily perform well finding corresponding regions or points in positions that have not been detected previously. To solve this problem we propose the determination of multiple geometric models in two and three views evaluating their corresponding error and compensating for that error in the estimation of the correspondence of the point in other views. In this way we show that by means of our algorithm it is possible to determine the correspondence with high precision, and moreover in any position in the image.

Our evaluation considers three types of images: images with external scenery and illumination, test objects with artificial illumination, and a sequence of images of bottle necks generated in the previous chapter. In the first two groups of images we determined the performance of the algorithm carrying out point-to-point correspondence in random positions whose real correspondence is known. In the last test group we determined the correspondence only in potential defects as a method to extend the AMVI model in product inspection. The results show that it is possible to determine the correspondence provided the algorithm has sufficient base points in correspondence dispersed in the image, in such a way that each partial model has an error close to the minimum. In other cases the algorithm has been effective to find the correspondence with performance close to the optimum.

Paper #5
On solving the point-to-point correspondence problem using multiple geometrical solutions

Miguel Carrasco and Domingo Mery
mlcarras@puc.cl - dmery@ing.puc.cl

Abstract

Determining point-to-point correspondence in multiple images is a complex problem because of the multiple geometric and photometric transformations and/or occlusions that the same point can undergo in corresponding images. Those transformations are generated either by the intrinsic motion of the point or by the movement of the camera over the object. Different approaches have been proposed to solve this problem, of which those methods that use the analysis of invariant features, which can be applied to sequences with large variations, are particularly important. This paper presents a method of point-to-point correspondence analysis based on the combination of two techniques: (1) correspondence analysis through similarity of invariant features, and (2) combination of multiple partial solutions through bifocal and trifocal geometry. This method is quite novel because it allows the determination of point-to-point geometric correspondence by means of the intersection of multiple partial solutions that are weighted through the MLESAC algorithm. The main advantage of our method is the extension of the algorithms based on the correspondence of invariant descriptors, generalizing the problem of correspondence to a geometric model in multiple views. The method has been evaluated in three types of sequences: indoor, outdoor, and industrial images. In the indoor sequence we got an F-score $= 87\%$ at a distance of less than 2 pixels. This low performance is due to the lower dispersion of the correspondences to determine the robust geometric model. In the outdoor and industrial sequences we got an F-score $= 97\%$ at a distance of less than 1 pixel, mainly because of the greater number of correspondences and of the better dispersion of points in corresponding images. These results show the effectiveness of the method in a wide range of applications.

Index Terms

computer vision, multiple view geometry, correspondence problem,tracking

1 Introduction

The point-to-point correspondence analysis between two or more images made of the same scene is a very relevant problem in the computer vision community. Problems such as 3D reconstruction,

Miguel Carrasco and Domingo Mery are with the Computer Engineering Department at Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860 (143), Santiago, Chile.

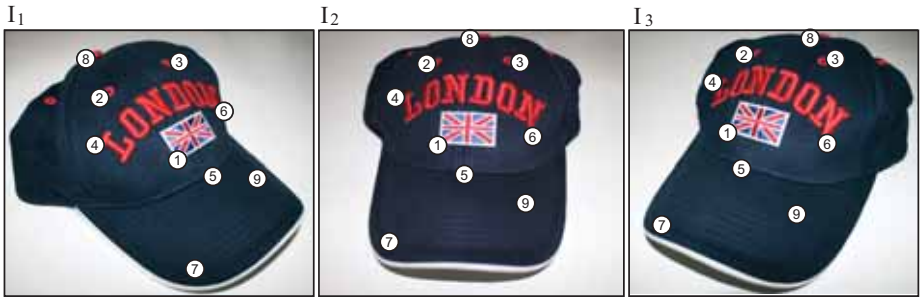


Fig. 1. Sequence of nine corresponding points $\{1,\ldots,9\}$ in the $\mathbf{I}_1$, $\mathbf{I}_2$,$\mathbf{I}_3$ views respectively.

robotic navigation, tracking in multiple views, estimation of transformation matrices and homographies, among others, are some of the applications that require the precise determination of correspondences. Correspondence analysis consists basically in determining a set of points in an image such that they are identified as the same in other images of the same scene. This situation is described clearly in Fig. 1, which shows nine corresponding points in three images of the same object. To human beings this problem does not represent any difficulty; however, determining these correspondences computationally is not a simple task. We must consider that corresponding points can undergo various transformations depending on the points of view from which they have been captured. This is due to the geometric and/or photometric transformations caused by the motion of the object as well as by the movement of the camera with respect to the object. To increase the problem's complexity, it is possible for other points to have a texture and color similar to that of the point whose correspondence we want to determine, increasing the complexity of the task of discriminating among possible corresponding pairs and triplets.

Different approaches for solving the correspondence matching have been developed over the last 30 years. Some of them are, for example, methods based on the analysis of invariant descriptors (Bay, Ess, Tuytelaars, & Gool, 2008; Lowe, 2004; Bosch, Zisserman, & X., 2007), estimation of affine transformations, homographies and estimation of perspective transformations (Caspi & Irani, 2000; Fitzgibbon, 2003), epipolar geometry analysis (Romano, 2002; Vidal, Ma, Soatto, & Sastry, 2006), and methods based on optical flow (Lucas & Kanade, 1981; Barron, Fleet, & Beauchemin, 1994). In general, all these methods differ in the type of motion of the objects contained in a video sequence, or in the simplest case through correspondence between two images. If the scene is static and there is no continuous change of the camera's position, the problem is reduced mainly to the analysis of the epipolar geometry for two images through stereo vision (Scharstein & Szeliski, 2002). On the other hand, if the scene is dynamic and the objects undergo small displacements, the differential techniques through optical flow have been shown to be an efficient way of determining the correspondences. Unfortunately, the latter are not designed for extensive displacements.

In spite of the large number of methods designed to solve this problem, the correspondence of images with very wide viewing angles has not been solved completely. This problem is commonly found when two or more independent cameras located at different positions and with wide perspective

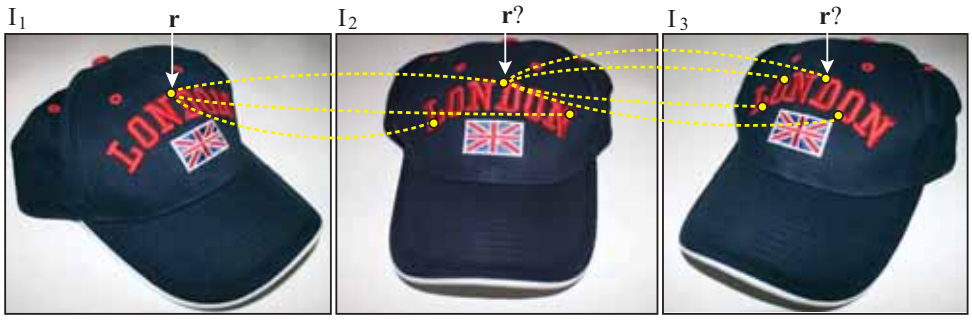


Fig. 2. General problem of correspondences in multiple views. Main objective: determining the correspondence of point **r** of image $\mathbf{I}_1$ in corresponding images $\mathbf{I}_2$ e $\mathbf{I}_3$.

angles are used (Caspi, Simakov, & Irani, 2006). On the other hand, this problem can be considered in a monocular way in a video sequence when an inter-frames correspondence is used (Carrasco & Mery, 2006). In such cases, the position of the objects and the camera vary in time, causing the objects of the sequence to be in very different positions. In the latter cases it is common to use techniques based on the analysis of invariant descriptors. Thanks to the invariance it is possible to solve and generalize the point-to-point correspondence problem by providing an extension of the methods based on stereo vision. That correspondence takes place as a function of the points of interest detected previously by some algorithm for detecting regions of interest (Moreels & Perona, 2007; Bhat et al., 2006). However, when the point of interest does not correspond to a point detected by the current saliency techniques (Kadir, Zisserman, & Brady, 2004; Matas, Chum, Urban, & Pajdla., 2002; Mikolajczyk & Schmid., 2004; Tuytelaars & Mikolajczyk, 2007), how can we determine its corresponding pair or triplet in the other images? In this case, the previous methods do not ensure finding a correct correspondence because they are designed to maximize their performance only in the regions of interest detected by the method, and not necessarily in other regions. This problem occurs commonly when the image sequence has a low signal/noise ratio, so invariant algorithms will not perform well due to the appearance of many false alarms.

To avoid the problems of the correspondence methods mentioned previously, in this research we propose a new method to determine the point-to-point correspondence in any kind of image, particularly when the displacement angles are wide. Furthermore, since we use a geometric model that is independent of the objects, it is possible to determine the position of corresponding points in those views in which the point may be occluded. Graphically, we propose to solve the problem of Fig. 2. Given a point **r** in the image $\mathbf{I}_1$, the objective is to determine a corresponding point in the image $\mathbf{I}_2$. In fact, only one possibility is correct, however it is common to note that other regions can be candidates for that correspondence if only an analysis of characteristics is used, or else corresponding point may not have appeared due to possible occlusions. Using a third image $\mathbf{I}_3$, the problem must consider that the correspondence for the first two images is solved.

In contrast with existing methods, our method is designed using two techniques known for their high performance: (1) correspondence through invariant descriptors in multiple views, and (2)

point-to-point correspondence through epipolar geometry based on the correspondences determined in the previous step. The first step is necessary for describing the geometric model of the second step through epipolar analysis. An advantage of epipolar geometry is that it can describe very precisely the geometric relation of the scene if 3D points are used to determine the transfer functions of a 3D point to the 2D plane (Hartley & Zisserman, 2000). However, due to the uncalibrated and dynamic nature normally present in multiple images, this solution is impracticable. Therefore, the normal procedure is to use points in correspondence (detected previously) to generate the geometric transfer model. Usually that estimation uses a robust correspondence selection process to minimize the re-projection error in the following views (Fischler & Bolles, 1981; Zhang, Deriche, Faugeras, & Luong, 1995; Torr & Zisserman, 2000). In general, most of the estimation methods end when the objective function finds the correspondence set that generates the smallest re-projection error. These methods determine the error of each random subset and choose the one that gets the smallest error of all the sets analyzed after a given number of iterations. During the process, multiple intermediate solutions with errors greater than the minimum are disregarded. This process is reasonable when there is a large number of incorrect correspondences and we get large errors; however, if a large percentage of the correspondences is correct, the errors can be similar to the minimum error and there is no justification to reject those solutions. For that reason, in this paper we propose to use the best solutions, i.e., not only the best solution, but also the following solutions (with errors very similar to the minimum) in order to increase the performance of the geometric correspondence model. The central idea of our work is to estimate a corresponding point on the second view from epipolar lines obtained from the best estimations of the epipolar geometry for these views, and not be limited to a single solution. This idea is extrapolated to the three-view geometry.

In the following sections we detail our methodology for estimating the correspondence matrices in multiple views in uncalibrated sequences. The rest of the document is organized in the following sections: section 2 includes a description of the proposed method; section 3 includes the experiments and results; and finally, section 4 presents the conclusions and future work.

2 PROPOSED METHODOLOGY

As far as the authors are aware, all the search methods for the fundamental matrix and trifocal tensors (Hartley & Zisserman, 2000) have the purpose of finding the best model generated from a random set of pairs in correspondence through an error minimization process (Fischler & Bolles, 1981; Zhang et al., 1995; Torr & Zisserman, 2000; Torr, 2002). This process can take place, for example, by means of a sampling consensus known as RANSAC (Fischler & Bolles, 1981), or the likelihood maximization in MLESAC by random sampling (Torr & Zisserman, 2000). Both methods, as well as the improvements proposed by Torrdoff and Murray (Tordoff & Murray, 2005), have been shown to be efficient methods for finding the fundamental matrices and perspectives in problems of computer vision. In the case of two views, the objective of the minimization process is to determine an epipolar single line in order to find an optimum epipole (O. D. Faugeras, 1993; Romano, 2002). However, these methods have been designed for problems in which there is a

considerable number of erroneous correspondences, so the random search for hypotheses has the objective of determining the quality of each selected hypothesis and in that way reduce the selection of erroneous correspondences (Torr, 2002). But what happens when there is a large number of correctly estimated correspondences? Is the best hypothesis the only solution that can be used? To answer these questions, below we present a new method for determining the point-to-point correspondence in two and three views in a geometric way.

2.1 Correspondence in two views

One of the most widely studied problems in computer vision is the geometric relation that exists between two corresponding images. A first step to solve this problem is to determine a set of point-to-point correspondences that estimate the geometric relations present in both images. In this section we will detail a new method that improves substantially the point-to-point correspondence contained in both images through a geometric formulation. In general, the problem of analysis in two views consists of how to determine the geometric relations of a 3D point and its projections on 2D planes. In what follows we will introduce the notation that relates the points in both images and the geometry that defines them. First, let $\mathbf{P}$ be a point in 3D space. In our example, point $\mathbf{P}$ is located in the upper corner of the 3D cube of Fig. 3. Second, let $\mathbf{C}_1$ and $\mathbf{C}_2$ be the optical centers of two cameras located at different viewpoints. For the following analysis, assume that we capture an image from the optical center $\mathbf{C}_1$, which generates image $\mathbf{I}_1$. Also, if we capture an image from the optical center $\mathbf{C}_2$, we generate image $\mathbf{I}_2$. According to this configuration, if we project a ray from center $\mathbf{C}_1$ to point $\mathbf{P}$, point $\mathbf{r}$ is generated on the 2D plane of image $\mathbf{I}_2$. Similarly, if we project a ray from center $\mathbf{C}_2$ to point $\mathbf{P}$, point $\mathbf{m}$ is generated, defined on the 2D plane of image $\mathbf{I}_2$. This relation implies that both rays intersect at a single point $\mathbf{P}$ defined in 3D space and its projections are on the $\mathbf{I}_1$ and $\mathbf{I}_2$ planes. In this way, points $\mathbf{r}$ and $\mathbf{m}$ correspond to a projection of point $\mathbf{P}$ generated from the optical centers $\mathbf{C}_1$ and $\mathbf{C}_2$. In the ideal case both points are corresponding, since they were generated from a single point, in this case point $\mathbf{P}$. On the contrary, if we do not know the existence of point $\mathbf{P}$ we cannot assure that the correspondence is true. The latter situation is what normally occurs in point-to-point correspondence problems. In what follows we will denote by $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$, when points $\mathbf{r}$ and $\mathbf{m}$ are corresponding, and as $\{\mathbf{r} \mapsto \mathbf{m}\}$ when the relation is hypothetical, i.e., we do not know if the relation is true or false and we want to find out.

A conventional way of proving the relation between points $\mathbf{r}$ and $\mathbf{m}$ is through the *fundamental matrix* $\mathbf{F}$ (Hartley & Zisserman, 2000; Romano, 2002). Formally, the fundamental matrix encapsulates the intrinsic geometry of two views, called *epipolar geometry*. For its determination it is necessary to know a minimum set of correspondences in both views. The main relation that establishes it is: given a pair of $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$ correspondences, they always satisfy the following epipolar restriction:

$$\mathbf{m}^{\top} \cdot \mathbf{F} \cdot \mathbf{r} = 0,$$

Unfortunately, this relation is valid for all the points that are found at the intersection of the

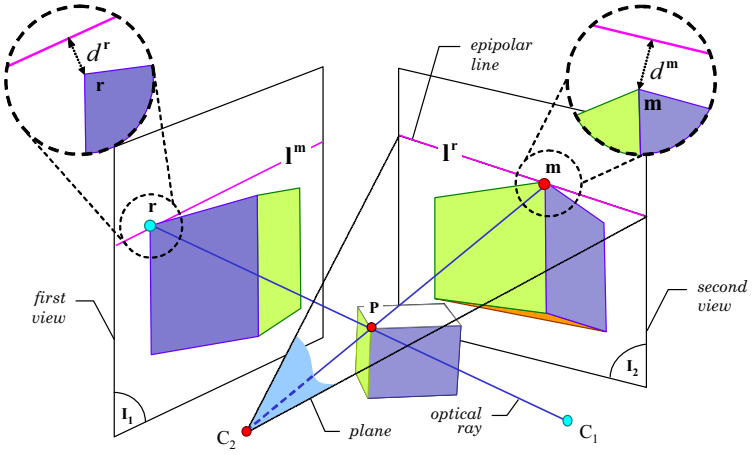


Fig. 3. General epipolar geometry of a 3D object and its projections.

projection plane of the optical center $\mathbf{C}_2$ and the $\mathbf{I}_2$ plane. The line generated by that intersection is known as *epipolar line* (Hartley & Zisserman, 2000). Since point $\mathbf{r}$ belongs to the plane of the optical center $\mathbf{C}_2$, we say that the epipolar line in the second view $\mathbf{I}_2$ is correspondent with point $\mathbf{r}$ in the first view $\mathbf{I}_1$. According to this analysis, it is not possible to determine a bi-univocal relation between points $\mathbf{r}$ and $\mathbf{m}$ using only an epipolar line, i.e., it is not possible to determine the position of point $\mathbf{m}$ from point $\mathbf{r}$; therefore, using the previous notation, we say that $\{\mathbf{r} \mapsto \mathbf{m}\}$ does fulfill at least the epipolar restriction.

Various methods for estimating the fundamental matrix have been developed in recent years, e.g. (Hartley & Zisserman, 2000; Chen, Wu, Shen, Liu, & Quan, 2000; Bartoli & Sturm, 2004). Regardless of the method for estimating the fundamental matrix, once it is determined it is possible to calculate the epipolar line described above, as shown in Fig. 3. Formally, let $\mathbf{l}^{\mathbf{r}}$ be the epipolar line of point $\mathbf{r}$ located in view $\mathbf{I}_2$, defined as $\mathbf{l}^{\mathbf{r}} = \mathbf{F} \cdot \mathbf{r}$. Because of the properties of the fundamental matrix, to carry out the transformation in the first view it is sufficient to transpose it. In this way, let $\mathbf{l}^{\mathbf{m}}$ be the epipolar line of point $\mathbf{m}$ in the first view, defined as $\mathbf{l}^{\mathbf{m}} = \mathbf{F}^{\top} \cdot \mathbf{m}$. This relation does not mean that the epipolar line will always be visible, but at least it will be when both points are visible in both views if the fundamental matrix is valid.

Let us assume that points $\mathbf{r}$ and $\mathbf{m}$ are corresponding. Therefore, they must be on epipolar lines $\mathbf{l}^{\mathbf{r}}$ and $\mathbf{l}^{\mathbf{m}}$ because they belong to the same plane. That is, $\mathbf{l}^{\mathbf{m}} \cdot \mathbf{r} = 0$ and $\mathbf{l}^{\mathbf{r}} \cdot \mathbf{m} = 0$. However, in practice the measurements of both views are not precise, and this implies that the epipolar lines do not necessarily intersect the points in correspondence. Mathematically, this means that $\mathbf{l}^{\mathbf{m}} \cdot \mathbf{r} \neq 0$ and $\mathbf{l}^{\mathbf{r}} \cdot \mathbf{m} \neq 0$ (Hartley & Zisserman, 2000). This error is reflected in the Euclidian distances $d^{\mathbf{r}}$ and $d^{\mathbf{m}}$ between the real point and the epipolar line, in which $d^{\mathbf{r}} > 0$ and $d^{\mathbf{m}} > 0$ (Fig. 3). Both distances should be minimal so that the projections are correct. To minimize that error it is necessary to determine a set of correspondences that minimize a re-projection error, i.e., minimize the distance between the real position of the point and the position of the projected epipolar line. Normally

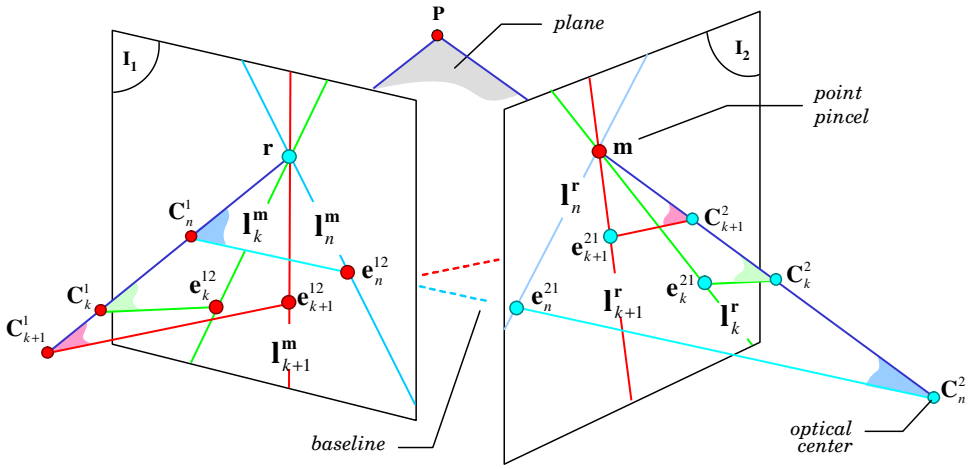


Fig. 4. Family of n multiple epipolar lines in two views from multiple epipoles.

the correspondence selection methods use some distance measure or probabilistic value to carry out that minimization. In some cases the error is generated by optical distortions belonging to the lenses or by Gaussian noise present in the acquisition of the coordinates in correspondence. As a result, slight errors in the determination of the correspondences increase the degradation of the geometric model.

An important point related to the estimation of the fundamental matrix is its dependence with respect to the set of correspondences used; i.e., for every set of correspondences a new fundamental matrix is determined. Even when the fundamental matrices are different, all of them remain valid provided $|\mathbf{F}| = 0$. However, every fundamental matrix has associated with it a level of error due to the inaccuracies of the set of correspondences used. In spite of this error, the use of multiple fundamental matrices has two important advantages. (1) Every new fundamental matrix defines a new epipole position in the $\mathbf{I}_1$ and $\mathbf{I}_2$ planes. (2) The intersection of the epipole and the hypothetical point in correspondence ($\mathbf{r}$ or $\mathbf{m}$) generates a new epipolar line. Taking into account the two previous properties, let us assume that we choose k sets in correspondence, where $k \in [1,\ldots,n]$ and n is the maximum number of sets in correspondence. According to Fig. 4 the $\mathbf{e}_k^{12}$ and $\mathbf{e}_k^{21}$ epipoles are defined as the points of intersection between the baseline of the optical centers $\mathbf{C}_k^1$ and $\mathbf{C}_k^2$, and the $\mathbf{I}_1$ and $\mathbf{I}_2$ planes, respectively. In this way, for every image there is an epipole, even though it is not necessarily visible in the plane. In this case, for the model proposed in Fig. 4 we assume that the position of point $\mathbf{P}$ is fixed.

To illustrate the process of estimation of correspondences in two views, we will assume that given a point $\mathbf{r}$ in the first view, there is a corresponding point in the second view. Since we do not know that correspondence, in our example we will assume that there are three hypothetical corresponding points, that we will call $\mathbf{m}$, $\mathbf{n}$, and $\mathbf{p}$. As shown in Fig. 5, for the first set of correspondences the epipolar line $\mathbf{l}_1^{\mathbf{r}}$ intersects points $\mathbf{m}$, $\mathbf{n}$, and $\mathbf{p}$ in the second view $\mathbf{I}_2$. Therefore, let Θ be the set of hypothetical correspondences, where $\Theta = \{\{\mathbf{r} \mapsto \mathbf{m}\}, \{\mathbf{r} \mapsto \mathbf{n}\}, \{\mathbf{r} \mapsto \mathbf{p}\}\}$. Our objective is to

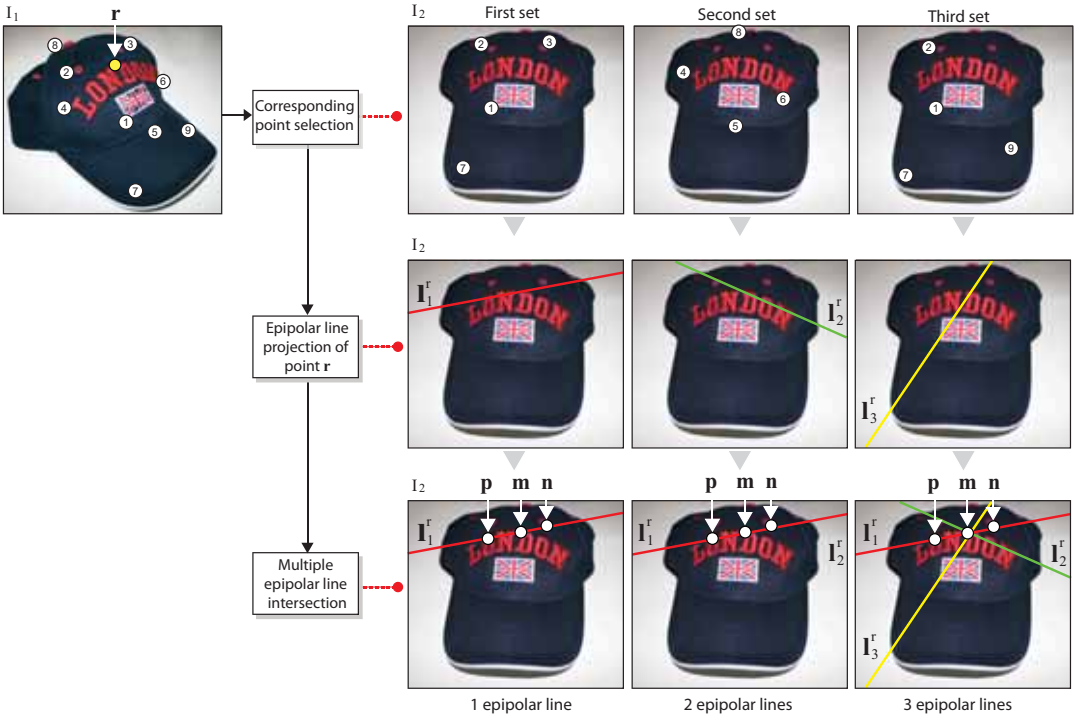


Fig. 5. New epipolar lines are created when a new set of corresponding points is created. In the example, line $\mathbf{l}_1^{\mathbf{r}}$ was obtained with set points {1,2,3,7}, line $\mathbf{l}_2^{\mathbf{r}}$ with set points {4,5,6,8}, and finally, line line $\mathbf{l}_3^{\mathbf{r}}$ with set points {1,2,7,9}.

determine a single correct pair of set Θ; i.e., select the $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$ pair and consequently discard the incorrect pairs $\{\mathbf{r} \mapsto \mathbf{n}\}$ and $\{\mathbf{r} \mapsto \mathbf{p}\}$. Based on the above discussion, if we intersect two epipolar lines $\mathbf{l}_1^{\mathbf{r}}$ and $\mathbf{l}_2^{\mathbf{r}}$, –both generated by two different subsets of correspondences– it is clearly seen that line $\mathbf{l}_2^{\mathbf{r}}$ is at a considerable distance from the correspondences $\mathbf{n}$ and $\mathbf{p}$. Similarly, a third epipolar line $\mathbf{l}_3^{\mathbf{r}}$ intersects the two previous ones at point $\mathbf{m}$ because the set of projected epipolar lines of point $\mathbf{r}$ intersect only one corresponding point in the second view, which in this case is point $\mathbf{m}$, generating an *point pencil*. That effect is repeated in both images, as shown by the model of Fig. 4 and Fig. 5.

