



Escuela Politécnica Superior

Departamento de Tecnología Electrónica y de las Comunicaciones

**CONTRIBUTIONS TO HAND GESTURE RECOGNITION ON THE BASIS OF
RANGE DATA**

PhD Thesis written by
Javier Molina Vela
under the supervision of
Prof. José María Martínez Sánchez

Madrid, October 2012

Copyright © 2012 Javier Molina Vela

All rights reserved. No part of this work may be reproduced, stored, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without prior permission. All trademarks are acknowledged to be the property of their respective owners.

Department: Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid, Spain

PhD Thesis: Contributions to Hand Gesture Recognition on the basis of Range Data

Author: **Javier Molina Vela**
Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)

Supervisor: **Jose María Martínez Sánchez**
Doctor Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)
Universidad Autónoma de Madrid , Spain

Year: 2012

Comittee: President: **Montserrat Pardàs Feliu**
Doctora Ingeniero de Telecomunicación
(Universitat Politècnica de Catalunya)
Universitat Politècnica de Catalunya

Secretary: **Jesús Bescós Cano**
Doctor Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)
Universidad Autónoma de Madrid , Spain

Vocal 1: **Luis Salgado Álvarez de Sotomayor**
Doctor Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)
Universidad Autónoma de Madrid , Spain

Vocal 2: **Fernando Díaz de María**
Doctor Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)
Universidad Carlos III de Madrid , Spain

Vocal 3: **Francisco Morán Burgos**
Doctor Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)
Universidad Politécnica de Madrid , Spain



The work described in this Thesis was carried out within the Video Processing and Understanding Lab at the Dept. of Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid (from 2008 to 2012). This work has been supported by the Spanish Administration agency CDTI under project CENIT- VISION 2007-1007.

To my family.

Acknowledgments

First of all I want to express my gratitude to my parents, which have been able to understand the problems associated to writing a thesis. And specially to my dad, who made me engineer before I knew what that meant.

Special thanks to my advisor, José M. Martínez, for his useful contributions, and to Jesús Bescós for his observations to the papers that resulted of this work.

This work would not have been the same without the great environment in which it was conceived, absolutely defined by the people of the lab: Victor Valdés for the right observation in the right moment, Luis Herranz for his dissertations, Fabri for being such a good discussion issue ;), Alvaro García I for introducing me to Weka, Pajuelo and Laura for making me feel as an advisor, Alvaro García II, Marcos, Juan Carlos, Victor Fernández, Fabri, Miguel Ángel, Anita, Ignacio... Also thanks to the NETS components: Myriam, Ivan, David, Mariano, Pablo... and to Jorge for his advice. To Oscar for letting me know about the vacancy in the Lab, and to Susy for teaching me the meaning of 'tikitiki'.

Special mention to the “flat mates” that have supported me and understood my changes of mood during these years: Chainy, Ramos, mikiroi, feerrr, danialón, maripa, pelu, maura...

For sure I am forgetting a lot people, I am sure they will understand, because they already know how grateful I am.

Javier Molina, October 2012

Abstract

The use of hand gestures offers an alternative to the commonly used human computer interfaces, providing a more intuitive way of navigating among menus and multimedia applications. This work presents contributions to hand gesture recognition on the basis of range data. Firstly, a real and synthetic benchmarking dataset is compiled, providing with a comparison framework for hand gesture recognition systems. A novel collection of critical factors of utility when designing a hand gesture dataset is proposed, analyzing the State Of Art solutions attending to these criteria. In terms of gesture scalability, apart from the method for synthetic generation of hands range data included in the mentioned dataset, the concept of synthetic subject is introduced, improving classification results by the variation of several intra-hand parameters. The representativity of the synthetically generated collection is demonstrated with an evaluation scheme in which the learning stage is fed with synthetic data while the evaluation is done with real subjects recordings.

Two examples of gesture recognition systems are presented in this work, making use of two of the dictionaries proposed in the dataset:

- The first one is a framework for generic hand gesture recognition which performs hand segmentation as well as a low-level extraction of potentially relevant features which are related to the morphological representation of the hand silhouette. Classification based on these features discriminates between a set of possible Static Hand Postures (SHPs) which results, combined with the estimated motion pattern of the hand, in the recognition of Dynamic Hand Gestures (DHGs).
- The second recognition system is focused on motion-based hand gestures. The hand translations are modelled on the basis of a novel human arm model, which is able to represent different motion patterns at different speeds.

Both systems work in real-time, allowing practical interaction between user and application.

Resumen

El uso de gestos manuales ofrece una alternativa a los interfaces hombre-máquina más comunes, proporcionando una manera más intuitiva de navegar a través de menús y aplicaciones multimedia. Este trabajo presenta contribuciones en el ámbito de reconocimiento de gestos manuales tomando como entrada información de profundidad de la escena. Inicialmente se presenta una colección de videos e imágenes de profundidad asociada a distintos diccionarios de gestos manuales: los videos son capturas reales, mientras que las imágenes son generadas sintéticamente. Este contenido constituye un marco para la comparación de distintas soluciones de reconocimiento de gestos manuales. Junto a esta colección se proponen una serie de factores críticos a considerar a la hora de recopilar contenido asociado a gestos manuales, evaluando su incidencia en las distintas colecciones disponibles en el Estado del Arte. En términos de escalabilidad, además de la solución para la generación de datos sintéticos utilizada en la colección propuesta, se propone el concepto de usuario sintético, que es el resultado de introducir variaciones en los parámetros que definen la mano sintética utilizada en el proceso de creación de contenido artificial. La representatividad de la colección sintética queda demostrada con un esquema de evaluación en el que el entrenamiento se realiza con esta misma, mientras la evaluación se realiza con contenido de usuarios reales.