Theoretically, every new epipolar line improves the precision of corresponding point. However, in practice there is no single intersection point because of the uncalibrated nature of corresponding points used to formulate the geometric model, giving rise to an error in the estimation of the fundamental matrix. According to this analysis, one of the main problems in the estimation of the epipolar lines consists of determining their error level. Clearly, not all the epipolar lines have the same error, and for that reason we designed a method to determine the error associated with the Euclidian distance of each epipolar line. For the following analysis we will introduce the distance notation between the hypothetical point with respect to the epipolar line in the second view. Let

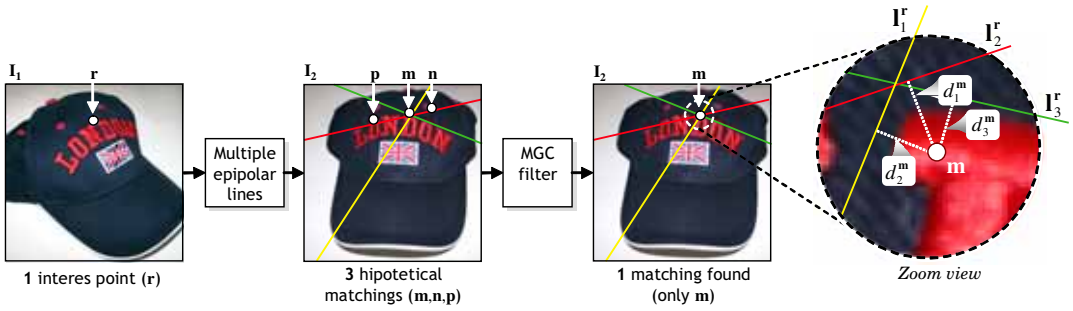


Fig. 6. The Multiple Geometric Correspondence (MGC) filter allows the determination of the point-to-point correspondence and distinguishing incorrect correspondences {p, n}.

$d_k^{\mathbf{m}}$ be the Euclidian distance between the $\mathbf{m}$-th point of the second view and the epipolar line $\mathbf{l}_k^{\mathbf{r}}$, where $\mathbf{r}$ is the r-th point of the first view. The distance $d_k^{\mathbf{m}}$ is defined as

$$d_k^{\mathbf{m}} = \frac{|\mathbf{m}^\top \mathbf{F}_k \mathbf{r}|}{\sqrt{(\mathbf{F}_k \mathbf{r})_1^2 + (\mathbf{F}_k \mathbf{r})_2^2}} \tag{1}$$

where $(\mathbf{F}_k \mathbf{r})_i$ is the i-th component of vector $\mathbf{F}_k \mathbf{r}$. As mentioned earlier, our objective is to find the correspondence $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$ of the set Θ. However, we do not know the estimation error of each epipolar line. To make that estimation we will use the MLESAC algorithm, proposed by Torr and Zisserman (Torr & Zisserman, 2000). The objective of this process is to re-estimate the Euclidian distances $d_1^{\mathbf{m}}$, $d_2^{\mathbf{m}}$ and $d_3^{\mathbf{m}}$ weighting the error of each epipolar line, a process that will be described below.

First we will introduce some previous concepts of the MLESAC algorithm (Torr & Zisserman, 2000) to give greater clarity to the reader. MLESAC is a robust estimation algorithm to establish the point correspondences in multiple views, generalizing the RANSAC estimator (Fischler & Bolles, 1981). According to Torr and Zisserman, MLESAC performs better than RANSAC mostly because it is based on minimizing the likelihood error instead of maximizing the number of correspondences. In our proposal MLESAC is an intermediate step in the error estimation process because the error estimated by MLESAC later allows weighting the individual error of each epipolar line. One of the main advantages of MLESAC is that it is designed considering that the error $P(e)$ is a mixture of Gaussians and uniform distributions, where e is the error of the estimation of the fundamental matrix such that

$$P(e) = \left(\gamma \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e^2}{2\sigma^2}\right) + (1-\gamma)\frac{1}{v}\right) \tag{2}$$

where γ is a mixing parameter, v is an a priori constant that indicates the distribution of the data, and σ is the standard deviation of the error in each coordinate. Parameters γ and v are not known, but they can be estimated by means of the EM (Dempster, Laird, & Rubin, 1977) algorithm. In this way, the objective function is to minimize the log-likelihood of the error, which in our case is the

distance $d_k^{\mathbf{m}}$ between a point and the epipolar line, and therefore

$$-L_k = -\sum_k \left(\gamma \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{(d_k^{\mathbf{m}})^2}{2\sigma^2} \right) + (1-\gamma)\frac{1}{\nu} \right) \tag{3}$$

One of the main advantages of estimating the error by means of the MLESAC estimator is that the inliers, or correct correspondences, have a high weight, in contrast with the RANSAC algorithm, in which only the outliers are considered in the cost function. We had previously mentioned that γ and ν are not known. For completeness, we now indicate how they are estimated. Assuming that there are k sets of correspondences, let η_k, where $\eta_k = 1$ if the correspondence is correct, i.e., $d_k^{\mathbf{m}} = 0$, and $\eta_k = 0$ if the k correspondence is incorrect. The EM algorithm considers that η_k is an unknown value, and therefore it takes the following steps for its estimation: (1) it generates an initial value for γ, (2) it estimates the η_k value using the initial γ estimation, and (3) it makes an estimation of γ from the new estimated value η_k, and returns to step (2). The process is repeated until it converges.

As mentioned before, the MLESAC algorithm uses EM for estimating the γ and ν parameters and the probability that a putative selection will be an inlier or an outlier. For this, let p_k be the likelihood of distance $d_k^{\mathbf{m}}$ when it is an inlier, and p_o the likelihood of distance $d_k^{\mathbf{m}}$ when it is an outlier. Consequently, given the initial value of $\gamma = \frac{1}{2}$, the probabilities p_k and p_o are estimated according to

$$p_k = \gamma \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^2 \exp\left(-\frac{d_k^{\mathbf{m}2}}{2\sigma^2} \right) \tag{4}$$

$$p_o = (1-\gamma)\frac{1}{\nu} \tag{5}$$

Once the probabilities p_k and p_o have been estimated from the initial value γ, the following step is to re-estimate the $P(\eta_k = 1|\gamma)$ value according to

$$P(\eta_k = 1|\gamma) = \frac{p_k}{p_k + p_o}, \tag{6}$$

and finally, in the phase called '*maximization*' of step (3), the value γ is re-estimated according to the updated mixture of the probabilities p_k and p_o,

$$\gamma = \frac{1}{n}\sum_k \left(\frac{p_k}{p_k + p_o} \right) \tag{7}$$

Normally three iterations are needed for the algorithm to converge. Recall that MLESAC uses the random selection of random solutions; in our case the random solutions generate the set of epipolar lines. In this way the estimation of the log-likelihood of the k-th hypothesis of each epipolar line allows us to weight correctly the real distance $d_k^{\mathbf{m}}$. To make that estimation we will use the partial values of the log-likelihood (L_k) and in that way we weight the distance $d_k^{\mathbf{m}}$ according to the following formulation:

Algorithm 1 : *Bifocal Geometric Correspondence* (BIGC) algorithm in two views.

1: Determine n sets in correspondence in two views. These sets are known or estimated in a process that can be off-line or automatic by means of the analysis of correspondences; for example, with SIFT (Lowe, 2004) or SURF (Bay et al., 2008).
2: Determine the fundamental matrix $\mathbf{F}_k$, for k sets in correspondence, where $k < n$.
3: Determine the epipolar line $\mathbf{l}_k^{\mathbf{r}}$ of point $\mathbf{r}$ in the first view.
4: Determine the error associated with each epipolar line $\mathbf{l}_k^{\mathbf{r}}$ with the MLESAC algorithm and re-estimate the real distance $\tilde{d}_k^{\mathbf{m}}$ between the hypothetical correspondence and the epipolar line.
5: Assign the correspondence to point $\mathbf{m}$ provided that the restriction $\tilde{d}_k^{\mathbf{m}} < \varepsilon$ is fulfilled for all $\mathbf{m} \in \Theta$.

$$\tilde{d}_k^{\mathbf{m}} = d_k^{\mathbf{m}} \left(\frac{|min(L_k) - L_k| + 1}{\sum_k (L_k)} \right). \tag{8}$$

where $\tilde{d}_k^{\mathbf{m}}$ is a weighted distance that considers the error associated with each fundamental matrix. This procedure allows weighting and reestimating the distance of the epipolar lines according to the log-likelihood of the projection error with respect to the set of hypothetical points in the second view. Let us recall that each epipolar line is generated from the epipole estimation, and that is why the above procedure determines indirectly the error associated with each epipole. However, why is it relevant to make an estimation of the distance of the epipolar line with respect to a hypothetical correspondence in the second view? Remember that the random selection of correspondences is subject to an error. The estimation of the error allows weighting correctly the distance $d_k^{\mathbf{m}}$, increasing or decreasing it according to the size of its error. Therefore, to determine a correspondence, we determine the distance with respect to the set Θ. Finally, to identify the correspondence of point $\mathbf{r}$ the following relation must be satisfied:

$$\tilde{d}_k^{\mathbf{m}} < \varepsilon\,, \tag{9}$$

where ε is a distance measured in pixels. The final result allows the determination of which points are corresponding and which, depending on a threshold level, must be discarded. Fig. 6 presents an example of how the error estimation discards points $\mathbf{n}$ and $\mathbf{p}$ from the set of correspondences Θ. In particular, the Multiple Geometric Correspondence MGC filter' block is in charge of reestimating the distances. In the example, once point $\mathbf{m}$ is chosen, only the $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$ combination is possible. This correspondence pair is then used as starting point for the analysis with a third view in correspondence.

As shown in the previous steps, in spite of the errors existing in the estimation of the epipolar lines, their set allows the estimation of point-to-point correspondence. A complete description of the proposed methodology, which we call *Bifocal Geometric Correspondence* (BIGC), is presented in the 1 Algorithm.

2.2 Tracking in three views

In the previous section we described an algorithm for estimating the error of the $d_k^{\mathbf{m}}$ distances with respect to a hypothetical point in correspondence in the second view. The main idea was to model the error through the robust estimation of the MLESAC algorithm considering a multiple epipolar lines analysis. In this section we will make a similar analysis but considering three views. In relation to this point it is important to mention that the use of more views does not necessarily increase the performance, because it depends on the type of application in which the matching is inserted. Our hypothesis is that a third view can reduce the remaining false alarms because they have a lower probability of remaining in their relative position in three views.

The same as in the previous procedure, we propose to use re-projection from the estimation of multiple projections of hypothetical correspondences in two views. In this case the error must be estimated to weight correctly the distances of the re-projected point in the third view with respect to the positions of the hypothetical correspondences. Again, a selection of random sets of correspondences is used to determine the solution of each geometric model. Normally the RANSAC based algorithms would discard the intermediate solutions and would use the set with the smallest re-projection error. However, in our problem the estimation of the error generated by the set of random correspondences is relatively low. Therefore, it is feasible to use more than one equally valid set of correspondences. This process is similar to the estimation of multiple fundamental matrices.

Now we will discuss briefly the estimation of the geometric model with three views. The same as in two views, the analysis in three views allows modeling all the geometric relations generated in the 3D space, regardless of the structure contained in each image (Hartley & Zisserman, 2000; O. Faugeras, Luong, & Papadopoulo, 2001). An example of this is presented in Fig. 7, which shows the relation generated from the projection of a 3D point with respect to three bi-dimensional projection planes. Formally, let $\mathbf{P}$ be a homogeneous point defined in the 3D space. The projection of $\mathbf{P}$ on the $\mathbf{I}_1$, $\mathbf{I}_2$ and $\mathbf{I}_3$ planes generates the $\mathbf{r}$, $\mathbf{m}$ and $\mathbf{s}$ projections in each image, respectively. Even if the projection is not visible, either because the point is occluded or it is outside the camera's field of view, the geometric model is still valid for that point. To make that estimation it is necessary to determine the matrix called *Trifocal tensor* (Hartley & Zisserman, 2000). One of the great advantages of geometric modeling in three views, and particularly of the estimation of the trifocal tensors, is that it depends only on the motion between the views and on the internal parameters of the cameras, and it can be completely defined by the projection matrices of which it is composed. It is important to note that the trifocal tensors can be calculated through the correspondences of the images without a priori knowledge of the object. Therefore, our analysis is based on how to estimate the error of the projection matrices from the set of correspondences in three views.

Formally, the trifocal tensor [1] $\mathbf{T} = (T_t^{rs})$ is a $3 \times 3 \times 3$ matrix that codes the relative motion between the $\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3$ views. As already mentioned, one of its most relevant properties is that from the estimation of the tensor we can determine the position of a point $\mathbf{s}$ in the $\mathbf{I}_3$ plane using the positions of the correspondences $\{\mathbf{r} \mapsto \mathbf{m}\}$ of the first and second views, respectively, as shown in

1. See Hartley and Zisserman (Hartley & Zisserman, 2000) for details on the computation of the trifocal tensors.

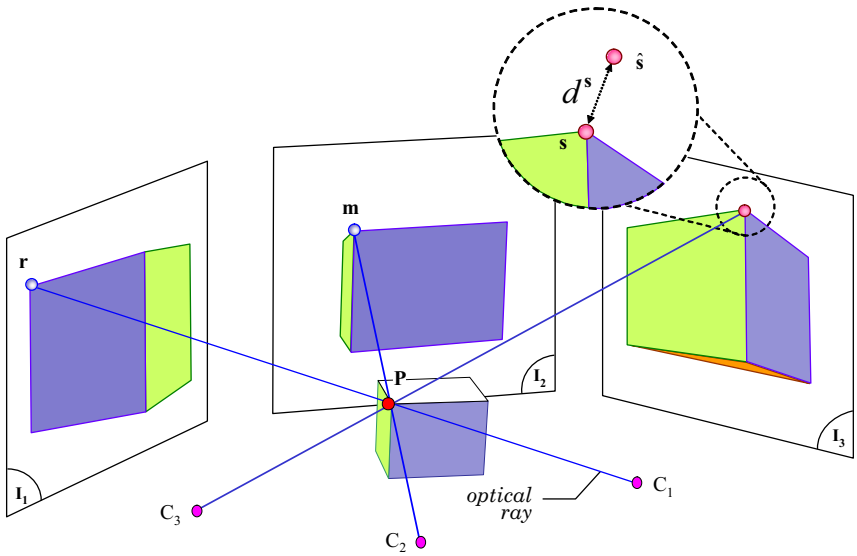


Fig. 7. Epipolar geometry of a 3D object using three views.

Fig. 7. The determination of the re-projection is defined in terms of the $\mathbf{r} = [x_1, y_1, 1]^\top$, $\mathbf{m} = [x_2, y_2, 1]^\top$ positions in homogeneous coordinates and of the tensor $\mathbf{T}$, derived from the first two trilinearities of Shashua (Shashua, 1995). In particular, we use re-projection by means of the point-line-point method proposed by (Hartley & Zisserman, 2000, pp.373). For that purpose, let $\hat{\mathbf{s}}$ be the projection of the trifocal tensor in the third view, defined as $\hat{\mathbf{s}} = [x_3, y_3, 1]^\top$.

Unfortunately, the estimation of the projected point $\hat{\mathbf{s}}$ is subject to an error that can be generated for two reasons: (1) The intrinsic error in the estimation of the tensors due to an incorrect choice of the set of correspondences, and (2) the correspondence error between the $\mathbf{r}$ and $\mathbf{m}$ pairs. The latter case is not visualized in Fig. 7, however most of the correspondences include that error. Even in the ideal case, when the tensors are relatively stable in uncalibrated sequences, there is always an error between the hypothetical correspondence $\mathbf{s}$ and the re-projected point $\hat{\mathbf{s}}$. For simplicity, we assume that the distance between these points is the Euclidian distance $d^{\mathbf{s}}$ of point $\mathbf{s}$, defined as

$$d^{\mathbf{s}} = \|\hat{\mathbf{s}} - \mathbf{s}\|. \tag{10}$$

Normally the process of estimating the trifocal tensor when some error metric is reduced, such as minimizing the distance $d^{\mathbf{s}}$ from multiple random solutions, or modeling the error as a probability distribution. In any case, for every estimation of the tensor there is a unique possible projection associated with each pair of correspondences $\{\mathbf{r} \mapsto \mathbf{m}\}$. Similarly, assuming that other random solutions are valid, our objective will consist in estimating the error of each re-projection. The error associated with each selection of the trifocal tensor will be used later to re-estimate the re-projection distances of point $\mathbf{s}$ in the third view in a similar way as the estimated error of the distance to the epipolar line.

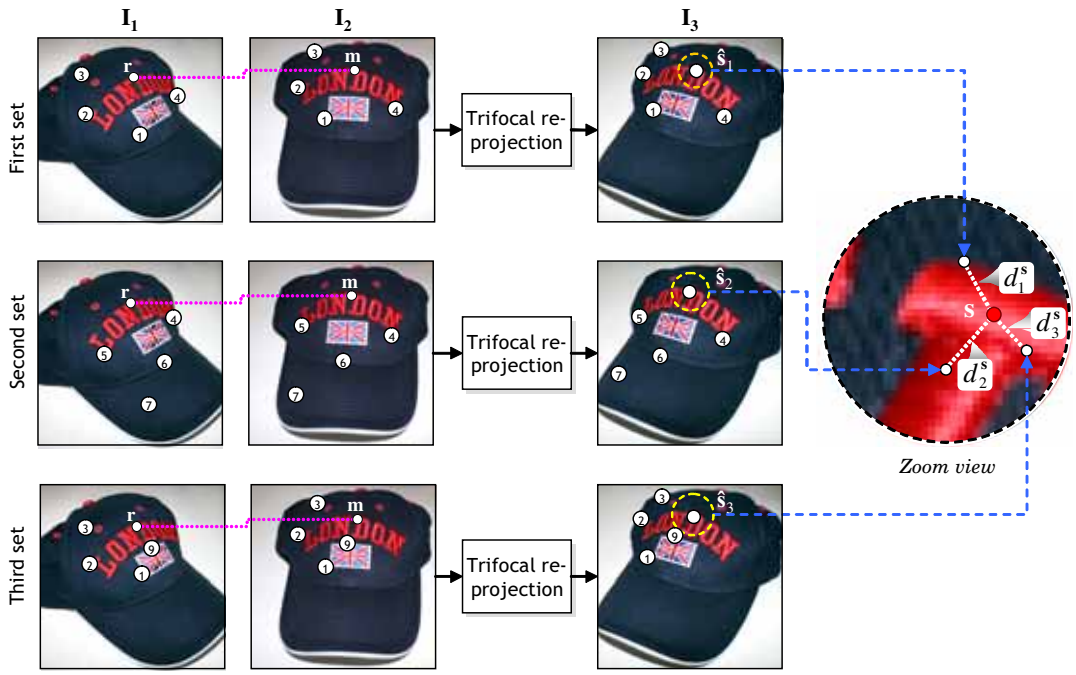


Fig. 8. Re-projection process of hypothetical points using a third view through the trifocal tensors. In the example, the projection of point $\hat{\mathbf{s}}_1$ was determined with the correspondences $\{1,2,3,4\}$, that of point $\hat{\mathbf{s}}_2$ with correspondences $\{4,5,6,7\}$, and that of point $\hat{\mathbf{s}}_3$ with correspondences $\{1,2,3,9s\}$.

As an example, in Fig. 8 we point out three sets of independent correspondences. Each subset generates a re-projection in the $\hat{\mathbf{s}}_1$, $\hat{\mathbf{s}}_2$, $\hat{\mathbf{s}}_3$ positions, corresponding to the re-projection of the tensor in the third view. Extending the above example, let k be the number of correspondences used. In this way, from each k subset it is possible to estimate the trifocal tensor $\mathbf{T}_k$. Each tensor is unique and independent of the previous one, provided the selection of the subsets is different. Assuming independence between n sets, let $\hat{\mathbf{s}}_k$ be the re-projection of the tensor $\mathbf{T}_k$ generated from the re-projection of the pair of points in correspondence $\mathbf{r}$ and $\mathbf{m}$.

As mentioned before, if we consider that the error associated with the estimation of each tensor is different, the re-projection distance should consider that error. The objective is to weight the $d_k^{\mathbf{s}}$ distances that have a smaller error and discard those tensors that have a large level of error. For this we use again the MLESAC algorithm with the purpose of estimating the error of each random solution. Even though the final objective is to estimate multiple re-projection points, it is first necessary to estimate the error associated with each trifocal tensor. For this we consider the re-projection distance assuming a Gaussian noise together with a noise with uniform distribution, where the error integrates the $d_k^{\mathbf{s}}$ distance for each k subset of correspondences used.

$$-L_k = -\sum_k \left(\gamma \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{(d_k^{\mathbf{q}})^2}{2\sigma^2} \right) + (1-\gamma)\frac{1}{v} \right) \tag{11}$$

Algorithm 2 : *Trifocal Geometric Correspondence* (TRIGC) algorithm in three views.

1: Determine n sets in correspondence in three views. These sets are known or estimated in a process that can be off-line, or automatic by means of the analysis of correspondences; for example, with SIFT (Lowe, 2004) or SURF (Bay et al., 2008).
2: Use the BIGC algorithm to determine pairs of point-to-point correspondence in the first and second views.
3: Determine k trifocal tensors $\mathbf{T}_k$, where $k < n$.
4: Determine the re-projection of the trifocal tensor $\mathbf{T}_k$ for each pair corresponding to step 2.
5: Determine the error associated with each trifocal tensor with the MLESAC algorithm and re-estimate the distance $\tilde{d}_k^{\mathbf{s}}$ between the hypothetical correspondence and the projected position.
6: Assign the correspondence with point $\mathbf{s}$ provided that the $\tilde{d}_k^{\mathbf{m}} < \varepsilon$ restriction is fulfilled for every pair $\{\mathbf{r} \mapsto \mathbf{m}\}$.

Estimation of the σ and v parameters is made again by the procedure described previously. Finally, the same as in the bifocal analysis, we weight the error associated with each trifocal tensor according to the log-likelihood estimated for each subset of correspondences. To weight the error we use the log-likelihood L_k of the k-th estimation of the tensor as:

$$\tilde{d}_k^{\mathbf{s}} = d_k^{\mathbf{s}} \left(\frac{|min(L_k) - L_k| + 1}{\sum_k (L_k)} \right) \tag{12}$$

where $\tilde{d}_k^{\mathbf{s}}$ corresponds to the weighted distance of the projection of the trifocal tensor and the hypothetical correspondence. Finally, to identify the correspondence in the third view it must satisfy the following relation:

$$\tilde{d}_k^{\mathbf{s}} < \varepsilon\,, \tag{13}$$

where ε is a distance in pixels. Remember that the weighting of the distance is related to the error associated with each tensor, so in this way the tensors with the largest error will have less influence on the re-projection distance. A complete description of the methodology in three views, called *Trifocal Geometric Correspondence* (TRIGC) is detailed in the 2 Algorithm.

3 Experimental Results

This section presents the experimental results generated with sequences of uncalibrated images in two and three views. We divided our experiments into three parts considering different types of images and different numbers of correspondence sequences. They correspond to: (1) Outdoor images: A set of 10 stereo images composed mainly of landscapes and walls, most of them supplied by the authors (Fig. 9), (2) Indoor Images: A set of 9 stereo images with test objects under ideal conditions generated in Kushal and Ponce (2006). (Fig. 11), (3) Industrial bottle images: A set of 120 images of bottle necks with manufacturing faults generated in Carrasco, Pizarro, and Mery (2008) (Fig. 13). In all the experiments we have considered two standard indicators (Olson, 2008): $r = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$ (recall) and $p = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$ (precision). TP is the number of *true positives* or correctly classified correspondences. FN is the number of *false negatives* or real correspondences not detected by our

algorithm. FP is the number of *false positives* or correspondences classified incorrectly. These two indicators can be joined in a single measure F-score $= \frac{2 \cdot p \cdot r}{p+r}$ (Olson, 2008). Ideally, one can expect that $r = 100\%, p = 100\%,$ and F-score $= 1$.

According to the statement made in the previous section, we evaluated the influence of parameter k when the number of solutions of the proposed method is varied. In the same way we evaluated the influence of the Euclidian distance ε. We recall that both parameters can be modified in combination. For that we separated the analysis varying each of them independently. Recall that the variations of parameter k increase the number of fundamental matrices and the number of trifocal tensors for two and three views, respectively. Also, parameter ε determines the Euclidian distance between the fundamental matrix and the position of corresponding hypothetical point (the case of two views Fig. 6), and the re-projection of the trifocal tensor and the hypothetical correspondence (in the case of three views Fig. 8). As we already established, the determination of a new solution of the geometric problem in two and three views implies the re-projection of a new geometric solution, restricting the search space for a correspondence.

Since our evaluation considers performance in two and three views, we have obtained different results in each case. This is due to two reasons: First, because in two views we determined the distance between each epipolar line with respect to the hypothetical point. Second, because in three views the distance ε is determined with respect to the point re-projected by the tensor. In the latter case we must consider that the reprojection requires correspondence in the first two views to determine the reprojection in the third view, making it necessary to have three corresponding images. In all the experiments we considered the average performance of the set of images. In the following subsections we will detail these aspects.

3.1 Outdoor images

The first test set is composed of 10 pairs of images with a resolution of 1200 $\times$ 800 pixels. The main existing geometric transformations are perspective, rotation, translation, and different degree scale (Fig. 9). This set consists of landscapes and walls in settings with natural lighting, showing a large number of regions in correspondence. According to the steps described in the 1 Algorithm, the first step consists of determining n sets of corresponding pairs. This process was performed with the SURF (Bay et al., 2008) algorithm, from which we selected the best k sets with the least projection error according to the MLESAC estimator. To evaluate the performance of the algorithm we determined 300 corresponding points in random positions within each pair of images in a process carried out off-line by means of the SURF algorithm. We then evaluated the ability of the algorithm to determine the correspondence by varying the $k \in [1, \ldots, 14]$ parameter and the $\varepsilon \in [0, \ldots, 10]$ parameter. Note that in the latter parameter the values are rounded.

Below we present the results according to the variations of the k and ε parameters. In the first case we analyze the influence of the k parameter keeping distance ε fixed. As seen in Fig. 10a, our best performance had an F-score$= 0.97$ at a discretized distance $\varepsilon = 0$, using the intersection of three fundamental matrices ($k = 3$). It is interesting to mention that as the ε parameter increases,

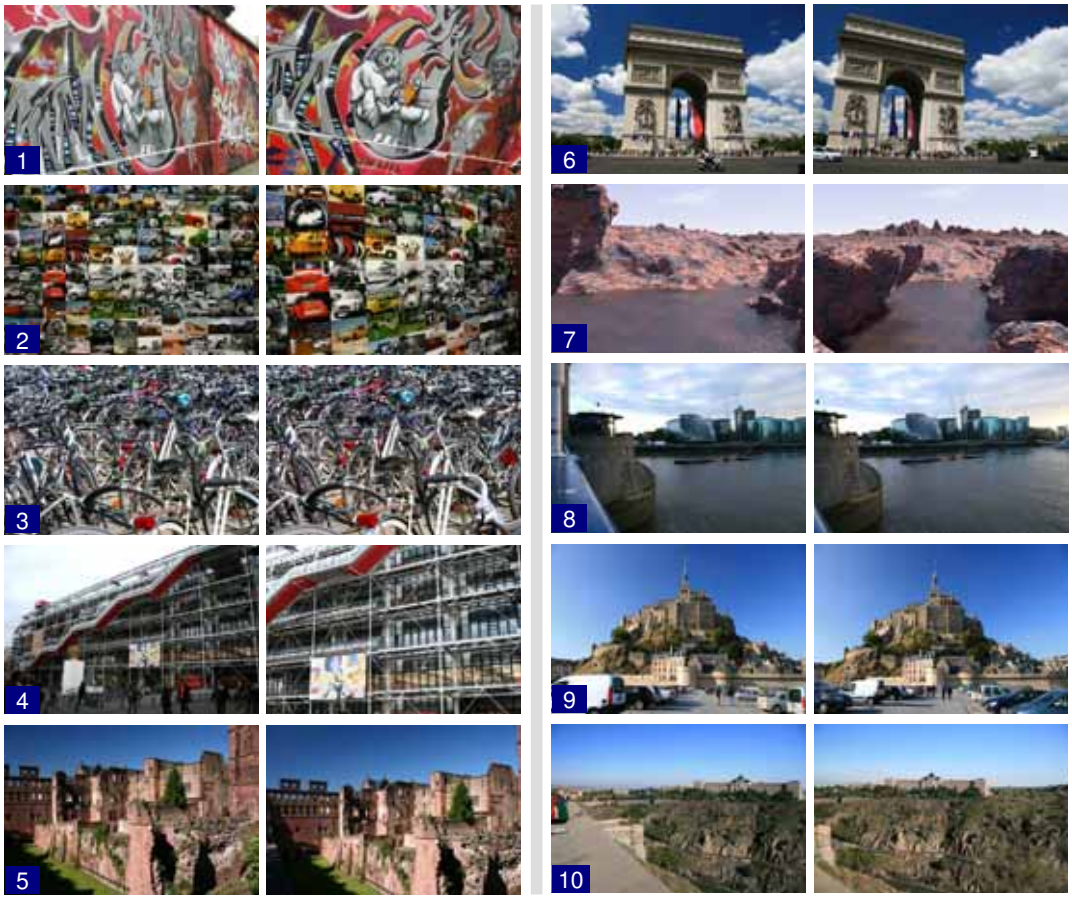
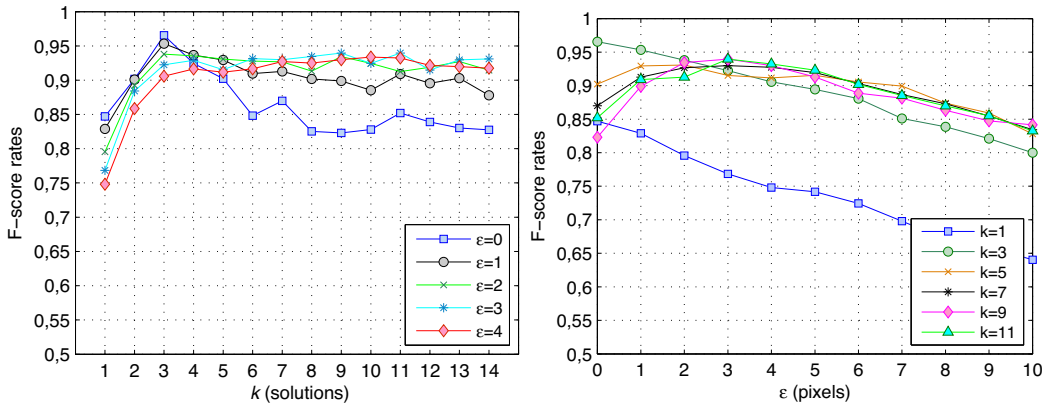


Fig. 9. Outdoor set of 10 stereo images

performance starts dropping. This indicates that the method is very precise in these kinds of images because there is a large number of correspondences. In the second case we analyze the influence of parameter ε keeping fixed the number of solutions k. According to the results obtained, we see a maximum performance at $k = 3$. On the contrary, an increase of this value decreases the performance of the algorithm because the projection error increases.

Remember that in the analyzed sequence there is a large number of correspondences in spite of the geometric transformations present in them. Therefore, these results indicate that it is possible to use and estimate geometric models in two images with high precision at a sub-pixel resolution. According to the performance indicated in Fig. 10a, after $k = 4$ there is no improvement in the performance for $\varepsilon > 4$.



(a) (b)

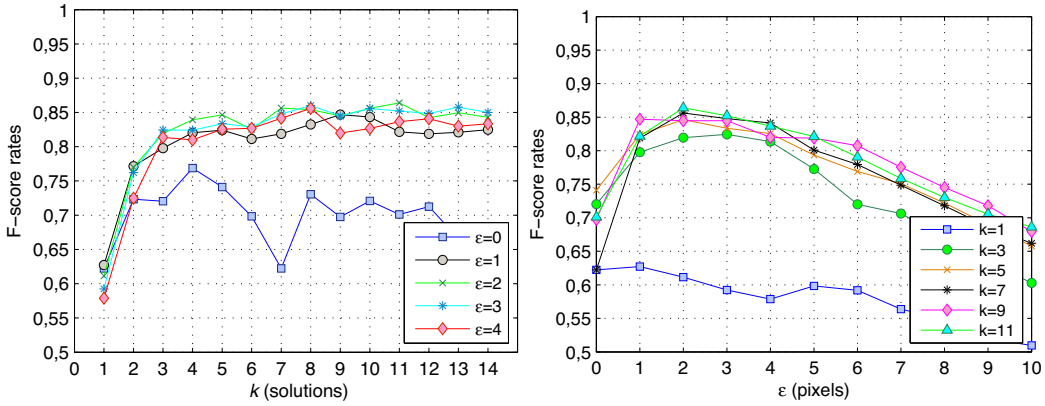
Fig. 10. Average performance of the set of outdoor images (a) Influence of the k parameter as the maximum tolerance distance in pixels ε varies (b) Influence of distance (ε) on the detection of correspondences for different numbers of solutions (k).

Fig. 11. Indoor set of 9 objects generated in Kushal & Ponce (2006).

3.2 Indoor images

The second test set is composed of 9 stereo images with a resolution of 600 $\times$ 900 pixels generated in Kushal and Ponce (2006) (Fig. 11). The transformations of this set correspond to changes in the points of views due to rotation on the vertical axis of each object. As a result, the determination of base correspondences is reduced to a smaller number, and further, they are less disperse in each pair of images. The same as in the previous case, we determined 300 corresponding points in random positions for each pair of images in a process carried out off-line by means of the SURF algorithm.

The same as in the previous case, we determined the results according to the variations of the k



(a) (b)

Fig. 12. Average performance of the set of indoor images (a) Influence of parameter k when the maximum tolerance distance in pixels ε is varied (b) Influence of distance (ε) in the detection of correspondences for different numbers of solutions (k).

and ε parameters. In the first case we analyze the influence of parameter k keeping distance ε fixed. As shown in Fig. 12a, the method has a maximum at $k = 11$; however, from $k = 4$ the performance tends to become stabilized considering a value $\varepsilon > 0$. In the second graph (Fig. 12b) it is clearly noted that the use of a fundamental matrix ($k = 1$) causes low performance, about 60%; on the contrary, the use of multiple solutions improves performance from 20% to 25%.

3.3 Industrial bottle images

The third test set contains 120 sequences of images of bottle necks with faults or regions with defects generated in Carrasco et al. (2008) (Fig. 13). Each sequence is composed of three images with an angle of rotation $\alpha = 15$. From the captured images we have extracted sub-images of 1000 $\times$ 250 pixels. The base correspondence was determined by means of markers outside the object that comply with the object's motion. In this case the objective of the point-to-point correspondence was to determine the trajectory of multiple faults in the sequence that must be detected to determine the quality of the bottle in a multiple view inspection process.

Evaluation in relation to the number of partial solutions: We now separate the results for two and three views, both detailed in Fig. 14a and Fig. 14b, respectively. (1) Two views: In this case the results indicate that the performance F-score is directly related to the increase of parameter k, becoming stabilized at $k = 9$. With respect to the influence of parameter ε on the performance F-score, we see an improve in performance as the distance between corresponding point and the epipolar line increases. It should be noted that performance becomes stabilized after $\varepsilon = 3$. These results imply that in two views, when the $k = 9$ combination is used, better performance is obtained when the distance between corresponding point is $\varepsilon = 4$. (2) Three views: In contrast with the two views, we got the best performance when a discretized distance $\varepsilon = 0$ was used. Similarly to the

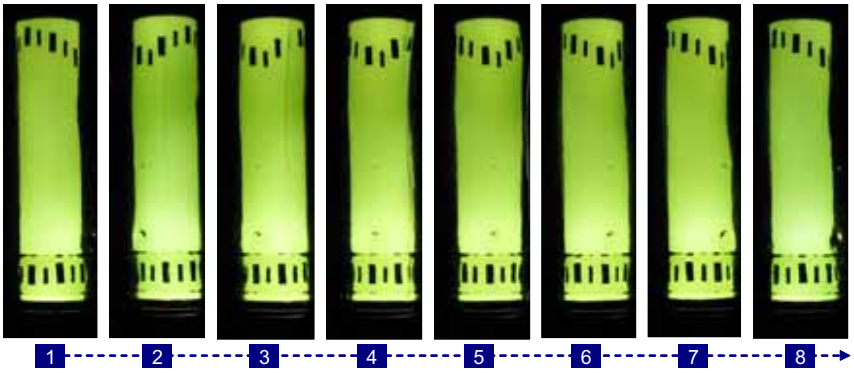


Fig. 13. Sequence of images of bottle necks for the tracking process as a quality control method.

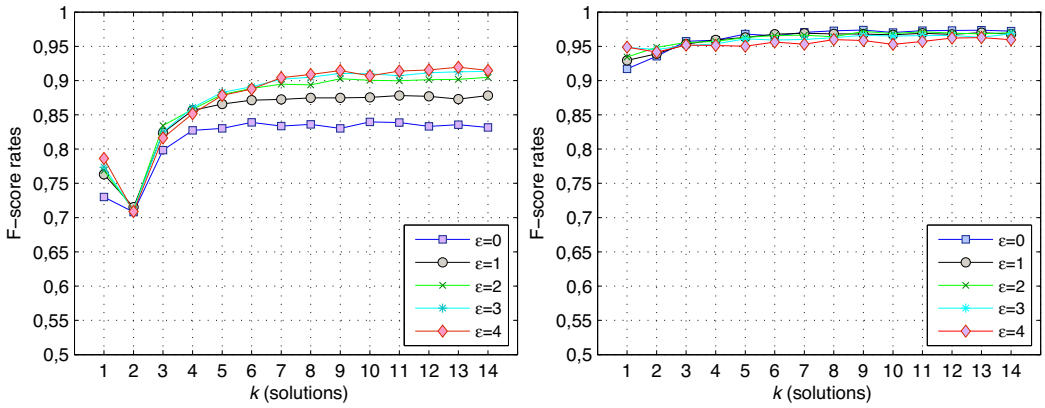


Fig. 14. Influence of parameter k as the maximum tolerance distance in pixels ε is varied. (a) Variation using two views, (b) Variation using three views.

two views, the $k = 9$ combination is repeated. These results indicate that at $\varepsilon = 0$ we get a trifocal correspondence with a performance F-score$= 0.97$. It is interesting to note that as parameter ε is increased, the performance of the method starts dropping. This effect is the opposite of that with two views.

As already mentioned, parameters k and ε modify the performance of the method differently for two and three views. Fig. 15 presents the maximum performance of the system using the optimum ε value for each variation of parameter k. It is seen that from five combinations ($k = 5$) using three views, the best distance remains at zero pixels. The results of this graph agree with those presented previously, because with $k = 9$ we get the best performance for two and three views.

Evaluation according to reprojection distance: In the previous evaluation we varied parameter k to determine its influence on the performance of the system of correspondences. In this case we evaluated the influence of the distance ε keeping parameter k fixed. Because of the large number

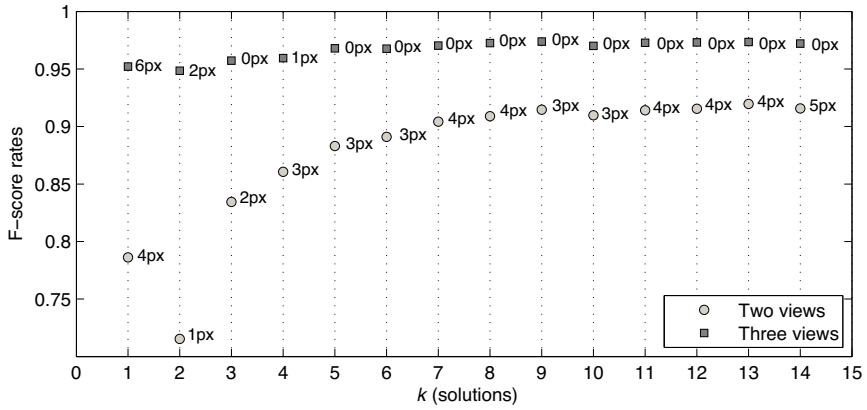


Fig. 15. Best performance of correspondence varying according to the number of solutions. The best ε value has been chosen in each performance curve.

of curves generated, we graphed only odd numbers of parameter k (Fig. 16). Below we detail the results for two and three views. (1) Two views: The results agree with the previous description on the number of intermediate solutions. When we set the parameter at $k = 1$, i.e., we use only one epipolar line, we get a maximum F-score equivalent to an F-score$= 0.78$ at a distance of 4 pixels ($\varepsilon = 4$). Taking $\varepsilon = 4$ and using a $k = 5$ combination, performance improves to an F-score$= 0.87$. Finally, with nine combinations ($k = 9$) performance is maximum with an F-score$= 0.91$. (2) Three views: Clearly there is an important difference when using a trifocal tensor versus multiple trifocal tensors. For example, when using a single trifocal tensor, the maximum performance is obtained at a distance of 6 pixels ($\varepsilon = 6$), with an F-score$= 0.95$. In contrast, when using nine combinations we get a performance F-score$= 0.97$ at a discretized distance $\varepsilon = 0$. The latter result shows the effectiveness of using the multiple intermediate solutions combination by the proposed method. The difference in performance between two and three views is accounted for by the greater number of false alarms in two views compared to three views, i.e., there is a greater number of false alarms with respect to the analysis with three views.

4 CONCLUSIONS

In this paper we have developed two important contributions. First, we presented a method that uses the intersection of multiple geometric solutions in two views and in three views to determine point-to-point correspondence. Second, for each geometric model we have determined the real distance with respect to corresponding point by means of the MLESAC estimator, in that way weighting the error associated with each intermediate solution. The main novelty of our proposal is the geometric methodology for solving the problem of the estimation of point-to-point correspondence, regardless of the angles of the points of view of the objects contained in the images and of the geometric transformations present in them. We call these algorithms Bifocal Geometric Correspondence (BIGC) for the correspondence in two views, and Trifocal Geometric Correspondence (TRIGC) in the case of three views.

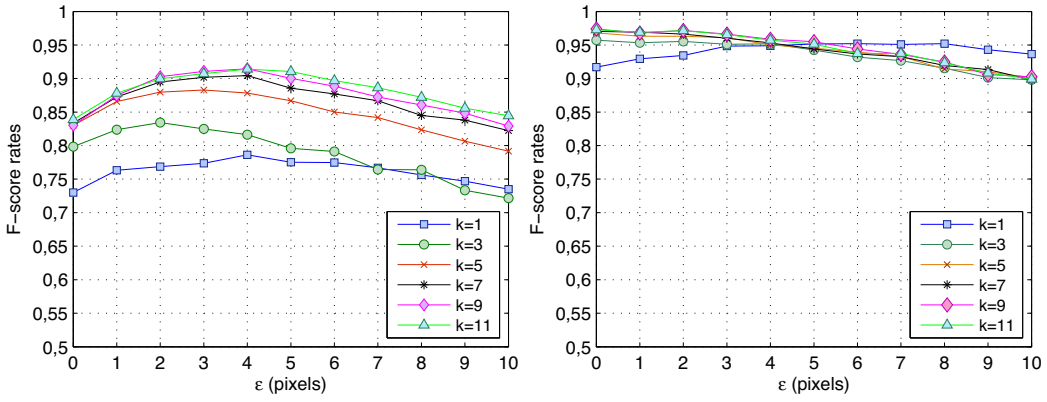


Fig. 16. Influence of distance (ε) on the detection of correspondences for different numbers of solutions (k). (a) In the case of two views, the larger the number of solutions the greater the system's performance, becoming stabilized above $k = 9$; (b) In the case of three views, the larger the number of solutions, the shorter the distance ε.

It is important to note that the point can be occluded in the following views, but its position remains valid because our method is based on a geometric model that defines the scene. We also show that the use of multiple random solutions makes it possible to improve the performance of the correspondence. Although our method starts from the basis that there is a set of points in previous correspondence necessary to determine the fundamental matrices and the trifocal tensors, it is designed to maximize the correspondences in specific regions of each image and not necessarily in a specific point that is not relevant to that method.

In the experiments performed we considered three sets of images: outdoor, indoor, and industrial images. For the first two we used a correspondence through invariant characteristics of SURF (Bay et al., 2008) with the purpose of constructing multiple geometric solutions in two views. The results obtained with these sets indicate that the BIGC algorithm was capable of determining point-to-point correspondence precisely, with a performance F-score$= 97\%$ in stereo images at a discretized distance $\varepsilon = 0$ pixels for outdoor images. This performance decreases in the case of indoor images to an F-score$= 87\%$ with $k = 11$ using a distance of $\varepsilon = 2$. This result is due in part to the lower dispersion of correspondences in pairs of images. It is interesting to note that the algorithm required a greater number of solutions in correspondences to increase its performance. In the last experiment, the base correspondence was determined according to the relation of external markers that comply with the motion of the object. In this case the best performance with the TRIGC algorithm was an F-score$= 97\%$ at a discretized distance of $\varepsilon = 0$ pixels in a sequence of three views. This means that the correspondence in three views has a sub-pixel resolution.

For all the images analyzed, regardless of the base method of correspondence used, we showed that the point-to-point correspondence can be generated through a multiple geometric relation between two and three views. An important characteristic of our method is that it can be used in

sequences of images that have a low signal-to-noise ratio. In those cases invariant algorithms will not achieve a good performance due to the appearance of many false alarms. Our method, on the other hand, can solve this problem due to its geometric formulation, as was mentioned in the set of industrial images. In relation to this last point we stress that out method can be applied as support of industrial control to follow-up faults in uncalibrated sequences, among other applications.

Acknowledgment: This work was partially supported by the Grant No. ACT-32 sponsored by Anillos-Conicyt-Chile

REFERENCES

Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1994, Feb.). Performance of optical flow techniques. *International Journal of Computer Vision*, *12*(1), 43–77.

Bartoli, A., & Sturm, P. (2004, March). Nonlinear estimation of the fundamental matrix with minimal parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(3), 426–432.

Bay, H., Ess, A., Tuytelaars, T., & Gool, L. (2008). Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, *110*(3), 346–359.

Bhat, P., Zheng, K., Snavely, N., Agarwala, A., Agrawala, M., Cohen, M., et al. (2006). Piecewise image registration in the presence of multiple large motions. In *Ieee conference on computer vision and pattern recognition* (Vol. 2, pp. 2491–2497). IEEE.

Bosch, A., Zisserman, A., & X., M. (2007, July 9–11). Representing shape with a spatial pyramid kernel. In ACM (Ed.), *Proceedings of the 6th acm international conference on image and video retrieval (civr)* (pp. 401 – 408). Amsterdam, The Netherlands: ACM.

Carrasco, M., & Mery, D. (2006). Automated visual inspection using trifocal analysis in an uncalibrated sequence of images. *Materials Evaluation*, *64*(9), 900–906.

Carrasco, M., Pizarro, L., & Mery, D. (2008). Image acquisition and automated inspection of wine bottlenecks by tracking in multiple views. In *Proc. of the 8th int. conf. on signal processing, computational geometry and artificial vision (iscgav'08)* (pp. 82–89). Rhodes Island, Greece.

Caspi, Y., & Irani. (2000). A step towards sequence-to-sequence alignment. In *Ieee conference on computer vision and pattern recognition (cvpr)*. (pp. 682–689). Hilton Head Island, South Carolina: IEEE.

Caspi, Y., Simakov, D., & Irani, M. (2006). Feature-based sequence-to-sequence matching. *International Journal of Computer Vision*, *68*(1), 53–64.

Chen, Z., Wu, C., Shen, P., Liu, Y., & Quan, L. (2000). A robust algorithhm to estimate the fundamental matrix. *Pattern Recogn Letters*, *21*, 851–861.

Dempster, A. P., Laird, N., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B*, *39*, 1–38.

Faugeras, O., Luong, Q.-T., & Papadopoulo, T. (2001). *The geometry of multiple images: The laws that govern the formation of multiple images of a scene and some of their applications*. Cambridge MA, London: The MIT Press.

Faugeras, O. D. (1993). *Three-dimensional computer vision*. MIT Press.

Fischler, M., & Bolles, R. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Fitzgibbon, A. (2003, Dec.). Robust registration of 2d and 3d point sets. *Image and Vision Computing*, *21*(13–14), 1145–1153.

Hartley, R., & Zisserman, A. (2000). *Multiple view geometry in computer vision*. Cambridge, UK: Cambridge University Press.

Kadir, T., Zisserman, A., & Brady, M. (2004, May.). An affine invariant salient region detector. *Lecture Notes in Computer Science*, *1*(3021), 228–241.

Kushal, A., & Ponce, J. (2006). Modeling 3d objects from stereo views and recognizing them in photographs. In *European conference on computer vision* (Vol. 3952, pp. 563–574). Springer.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int J Comput Vis*, *60*(2), 91-110.

Lucas, B., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *In Proceedings of Image Understanding Workshop*, 674–679.

Matas, J., Chum, O., Urban, M., & Pajdla., T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *Procedings of british machine vision conference (bmvc)* (pp. 384–393).

Mikolajczyk, K., & Schmid., C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, *60*(1), 63–86.

Moreels, P., & Perona, P. (2007). Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, *73*(3), 263–284.

Olson, D., David L.; Delen. (2008). *Advanced data mining techniques*. Springer.

Romano, R. (2002). *Projective minimal analysis of camera geometry*. Phd. thesis, M.I.T., USA.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, *47*(1–3), 7–42.

Shashua, A. (1995). Algebraic functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*(8), 779–789.

Tordoff, B. J., & Murray, D. W. (2005, Oct.). Guided-mlesac: faster image transform estimation by using matching priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(10), 1523–1535.

Torr, P. (2002). Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision*, *50*(1), 35–61.

Torr, P., & Zisserman, A. (2000). Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, *78*, 138–156.

Tuytelaars, T., & Mikolajczyk, K. (2007). Local invariant feature detectors: A survey. *Computer Graphics and Vision*, *3*(3), 177-280.

Vidal, R., Ma, Y., Soatto, S., & Sastry, S. (2006, Jun). Two-view multibody structure from motion.

International Journal of Computer Vision, *68*(1), 7–25.

Zhang, Z., Deriche, R., Faugeras, O., & Luong, Q.-T. (1995). A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, *78*(1-2), 87–119.

Chapter 5

Prediction of user's intentions

5. PREDICTION OF USER'S INTENTIONS

This chapter presents an algorithm for the detection of intentions based on the extensive use of multiple images in correspondence. The objective is to determine the intention of human action in grasping tasks. To carry out this process we considered the human movement contained in a temporal block of images. Then, for each block, we analyzed the correspondence movement in relation to the last image.

The features extracted from the grasping movement were considered as observations of a Markovian HMM system (Rabiner, 1989). In this way the set of observations in time allows predicting the present state of the system. Another characteristic of our algorithm is its capacity to identify an object in the sequence of images. This process requires detecting previously different objects in the scene and storing a simplified version of each in a dictionary of objects. This information, together with the prediction of the state, were used again as observations in a second HMM system that predicts the user's final action.

This project is oriented at people suffering from altered motion, such as Parkinson or other diseases, in particular to assist in the grabbing movement by means of a robotic arm attached physically to a person's arm (orthosis). This work falls within the framework of the BRAHMA project[1], which is the fusion of multiple automatic systems (electronic, mechanical, and computer vision) for human manipulation and assistance in cases of replacement of an arm, or help in moving in case of motor diseases. We highlight the use of different types of cameras, in particular an eye-tracking device that follows the movement of the pupil in relation to the scene. Merging the information from the cameras on the body and the camera that determines the movement of the eye is of great interest and is currently an area of research.

The main relation between this application and the previous chapters is that the prediction model can be applied to detect defects in correspondence. The idea is that the potential defects that do not show the same movement with respect to corresponding points are considered false alarms. In the literature most inspection systems are based on calibrated systems in which the intrinsic and extrinsic camera parameters are known. Unfortunately, many calibrated systems require periodic re-calibration due to the random motion inherent to any industrial system, which generates motion of the object that is being inspected and/or of the position of the cameras. In this sense, the study of uncalibrated and/or self-calibrated systems is an important source of development for future research in the inspection system as well in the application that we present in this chapter.

[1] Further information in: http://brahma.robot.jussieu.fr/

Paper #6
Prediction of user's intention based on the eye-hand coordination

Miguel Carrasco and Xavier Clady
mlcarras@puc.cl - xavier.clady@upmc.fr

Abstract

The reach-to-grasp an object movement by a human hand comprises not only eye-hand coordination, but also a dynamic motion process that controls and plans the human body in order to grasp an object. To understand this behavior it is necessary to study the eye and hand movements simultaneously. This paper proposes a novel approach for detecting a reach-to-grasp movement by means of computer vision techniques. Our solution blends two viewpoints taken from the user's perspective. First, using an eye-tracker device from the user's head, and second, using a wearable camera from the user's hand perspective. We use the information from these two viewpoints, and we characterize multiple hand movements in conjunction with eye-gaze movements trough a Hidden-Markov Model framework. We show that combining these two sources allows predicting a reach-to-grasp movement and the object wanted with at least 90% of performance.

Index Terms

Gesture recognition, motion planning, image motion analysis, image motion detection.

1 Introduction

Human beings posses a highly developed ability to grasp objects under many different conditions, taking into account variations in position, location, structure, and orientation. This natural ability controlled by the human brain is called eye-hand coordination. Normally, the grasping movement is initiated time before a hand can reach an object, and is regulated by the interaction of several sensorimotor systems such as the visual system, the vestibular system, and proprioception working in conjunction with the head, eye, hand and arm control systems (Crawford, Medendorp, & Marotta, 2004). A central part of this activity occurs in different cortical and subcortical brain regions, with particular importance of the use of our underlying cognitive processes such as attention and memory. According to Flanagan and Lederman (2001), when we grasp an object, the information

Miguel Carrasco is with the Computer Engineering Department at Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860 (143), Santiago, Chile. Xavier Clady is with the Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie-Paris VI, Pyramide-T55/65 CC 173, 4 Place Jussieu, 75005 Paris. France.

perceived by our sensory signals is the result of our preconceived ideas of the object's shape, and the interpretation of the perceived; that is, our brain uses memory representations and visual information simultaneously to grasp objects. We infer that this dynamic activity is an exploratory search, attempting to find the best solution for planning movements to a target and controlling the user's hand, not like the use of active sensors with the purpose of capturing data.

Researchers in many fields have been studying this process for many years, trying to understand the brain mechanism that controls this coordination. However, so far there is no a unique theory which explains this effectively, and furthermore, it is not completely understood (Hayhoe, Bensinger, & Ballard, 1998; Brouwer & Knill, 2007). In spite of this, we find that this field has been expanded to other subjects, especially to Human-Computer Interaction (HCI), within which there is special interest in the design of computer interfaces that take advantage of human interaction, and in particular human vision. A central part of these studies requires the utilization of eye-trackers as mean to get the user's gaze. In brief, an eye-tracker is a device that captures the eye's movements by looking at the corneal reflection to get the eye's position with regard to a reference frame. As a result, it recovers the 2D-position of what we are looking at, that is, the user's gaze.

Over the last years, several studies have successfully used eye-trackers to understand the underlying brain processes that control eye-hand coordination (Karn & Hayhoe, 2000; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Mrotek & Soechting, 2007; Brouwer & Knill, 2007), and more recently, new applications have been designed to help people with motor difficulties, e.g. (Perini, Soria, Prati, & Cucchiara, 2006). Clearly, this technology is providing new ways of increasing HCI by allowing us to capture human vision. However, so far eye-trackers do not succeed at inferring the user's intention. In our problem, the user's intention is the active conscious action with the goal to reach-to-grasp an object with a hand. Generally, we perform reach-to-grasp actions very rapidly and precisely; almost unconsciously because our brain resolves this complex coordination in a short time. A simple approach to get the user's intention –using an eye-tracker– is to gaze persistently at the same object for a long time, so the model may infer the object required using the coordinates of the eye gaze position. However, our eyes do not remain stationary in the same position for a long time. Normally, our eyes change their position constantly, actually at least three times per second or even more depending on the task's complexity. This natural movement, called *eye saccades*, is the way like our brain builds the user's world (Hayhoe et al., 1998). Conversely, the gaze remains stationary when the user initiates a reach-to-grasp action in a short period, and at the same time, the hand's trajectory remains stable toward the object (Roby-Brami, Bennisa, Mokhtaria, & Baraduca, 2000). These two features are essential to predict the user's intention.

This paper describes a novel approach to recognize the user's intention based on the visual information captured from the user himself. Our work contrasts with the classical methods for recognizing hand gestures. Generally, most motion recognition methods capture the user's movements by tracking body parts. Instead, we propose a system that captures the scene by using the user's gaze and reach-to-grasp movements; thus, implicitly we infer the user's intention. The system consists of two visual acquisition systems: an eye-tracker and a micro-camera. Today, with the advent of

fast micro-cameras, it is possible to use a camera without disturbing the user's interaction. This, combined with an eye-tracker, makes it possible to have two points-of-view at the same time, one on the user's head and other under the user's wrist. A general overview of the system is presented in Fig.1

As mentioned previously, the reach-to-grasp action has a particular period when the user initiates the movement toward an object. Based on such observation, our system exploits this feature by allowing us to detect the user's intention. This work is very challenging because many postures for reaching a target can be presented by the same person, increasing the complexity to characterize a unique representation of the grasping movement. Likewise, due to the high variability involved in each movement, several assumptions are need it in order to predict a movement at an early stage. This is the main point addressed in this paper.

To the best of our knowledge, there is no work reporting on a system that blends an eye-tracker and a camera under the user's wrist. The key of this configuration is the possibility of having another viewpoint as a third eye, so multiple viewpoints allow us to increase the probability of detecting the user's intention. As we will see in the following sections, the prediction of the user's intention can be used as a key factor within the HCI domain, such as potential tasks described below, for example.

1) Some diseases with progressive degenerative disorders of the central nervous system often alter motor skills, causing muscle rigidity, tremor, or slow movements, e.g., Parkinson's disease (Jankovic, 2008). Although some computer vision methods have been designed to understand movements in people with motorial disorders, as far as we can tell, so far there have been no studies attempting to understand human intentions in people with degenerative disorders. Therefore, to design a system that can predict the user's intention would allow researchers to understand human behavior, and therapists to apply rehabilitation protocols more efficiently.
2) Interactive robots have been used efficiently in the rehabilitation domain; nonetheless, the co-manipulation domain is still underexploited, mainly because of a lack of research in this area. Recently an important effort has been made to design an active orthosis for aiding people with arm disabilities through the BRAHMA project[1]. To carry out that system, it is critical to know the user's intentions, so the active orthosis can operate to control the user's arm.
3) During the last years several applications of intelligent robotic systems have been developed to help workers perform tasks more efficiently, e.g., attentive workbench (AWB) (Tamura, Sugi, Ota, & Arai, 2004, 2007). For those applications it is essential to detect the user's gestures on hands, fingers, eye-movements, pointing gestures, or a combination of them. As a result, the system performs real-time tasks like handing over assembly parts and removing finished products without explicit instructions (Sugi et al., 2006).

This study proposes the development of the first task, specifically designed to capture the user's intention in people with motorial disorders. Since this area is still unexplored, it is highly relevant

1. The BRAHMA project is currently being carried out by five French laboratories with the aim of developing advanced robotic technology to assist human upper limb motion. More information is available in http://brahma.robot.jussieu.fr/

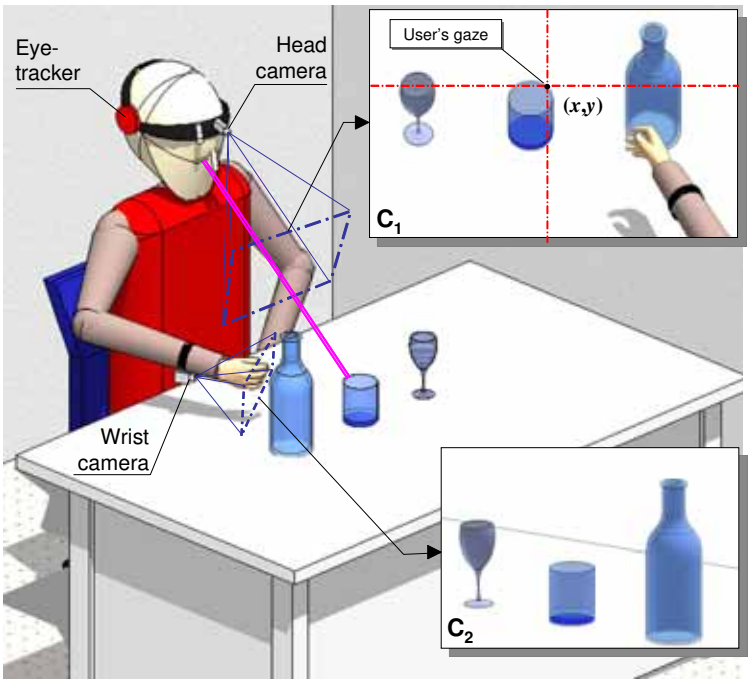


Fig. 1. Schematic view of the propose framework with two wearable cameras and an eye-tracker device. Camera C_1 on the user's head and Camera C_2 under the user's wrist share the same field-of-view (FOV) only when the user starts the reach-to-grasp movement.

to design a system in order to predict the object that the user wants to grasp once the movement has been initiated.

The rest of the paper is organized as follows: Section 2 discusses prior work on human gesture recognition; Section 3 explains our proposed method; Section 4 shows the experimental results; and finally, Section 5 presents our contributions and succinctly describes some ongoing and future work.

2 Related Work

Human activity recognition is currently becoming more and more important. In the last years a wide variety of applications have been developed, such as athletic performance analysis, surveillance, man-machine interfaces, entertainment systems, video conferences. Along this line, HCI is one of them, highly extended, especially due to interest in understanding and recognizing human behavior, as well as in designing computer interfaces that are more usable and receptive of the user's needs. This section shows the main approaches to detect human gestures using computer vision methods. Next, we briefly introduce the main paradigms for motion detection. For a comprehensive review see Gavrila (1999), or Aggarwal and Cai (1997).

For decades, the computer vision community has played an important role in understanding human motion by providing new ways to distinguish several human gestures. Generally, most

applications have been designed to recognize some parts of the human body in order to obtain better representations of them, such as arms, lips, hands, legs, face, and body movements or combinations of them. For that reason, much effort has been made to understand human motion in an overall sense, e.g., (Bobick & Davis, 2001; Kim, Kwak, & Ch, 2006; Shechtman & Irani, 2005); interpreting emotions by means of face movements (Cowie et al., 2001; Busso et al., 2004); or using the body language, e.g. (Achard, Qu, Mokhber, & Milgram, 2007; Yamato, Ohya, & Ishii, 1992). In general, the study of human motion by computer vision methods can be divided in three main approaches: *Passive*, *Active* and *Pointer* paradigms relative to the position of the camera around the user.

2.1 Passive

In this approach the camera is located in a fixed position, normally in front of the user. This means that the camera's field-of-view (FOV) remains constant. This approach has two main scenarios, whether the subject is captured with one stationary camera or from multiple perspectives in correspondence with multiple cameras (Aggarwal & Cai, 1997). The first scenario uses a single camera located in a stationary position. In order to record the user's actions, internal or external markers are used as features to be tracked such as pixels, lines, blobs, and regions. As a main advantage, this configuration uses the same spatial reference to resolve the matching problem in successive frames. The second scenario employs multiple cameras in correspondence with the user. In this case the main benefit is to increase the FOV of the scene; thus, if the subject disappears from one view, the system is capable of seeing it through another camera. However, this configuration is more complex, since, it is necessary to establish a feature correspondence from multiple viewpoints and coordinate them in the same spatial domain, e.g. (Dockstader & Tekalp, 2001). Both systems are suitable when we are interested in knowing the user's movements.

2.2 Active

In this approach the camera is not limited to its position and can interact with its environment, achieving a continuous representation of its surrounding. Bajcsy (1988) and Aloimonos (1990) introduced this paradigm to propose new models and control strategies in active perception systems. Active cameras were initially used to provide perception in autonomous robots; furthermore, they are being employed in human operators by adding a new sense of interaction with its world. Normally it uses external cameras and other devices attached to the human body; this last arrangement is called *wearable* because it is worn on the body. At present, wearable active cameras are offering new ways to increase human-computer interactions by allowing the user to gaze at the world and letting him/her move freely. From a computational perspective, this allows us to obtain a better representation of the user's surrounding and thus to infer the user's gestures. Readers may refer to (Mayol, Tordoff, & Murray, 2000; Davison, Mayol, & Murray, 2003; Kurata, Sakata, Kourogi, Kuzuoka, & Billinghurst, 2004; Campos, Mayol, & Murray, 2006) for further details of wearable vision systems.

2.3 Pointer

Passive and active approaches are useful to understand human motion; however, the major drawback of those approaches is that they are not designed to learn about the user's gaze. The *Pointer* paradigm is designed to overcome this disadvantage. It is based on the idea *what I am looking at is what I want*. Currently, the most widely used device to get the user's gaze is the eye-tracker. The eye-tracker allows tracking eyes movement by giving an estimated position of the user's gaze in real-time relative to an image frame, normally after an initial calibration. The system is composed of two head-mounted cameras: i) a camera that looks at the user's gaze. This camera has almost the same user's field-of-view (FOV), so it answers the first part of *what I am looking at*; and ii) a camera that captures eye movements by means of the corneal reflection; thus, it recovers the position of *what I want*. Initial studies of eye movements were designed to understand the observable surface of the eye when the user was reading (Jacob & Karn, 2003). Today they are widely employed in different areas such as psychology, product design, biology, cognitive-neuroscience, and computer vision. Only in recently they are being used for disabled people with the purpose of increasing the users' interactions with their environment, and thus overcome their motorial difficulties, e.g. (Perini et al., 2006). In general, we observe that this technology is opening new opportunities to understand visual perception from a cognitive perspective and to explain the inherent mechanisms that control eye-hand coordination. However, so far there is no clear consensus or a unified theory that can explain this process effectively (see Desmurget, Pelisson, Rossetti, and Prablanc (1998) for detailed discussions). Therefore, it is not possible to use a specific model that explains the procedure underlying eye-hand coordination. Additionally, this technology is not enough to infer the user's intention; required to predict the grasping movement.

2.4 Discussion

In general, most methods used to detect human motion have been designed using passive and active approaches. These methods have proved to be an effective means to represent the action that takes place in the scene, but unfortunately they cannot interpret the user's intentions, defined as the action to reach-to-grasp an object, because, they are not designed to capture the user's visual system. On the other hand, although the pointer approach has been designed to predict the user's gaze, it cannot differentiate the user's intention unless the user stays explicitly more time watching an object, obviously impractical in real-life situations. Besides, eye-trackers have not been designed to analyze the hand's trajectory toward an object because it does not have a direct vision over the hand. In general, the aim of recognizing human intentions is to predict the inherent intentions in people without explicit instructions. In other words, we aim to know when the user initiates a reach-to-grasp movement toward an unknown object. This last arrangement is the opposite of the classical approaches used in the passive paradigm due to the active camera position.

To overcome these drawbacks we develop a blend of the active and pointer paradigms. To our best knowledge, there is no work that attempts to predict human intentions by exploiting the eye-hand coordination relating these two approaches in combination. In Section 3 we address this problem

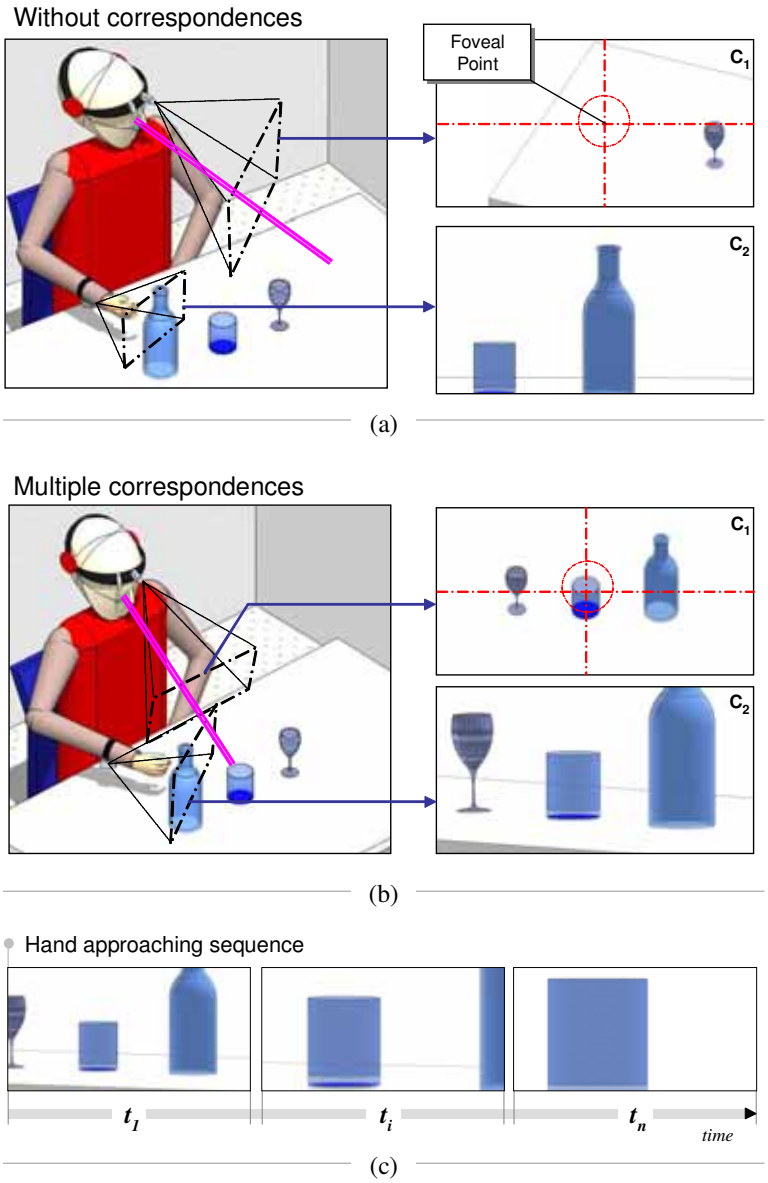


Fig. 2. User's posture when he/she is performing a reach-to-grasp action using a camera on the user's head and a camera under the user's wrist. (a) No correspondence is established when the user changes his Field-of-View (FOV) in relation to his hand; (b) multiple correspondences are established when two FOVs share a similar space; (c) hand approaching sequence using a camera under the user's wrist. At time t_1 multiple objects are detected, but later, at time t_i and t_n the FOV is almost filled due to the proximity between the hand and the object.

in greater detail.

3 PROPOSED METHOD

This section describes the proposed methodology to predict the user's intention when he/she is performing a reach-to-grasp action by means of the eye-hand coordination. As stated above, there are different approaches to detect human motion based on a combination of one or multiple cameras, and the camera position relative to the user and the scene. Namely, three main approaches have been proposed: passive, active, and pointer paradigms. The passive and active approaches are the most widely used in the human recognition domain, and currently, the pointer approach is being used as a method to exploit human vision. This paper proposes a blending of the active and pointer approaches by means of multiple correspondence analysis in multiple views, taking advantage of human vision and active perception.

The main idea is to use the same spatial domain without external markers on the objects in order to infer the user's intention, and thus, to detect the object wanted. To achieve this goal it is necessary to detect the grasping action at an early stage when the movement has been initiated. Normally, the reach-to-grasp movement is very fast, at least at 50 cm/s (Tamura, Sugi, Ota, & Arai, 2006), so the intention recognition should be detected before the hand can reach the object. Based on that observation, we propose to separate the analysis into two tasks. First, to perform a motion prediction through an HMM framework to detect initial hand movement, and second, to detect the target object using eye movements and a camera beneath the user's wrist. A general configuration is presented in Fig.2.

The system is composed of the following devices: 1) an eye-tracker device to get the position of the user's gaze; 2) a camera located on the user's head to capture the user's gaze, calibrated with the eye-tracker; and 3) a camera under the user's wrist to analyze the grasping movements. However, only two cameras are directly visualizing the user's scene (cameras C_1 and C_2 in Fig.2). These cameras are capturing two different FOVs, generally without correspondences (Fig.2a). When the cameras are pointing at the same object, and the eye-tracker gives a stable position on it, the correspondences are increased because the visual information has more correlated areas (Fig.2b).

Normally, the eyes are focused on an object for a small period before the user moves his hand; afterward, when the action has been initiated, the trajectory remains stable toward the object. This last statement does not imply that the movement has the same velocity, trajectory or direction; conversely, each grasping movement is highly variable and several assumptions are needed in order to reduce its complexity.

In summary, the main difference between this contribution and other approaches lies in the clever choice of using inter-frames motion generated when the user moves his hand toward an object. In contrast, other approaches are focused on inferring the user's intentions using body parts as means to infer human intentions. Next, we describe our prediction model used on the user's hand in order to infer the user's intention, and then we describe the blending model using the visual information of the eye-tracker.

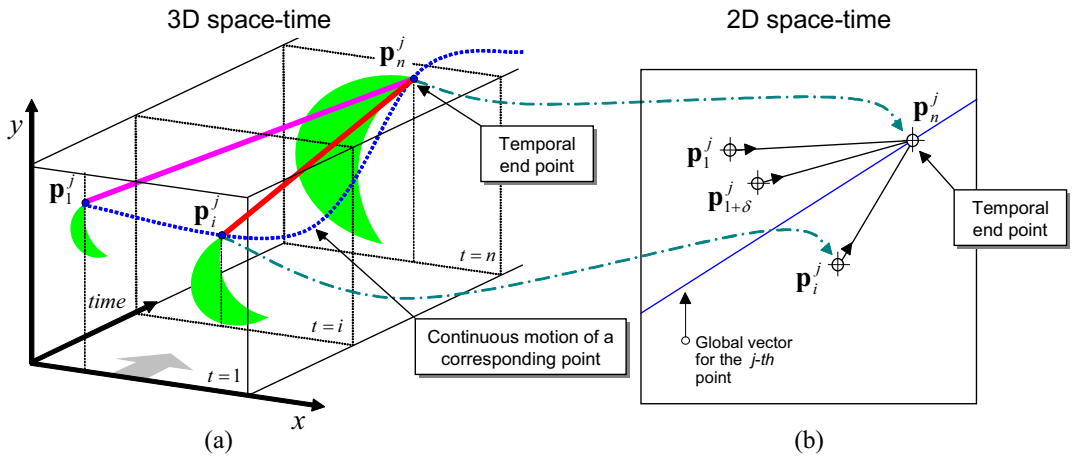


Fig. 3. Schematic view of the point correspondence in time-space. (a) Corresponding points in 3D time-space volume; (b) Corresponding points in 2D coordinates based on the 3D model.

3.1 Intention recognition approach

Intention recognition is performed using the appearance-base model. This situation is better illustrated in Fig.2c. As seen in the sequence, one can infer that the trajectory remains constant when the user initiates a movement toward a specific object. Accordingly, all objects in the scene start disappearing from the user's FOV until the hand has reached the desired object. Conversely, if the hand movement is stochastic, the probability that the user is performing a reach-to-grasp action is reduced because the motion descriptor does not have an approaching pattern. This last statement is the key point of hand recognition. The problem now is how to build a robust pattern of motion. To achieve this goal, a tracking analysis will be performed for solving the correspondence problem. Here we briefly review the main tracking methods for the sake of completeness (for a comprehensive review see Yilmaz, Javed, and Shah (2006).)

Tracking is defined as a problem of estimating the correspondence trajectory of an object while it moves around a scene (Yilmaz et al., 2006). In other words, a tracker algorithm infers the position of one or multiple objects contained in a video sequence based on its previous states. Normally appearance, shape, velocity or any kind of a priori information is used to limit the problem's complexity. Although many object tracking methods have been proposed, all of them share the following three main steps: First, selecting a specific object representation; second, determining a robust feature descriptor; and third, choosing an object detection algorithm to track an object along the video sequence based on the previous steps. The classical tracking approaches include methods based on point-tracking, kernel-tracking, and silhouette-tracking. Here we are interested in the point-tracking approach.

The main advantage of the point-tracking approach is to model motion without the object structure, that is, by analyzing the point-trajectory of each point independently of the object to which it belongs. The only relevant information of each point is the trajectory motion, not the object

by itself. Without this restriction, it is possible to analyze the trajectory of multiple points along time. Similar ideas have been treated by Shafique and Shah (2003) and Veenman, Reinders, and Backer (2001), by means of the development of strong heuristics to resolve the multi-frame correspondence problem, especially in the presence of occlusions, misdetections, entries, and exits of objects. Here, the point representation is not treated as a central problem, because the correspondence has been resolved by using invariant features, as will be shown below.

In general, whatever tracking method is selected, feature selection plays a critical role in tracking performance. Usually, two characteristics are required to have a good descriptor. First, the descriptor has to be highly invariant, that is, the features extracted for each point have to remain stable under different viewing conditions, geometric and photogrammetric deformations, and noise variance. Second, it has to be distinctive and unique, so each point can be distinguished clearly in the feature space. Many approaches have been proposed to find robust feature descriptors (Lowe, 1999; Mikolajczyk & C., 2004; Bay, Tuytelaars, & Gool, 2006; Ke & Sukthankar, 2004). In this category, the descriptor proposed by Lowe (Lowe, 2004), called SIFT, is one of those with the best performance. However, according to Bay et al. (2006), is is also more time-consuming. For this reason, here we use the SURF descriptor, proposed by Bay et al. (2006), mainly because of its robustness and speed against variations in scale and rotations.

In what follows we describe our proposed approach to detect hand intentions. There are two main steps involved: First, extracting a robust feature vector composed of the inter-frame motion for each slide window; and second, predicting the user's motion by means of an HMM-based framework. The features proposed have been designed to capture the motion itself, regardless of the objects contained in the video sequence, and without prior information of the object's distance, shape and/or orientation. Our aim is to characterize the motion based on primitive features like velocity, acceleration, orientation, and angle variation. A general overview of our model is presented in Fig.4.

I. Feature matching

Most tracking algorithms based on the appearance-base models (Yamato et al., 1992; Starner & Pentland, 1995; Achard et al., 2007) compute the object's trajectory by using a displacement difference between multiple frames. These methods are appropriate when the object's motion is smooth and without abrupt changes, as for example methods based on the estimation of optical flow (Barron, Fleet, & Beauchimen, 1994). Contrariwise, in our problem the hand motion is particularly fast when the user is performing a reach-to-grasp action, or too stochastic in other cases. Consequently, in neither case it is possible to apply this approach. For that reason, we propose to analyze the motion displacement between intermediate frames.

In a manner similar to that of the spatiotemporal methods described in (Shechtman & Irani, 2005; Niyogi & Adelson, 1994; Laptev & Lindeberg, 2003), our method uses the temporal slide window (TSW) approach extracted along the video sequence. As suggested by Shechtman and Irani (2005), each human action can induce a particular pattern despite differences in illumination, background,

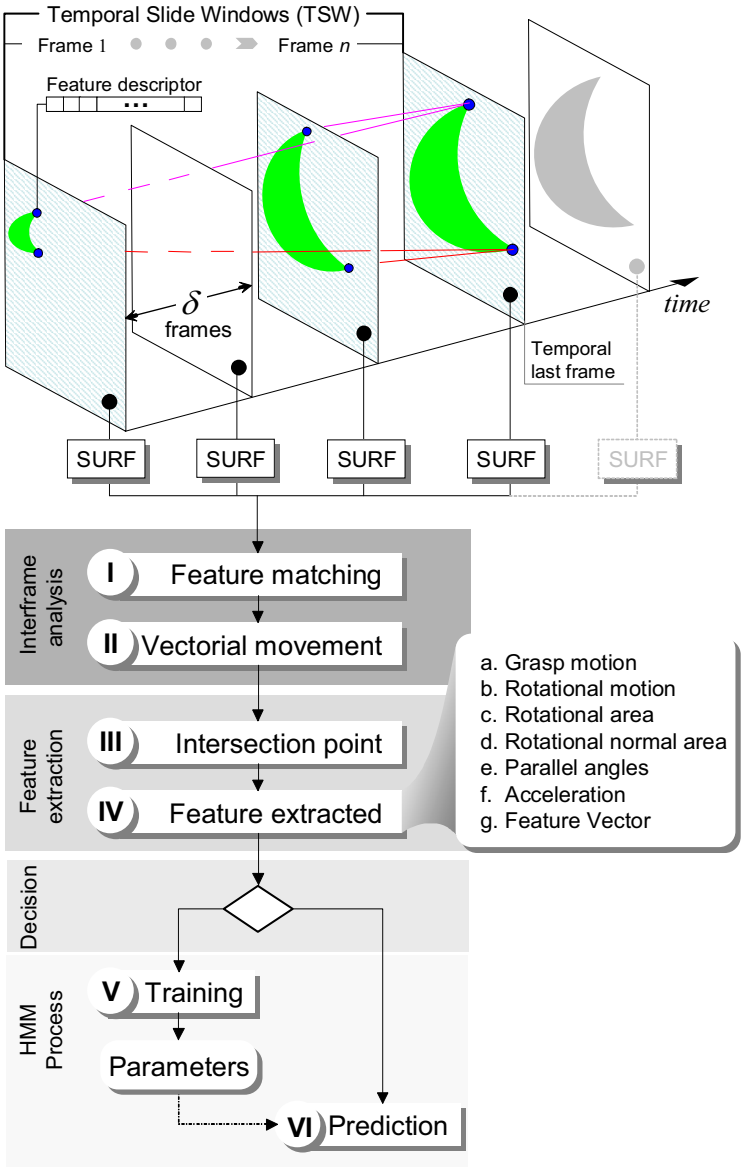


Fig. 4. Proposed hand intention recognition model based on the analysis of temporal slide windows (TSW) interspaced by δ-frames

color, or texture. Based on this idea, we propose to build a motion model of multiple corresponding points in relation to each last corresponding point frame. Unlike the current classical frame-to-frame correspondences, our method is able to estimate the overall motion from each slide window. This is schematically outlined in Fig.3.

First we compute invariant interest-points by means of the SURF algorithm (Bay et al., 2006).

This task is performed for each δ-frames contained on a TSW of the sequence, where $\delta \in \{1, \ldots, 5\}$. Assuming that the features extracted by SURF are more resilient to long variations, it is possible to relate multiple corresponding points along time with greater probability. For instance, let $\mathbf{p}_1^j = [x_1^j, y_1^j, 1]^\top$ be the position of the $j\text{-}th$ interest point in time $t = 1$ stored in homogenous coordinates. If this interest point is in correspondence with point $\mathbf{p}_n^j$ in time $t = n$ it must have a strong similarity between their features. Likewise, after δ-frames, point $\mathbf{p}_i^j$ is corresponding with $\mathbf{p}_n^j$ using the same similarity metric, where $i \in \{1, \ldots, n - \delta\}$. Although points $\mathbf{p}_1^j$ and $\mathbf{p}_{1+\delta}^j$ are corresponding to each other, here we are not interested in the motion between some small displacement. Contrariwise, we seek to compute the global motion, as depicted in Fig.4a. In the following analysis we consider the first slide window contained at time $t = 1$ and $t = n$.

Second, once the interest points for some δ-frames are extracted, we attempt to relate them. Namely, we try to find a vector that relates the $j\text{-}th$ point $\mathbf{p}_i^j \mapsto \mathbf{p}_n^j$ for all $i \in \{1, \ldots, n - \delta\}$. Here the key point is to relate multiple corresponding points with respect to the set of points extracted in the last frame of each TSW. If for some frames this relation does not exist, it is not relevant while a minimum number of correspondences has been established. As a result, we reduce the motion's complexity caused by the inter-frame approach (Shafique & Shah, 2003), and we also ensure a single correspondence between multiple frames.

Our feature matching procedure to relate two interest points is the following. First, we compute the distance of the feature vector $\mathbf{f}_i^j$ of the $j\text{-}th$ point against all feature vectors extracted in time $t = n$ as

$$\mathbf{F}_{i,n}^j(\omega) = \arccos\left(\frac{\mathbf{f}_i^j \cdot \mathbf{f}_n^\omega}{\|\mathbf{f}_i^j\| \|\mathbf{f}_n^\omega\|}\right) \text{ for all } \omega \in \Omega\,, \tag{1}$$

where $\Omega = \{1, \ldots, s\}$ is the set of interest points detected at time $t = n$, and $\mathbf{F}_{i,n}^j$ is the vector containing the angle-value for each point j with regard to point ω. Although the cosine similarity is useful for finding the most similar vector by seeking the lowest angle-value, in many cases this correspondence is incorrect because corresponding point does not exist in the last frame. To avoid this error, we use a procedure to reinforce a correct matching. Second, we extract the two lowest values of vector $\mathbf{F}_{i,n}^j$ defined as

$$d^{j,j'} = \mathbf{F}_{i,n}^j(j') \text{ and } d^{j,j''} = \mathbf{F}_{i,n}^j(j'')\,, \tag{2}$$

where $d^{j,j'}$ is the first lowest angle-value, and $d^{j,j''}$ is the second lowest angle-value of $\mathbf{F}_{i,n}^j$, respectively. Therefore, the link between $\mathbf{p}_i^j \mapsto \mathbf{p}_n^{j'}$ is established if the constraint

$$\frac{d^{j,j'}}{d^{j,j''}} < r \qquad \text{where } r \in [0, 1]\,, \tag{3}$$

is fulfilled. Parameter r is the ratio between the best two feature candidates j' and j'' in order to reduce the number of mismatches and to retain the maximum amount of correct matches. In other words, this criterion assures that the $j\text{-}th$point is matched with its nearest neighbor if this is much

closer than the second neighbor. Note that the point j' refers to an unknown point at time $t = n$; nonetheless, in case of a correct match $j' = j$, since it is the same point between the time $t = i$ and $t = n$.

This matching criterion is known as the Nearest-Neighbor with Distance Ratio (NNDR) (Lowe, 1999). In general, the NNDR criterion reduces the number of corresponding points when there are noise-points, and when no corresponding point exists. This last fault normally occurs when there is a fast motion sequence, as in our problem. According to Sidibe, Montesinos, and Janaqi (2007), although the NNDR criterion does not have the best performance, it has been selected because it is computationally less expensive.

II. Vectorial movement

Once a set of corresponding points contained in the TWS is established, we determine the motion vector for that point (see Fig.4). For instance, let $\mathbf{q}_{i,n}^{j}$ be a homogenous vector that crosses the points $\mathbf{p}_i^j \mapsto \mathbf{p}_n^{j'}$ defined as

$$\mathbf{q}_{i,n}^{j,j'} = \mathbf{p}_i^j \times \mathbf{p}_n^{j'} = [x_i^j, y_i^j, 1] \times [x_n^{j'}, y_n^{j'}, 1]. \tag{4}$$

The $\mathbf{q}_{i,n}^{j,j'}$ vector is established between time $t = i$ and time $t = n$ only for the $j\text{-}th$ point[2]; however, several vectors of the same point are required to establish the motion field along time. For this, we define the general motion of multiple vectors that reach point $\mathbf{p}_n^j$ as

$$\mathbf{Q}_{1\mapsto n}^{j} = \begin{bmatrix} \mathbf{q}_1^j \\ \vdots \\ \mathbf{q}_i^j \\ \vdots \\ \mathbf{q}_{n-\delta}^j \end{bmatrix} \tag{5}$$

Matrix $\mathbf{Q}_{1\mapsto n}^{j}$ defines the motion field of the $j\text{-}th$ point for all frames until time $t = n$, for each δ-frame. However, this procedure does not assure that in every δ-frame there is a correspondence, because high geometric and photometric distortions, or partial occlusions may be present in some frames, and therefore they do not satisfy the relation (3). In general, the best case occurs when the row-dimension has $(\frac{n-1}{\delta})$ rows; however, this is not always achieved. To ensure that the motion field is correct, we define a parameter ρ as the minimum number of rows in the matrix $\mathbf{Q}_{1\mapsto n}^{j}$. Hence, the number of $inliers$-rows in the motion field matrix can vary between $\rho \leq inliers \leq \frac{n-1}{\delta}$. Therefore, the motion field for the $j\text{-}th$ point is established if the constraint

$$inliers \geq \rho\,, \tag{6}$$

is fulfilled. Conversely, if this last constraint is not fulfilled, we discard the motion field for that

2. For simplicity, we have changed the $\mathbf{q}_{i,n}^{j,j'}$ notation to $\mathbf{q}_i^j$, assuming correct matching between the $j\text{-}th$ and $j'\text{-}th$ and between time $t = i$ and $t = n$

point. The next step is to derive only one vector that represents the motion of the $j\text{-}th$ point in time. For this, we map the angle of the $j\text{-}th$ feature point along all $inliers$-frames as

$$\mathbf{F}^{j}_{1\mapsto n} = \left[\mathbf{F}^{j}_{1,n}, \ldots, \mathbf{F}^{j}_{i,n}, \ldots, \mathbf{F}^{j}_{n-\delta,n}\right] \tag{7}$$

where $\mathbf{F}^{j}_{1\mapsto n}$ is a $(1 \times inlier)$ vector of feature angles extracted in different δ-frames for the $j\text{-}th$ point. In other words, each $\mathbf{F}^{j}_{i,n}$ angle weights the relative significance between the points $\mathbf{p}^{j}_{i} \mapsto \mathbf{p}^{j'}_{n}$. Thus, the smaller the angle between two vectors, the stronger is the relation of the same point. Conversely, when the angle is maximum, it can be considered as noise, even if the (3) condition is fulfilled. Based on such observation, we propose to represent each angle-value as a weight vector after a linear transformation. Hence, the $\mathbf{F}^{j}_{1\mapsto n}$ vector is transformed into a $\tilde{\mathbf{F}}^{j}_{1\mapsto n}$ vector used for weighting each motion vector such that

$$\tilde{\mathbf{F}}^{j}_{1\mapsto n} = 1 - \frac{\alpha \mathbf{F}^{j}_{1\mapsto n}}{\max\left(\mathbf{F}^{j}_{1\mapsto n}\right)}. \tag{8}$$

The $\tilde{\mathbf{F}}^{j}_{1\mapsto n}$ vector is a scale-value that gives more relevance to the smaller values. That is, the maximum value is zero, and the smallest value is maximum when $\alpha = 1$. Experimentally, α was fixed at 0.98 to use all vectors mapped in $\mathbf{F}^{j}_{1\mapsto n}$. However, the $\tilde{\mathbf{F}}^{j}_{1\mapsto n}$ vector is not correctly scaled. To determine a correct scale measure we compute $\mathbf{N}^{j}_{1\mapsto n}$ as

$$\mathbf{N}^{j}_{1\mapsto n} = \frac{\tilde{\mathbf{F}}^{j}_{1\mapsto n}}{\sum_{i=1}^{inlier} \tilde{\mathbf{F}}^{j}_{1\mapsto n}(i)}, \tag{9}$$

where $\sum_{i=1}^{inlier} \mathbf{N}^{j}_{1\mapsto n}(i) = 1$. The resultant $\mathbf{N}^{a}_{1\mapsto n}$ vector gives a correct measure of each angle value by taking into account the relative significance between the angles contained in $\mathbf{F}^{j}_{1\mapsto n}$. Finally, we compute the global vector of point $j\text{-}th$ as

$$\mathbf{v}^{j}_{1\mapsto n} = \mathbf{Q}^{j\top}_{1\mapsto n}\mathbf{N}^{j\top}_{1\mapsto n}. \tag{10}$$

where $\mathbf{v}^{j}_{1\mapsto n}$ is a (1×3) vector that maps all $\mathbf{Q}^{j}_{1\mapsto n}(k)$ vectors into a single one by giving more value to vectors with greater similarity, based on the weight feature vector encoded in $\mathbf{N}^{j}_{1\mapsto n}$. More precisely, $\mathbf{v}^{j}_{1\mapsto n}$ is a directional vector of the $j\text{-}th$ point, as shown in Fig.5a.

Additionally, we compute the normal directional vector with the aim of detecting rotational movements, as we shall see later. For this, let $\mathbf{q}_{\perp}{}^{j}_{i,n}$ be the normal vector between points $\mathbf{p}^{j}_{i} \mapsto \mathbf{p}^{j'}_{n}$ established between time $t = i$ and time $t = n$ for the $j\text{-}th$ in time, defined as

$$\mathbf{q}_{\perp}{}^{j,j'}_{i,n} = \begin{bmatrix} x^{j}_{i} - x^{j'}_{n} \\ y^{j}_{i} - y^{j'}_{n} \\ x^{j'}_{n} \cdot (x^{j'}_{n} - x^{j}_{i}) + y^{j'}_{n} \cdot (y^{j'}_{n} - y^{j}_{i}) \end{bmatrix}. \tag{11}$$

Based on this, let $\mathbf{Q}_{\perp}{}^{j}_{1\mapsto n}$ be the matrix of the normal motion field for the $j\text{-}th$ point, in a manner similar to (5). Therefore the normal global vector is as follows

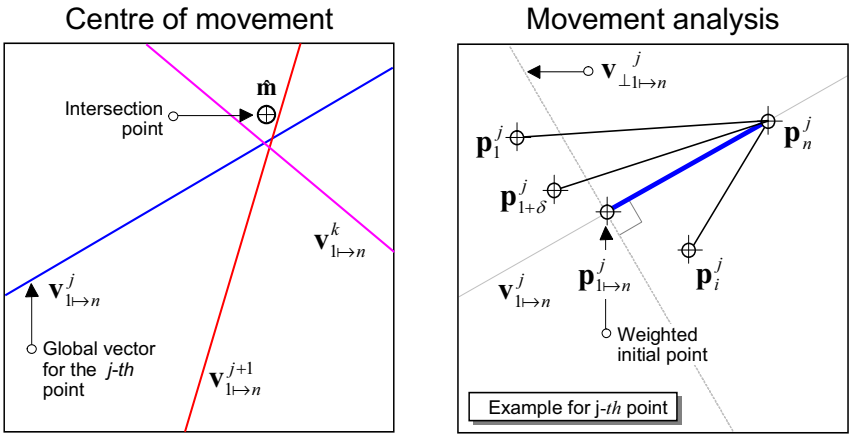


Fig. 5. (a) Multiple lines converge at one point when a reach-to-grasp movement is performed, (b) Once a central point is established, a movement analysis toward that point is performed

$$\mathbf{v}_{\perp 1\mapsto n}^{\ j} = \mathbf{Q}_{\perp 1\mapsto n}^{j\top}\mathbf{N}_{1\mapsto n}^{j\top}. \tag{12}$$

Note that $\mathbf{v}_{\perp 1\mapsto n}^{\ j}$ was computed in the same way as $\mathbf{v}_{1\mapsto n}^{j}$, however in this case the $\mathbf{Q}_{\perp 1\mapsto n}^{\ j}$ matrix is composed of an array of normal vectors, as shown in Fig.5b.

III. Intersection point

For the sake of simplicity, the last procedure has considered the motion of the j-th point. We now turn to the problem of estimating the intersection point of multiple points in correspondence. Suppose we have determined multiple $\mathbf{v}_{1\mapsto n}^{\Theta}$ vectors, where $\Theta = \{1,\ldots,j,\ldots,k\}$ is the set of interest points detected between time $t=1$ and $t=n$, and k is the last point in correspondence that satisfies the relation (6), as shown in Fig.5a. For this, let $\mathbf{A}_{1\mapsto n}^{\Theta}$ be a $(k\times 3)$ matrix that encodes all motion vectors as

$$\mathbf{A}_{1\mapsto n}^{\Theta} = \begin{bmatrix} \mathbf{v}_{1\mapsto n}^{1} \\ \vdots \\ \mathbf{v}_{1\mapsto n}^{j} \\ \vdots \\ \mathbf{v}_{1\mapsto n}^{k} \end{bmatrix} = \begin{bmatrix} a_1 & b_1 & c_1 \\ \vdots & \vdots & \vdots \\ a_j & b_j & c_j \\ \vdots & \vdots & \vdots \\ a_k & b_k & c_k \end{bmatrix}. \tag{13}$$

The next step is to estimate the central point using the vectors contained in $\mathbf{A}_{1\mapsto n}^{\Theta}$. Experimentally, when the reach-to-grasp movement has been initiated, multiple vectors cross over one common point, called *intersection point*. This situation is better illustrated in Fig.5. To estimate the position of the unknown intersection point, we formulate a non-homogeneous system of linear equations,

described as follows

$$\underbrace{\begin{bmatrix} a_1 & b_1 & c_1 \\ \vdots & \vdots & \vdots \\ a_j & b_j & c_j \\ \vdots & \vdots & \vdots \\ a_k & b_k & c_k \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{H}} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}}_{\mathbf{b}}. \tag{14}$$

Changing the notation in matrix terms, (14) can be expressed as

$$\mathbf{H}\mathbf{m} = \mathbf{b}\,,$$

where $\mathbf{H}$ is an overdetermined matrix coefficient of $\mathbf{A}^{\Theta}_{1\mapsto n}$ vectors; because $k \geq \rho$; $\mathbf{m} = [x, y, 1]^\top$ is the vector of unknown (x, y), and $\mathbf{b} = [0, \ldots, 1]^\top$ is the vector of the right hand side solution of the linear system. Since the intersection of vectors does not have a unique intersection point, here we aim to find a vector $\hat{\mathbf{m}}$ such that $\|\mathbf{H}\mathbf{m} - \mathbf{b}\|$ is minimum. A trivial solution of this problem is solved by means of the Least Square (LS)-solution, that is

$$\hat{\mathbf{m}} = \left(\mathbf{H}^\top\mathbf{H}\right)^{-1}\mathbf{H}^\top\mathbf{b}.$$

Nevertheless, if the $\mathbf{H}^\top\mathbf{H}$ product is ill-conditioned, the estimated LS amplifies the errors, giving an inaccurate position of the intersection point. Hence, we use an orthogonal solution because it is numerically more stable. In particular, we use the $\mathbf{QR}$ transformation (Higham, 1996).

The $\mathbf{QR}$ decomposition applied to the $\mathbf{H}$ matrix generates an orthogonal decomposition in terms of an orthogonal matrix $\mathbf{Q}$ and the upper triangular matrix $\mathbf{R}$ such as $\mathbf{H} = \mathbf{QR}$. Therefore, the solution for the non-homogeneous system, using the $\mathbf{QR}$ transformation is

$$\hat{\mathbf{m}} = \mathbf{R}^{-1}\left(\mathbf{Q}^\top\mathbf{b}\right). \tag{15}$$

Finally, since $\hat{\mathbf{m}} = [\hat{m}_1, \hat{m}_2, \hat{m}_3]$ is in homogenous coordinates, the intersection point defined in the (x, y)-plane is $\hat{\mathbf{m}}_{x,y} = (\hat{m}_1/\hat{m}_3, \hat{m}_2/\hat{m}_3)$. Once the intersection point is established, we seek to compute the *normal intersection point* defined as the intersection of all normal vectors $v_{\perp}{}^{\Theta}_{1\mapsto n}$. Based on the above procedure, from (14) to (15), first we define the $\mathbf{A}_{\perp}{}^{\Theta}_{1\mapsto n}$ matrix of all normal vectors contained in Θ as

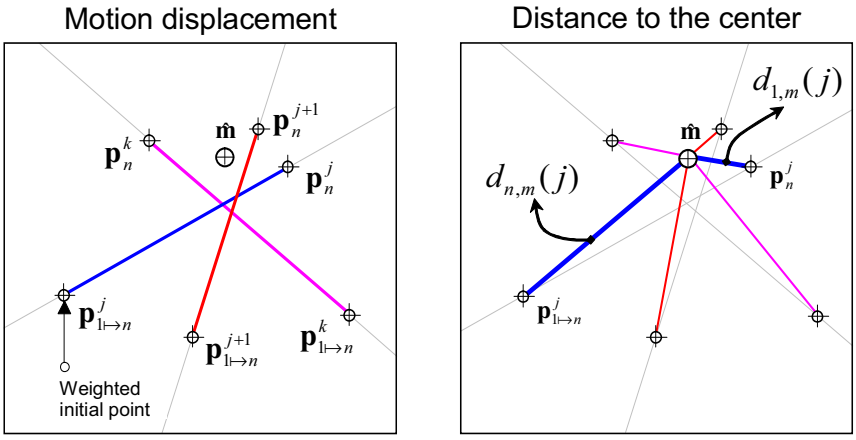


Fig. 6. Analysis of multiple points. (a) Motion of each initial and final trajectory point, (b) Distance to the intersection point

$$\mathbf{A}_{\perp}{}^{\Theta}_{1\mapsto n} = \begin{bmatrix} \mathbf{v}_{\perp}{}^{1}_{1\mapsto n} \\ \vdots \\ \mathbf{v}_{\perp}{}^{j}_{1\mapsto n} \\ \vdots \\ \mathbf{v}_{\perp}{}^{k}_{1\mapsto n} \end{bmatrix}. \tag{16}$$

Then, changing the matrix terms notation of (3.1), the problem of estimating the normal intersection point can be expressed as

$$\mathbf{H}'\mathbf{m}_{\perp} = \mathbf{b}', \tag{17}$$

where $\mathbf{m}_{\perp}$ is a non-homogenous vector that encodes the intersection point of normal vectors in correspondence. Using the $\mathbf{QR}$ transformation applied to the $\mathbf{H}'$, matrix such as $\mathbf{H}' = \mathbf{Q}'\mathbf{R}'$, the normal intersection point is defined as follows,

$$\hat{\mathbf{m}}_{\perp} = \mathbf{R}'^{-1}\left(\mathbf{Q}'^{\top}\mathbf{b}'\right). \tag{18}$$

IV. Featured extracted

Below is an explanation of eight features proposed to predict different hand motions, namely, approaching, distancing, rotational and translational invariant movements. Recall that at this stage we are not interested in detecting the object itself or detecting reach-to-grasp movements. The proposed features are used to predict the motion and later, with an eye-tracker, the object estimation and the user's intentions will be determined.

a. *Grasp motion*: The first two features proposed are related to the grasping action. In general, the grasping motion can be split into two different events. *Zoom-in*: when the hand is going to reach an object; and *Zoom-out*: when the hand is moving away from an object[3]. Here, we propose a simple procedure to infer whether a hand is reaching an object or not based on the intersection point estimated in (15), and the motion transition along the TSW.

Let $\mathbf{P}^{j}_{1\mapsto n}$ be a $(inliers \times 3)$ matrix representing the 2D position in time $[1,\ldots,n]$ for each δ-frame; computed in the same way as matrix $\mathbf{Q}^{j}_{1\mapsto n}$.

$$\mathbf{P}^{j}_{1\mapsto n} = \begin{bmatrix} \mathbf{p}^{j}_{1} \\ \vdots \\ \mathbf{p}^{j}_{i} \\ \vdots \\ \mathbf{p}^{j}_{n-\delta} \end{bmatrix} \tag{19}$$

Namely, the $\mathbf{P}^{j}_{1\mapsto n}$ matrix codes the motion of the j-th point until the last $(n-\delta)$ frame. Based on (10), we remap the motion points taking into account the variations in their features matching as

$$\mathbf{p}^{j}_{1\mapsto n} = \mathbf{P}^{j\top}_{1\mapsto n}\mathbf{N}^{j\top}_{1\mapsto n}\,, \tag{20}$$

where $\mathbf{p}^{j}_{1\mapsto n}$ is a weighted mean position within the vector $\mathbf{v}^{j}_{1\mapsto n}$ as is illustrated in Fig.5b. Extending this procedure for all Θ-points that satisfy the condition (3), let $\mathbf{p}^{\Theta}_{1\mapsto n}$ be the motion of all points in the TSW $[1,\ldots,n]$, and let $\mathbf{p}^{\Theta}_{n}$ be the final position of each point, defined as,

$$\mathbf{p}^{\Theta}_{1\mapsto n} = \begin{bmatrix} \mathbf{p}^{1}_{1\mapsto n} \\ \vdots \\ \mathbf{p}^{j}_{1\mapsto n} \\ \vdots \\ \mathbf{p}^{k}_{1\mapsto n} \end{bmatrix}, \quad \text{and} \quad \mathbf{p}^{\Theta}_{n} = \begin{bmatrix} \mathbf{p}^{1}_{n} \\ \vdots \\ \mathbf{p}^{j}_{n} \\ \vdots \\ \mathbf{p}^{k}_{n} \end{bmatrix} \tag{21}$$

Since vector $\mathbf{p}^{\Theta}_{1\mapsto n}$ codes the initial weighted position, let d_1 be the Euclidean distance of each $\mathbf{p}^{\Theta}_{1\mapsto n}$ vector in relation to the intersection point $\hat{\mathbf{m}}$, and let d_n be the Euclidean distance of each final position $\mathbf{p}^{\Theta}_{n}$ in relation to same intersection point $\hat{\mathbf{m}}$ as,

$$d_{1,m}(j) = \|\mathbf{p}^{\Theta}_{1\mapsto n}(j) - \hat{\mathbf{m}}\|\,, \quad d_{n,m}(j) = \|\mathbf{p}^{\Theta}_{n}(j) - \hat{\mathbf{m}}\| \tag{22}$$

The Euclidean distance $d_{1,m}$ and $d_{n,m}$ represents the temporal movement around the intersection point $\hat{\mathbf{m}}$, as shown in Fig.6b. Since we know the estimated position of the initial, final, and intersection points, the next step is to determine whether the motion is toward the center or

3. Note that if the object has been grasped, the second event is totally irrelevant. Nevertheless, this analysis allows us to infer more precisely the zooming-in motion.

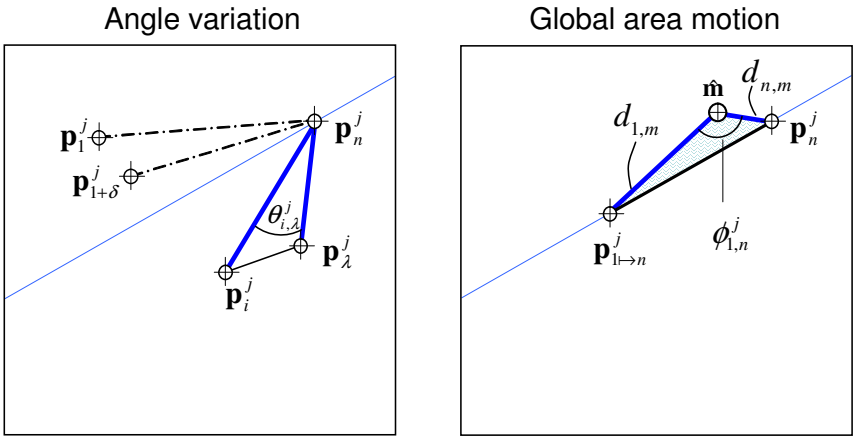


Fig. 7. (a) Temporal angle variation, (b) Global area motion between weighted mean position and last point contained in each time-window.

not. Based on these values, we define the function $v(j)$ as the number of nearest points to the intersection point, as follows:

$$v(j) = \begin{cases} 1 & \text{if } d_{n,m}(j) \geq d_{1,m}(j) \\ 0 & \text{otherwise.} \end{cases}$$

Finally, from $v(j)$, we extract the mean f_1 and the second central moment f_2, as follows:

$$f_1 = \mu(v) \tag{23}$$

$$f_2 = \sigma^2(v)\,, \tag{24}$$

where $\mu(\cdot)$ is the mean and $\sigma^2(\cdot)$ is the variance. The above features indicate that the movement is toward an object if $f_1 \mapsto 1$; and conversely, the movement is against an object if $f_1 \mapsto 0$. To confirm this prediction, the variance σ^2 should be low in any case.

b. *Rotational motion*: The rotational motion feature gives a temporal variation of each point in correspondence. The main idea is to capture rotational movements independently of their turning direction, and thus, to compute the angular velocity of each point.

Suppose that the link between $\mathbf{p}_i^j \mapsto \mathbf{p}_n^j$ and $\mathbf{p}_\lambda^j \mapsto \mathbf{p}_n^j$ exists. Therefore, s_i^j and s_λ^j are two consecutive slopes of the j-th point separated by λ-frames, respectively, defined as,

$$s_i^j = \frac{y_i^j - y_n^j}{x_i^j - x_n^j}, \qquad s_\lambda^j = \frac{y_\lambda^j - y_n^j}{x_\lambda^j - x_n^j}.$$

Since both points are aiming at the last point p_n^j in time $t = n$, as depicted in Fig. 7a; therefore, by transitivity, this also implies that $\mathbf{p}_i^j \mapsto \mathbf{p}_\lambda^j$, where $t_\lambda > t_i$. Thereby, the angle between these

consecutive slopes is

$$\theta_{i,\lambda}^{j} = \arctan\left|\frac{s_i^j - s_\lambda^j}{1 + s_i^j s_\lambda^j}\right|,$$

Based on this result, we calculate the angular velocity ω between p_i^j and p_λ^j in order to compute the motion variation over time, defined as

$$\omega_{i,\lambda}^{j} = \frac{\triangle\theta_{i,\lambda}^{j}}{\triangle t_{i,\lambda}},$$

for all $i = 1, \ldots, inliers$, where $\triangle t_{i,\lambda}$ is the time difference between two consecutive frames. Normally, $\lambda = i + \delta$; however, if condition (3) is not satisfied for every δ-frame, the time difference between two points is increased. For this reason we consider that λ varies according to the time difference between two consecutive temporal frames; that is $\lambda \geq i + \delta$. Clearly, the angular velocity ω is a useful feature to estimate the motion rate of each point along the TSW. Combining the above value with the Euclidean distance between points $\mathbf{p}_i^j$ and $\mathbf{p}_\lambda^j$ we propose an invariant feature that distinguishes rotational and translational movements as follows

$$f_3 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{inlier}\sigma^2(\omega_{i,\lambda}^{j})}{\sum_{j=1}^{k}\sum_{i=1}^{inlier}\sigma^2(\|\mathbf{p}_i^j - \mathbf{p}_\lambda^j\|)}, \tag{25}$$

The f_3 feature tends to zero when the movement is translational. This happens when the hand is moving constantly in the same direction, regardless of its angle direction. Thus, the variance of the Euclidean distance is high and the variance of the angular velocity is low. Conversely, when the motion is rotational, f_3 tends to be greater than one. Accordingly, the variance of the Euclidean distance as well as the variance of the angular velocity tend to be low, since every point undergoes the same angular rotation.

c. *Rotational area*: Along the same line as the above feature, we propose to compute the area covered by the central intersection point (15), the weighted mean position (20), and the final end position of each point as a measure to compute variations of the area over time, as shown in Fig.7b. More formally, let $\phi_{1\mapsto n}^{j}$ be the angle between $\mathbf{p}_{1\mapsto n}^{j}$ and p_n^j on point $\hat{\mathbf{m}}$. The log-area variation of multiple points along the TSW is the following,

$$f_4 = \log\left(\frac{1}{2k}\sum_{j=1}^{k} d_{1,m}(j) d_{n,m}(j)\sin(\phi_{1,n}^{j})\right) \tag{26}$$

where $d_{1,m}(j)$ and $d_{n,m}(j)$ are the Euclidean distance with respect to point $\hat{\mathbf{m}}$ of the j-th point, estimated previously in (22) contained in Θ. The above feature computes the relative area between the camera and the hand. In general, its variation over time is a useful way to estimate if the motion is toward an object or not. Here we compute $\log$-area in order to reduce its scale variation.

c. *Rotational Normal variation*: When the movement is purely rotational, the intersection point

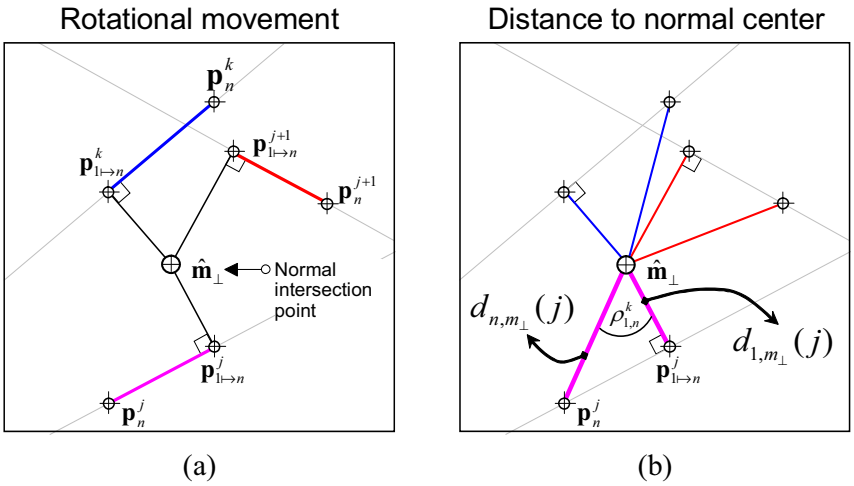


Fig. 8. Analysis of multiple points. (a) Motion of each initial and final trajectory point with respect to the intersection point, (b) Distance to the normal intersection point

$\hat{\mathbf{m}}$ does not represent the real center of motion. In a manner similar to that of the last feature, we first propose to compute the Euclidean distance to the normal intersection point $\hat{\mathbf{m}}_\perp$, defined as

$$d_{1,m_\perp}(j) = \|\mathbf{p}^{\Theta}_{1\mapsto n}(j) - \hat{\mathbf{m}}_\perp\|, d_{n,m_\perp}(j) = \|\mathbf{p}^{\Theta}_{n}(j) - \hat{\mathbf{m}}_\perp\| \tag{27}$$

where $d_{1,m_\perp}(j)$ and $d_{1,m_\perp}(j)$ are the temporal distance of the j-point around the normal intersection point $\hat{\mathbf{m}}_\perp$. Second, we compute the angle $\rho^j_{1,n}$ between $\mathbf{p}^j_{1\mapsto n}$ and $\mathbf{p}^j_n$ with respect to the point $\hat{\mathbf{m}}_\perp$, as shown in Fig.8b. Using the above values, we compute the $\log$-area of the rotational normal movement as,

$$f_5 = \log\left(\frac{1}{2k}\sum_{j=1}^{k} d_{1,m_\perp}(j) d_{n,m_\perp}(j)\sin(\rho^j_{1,n})\right) \tag{28}$$

Normally this variation is high when the motion is not rotational because the intersection of normal vectors does not exist. However, when the motion starts becoming rotational there is a point $\hat{\mathbf{m}}$ that intersects all normal vectors $v_{\perp}{}^{\Theta}_{1\mapsto n}$. Consequently, all points have the same spin angle and a similar variation.

A consequence of the above result is that we have obtained two angle variations, namely the angle variation $\rho^j_{1,n}$ for the j point with respect to the normal intersection point $\hat{\mathbf{m}}_\perp$, and the angle variation $\phi^j_{1,n}$ with respect to the intersection point $\hat{\mathbf{m}}$. Combining both angles in one feature allows us to get a variation of motion over time, defined as follows

$$f_6 = \frac{\sum_{j=1}^{k}\phi^j_{1,n}}{\sum_{j=1}^{k}\rho^j_{1,n}} \tag{29}$$

For example, for rotational movements, f_6 remains constant with a low value over time. In

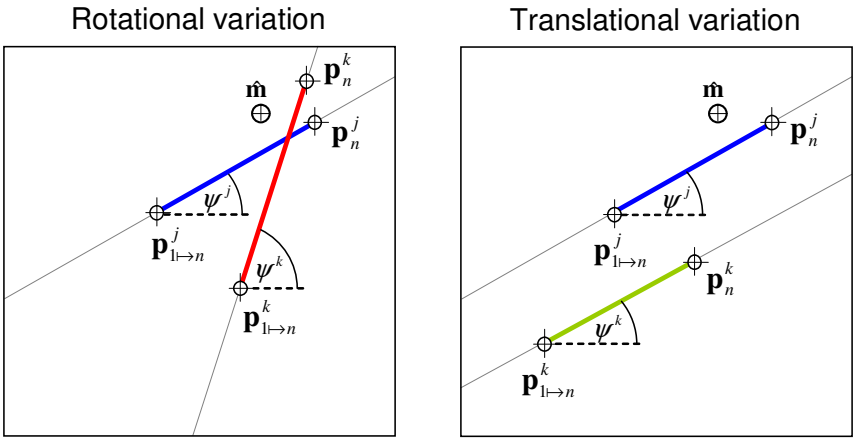


Fig. 9. (a) Different angles are found when motion is rotational (b) Similar angles are found when motion is purely translational

the case of linear movements, f_6 tends to be high, and finally, for zooming-in or zooming-out movements, f_6 varies according to each movement growing or declining, respectively

d. *Parallel angles*: The parallel angles gives the relative variation between the angles of each weighted mean position and its end position, as is illustrated in Fig.9. The key point of this feature is to detect only translational movements, regardless of the angular direction and orientation of the movement. We compute the angle variation ψ^j of each j- points as

$$\psi^j = \arctan \left| \frac{y^j_{1 \mapsto n} - y^j_n}{x^j_{1 \mapsto n} - x^j_n} \right|$$

where $\mathbf{p}^j_{1 \mapsto n} = [x^j_{1 \mapsto n}, y^j_{1 \mapsto n}]$ is the weighted mean position described previously, and $\mathbf{p}^j_n = [x^j_n, y^j_n]$ is the end position. Combining the above angle and the Euclidean distance, we compute the parallel variation as follows

$$f_7 = \log \left(\frac{\sum_{j=1}^{k} (\| \mathbf{p}^j_{1 \mapsto n} - \mathbf{p}^j_n \|)}{k \sigma^2(\psi)} \right) \tag{30}$$

In contrast to the f_3 feature, the above feature tends to zero when the motion is rotational and is only high when the motion is purely translational, because the angle variation is very low and the distance of each weighted point remains constant. In all other cases the angle variation is high, and the Euclidean distance varies according to the type of movement.

e. *Acceleration*: The last feature computes the acceleration of each corresponding point encoded in the $\mathbf{P}^{\Theta}_{1 \mapsto n}$ vector. Unlike the time, $\triangle t_{i,\lambda}$ relates the time difference between two consecutive frames; here we compute the time difference with respect to the last temporal frame $\mathbf{p}^{\Theta}_n$ for each

point contained in the Θ set; in other words, we compute $\triangle t_{i,n}$ in order to normalize the velocity to a single unit of time. For instance, let v_x^j and v_y^j be the temporal velocity with respect to point $\mathbf{p}_n^j$ taking into account the temporal difference $t_{i,\lambda}$ as follows

$$v_x^j(i) = \frac{x_n^j - x_i^j}{\triangle t_{i,n}} \qquad v_y^j(i) = \frac{y_n^j - y_i^j}{\triangle t_{i,n}} \tag{31}$$

Based on the above results, the acceleration of the $j\text{-}th$ point in time $t = i$ is defined by

$$a_x(i) = \frac{\sum_{j=1}^{k} v_x^j(i) - \sum_{j=1}^{k} v_x^j(i-\lambda)}{k \triangle t_{i,\lambda}} \tag{32}$$

$$a_y(i) = \frac{\sum_{j=1}^{k} v_y^j(i) - \sum_{j=1}^{k} v_y^j(i-\lambda)}{k \triangle t_{i,\lambda}} \tag{33}$$

In the above case, we compute the time $\triangle t_{i,\lambda}$ because we seek the relative acceleration between consecutive frames. Based on these results, we propose the following feature to quantify the global acceleration as

$$f_8 = \frac{\sigma^2(a_x)}{\sigma^2(a_x) + \sigma^2(a_y)}\,, \tag{34}$$

where a_x and a_y are two vectors containing the relative acceleration from each slide window.

f. *Feature vector*: In the previous steps we have proposed eight feature descriptors that encode different aspects of the reach-to-grasping movements, namely rotational acceleration, linear acceleration, angle variation, area variation and motion direction. These features are later used as an input for an HMM system as shown in the following section.

For simplicity, the above analysis has considered a TSW in time $[t = 1, \ldots, t = n]$. Thus, the first feature vector $\mathbf{o}_1$ is composed as follows:

$$\mathbf{o}_1 \equiv \mathbf{o}_{1 \mapsto n} = [f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8]^\top\,, \tag{35}$$

but to infer the user's intention it is necessary to get multiple TSW. Recall that each TSW is composed by a sequence of δ frames, as shown in Fig.4. Therefore, a movement is represented by a sequence of slide windows, each one composed of eight features,

$$\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]\,, \tag{36}$$

where T is the total frame number of the video sequence and $\mathbf{O}$ is the observed symbol sequence.

V. Training HMM for recognition

Below, we briefly describe the main component of an HMM-based system used to recognize the user's intentions. For a comprehensive review we refer to Rabiner (Rabiner, 1989). HMM is a type of stochastic signal model composed of a Markov Chain whose states cannot be observed directly,

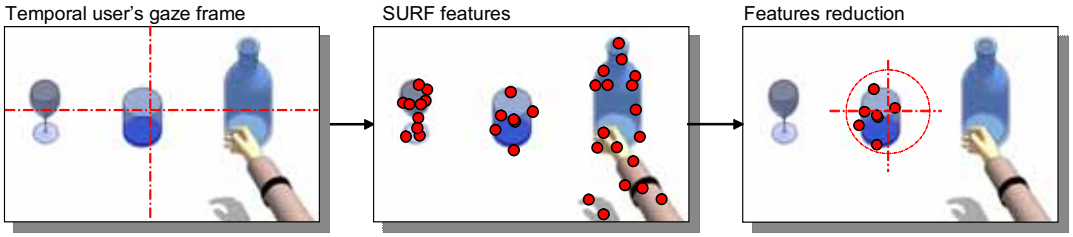


Fig. 10. (a) temporal user's gaze recorded by the head camera; (b) SURF extraction on this frame; (c) reduction of features extracted by SURF using a ratio centered from the user's gaze

but can be observed through the sequence of observations. Currently, HMMs have been used in a wide range of applications, especially in those where it is necessary to deal with time-series with spatial temporal variabilities, such as, for example, intention and gesture recognition (e.g. (Yamato et al., 1992; Starner & Pentland, 1995; He & Kundu, 1991; Achard et al., 2007)).

More formally, an HMM is composed of a number of N-states $\{S_1, S_2, \ldots, S_N\}$ connected by transitions, where each transition has an associated probability defined by matrix A; an emission distribution probability, or the probability of emitting an observation given a state, defined by matrix B; and an initial state distribution $\pi = \{\pi_i\}$. That is, using a compact notation an HMM is fully specified by the $\lambda = (A, B, \pi)$ triplet where

- $A = \{a_{ij}\}$ where $a_{ij} = Pr(q_{t+1} = S_j | q_t = S_i), 1 \leq i, j \leq N$ is the state transition probability distribution, and q_t represent the state at time t.
- $B = \{b_1(\mathbf{o}), b_2(\mathbf{o}), \ldots, b_N(\mathbf{o})\}$ correspond to the observation probability for each state. In our problem, observations are modeled with a Gaussian distribution $b_j(\mathbf{O}) = N(\mathbf{o}, \mu_j, \sigma_j)$ where $\mathbf{o}$ is the feature vector extracted in the last step.
- $\Pi \equiv \{\pi_1, \pi_2, \ldots, \pi_N\}$ where $\pi_i = p(q_1 = S_i), 1 \leq i \leq N$ is the initial state distribution.

Based on the above parameters, the problem is to classify each class defined as a particular user's intention. First we create an HMM for each category using the well known Forward-Backward algorithm (Rabiner, 1989) in order to find the best parameters for each HMM. This is a generalized Expectation-Maximization (EM) algorithm that maximizes the probability of observation sequence given each HMM model for all training sequences.

VI. Prediction HMM for recognition

Once the HMM parameters are established, our goal is to recognize an observed symbol sequence as a particular class or user's intention. Suppose that each λ_i where $i = 1, \ldots, C$, is a model parameter defined for i-class on C classes. Given a sequence of observations $\mathbf{O}$, we calculate $p(\mathbf{O}|\lambda_i)$ for each HMM λ_i and we choose the class with the maximum probability as:

$$class = arg \max_i (p(\mathbf{O}|\lambda_i)). \tag{37}$$

3.2 Integrating eye-tracker information

In the previous sections we have proposed a method for detecting the user's hand intentions by studying hand movements regardless of the objects contained in the scene. This task was performed by an HMM learning process. This section describes how the user's gaze and hand motion improve the detection of the user's intentions, i.e., detecting the object wanted and what the user wants to do. The key information given by an eye-tracker is the relative position with respect to the user's gaze. This information is fundamental because the eye positions over an object has a direct relation with the object wanted. This is because fixations seem to be stable for a period until the object has been grasped (Mrotek & Soechting, 2007; Brouwer & Knill, 2007; Crawford et al., 2004; Hayhoe et al., 2003), and after that fixations are not needed anymore. We recall that when the users initiates his movement toward an object, the probability that the object wanted is the same that the user wants is very high. This is the main idea that will allow us to detect the user's intention. An example of this process is shown in Fig.2b and an approaching sequence in Fig.2c.

In general, eye-trackers provide two sources of information. First, a video sequence with the user's gaze point-of-view; and second, the relative position (x, y) with respect to the above user's gaze sequence. Both sources are synchronized and calibrated, so the relative position of each frame is known. The objective here is to detect when a saccade has begun, and mainly to differentiate the object wanted with respect to other objects in the scene. Our approach is to combine both sources, i.e., to use the user's gaze not only as a point detector but also as a region detector. An example of this process is shown in Fig.10. For this we propose to reduce the number of features that must be searched by discarding all features outside the relative position, a method known as codebook vector quantization (Gersho & Gray, 1992).

In what follows we describe a general process to detect grasping intentions, and the object wanted. This process has been done in two steps. First, we previously classified each object involved in the sequence in an off-line process in order to generate a codebook, composed only of stable features. Second, we used the saccades, the object's stability over time, and the hand's likelihood to build another HMM that allows us to detect when the user is performing a grasping movement. Recall that in the first step it is necessary to know only the object detected, while other features like shape, size, distance and/or orientation are not relevant.

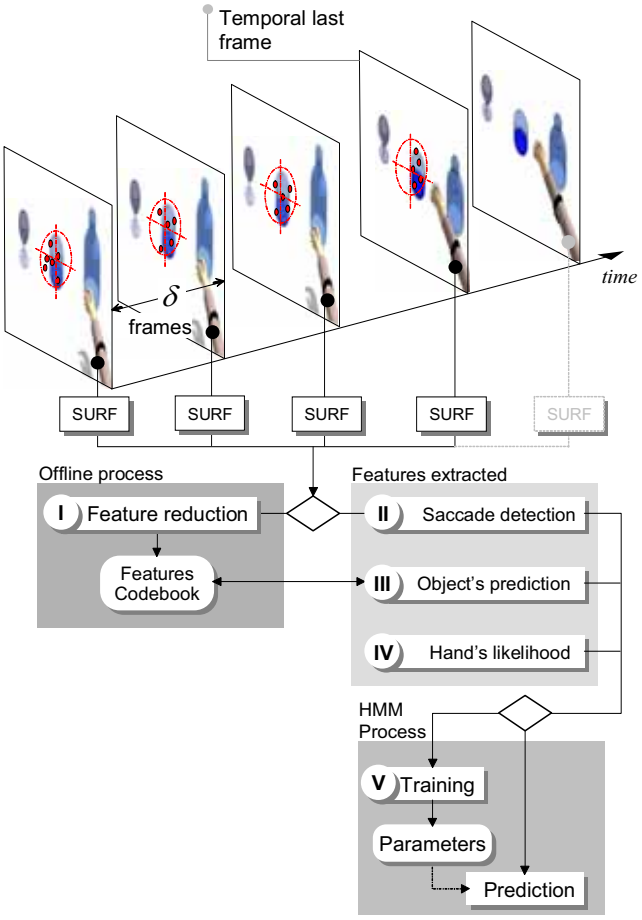


Fig. 11. Proposed user's intention recognition model

I. Features reduction

The main idea of this task is to find more resilient features over time in the video sequence for each object of interest. Our solution uses a similar method proposed by Sivic and Zisserman (Sivic & Zisserman, 2009) to build a visual vocabulary. The key idea is that only few descriptors can be seen many times over the sequence, and therefore more resilient features are used later to classify an object against a new video sequence.

In order to find these resilient features, here we build a codebook using the Mahalanobis distance used as a distance function. For example the distance between two descriptors f_1 and f_2 is given by

$$d(f_1, f_2) = \sqrt{(f_1 - f_2)^\top \Sigma^{-1} (f_1 - f_2)} \tag{38}$$

where the covariance matrix Σ is determined by computing the covariances for descriptors of the

same class within several frames interspaced by δ-frames. According to Sivic and Zisserman (2009), the main advantage of the Mahalanobis distance over the Euclidean distance is that the first one gives less value to noise components of the 128 vector and also de-correlates the components. Although their used SIFT descriptors (Lowe, 2004), the same behavior is valid for the SURF descriptor.

There are many ways to create a codebook (Gersho & Gray, 1992). Here we use a simple method to compute one. First, we extract random frames from a video sequence using the user's gaze; second, we classify each feature as part of an object; and third, we explore all other features space using the Mahalanobis distance in order to create a codebook using the Vector Quantization (VQ) algorithm. Once the VQ features of each object are extracted, we define the $\mathbf{F_n}$ matrix as the codebook of n-objects of interest classified in a test video sequence where

$$\mathbf{F_n} = \begin{bmatrix} class_1 & f_1^1 \\ class_1 & f_1^2 \\ \vdots & \\ class_i & f_i^j \\ class_i & f_i^{j+1} \\ \vdots & \\ class_n & f_n^k \end{bmatrix} \tag{39}$$

where $class_i$ belongs to the i-object classified in the video sequence, and f_i^j is the j feature vector that belongs to the same object after applying a VQ process.

II. Saccade detection

Many studies have shown that fixations are stable for a period before starting a grasping movement, as mentioned previously. Conversely, saccades are not stable in a same position. Since an eye-tracker provides the (x, y) position of the eye motion, we compute the velocity rate over a TSW defined as,

$$v_x(i) = \frac{x_{i+1} - x_i}{\triangle t_{i,i+1}} \quad v_y(i) = \frac{y_{i+1} - y_i}{\triangle t_{i,i+1}} \tag{40}$$

for all $i = 1, \ldots, n.$, where v_x and v_y are temporal velocity rates with regard to the eye's positions. Based on the above result we propose the following feature to quantify the global velocity as

$$h_1 = \sigma(v_x) + \sigma(v_y) \tag{41}$$

Normally this feature has a low value when we are fixing, and high values for saccades.

From a cognitive standpoint, always some amount of information is retained from saccades (Brouwer & Knill, 2007). However, experimentally we can not differentiate an object from another in this period, so when a saccade is detected our system cannot infer the object wanted. Therefore,

we are not performing a grasping movement.

III. Object prediction

After creating a codebook for all objects contained in the user's scene, we extract new features from another video sequence containing all the previously analyzed objects. The key idea is that some features are closer to a specific object contained in the codebook. To increase the probability of classifying an object correctly, several features contained in the same TSW have been extracted, as shown in Fig.11. Therefore, we increase the probability that a feature is correctly classified. Here we used the cosine angle distance to perform the matching between an unknown feature vector and a known feature vector contained in the codebook.

$$h_2 = \max_i \left(class(angle(f_i, \mathbf{F_n})) \right), \tag{42}$$

for all $i = 1, \ldots, p$, where f_i is an unknown feature vector extracted from the head video sequence; p is the number of vectors contained in a TSW, and $\mathbf{F_n}$ are the feature vectors contained in the codebook.

IV. Hand prediction

In the last section we have described a set of features to detect hand intentions based on an HMM framework. Normally the outcome of this process is defined by choosing the maximum class as $class = arg\max_i(p(\mathbf{O}|\lambda_i))$. However, in some situations the maximum posterior probability could be incorrect when the ratio between the outcome of other probabilities is low. For that reason here we use the probability of each class, given by,

$$\begin{aligned} h_3 &= p(\mathbf{O}|\lambda_1) \\ h_4 &= p(\mathbf{O}|\lambda_2) \\ h_5 &= p(\mathbf{O}|\lambda_3) \\ h_6 &= p(\mathbf{O}|\lambda_4) \end{aligned} \tag{43}$$

where $p(\mathbf{O}|\lambda_i)$ is the probability to have detected the i-action in the TSW.

V. HMM for recognition

In the above steps we have defined six h_i feature descriptors. These features have been designed to detect grasping movements using an HMM framework, as shown below. The main reason for combining information about hand intentions, eye positions, and object stability is that for only a time-delay, when the user is performing a grasping motion, eye fixations are high, the object is always the same, and the hand motion is performed toward the object wanted.

In the same way as the last HMM was defined, we define a new feature vector contained in a TSW as

$$\mathbf{o}_1 \equiv \mathbf{o}_{1 \mapsto n} = [h_1, h_2, h_3, h_4, h_5, h_6]^\top, \tag{44}$$

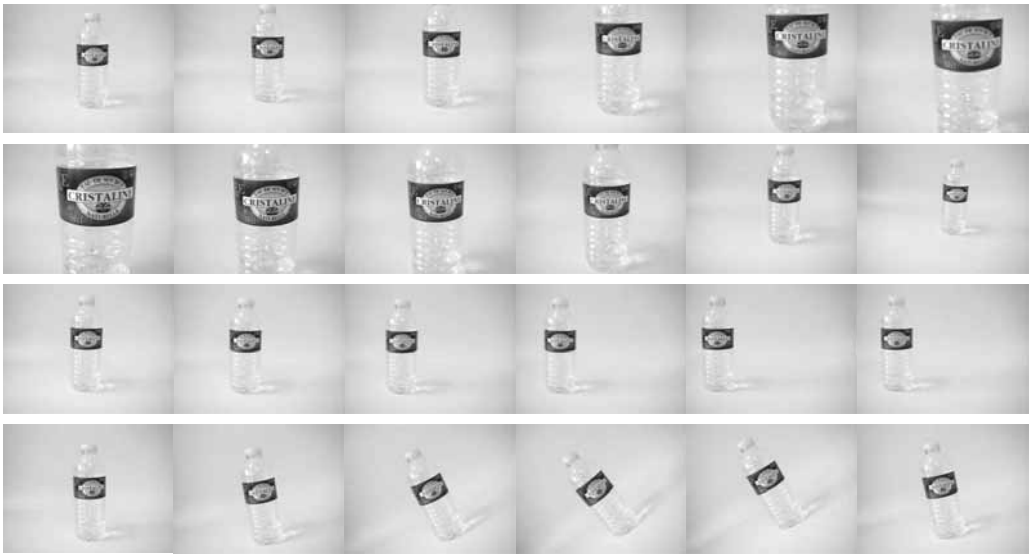


Fig. 12. Real image sequence with one object performing four actions. (a) zoom in, (b) zoom out, (c) translational, and (d) rotary movements

where $\mathbf{o}_1$ is defined between time $[t = 1, \ldots, t = n]$ for the first TSW. Finally the observed symbol sequence is defined as

$$\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T], \tag{45}$$

4 Experimental results

This section presents the results of two experiments carried out with 1) a camera under the wrist and 2) an eye-tracker[4] in parallel with a camera under the wrist based on the proposed framework as shown in Fig.1. In the first experiment (*Exp.1*) we describe the results of an HMM framework in order to make a motion prediction without markers on the objects. In the second experiment (*Exp.2*), we combine the prediction of the hand motion with the user's gaze position captured by an eye-tracker device in order to predict grasping motions and the object wanted.

I. Experiment 1

We defined four movements regardless of the objects involved in the sequence such as: zoom in, zoom out, translational and rotary movements. Since each movement is valid in a sequence of frames, our goal is to detect if each movement has been correctly predicted as the real movement performed by the user. In this experiment we use an HMM system for predicting the user's intentions. In order to build an HMM we performed an action several times using one object in the scene. The goal is that providing more testing sequences, for each class, we can increase the probability of classifying correctly an unknown sequence. In our experiments we used video sequences at 30

4. We employed an ASL Eye-Trac 6 to capture the user's gaze. For further details, refers to http://www.a-s-l.com/ site/ Products/ EYETRAC6Series/ tabid/ 56/ Default.aspx

fps digitalized into 320x200 pixel with 256 gray-level images. An example of the video sequence is shown in Fig.12.

To evaluate the performance, we consider that an action is correctly classified if the motion contained in each TSW has been predicted correctly. Additionally, the system must be independent of the objects contained in the scene. In general, the performance of an HMM varies according to the data employed for testing. Therefore, in our experiments we used the cross-validation method with $k = 10$. Here we are not interested in evaluating the performance of test data used for training the HMM. Our goal is to evaluate the performance in videos with other objects. For this reason we have tested each HMM on five different objects performing each particular action with one object at a time. Namely a cup, bottle, mug, box, deodorant. As we will show later, an HMM can be useful to predict the movement in sequences even with multiple objects.

Our solution uses TSWs with the aim of analyzing temporal motion contained in this period. Using this idea we can reduce the number of frames analyzed by about 67%, because each TSW is interspaced by δ-frames with $\delta = 3$. As shown in Table 4, 7.131 TSWs were analyzed from 21.544 frames of five video sequences. In order to evaluate the performance of the trained HMM, we have classified manually each of the 7.131 TWSs. With respect to the data for training, we have classified 1.466 TSWs from a mug without markers on the surface.

The performance of each HMM using different training sets shows that the Zoom-out movement has, on the average, the best performance, close to 90%, as shown in Fig.13. Also, the lowest performance has been detected in the Zoom-in movement, because it is normally incorrectly classified as a rotary movement. Along the same line, this performance can vary according to the object analyzed. For example, in the case of the bottle, the performance was lower because the SURF algorithm was unable to detect a large number of descriptors. Therefore, with fewer descriptors we cannot build a robust TSW. On the other hand, the analyzed mug had the best performance because of the large number of descriptors used (Fig.14b).

In our experiments we also used the best HMM generated with the cross validation method. For this task we have selected the best performance of each action using the best F-Score and $\mathbf{TPR}$ performance as a criterion. The results shows that we can increase the performance by 2% with the best F-Score and by more than 4% with the best $\mathbf{TPR}$, as shown Fig.14b-c.

TABLE 1
TSW analyzed over each hand motion video

Object	Zoom in-out		Rotation motion		Translational motion	
	Frames	TSW	Frames	TSW	Frames	TSW
Cup	1876	622	1226	405	1051	347
Bottle	1894	628	1211	401	726	240
Mug II	2231	739	1221	404	1016	336
Box	2393	792	1209	400	942	312
Deodorant	2414	799	1200	397	934	309
$\sum$	10808	3580	6067	2007	4669	1544
Mug I (†)	2292	759	1208	400	928	307

†: training data

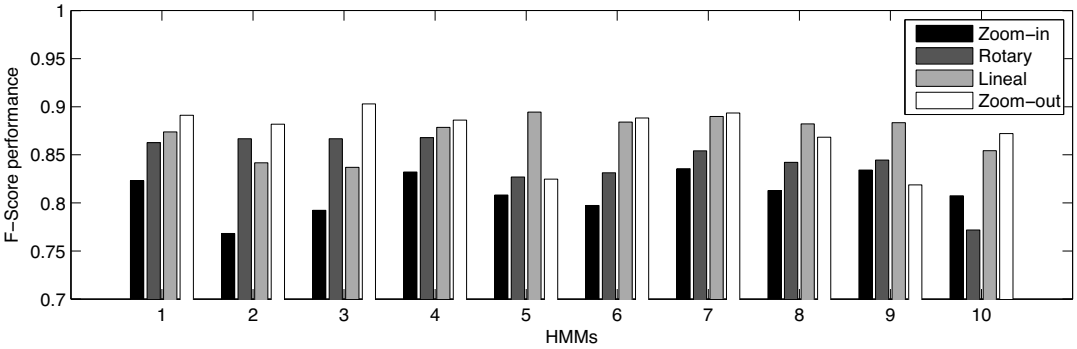


Fig. 13. Average performance of the F-Score over ten HMMs using five objects

I. Experiment 2

In the second experiment we aim at combining the user's hand intentions with the user's gaze position. As stated below, when the user wants to grasp an object there is a time-delay in which he/she can acquire the object in his FOV. In this period there are fixations around the object at one or at multiple points. Only after that, the user can move his hand toward the object wanted. In our experiments we assume that the object wanted is always the same in both views when the user initiates a grasping movement. The main reason is that it is always necessary to carry out fixations before grasping an object, as was described before.

Combining the user's gaze and hand movements allows us to increase the probability of predicting the grasping intention. Generally the hand camera and the head camera are not pointing at the same object all the time. In fact, the grasping movement can be detected only for a small fraction of time. That is why this task is complex. An example of this situation is illustrated in Fig.15. As we can see, when the grasping movement has been initiated, both views share some part of the object. Although in some cases only little information is shared, it is not relevant as long as we can predict the motion with the hand camera and detect the object with the head camera. Additionally, the main idea of using TSWs is that they can allows us to have information about the object even

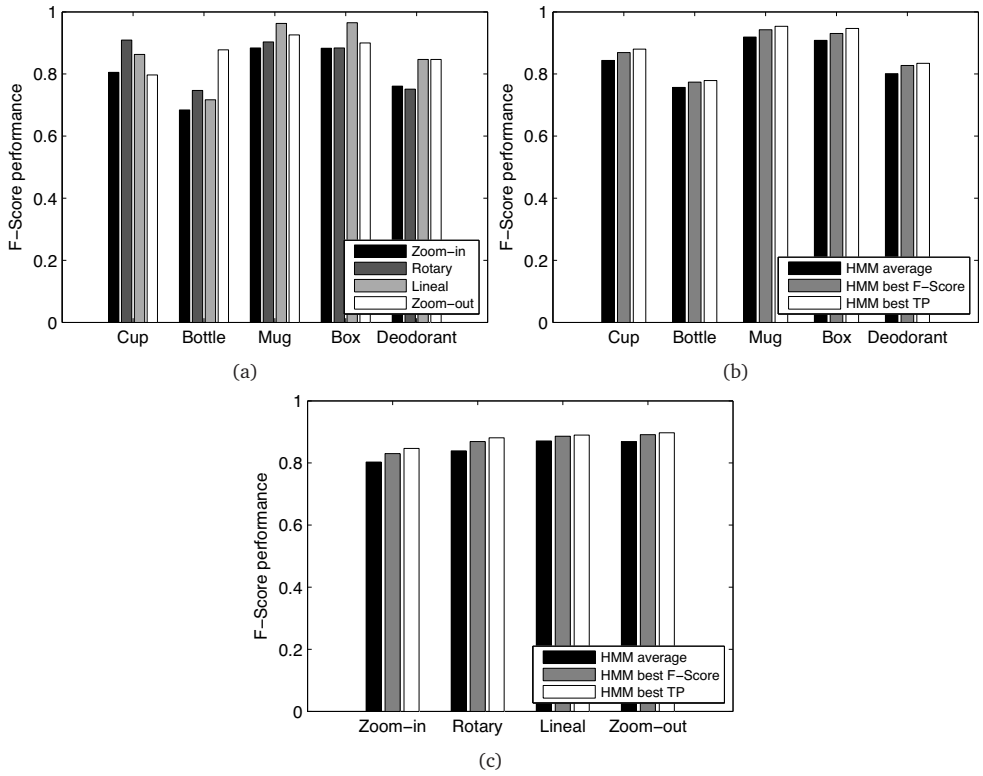


Fig. 14. (a) Average performance for each action using all HMMs; (b) Average performance for all actions on each object using three HMM parameters; (c) Average performance for all objects on each action using three HMM parameters

if it is totally occluded for a small period.

To evaluate the performance of the grasping prediction and the object recognition, we create a synchronized video with 404 TSWs containing four objects on it. Namely a mug, a keys, a box and a personal card. All objects are separated about 15cm from each other. With regard to the setup, two tasks were performed previously: First, an HMM using a video with four objects in multiple positions, and second, a $\mathbf{VQ}$ for each object. Although in this experiment we have multiple objects in the same scene, here we used an HMM trained with only one object. The main idea is that the hand's predictor can be able to predict the motion independently of the objects contained in the scene. In particular, we have employed the best HMM trained in the Exp.1 which is the HMM with higher TP rate (Fig.14b -HMM best TP). Generally, the hand's motion is very dynamical. For this reason it is necessary to detect with highly probability the Zoom-in gesture because a high value implies a reach-to-grasp movement. Also, the saccade detector can detect whether the user is watching or not an object. When this value is high, the saccade is detected. Taking into account the new features required to build the second HMM, the performance shows that it is possible to detect a grasping intention with $\mathbf{TPR} = \mathbf{92.5}\%$ and fixations with $\mathbf{TPR} = \mathbf{83.3}\%$, as shown in

TABLE 2
Performance of the grasp intention

Class	Classified as		Performance	
	Fixation	Grasping	**TPR**	**FPR**
Fixation	270	54	83.3%	1.9 %
Grasping	6	74	92.5%	16.7 %

TABLE 3
Performance of the object recognition

Class	Classified as				Performance	
	Mug	Keys	Box	Card	**TPR**	**FPR**
Mug	119	1	2	1	96.7%	0.7%
Keys	1	128	3	4	94.1%	2.2%
Box	0	2	74	1	96.1%	1.5%
Card	1	3	0	64	94.1%	1.8%
Mean					95.3%	1.6%

Table 4. Nevertheless, there is a big rate of $\mathbf{FPR}$, mainly because the system has been designed to predict a reach-to-grasp movements. With regard to the performance of the object recognition, we obtained, in average, a performance of $\mathbf{TPR} = \mathbf{95.3}\%$ and $\mathbf{FPR} = \mathbf{1.6}\%$, as shown in Table 4. This shows that the codebook constructed by the $\mathbf{VQ}$ method can be an efficient method to detect resilient features.

5 CONCLUSIONS

The main contribution of this work lies into performing a fusion system between the user's gaze and a hand motion estimation. This results can be applied on the project BRAHMA as a method to predict the user's intention. Specially on people with neural degenerative disorders. In these, the control movements are altered causing motion tremor, slow movements, etc. Despite the visual functions on these people have not been altered, the control system can not be able to plan a correct movement without any disruption.

In our experiments we show that it is possible to detect the hand's intentions using only the objects contained in the scene and without special markers on the objects' surface (*Exp.1*). However, the performance of this task varies according to the points of interest detected by the SURF algorithm. Similarly, our method can predict the user's grasping intention and the object placed in the scene by using a blending between two channels of information (*Exp.2*). Even if the object has been occluded, the system is able to identify it because our approach uses a combination of frames called Temporal Slide Windows (TSW). This approach can allow us to increase the temporal features of the same object in multiple frames. Consequently, the paradigm of the frame-by-frame tracking can be effectively replaced by a TSW approach. Although in our experiments the objects analyzed were limited, these results are very promising because we used a limited number of resilient features

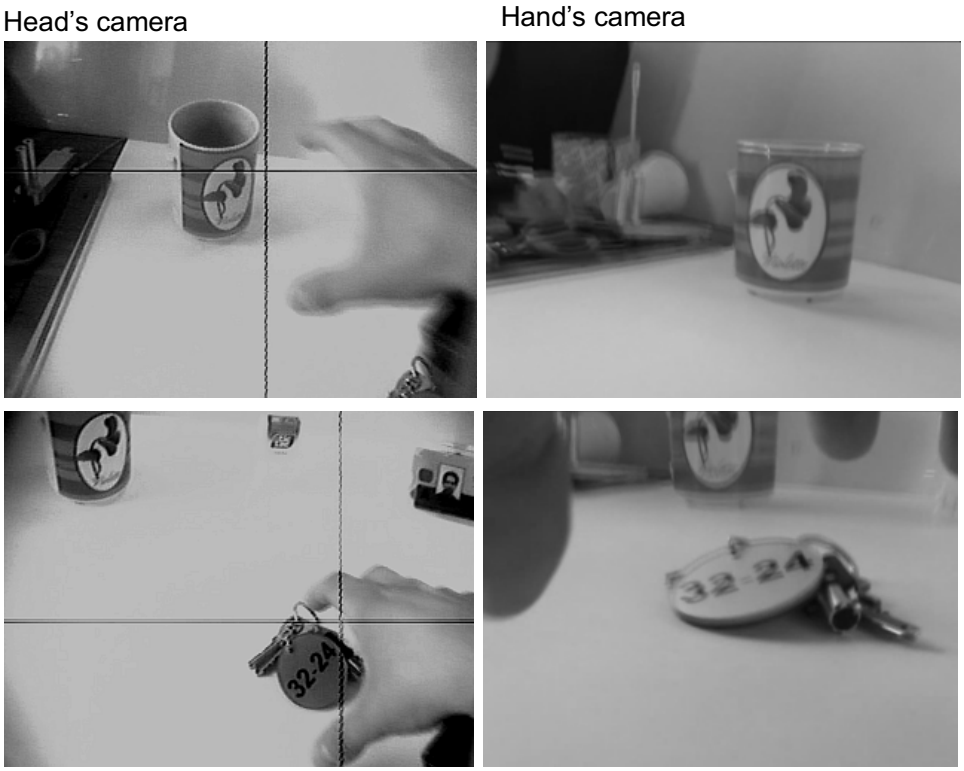


Fig. 15. Real shots captured from head camera and hand camera

and this task was conducted by searching for a match between both views. Our future work is to improve the performance of the grasping intention by reducing the **FPR** value and to evaluate our method on more objects, for example by using a set of multiple objects of the same class.

REFERENCES

Achard, C., Qu, X., Mokhber, A., & Milgram, M. (2007). Action recognition with semi-global characteristics and hidden markov models. In *Conference on advanced concepts for intelligent vision systems (acivs)*.

Aggarwal, J., & Cai, Q. (1997). Human motion analysis: A review. *Computer Vision and Image Understanding*, *73*, 90–102.

Aloimonos, J. (1990). Purposive and qualitative active vision. In *10th international conference on pattern recognition* (Vol. 1, pp. 346–360).

Bajcsy, R. (1988, Aug). Active perception. *Proceedings of the IEEE*, *76*(8), 966–1005.

Barron, J., Fleet, D., & Beauchimen, S. (1994). Performance of optical flow techniques. *International Journal of Computer Vision*, *12*(1), 43–77.

Bay, H., Tuytelaars, T., & Gool, L. (2006, May). Surf: Speeded up robust features. In *Proceedings of the 9th european conference on computer vision.*

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 257–267.

Brouwer, A.-M., & Knill, D. C. (2007, 6). The role of memory in visually guided reaching. *Journal of Vision*, *7*(5), 1–12.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., et al. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on multimodal interfaces* (pp. 205–211). State College, PA, USA: ACM.

Campos, T. de, Mayol, W., & Murray, D. (2006, Oct.). Directing the attention of awearable camera by pointing gestures. In *Proceedings of the 19th brazilian symposium on computer graphics and image processing, 2006. sibgrapi'06.* (pp. 179–186).

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1).

Crawford, J., Medendorp, W., & Marotta, J. (2004). Spatial transformations for eye–hand coordination. *Journal of Neurophysiology*, *92*, 10–19.

Davison, A. J., Mayol, W. W., & Murray, D. W. (2003). Real-time localisation and mapping with wearable active vision. In *Proceedings of the 2nd ieee and acm international symposium on mixed and augmented reality (ismar'03)* (pp. 18–27).

Desmurget, M., Pelisson, D., Rossetti, Y., & Prablanc, C. (1998). Neurosciences and biobehavioral review. *From eye to hand: planning goal-directed movements*, *22*(6), 761–788.

Dockstader, S., & Tekalp, A. (2001). Multiple camera tracking of interacting and occluded human motion. In *Proceedings of the ieee* (Vol. 89, pp. 1441–1455).

Flanagan, J., & Lederman, S. (2001). Neurobiology: Feeling bumps and holes. *Nature*, *412*, 389–391.

Gavrila, D. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, *73*.

Gersho, A., & Gray, R. M. (1992). *Vector quantization and signal compression*. Kluwer Academic Press.

Hayhoe, M., Bensinger, D., & Ballard, D. (1998). Task constraints in visual working memory. *Vision Research*, *38*, 125–137.

Hayhoe, M., Shrivastava, A., Mruczek, R., & Pelz, J. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*, 49–63.

He, Y., & Kundu, A. (1991, 14–17 April). Shape classification using hidden markov model. In *Proc. international conference on acoustics, speech, and signal processing icassp-91* (pp. 2373–2376).

Higham, N. (1996). *Accuracy and stability of numerical algorithms*. SIAM.

Jacob, R., & Karn, K. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises (section commentary). In J. Hyona, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573–605). Elsevier Science.

Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, *79*(4), 368–376.

Karn, K., & Hayhoe, M. (2000). Memory representations guide targeting eye movements in a natural task. *Visual Cognition*, *7*(6), 673-704.

Ke, Y., & Sukthankar, R. (2004, 27 June–2 July). Pca-sift: a more distinctive representation for local image descriptors. In *Proc. ieee computer society conference on computer vision and pattern recognition cvpr 2004* (Vol. 2, pp. II-506–II-513).

Kim, K., Kwak, K., & Ch, S. (2006). Gesture analysis for human-robot interaction. In *Proceedings of the 8th international conference in advanced communication technology, 2006. (icact'06)* (Vol. 3, pp. 1824–1827).

Kurata, T., Sakata, N., Kourogi, M., Kuzuoka, H., & Billinghurst, M. (2004). The advantages and limitations of a wearable active camera/laser in remote collaboration. In *Proceedings of the computer supported cooperative work, cscw'04*. Chicago, IL.

Laptev, I., & Lindeberg, T. (2003, 13–16 Oct.). Space-time interest points. In *Proc. ninth ieee international conference on computer vision* (pp. 432–439).

Lowe, D. G. (1999). Object recognition from local scaleinvariant features. In *International conference on computer vision* (pp. 1150–1157). Corfu, Greece.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91–110.

Mayol, W. W., Tordoff, B., & Murray, D. W. (2000). Towards wearable active vision platforms. In *in ieee sys. man and cybernetics conf* (pp. 1627–1632).

Mikolajczyk, K., & C., S. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, *60*(1), 63–86.

Mrotek, L. A., & Soechting, J. (2007). Target interception: Hand–eye coordination and strategies. *The Journal of Neuroscience*, *27*(27), 7297-7309.

Niyogi, S. A., & Adelson, E. H. (1994, 21–23 June). Analyzing and recognizing walking figures in xyt. In *Proc. cvpr '94. ieee computer society conference on computer vision and pattern recognition* (pp. 469–474).

Perini, E., Soria, S., Prati, A., & Cucchiara, R. (2006). Facemouse: A human-computer interface for tetraplegic people. In *Eccv workshop on hci 2006* (Vol. LNCS 3979, pp. 99–108).

Rabiner, L. R. (1989, Feb.). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Roby-Brami, A., Bennisa, N., Mokhtaria, M., & Baraduca, P. (2000, March). Hand orientation for grasping depends on the direction of the reaching movement. *Brain Research*, *869*, 121–129.

Shafique, K., & Shah, M. (2003, 13–16 Oct.). A non-iterative greedy algorithm for multi-frame point correspondence. In *Proc. ninth ieee international conference on computer vision* (pp. 110–115).

Shechtman, E., & Irani, M. (2005, 20–25 June). Space-time behavior based correlation. In *Proc. ieee computer society conference on computer vision and pattern recognition cvpr 2005* (Vol. 1,

pp. 405–412).

Sidibe, D., Montesinos, P., & Janaqi, S. (2007, March). Fast and robust image matching using contextual information and relaxation. In *2nd international conference on computer vision theory and applications, visapp*. Barcelona, Spain.

Sivic, J., & Zisserman, A. (2009, April). Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(4), 591–606.

Starner, T., & Pentland, A. (1995, 21–23 Nov.). Real-time american sign language recognition from video using hidden markov models. In *Proc. international symposium on computer vision* (pp. 265–270).

Sugi, M., Tamura, Y., Ota, J., Arai, T., Takamasu, K., & Suzuki, H. (2006). Implementation of human supporting production system "attentiveworkbench". In *Sice-icase, 2006. international joint conference* (pp. 1270–1273).

Tamura, Y., Sugi, M., Ota, J., & Arai, T. (2004). Deskwork support system based on the estimation of human intentions. In *Proceedings of the 13th ieee international workshop on robot and human interactive communication*.

Tamura, Y., Sugi, M., Ota, J., & Arai, T. (2006, Sept.). Prediction of target object based on human hand movement for handing-over between human and self-moving trays. In *Proc. 15th ieee international symposium on robot and human interactive communication roman 2006* (pp. 189–194).

Tamura, Y., Sugi, M., Ota, J., & Arai, T. (2007, Nov. 2). Estimation of user's intention inherent in the movements of hand and eyes for the deskwork support system. In *Ieee/rsj int'l conf. on intelligent robots and systems (iros 2007)* (pp. 3709–3714). San Diego, CA, USA.

Veenman, C. J., Reinders, M. J. T., & Backer, E. (2001, Jan.). Resolving motion correspondence for densely moving points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(1), 54–72.

Yamato, J., Ohya, J., & Ishii, K. (1992, 15–18 June). Recognizing human action in time-sequential images using hidden markov model. In *Proc. cvpr '92. ieee computer society conference on computer vision and pattern recognition* (pp. 379–385).

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *ACM Comput. Surv.*, *38*(4), 13.

Chapter 6

Conclusions

6. CONCLUSION AND FUTURE RESEARCH

This section presents the general conclusions of all the papers that make up this thesis. The main characteristics of the implemented methods, the relation between the specific objectives and the work done, and finally the work to be done in the future a are detailed below.

6.1. General remarks

This thesis, beside exploring new techniques and algorithms oriented to computer vision, have shown that the same principle can be exploited and developed in completely different areas. Although superficially the applications have different objectives, they all go through the same steps to process visual information. These steps are: (1) determining point-to-point correspondence in multiple views; (2) determining the geometric and/or dynamic relation of the correspondences; and (3) inferring new relations of the movement on unknown points using the model(s) from the previous step. The main variations correspond to the way in which we determine the correspondence; the calculation of the dynamics and the geometry between corresponding points; and finally how we use the inference of the motion to characterize the movement of the points.

The papers presented in the body of this thesis attempt to solve these problems by designing new methodologies and/or improving present methods. This process has allowed the evolution a substantial improvements of the initial designs of each topic, allowing each paper to make use of the results of previous research. As pointed out in the introduction, the evolution of the algorithms is transverse throughout the thesis. An example of this process has been the evolution of the uncalibrated AMVI methodology, which allows the evaluation of the quality of an object in multiple views by tracking potential defects. Initially the metod was designed to inspect automobile wheel rims. Then we submitted a functional prototype that allows inspecting bottle necks using the same methodology. This prototype has the peculiarity of implementing a method of internal illumination that projects the defects clearly on the CCD receptor. On the other hand, making use of the properties of the multiple views model, we designed a method for the detection of human intentions. In this case it was not necessary to determine the tracking of points, but rather the objective was to analyze the performance of corresponding points.

On the other hand, the use of invariant correspondences of the human-computer interaction system was a key idea for determining the correspondence of hypothetical defects in the wine-bottle neck inspection system. This relation between one paper and another is again present in the point-to-point correspondence algorithm, which uses the estimation of multiple geometric models influenced by the uncalibrated AMVI method, together with a system of invariant correspondences present in the human intention recognition algorithm. The combination of these two ideas led to the development of a new geometric method of correspondence that is independent of the objects contained in the scene. One of the greatest advantages is that it allows the determination of correspondences in

spite of the geometric and photometric transformations and possible existing occlusions since it is based on the geometry of the scene.

The main conclusions in relation to the transverse topics present in the papers are detailed below.

- The algorithms for invariant feature extraction have been shown to be an effective tool for determining multiple points in correspondence in spite of the transformations existing in the sequence of images. These algorithms can be used either to extract invariant features and evaluate the similarity of potential defects or to determine the point-to-point correspondence with the purpose of constructing a set of correspondences. Taking advantage of this potential, we have extended this topic to geometric correspondence, which can be used indistinctly in different types of images. However, these algorithms fail when the signal-to-noise ratio is low, as normally present in X-ray images, which implies a lower dispersion of the points on the image and a degradation in the estimation of the geometric and dynamic model of the movement. In these cases it is better to use algorithms that determine the correspondence as a function of the structures that compose the images. Invariant algorithms such as those of the moments of Hu (1962), or the moments of Flusser and Suk (1993) have been shown to be efficient methods for finding those correspondences. Later it is necessary to make an analysis of the characteristics belonging to the structures, mainly at their edges.

- Normally, the algorithms for estimating geometric models are based on the minimization of the re-projection error of a set of hypothetical correspondences. Those methods are efficient because they search a given number of times as a function of the noise present in the set. However, when the correspondence error is close to the minimum it is possible to use other hypotheses, increasing the general performance of the system. This is precisely the main contribution of the point-to-point geometric correspondence algorithm. On the other hand, when the movement is too accelerated, the use of geometric or optical flow models does not yield optimum results because they are not designed to predict the position of the object when considerable acceleration takes place. This happens normally with accelerated movements such as the human grasping movement, which is determined by a high acceleration at the beginning of the action and a deceleration at the end. In this case we showed that the dynamic features such as speed, angular acceleration, angular rotation, and parallel movements allow the determination and characterization of human movements with high precision.

- The algorithms that estimate the geometric re-projection, such as the fundamental matrix and trifocal tensors, allow the determination of a line or a re-projected point, respectively. We showed that the use of multiple fundamental matrices generated from multiple epipoles is an effective tool to estimate correspondence in two views. This result

is quite novel because, as far as we know, there is no method that carries out a similar process of correspondence in two views with bifocal geometry. To carry out this process it is necessary to determine the error associated with each partial solution, because not all the solutions have the same error. We have extended this same process to geometric analysis in three views. On the other hand, we showed that the set of dynamic features, such as the grasping movement, can be used as observations of a markovian system. It is important to point out that not all kinds of movements have the same performance. For example, the grasping movement can be composed of an approach together with a slight angular rotation. To decrease the model's error we limit our analysis to simple and differentiated types of movements.

6.2. Specific achievements

To clarify the results of each topic in relation to the specific objectives stated at the beginning of the thesis, the achievements based on each developed area are detailed below. In the topic of automatic visual inspection we dealt in detail with the multiple aspects that cover the AMVI methodology considering the analysis of invariant features, correspondence algorithms, and the construction of a prototype for inspecting bottles. In the topic of point-to-point correspondence we dealt with the selection of partial solutions algorithms as a means to determine the error associated with each intermediate solution. Finally, in the topic of prediction of human intentions we again used invariant correspondence algorithms to construct the vectorial model of the grasping movement. The main achievements are detailed below.

6.2.1. Automatic visual inspection

- In relation to the segmentation algorithms, we used the *valley-emphasis* (Hui-Fuang, 2006) algorithm to perform a robust binarization, the segmentation of differences in radiographic images algorithm (Carrasco & Mery, 2004), and the CLP profile search algorithm (Mery, 2003a).
- For the extraction of invariant features we used the following methods: Hu (1962), SFSK (Flusser & Suk, 1993), FSKS (Flusser et al., 1996), GPSO-PSO-GPD (Mindru et al., 2004), SURF (Bay et al., 2008), SIFT (Lowe, 2004), PHOG (Bosch et al., 2007), and non-invariant characteristics such as CLP (Mery, 2003a), gray average, average of the derivative, contrast deviation, difference between maximum and minimum intensity level (Gonzalez & Woods, 2008), and co-occurrence matrix (Haralick, Shanmugam, & Dinstein, 1973).
- For the analysis of characteristics we designed an intermediate classification algorithm based on the combination of the Branch & Bound (Somol et al., 2004), and Take-L, plus-R (Duda et al., 2001; Kudo & Sklansky, 2000) algorithms. The separation was made by

means of Fisher's discriminant (Stearns, 1976). Finally, training of the geometric system was performed with the RANSAC algorithm (Fischler & Bolles, 1981).

- In relation to the design with external markers, we built a prototype with an internally backlighted object together with external markers in a given configuration in order to facilitate the analysis of correspondences. The prototype is novel because it allows the reflection with high precision of the real flaws of the object that is being inspected.

- In relation to the design of the inspection prototype, we designed an algorithm to extract and analyze a set of potential defects by means of an uncalibrated model of tracking in multiple views. The simultaneous rotation of the bottle and the object with markers was the main achievement of the prototype, since both turn simultaneously on their vertical axis.

- In relation to the correspondence of regions algorithms, we designed a new algorithm that improves the NNDR method (Mikolajczyk & Schmid, 2004) by 10% through a bidirectional system called*bidirectional*-NNDR (bNNDR).

- In relation to the algorithms for the reduction of potential defects, we presented the Intermediate Classifier Block (ICB) algorithm, which allows a reduction of the number of potential defects in sequence by means of the analysis of properties.

6.2.2. Point-to-point correspondence

- In relation to the base correspondence algorithm, we used invariant SURF algorithm (Bay et al., 2008) combined with the NNDR algorithm (Mikolajczyk & Schmid, 2004).

- In relation to the selection of partial solutions algorithms, we used the MLESAC algorithm (Torr & Zisserman, 2000) in two or three views with the purpose of determining the error associated with each geometric solution.

- In relation to the determination of the best solution, we developed the BIGC and TRIGC algorithms which apply the philosophy of using all the solutions and weighting them based on their degree of error.

6.2.3. Prediction of user's intentions

- In relation to the state of the art in the hand-eye coordination tasks we presented an extensive review of the literature on this topic and the factors that influence that coordination, which provided the foundation of the methodology used in our predictive model.

- In relation to the hand-eye coordination algorithm, we developed a predictive algorithm that can predict the human grasping movement in a closed set of possible movements.

- In relation to the identification of the object to be grasped, we developed an object identification algorithm based on the construction of codebooks based on the system proposed by Gersho and Gray (1992).

- In relation to the system of merging information, we determined a set of movement characteristics based on visual flow when performing the grasping gesture together with the movement of the eyes. In this way it was possible to determine a probable state by means of a predictive model HMM (Rabiner, 1989).

6.3. Future Research Topics

Due to the extent of the algorithms and techniques used in this thesis, the possibility of expanding and improving the proposed algorithms is open. Below we detail some methods that remain to be evaluated as extensions of the methods already developed.

- Evaluate the performance of the intermediate classifier block (ICB) algorithm with invariant features (paper #2). The objective of the ICB was to search in space those features that are separable in such a way that only those regions of space that contain potential defects were analyzed again in more views. Performance in regions with invariant features remain to be determined.

- Use the geometric correspondence algorithm to track defects in the sequence of X-ray images. The algorithm presented in Chapter 3 was performed by the combined analysis of the features extracted from each defect and a simple geometric analysis. However, the geometric correspondence algorithm of Chapter 4 has proved to be efficient in images with low signal-to-noise ratio, and for that reason an evaluation of these types of images is necessary.

- Evaluate the correspondence of the tracking of invariant features versus the algorithm of geometric correspondence. In previous analysis we determined correspondence as a function of geometric methods combined with invariant methods. Analysis of each one independently remains to be determined.

- Evaluate the bNNDR algorithm as a method of feature correspondence in multiple views. As we showed in Chapter 4, the bNNDR method performs better by an average of 10% than the NNDR algorithm, however this was used as a method of correspondence in potential defects. We believe that its use can improve the performance of the geometric point-to-point correspondence system or the vectorial model of human intention system.

- Evaluate the human detection algorithm as a means to determine the quality of an object. The solution implemented by the dynamic model considers the correspondence for any point that is contained in the scene, while the AMVI model uses only known correspondences in some positions. Therefore, extending the dynamic model to the AMVI model is

a logical process. The greatest advantage would be to carry out the inspection on objects without knowing their structure, thereby ensuring the independence of the uncalibrated AMVI in relation to the object.

REFERENCES

Achard, C., Qu, X., Mokhber, A., & Milgram, M. (2007). Action recognition with semi-global characteristics and hidden markov models. In *Conference on advanced concepts for intelligent vision systems (acivs)*. Springer.

Adams, J. (1981). Do cognitive factors in motor performance become nonfunctional with practice? *Journal of Motor Behaviour*, *13*, 262-273.

Aggarwal, J., & Cai, Q. (1997). Human motion analysis: A review. *Computer Vision and Image Understanding*, *73*, 90–102.

Aloimonos, J. (1990). Purposive and qualitative active vision. In *10th international conference on pattern recognition* (Vol. 1, pp. 346–360). IAPR press.

Arulampalam, M., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filter for online nonlinear/non-gaussian bayesian tracking. *IEEE Trans. on Signal Processing*, *50*(2), 174–188.

Bajcsy, R. (1988, Aug). Active perception. *Proceedings of the IEEE*, *76*(8), 966–1005.

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66-80.

Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1994, Feb). Performance of optical flow techniques. *International Journal of Computer Vision*, *12*(1), 43–77.

Bartels, R., Beatty, J., & Barsky, B. (1998). *Bezier curves. an introduction to splines for use in computer graphics and geometric modelling* (Vol. Ch.10). San Francisco, CA.: Morgan Kaufmann.

Bay, H., Ess, A., Tuytelaars, T., & Gool, L. (2008). Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)*, *110*(3), 346–359.

Beardsley, P., Murray, D., & Zisserman, A. (1992). Camera calibration using multiple images. In G. Sandini (Ed.), *Proceedings european conference on computer vision (eccv-92)* (pp. 312–320). Springer.

Bhat, P., Zheng, K., Snavely, N., Agarwala, A., Agrawala, M., Cohen, M., et al. (2006). Piecewise image registration in the presence of multiple large motions. In *Ieee conference on computer vision and pattern recognition* (Vol. 2, pp. 2491–2497). IEEE Computer Society Press.

Bobick, A., & Davis, J. (1996, Dec). Real-time recognition of activity using temporal templates. In *Proc. 3rd ieee workshop on applications of computer vision wacv '96* (pp. 39–42). IEEE Computer Society.

Bobick, A., & Davis, J. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 257–267.

Bosch, A., Zisserman, A., & X., M. (2007, Jul). Representing shape with a spatial pyramid kernel. In ACM (Ed.), *Proceedings of the 6th acm international conference on image and video retrieval (civr)* (pp. 401 – 408). Amsterdam, The Netherlands: ACM.

Brouwer, A.-M., & Knill, D. C. (2007). The role of memory in visually guided reaching. *Journal of Vision*, *7*(5), 1–12.

Buneo, C., Jarvis, M., Batista, A., & Andersen, R. (2002). Direct visuomotor transformations for reaching. *Nature, 416*(6881).

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., et al. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on multimodal interfaces* (pp. 205–211). State College, PA, USA: ACM.

Campos, T. de, Mayol, W., & Murray, D. (2006, Oct). Directing the attention of awearable camera by pointing gestures. In *Proceedings of the 19th brazilian symposium on computer graphics and image processing, 2006. sibgrapi'06.* (pp. 179–186). IEEE Computer Society Press.

Carrasco, M., & Mery, D. (2004). Segmentation of welding discontinuities using a robust algorithm. *Materials Evaluation*, *62*(11), 1142–1147.

Carrasco, M., & Mery, D. (2006). Automated visual inspection using trifocal analysis in an uncalibrated sequence of images. *Materials Evaluation*, *64*(9), 900–906.

Carrasco, M., Pizarro, L., & Mery, D. (2008). Image acquisition and automated inspection of wine bottlenecks by tracking in multiple views. In *Proc. of the 8th int. conf. on signal processing, computational geometry and artificial vision (iscgav'08)* (pp. 82–89). Rhodes Island, Greece: WSEAS Press.

Caspi, Y., & Irani. (2000). A step towards sequence-to-sequence alignment. In *Ieee conference on computer vision and pattern recognition (cvpr)*. (pp. 682–689). Hilton Head Island, South Carolina: IEEE Computer Society Press.

Caspi, Y., Simakov, D., & Irani, M. (2006). Feature-based sequence-to-sequence matching. *International Journal of Computer Vision*, *68*(1), 53–64.

Chen, Z., Wu, C., Shen, P., Liu, Y., & Quan, L. (2000). A robust algorithhm to estimate the fundamental matrix. *Pattern Recogn Letters*, *21*, 851–861.

Chin, R., & C.A., H. (1982). Automated visual inspection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *4*(6), 557-573.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1).

Crawford, J., Medendorp, W., & Marotta, J. (2004). Spatial transformations for eye–hand coordination. *Journal of Neurophysiology*, *92*, 10–19.

Davison, A. J., Mayol, W. W., & Murray, D. W. (2003). Real-time localisation and mapping with wearable active vision. In *Proceedings of the 2nd ieee and acm international symposium on mixed and augmented reality (ismar'03)* (pp. 18–27). IEEE Computer Society Press.

Desmurget, M., Pelisson, D., Rossetti, Y., & Prablanc, C. (1998). Neurosciences and biobehavioral review. *From eye to hand: planning goal-directed movements*, *22*(6), 761–788.

Dockstader, S., & Tekalp, A. (2001). Multiple camera tracking of interacting and occluded human motion. *Proceedings of the IEEE, 89*(10), 1441–1455.

Dodge, R., & Cline, T. (1901). The angle velocity of eye movements. *Psychological Review*, *8*, 145–157.

Donkelaar, P., Lee, J.-H., & Drew, A. (2000). Transcranial magnetic stimulation disrupts eye-hand interactions in the posterior parietal cortex. *The Journal of Neurophysiology*, *84*(3), 1677–1680.

Drury, C. (1992). Inspection performance. In (pp. 2282–2314). New York, NY, USA: John Wiley and Sons.

Drury, C., Saran, M., & Schultz, J. (2004, Jan). *Effect of fatigue, vigilanc, environment on inspectors performing fluorescent penetrant and/or magnetic particle inspection* (Interim Report). Federal Aviation Administration William J. Hughes Technical Center: University at Buffalo.

Duda, R. O., Hart, P. E., , & Stork, D. G. (2001). *Pattern classification* (Second Ed. ed.). New York, NY, USA: John Wiley and Sons.

Engel, K. C., Flanders, M., & Soechting, J. F. (2002, Jul). Oculocentric frames of reference for limb movement. *Archives Italiennes de Biologie*, *140*(3), 211–219.

Fischler, M., & Bolles, R. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Fitts, P. M., Jones, R. E., & Milton, J. L. (1950). Eye movements of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review*, *9*(2), 24–29.

Fitzgibbon, A. (2003, Dec). Robust registration of 2d and 3d point sets. *Image and Vision Computing*, *21*(13–14), 1145–1153.

Flanagan, J., & Lederman, S. (2001). Neurobiology: Feeling bumps and holes. *Nature*, *412*, 389–391.

Flusser, J., & Suk, T. (1993). Pattern recognition by affine moment invariants. *Pattern Recogn*, *26*(1), 167-ñ174.

Flusser, J., Suk, T., & Saic, S. (1996). Recognition of images degraded by linear motion blur without restoration. *Computing. Supplement*, *11*, 37–51.

Gersho, A., & Gray, R. M. (1992). *Vector quantization and signal compression*. Kluwer Academic Press.

Gonzalez, R., & Woods, R. (2008). *Digital image processing* (3rd ed.). Prentice Hall.

Gumustekin, S. (2004, April). A visual inspection system using a single camera and mirrors. In *Signal processing and communications applications conference, 2004. proceedings of the ieee 12th* (p. 257-260). IEEE Computer Society.

Gupta, A., & O'Malley, M. (2006). Design of a haptic arm exoskeleton for training and rehabilitation. *IEEE/ASME Transactions onMechatronics*, *11*(3), 280–289.

Haralick, R., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. on Systems, Man, and Cybernetics*, *SMC-3*(6), 610-621.

Hartley, R., & Zisserman, A. (2000). *Multiple view geometry in computer vision*. Cambridge, UK: Cambridge University Press.

Hayhoe, M., Bensinger, D., & Ballard, D. (1998). Task constraints in visual working memory. *Vision Research*, *38*, 125–137.

Hayhoe, M., Shrivastava, A., Mruczek, R., & Pelz, J. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*, 49–63.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Trans Info Theory*, *IT*(8), 179–187.

Hui-Fuang, N. (2006, Oct). Automatic thresholding for defect detection. *Pattern Recognition Letters*, *27*(14), 1644-1649.

Humayun, M. S., Juan Jr., E. de, Weiland, J. D., Dagnelie, G., Katona, S., Greenberg, R., et al. (1999). Pattern electrical stimulation of the human retina. *Vision Research*, *39*, 2569–2576.

Jacob, R., & Karn, K. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573–605). Elsevier Science.

Jacob, R., Raina, S., Regunath, S., Subramanian, R., & Gramopadhye, A. (2004). Improving inspector's performance and reducing errors - general aviation inspection training systems (gaits). In *In proceedings of the human factors and ergonomics society annual meeting proceedings.* Human Factors and Ergonomics Society.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000, Jan). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37.

Jarrasse, N., Paik, J., Pasqui, V., & Morel, G. (2008, May). How can human motion prediction increase transparency? In *Robotics and automation, 2008. icra 2008. ieee international conference on robotics and automation, 2008. icra* (pp. 2134–2139). Pasadena, CA, USA: IEEE Computer Society Press.

Jarvis, J. F. (1980). Visual inspection automation. *Computer*, *13*(5), 32–38.

Jessell, T., Schwartz, J., & Kandel, E. (2000). *Principles of neural science* (4th ed.). McGraw-Hill Medical.

Johansson, R., Westling, G., Bäckström, A., & Flanagan, J. (2001). Eye-hand coordination in object manipulation. *The Journal of Neuroscience*, *21*(17), 6917-6932.

Kadir, T., Zisserman, A., & Brady, M. (2004, May). An affine invariant salient region detector. *Lecture Notes in Computer Science*, *1*(3021), 228–241.

Kiguchi, K., & Fukuda, T. (2004). A 3dof exoskeleton for upper-limb motion assist-consideration of the effect of bi-articular muscles. In IEEE (Ed.), *Ieee international conference on robotics and automation icra* (Vol. 3, pp. 2121–2429). IEEE Computer Society Press.

Kim, K., Kwak, K., & Ch, S. (2006). Gesture analysis for human-robot interaction. In *Proceedings of the 8th international conference in advanced communication technology, 2006. (icact'06)* (Vol. 3, pp. 1824–1827). IEEE Computer Society Press.

Kita, Y., Highnam, R., & Brady, M. (2001). Correspondence between different view breast X-rays using curved epipolar lines. *Computer, Vision and Understanding*, *83*(1), 38-56.

Krebs, H., Hogan, N., Aisen, M., & Volpe, B. (1998). Robot-aided neurorehabilitation. *IEEE Transactions on rehabilitation engineering*, *6*.

Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *The Journal of the Pattern Recognition Society*, *33*, 25–41.

Kumar, A. (2008, Jan). Computer-vision-based fabric defect detection: A survey. *IEEE Transactions on Industrial Electronics*, *55*(1), 348–363.

Kurata, T., Sakata, N., Kourogi, M., Kuzuoka, H., & Billinghurst, M. (2004). The advantages and limitations of a wearable active camera/laser in remote collaboration. In *Proceedings of the computer supported cooperative work, cscw'04*. Chicago, IL: ACM.

Land, M., & Fernald, R. (1992). The evolution of eyes. *Annual Review of Neuroscience*, *15*, 1–29.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*, 1311–1328.

LeBeau, C. (1991, March). Machine vision platform requirements for successful implementation and support in the semiconductor assembly manufacturing environment. In B. G. Batchelor & F. Waltz (Eds.), *Machine vision systems integration in industry* (Vol. 1386, pp. 228–231). SPIE press.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int J Comput Vis*, *60*(2), 91-110.

Lucas, B., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *In Proceedings of Image Understanding Workshop*, 674–679.

Luck, S., Girelli, M., McDermott, M., & Ford, M. (1997). Bridging the gap between monkey neurophysiology and human perception: An ambiguity resolution theory of visual selective attention. *Cognitive Psychology*, *33*, 64–87.

Mackworth, N. H., & Thomas, E. L. (1962). Head-mounted eye-marker camera. *Journal of the Optical Society of America*, *52*.

Malamas, E., Petrakis, E. G., & Zervakis, M. (2003). A survey on industrial vision systems, applications and tools. *Image and Vision Computing*, *21*(2), 171–188.

Martin, J. (2005). Basic stamp syntax and reference manual. *Parallax USA*. (accessed in 2009, www.parallax.com/dl/docs/prod/stamps/web-BSM-v2.2.pdf)

Matas, J., Chum, O., Urban, M., & Pajdla., T. (2002). Robust wide baseline stereo from maximally stable extremal regions. In *Procedings of british machine vision conference (bmvc)* (pp. 384–393). BMVA.

Mayol, W. W., Tordoff, B., & Murray, D. W. (2000). Towards wearable active vision platforms. In *Ieee sys. man and cybernetics conference* (pp. 1627–1632). IEEE Computer Society.

McConkie, G., & Currie, C. (1996). Visual stability across saccades while viewing complex pictures. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 563–581.

Mery, D. (2003a). Crossing line profile: a new approach to detecting defects in aluminium castings. *LNCS*, *2749*, 725–732.

Mery, D. (2003b). Exploiting multiple view geometry in x-ray testing: Part I, theory. *Materials Evaluation*, *61*(11), 1226–1233.

Mery, D., & Carrasco, M. (2005). Automated multiple view inspection based on uncalibrated image sequence. *LNCS*, *3540*, 1238–1247.

Mery, D., & Carrasco, M. (2006). Advances on automated multiple view inspection. *LNCS*, *4319*, 513–522.

Mery, D., & Filbert, D. (2002). Automated flaw detection in aluminum castings based on the tracking of potential defects in a radioscopic image sequence. *IEEE Trans. Robotics and Automation*,

18(6), 890-901.

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, *60*(1), 63–86.

Mikolajczyk, K., & Schmid, C. (2005, Oct). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(10), 1615–1630.

Mindru, F., Tuytelaars, T., Van Gool, L., & Moons, T. (2004). Moment invariants for recognition under changing viewpoint and illumination. *Comput Vis Image Underst*, *94*(1-3), 3–27.

Mital, A., Govindaraju, M., & Subramani, B. (1998). A comparison between manual and hybrid methods in parts inspections. *Integrated Manufacturing Systems*, *9*(6), 344–349.

Mitchell, T. (1997). *Machine learning*. Boston: McGraw-Hill.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. *IEEE Trans. On Neuronal Networks*, *13*, 3–14.

Moreels, P., & Perona, P. (2007). Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, *73*(3), 263–284.

Mrotek, L. A., & Soechting, J. (2007). Target interception: Hand–eye coordination and strategies. *The Journal of Neuroscience*, *27*(27), 7297-7309.

Newman, T. S., & Jain, A. K. (1995). A survey of automated visual inspection. *Computer Vision and Image Understanding*, *61*(2), 231–262.

Nirmalan, P. K., Tielsch, J. M., Katz, J., Thulasiraj, R. D., Krishnadas, R., Ramakrishnan, R., et al. (2005). Relationship between vision impairment and eye disease to vision-specific quality of life and function in rural india: The aravind comprehensive eye survey. *Investigative Ophthalmology and Visual Science*, *46*, 2308–2312.

Perini, E., Soria, S., Prati, A., & Cucchiara, R. (2006). Facemouse: A human-computer interface for tetraplegic people. In *Eccv workshop on hci 2006* (Vol. LNCS 3979, pp. 99–108). Springer.

Perry, J., & Rosen, S., J. andBurns. (2007). Upper-limb powered exoskeleton design upper-limb powered exoskeleton design upper-limb powered exoskeleton design. *IEEE/ASME Transactions on Mechatronics*, *12*(4), 408–417.

Pizarro, L., Mery, D., Delpiano, R., & Carrasco, M. (2008). Robust automated multiple view inspection. *Pattern Analysis and Applications*, *11*(1), 21–32.

Polana, R., & Nelson, R. (1994, Nov). Low level recognition of human motion (or how to get your man without finding his body parts). In *Proc. ieee workshop on motion of non-rigid and articulated objects* (pp. 77–82). IEEE Computer Society.

Rabiner, L. R. (1989, Feb). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rizzolatti, G., Fogassi, L., & Gallese, V. (1997). Parietal cortex: from sight to action. *Current Opinion in Neurobiology*, *7*, 562–567.

Romano, R. (2002). *Projective minimal analysis of camera geometry*. Phd. thesis, M.I.T., USA.

Rosenfeld, A. (1969). *Picture processing by computer*. New York: Academic Press.

Sabra, A. (1989). *The optics of ibn al-haytham*. Univeristy of London: The Warburg Institute.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, *47*(1–3), 7–42.

Shafique, K., & Shah, M. (2003, Oct). A non-iterative greedy algorithm for multi-frame point correspondence. In *Proc. ninth ieee international conference on computer vision* (pp. 110–115). IEEE Computer Society Press.

Shechtman, E., & Irani, M. (2005, Jun). Space-time behavior based correlation. In *Proc. ieee computer society conference on computer vision and pattern recognition cvpr 2005* (Vol. 1, pp. 405–412). IEEE Computer Society.

Somol, P., Pudil, P., & Kittler, J. (2004, Jul). Fast branch & bound algorithms for optimal feature selection. *IEEE Transactions on Pattern Analisis and Machine Intelligence*, *26*, 900–912.

Spencer, F. (1996, Sep). *Visual inspection research project report on benchmark inspections* (Technical Report No. DOT/FAA/AR-96/65). Office of Aviation Research Washington, D.C. 20591: U.S. Department of Transportation, Federal Aviation Administration,Washington, DC.

Spicer, P., Bohl, K., Abramovich, G., & Barhak, J. (2006, Feb). Robust calibration of a reconfigurable camera array for machine vision inspection (RAMVI): Using rule-based colour recognition. In *Proc. of the 1st international conference on computer vision theory and applications ultrasonics symposium (visapp)* (pp. 131–138). Setúbal, Portugal.

Starner, T., & Pentland, A. (1995, Nov). Real-time american sign language recognition from video using hidden markov models. In *Proc. international symposium on computer vision* (pp. 265–270). IEEE Computer Society.

Starr, M., & Rayner, K. (2001). Eye movements during reading: some current controversies. *Trends in Cognitive Sciences*, *5*(4), 156–163.

Stearns, S. (1976). On selecting features for patterns classifiers. In *Iapr international conference on pattern recognition* (pp. 71–75). IAPR press.

Sternberg, R. (2003). *Cognitve psychology* (3rd ed.). Thomson Wadsworth.

Sugar, T., He, J., Koeneman, E., Koeneman, J., Herman, R., Huang, H., et al. (2007). Design and control of rupert: A device for robotic upper extremity repetitive therapy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *15*(3), 336–346.

Tbakhi, A., & Amr, S. (2007). Ibn al-haytham : Father of modern optics. *Annuals of Saudi Medicine*, *27*(6), 464–467.

Torr, P., & Zisserman, A. (2000). Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, *78*, 138–156.

Tuytelaars, T., & Mikolajczyk, K. (2007). Local invariant feature detectors: A survey. *Computer Graphics and Vision*, *3*(3), 177-280.

Veenman, C. J., Reinders, M. J. T., & Backer, E. (2001, Jan). Resolving motion correspondence for densely moving points. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(1), 54–72.

Vidal, R., Ma, Y., Soatto, S., & Sastry, S. (2006, Jun). Two-view multibody structure from motion. *International Journal of Computer Vision*, *68*(1), 7–25.

Wu, S.-Y., Hennis, A., Nemesure, B., & Leske, M. C. (2008). Impact of glaucoma, lens opacities, and cataract surgery on visual functioning and related quality of life: The barbados eye studies. *Investigative Ophthalmology and Visual Science*, *49*, 1333–1338.

Yamato, J., Ohya, J., & Ishii, K. (1992, Jun). Recognizing human action in time-sequential images using hidden markov model. In *Proc. ieee computer society conference on computer vision and pattern recognition (cvpr'92)* (pp. 379–385). IEEE Computer Society.

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *ACM Comput. Surv.*, *38*(4), 13.

Yilmaz, A., Li, X., & Shah, M. (2004). Contour based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Trans. Pattern Analysis and Machine Intelligence*, *26*(11), 1531–1536.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, *22*(11), 1330-1334.

Appendix

APPENDIX A. ADDITIONAL RESOURCES

A.1. Estimation of the Fundamental Matrix

Two points in homogeneous coordinates $\mathbf{m}_p = [x_p, y_p, 1]^\top$ and $\mathbf{m}_q = [x_q, y_q, 1]^\top$ in the views p and q, respectively, are in correspondence if the epipolar constraint holds

$$\mathbf{m}_q^\top \, \mathbf{F}_{pq} \, \mathbf{m}_p = 0 \,, \tag{A.1}$$

where

$$\mathbf{F}_{pq} = \begin{bmatrix} f_1 & f_2 & f_3 \\ f_4 & f_5 & f_6 \\ f_7 & f_8 & f_9 \end{bmatrix} \,,$$

is the fundamental matrix. Its estimation is performed using the set of correspondences between the views p, q computed thanks to the artificial markers. Additionally, a subset with the best correspondences can be obtained using the Random Sample Consensus (RANSAC) (Fischler & Bolles, 1981) technique. To estimate the fundamental matrix we employ the method proposed by Chen et al. (2000) with normalized 2-D coordinates (Hartley & Zisserman, 2000). Making use of the fact that $rank\,(\mathbf{F}_{pq}) = 2$, Chen et al. (2000) define the *biepipolar constraint* as

$$\mathbf{A}\,\mathbf{f} = 0 \,, \tag{A.2}$$

where $\mathbf{f} = (f_1, f_2, f_4, f_5)^\top$ and the $N \times 4$ matrix

$$\mathbf{A} = \begin{bmatrix} (x_1 - \alpha)(x'_1 - \alpha') & \dots & (x_N - \alpha)(x'_N - \alpha') \\ (x_1 - \alpha)(y'_1 - \beta') & \dots & (x_N - \alpha)(y'_N - \beta') \\ (y_1 - \beta)(x'_1 - \alpha') & \dots & (y_N - \beta)(x'_N - \alpha') \\ (y_1 - \beta)(y'_1 - \beta') & \dots & (y_N - \beta)(y'_N - \beta') \end{bmatrix}^\top \,, \tag{A.3}$$

with N being the total number of correspondences available. In our experiments we have generated $N = 1000$ corresponding points between for every pair of views. Disregarding the trivial solution $\mathbf{f} = 0$ in (A.2), it must hold that $\det(\mathbf{A}) = 0$. As a consequence we can first estimate the vector of parameters $\Gamma = (\alpha, \beta, \alpha', \beta')^\top$ and subsequently the unknown vector $\mathbf{f}$. Afterwards, the remaining entries of the fundamental matrix are computed as follows

$$
\begin{aligned}
f_3 &= -\alpha' f_1 - \beta' f_2 \\
f_6 &= -\alpha' f_4 - \beta' f_5 \\
f_7 &= -\alpha f_1 - \beta f_4 \\
f_8 &= -\alpha f_2 - \beta f_5 \\
f_9 &= -\alpha'\alpha f_1 - \alpha'\beta f_4 + \beta'\alpha f_2 + \beta'\beta f_5 \,.
\end{aligned} \tag{A.4}
$$

Let $\mathbf{B}_i$ be the i-th row vector of $\mathbf{A}$

$$
\mathbf{B}_i(\alpha, \beta, \alpha', \beta') = \begin{bmatrix} (x_i - \alpha)(x_i' - \alpha') \\ (x_i - \alpha)(y_i' - \beta') \\ (y_i - \beta)(x_i' - \alpha') \\ (y_i - \beta)(y_i' - \beta') \end{bmatrix} . \tag{A.5}
$$

To estimate the vector of parameters Γ we need to select four row vectors (i.e. four correspondences) from $\mathbf{A}$ such that

$$
\mathbf{B}_{ijkl} := \det\left([\mathbf{B}_i \; \mathbf{B}_j \; \mathbf{B}_k \; \mathbf{B}_l]^\top\right) = 0 \,. \tag{A.6}
$$

The vector of parameters is thus obtained as

$$
\Gamma = \arg \min_{\alpha,\beta,\alpha',\beta'} \sum^{|R|} |\mathbf{B}_{ijkl}| \,, \tag{A.7}
$$

where $|R|$ is the cardinality of the set R of all possible combinations of four row vectors. Note that this optimization problem is unfeasible with the number N of correspondences we have. Instead, we solve it just by randomly selecting 20% of the correspondences.

Knowing the vector of parameters Γ we now solve (A.2) for $\mathbf{f}$. Consider the singular value decomposition of the matrix $\mathbf{A}$

$$
\mathbf{A} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^\top \,, \tag{A.8}
$$

where $\mathbf{D}$ is a diagonal matrix of the same dimension as $\mathbf{A}$, with nonnegative diagonal elements in decreasing order, and unitary matrices $\mathbf{U}$ and $\mathbf{V}$. The solution of (A.2) corresponds to the last column vector of $\mathbf{V}$ associated with the smallest singular value of $\mathbf{D}$. With Γ, $\mathbf{f}$ and (A.4) the fundamental matrix $\mathbf{F}_{pq}$ is fully determined.