Se presentan dos ejemplos de aplicación de sistemas de reconocimiento de gestos manuales, que hacen uso de dos de los diccionarios propuestos en la colección:

- El primero reconoce gestos de distinta naturaleza, previa segmentación basada en información de profundidad, y en base a una descripción de bajo nivel relacionada con información morfológica del contorno de la mano. La posterior clasificación discrimina entre un diccionario de poses estáticas de mano para luego, en combinación con la trayectoria estimada de la mano, realizar una separación entre los gestos dinámicos..
- El segundo sistema se centra en gestos manuales basados en la trayectoria realizada. En esta aproximación se modelan distintas trayectorias mediante un modelo sintético de brazo humano, que es configurado para la consecución de patrones de movimiento de distinto tipo y con distintas velocidades.

Ambos sistemas son capaces de trabajar en tiempo real, permitiendo así la interacción práctica entre hombre y máquina.

Contents

I	Introduction	1
1	Introduction	3
1.1	Motivation	3
1.2	Related work	4
1.3	Structure of the document	6
II	Dataset	9
2	A natural and synthetic corpus for benchmarking of hand gesture recognition systems	11
2.1	Introduction	11
2.1.1	Motivation	11
2.1.2	Critical factors involved in hand gesture datasets	12
2.1.3	Dataset design	14
2.2	Generation of syntethic hand poses	15
2.2.1	Hand pose definition	16
2.2.2	Depth image generation	18
2.3	Dataset generation	20
2.3.1	Dictionary selection	20
2.3.2	Recording of natural gestures	23
2.3.3	Generation of synthetic gestures	23
2.4	Discussion	23
2.4.1	Estimating the separability of pose-based gestures	24
2.4.2	Validation of the scalability approach	25
2.5	Conclusions	26
3	A Method for enhancing gesture scalability	29
3.1	Introduction	29

3.2	Generation of synthetic users profiles	30
3.2.1	Previous work	30
3.2.2	Hand parametrization: user profiles	30
3.2.3	Sets of user profiles	31
3.3	System validation	31
3.3.1	Dataset	31
3.3.2	Experimental setups	33
3.3.3	Results	33
3.4	Conclusions	34

III Recognition systems 37

4 Simple, compound and motion-based hand gesture recognition using static and dynamic models 39

4.1	Introduction	39
4.2	Related work	40
4.3	System overview	41
4.4	Hand gesture recognition approach	42
4.4.1	Introduction	42
4.4.2	Static hand posture recognition	43
4.4.2.1	Dictionary of static hand postures	43
4.4.2.2	Learning static hand postures	44
4.4.3	Dynamic hand gesture recognition	45
4.4.3.1	Dictionary of dynamic hand gestures	45
4.4.3.2	Detecting the dynamic hand gestures	49
4.5	Experiments	56
4.5.1	Experimental setup	56
4.5.2	Results	57
4.5.3	Computational cost	59
4.6	Conclusions	59

5 Motion-based hand gesture recognition using synthetic trajectories 61

5.1	Introduction	61
5.2	Related work	62
5.3	System overview	63
5.4	Hand gesture recognition approach	64
5.4.1	Introduction	64

5.4.2	Dataset	64
5.4.3	Motion pattern modelling	65
5.4.4	Motion pattern definition	67
5.4.5	Motion pattern capturing	68
5.4.6	Patterns comparison	68
5.5	Experiments	69
5.5.1	Experimental setup	69
5.5.2	Results	69
5.5.3	Computational cost	71
5.6	Conclusions	72
IV	Conclusions	75
6	Conclusions and future work	77
6.1	Summary of achievements	77
6.2	Future work	78
	Bibliography	80
V	Appendixes	89
	Appendix A: Confussion matrixes for synthetic training schemes	91
	Appendix B: Hand Descriptor	109
	Appendix C: Glossary	113
	Appendix D: Publications	115
	Appendix D: Conclusiones y trabajo futuro	117

List of Figures

2.1.1 Incidence of <i>pose issues</i>	15
2.2.1 Hand model set-ups.	17
2.2.2 Synthetically generated range data images with different points of view.	18
2.2.3 Synthetic depth image generation from hand model set-up. 'i' <i>extra</i> points, '*' <i>joint</i> points, 'o' <i>auxiliuar</i> points.	19
2.3.1 Captures from compiled dictionaries. First row of images of real users performing <i>static pose videos</i> . Second row of synthetic images.	21
2.3.2 Temporal evolution of motion-based and compound gestures.	22
2.4.1 Accuracy comparison for different descriptors and synthetic train setups.	26
3.2.1 Hand Model, the joints are denoted with *.	30
3.3.1 Accuracy comparison for different evaluation schemes. BL refers to Base Line, these accuracy rates are obtained assuming the gestures of each dictionary equiprobables.	35
4.3.1 System Overview.	41
4.4.1 Depth Segmentation: the captured depth image of a hand and its segmentation	43
4.4.2 Execution of DHG Fist performed by training user 3	46
4.4.3 Examples of static DHGs: its recognition just relies on SHP recognition and hand stillness.	47
4.4.4 Examples of simple DHGs involving motion, which requires SHP recognition and a estimation of the hand motion pattern. Notice that these two gestures are the same as N and W in Figure 2.3.2a.	48
4.4.5 Stability of the Y coordinate evolution for each of the characteristic points. Three types of DHGs, perfomed by training users, are evaluated: MenuClose (1st row), Simple Static Gestures(2nd row) and MenuOpen (3rd row).	52
4.4.6 Y coordinate trajectory characterization	53
4.4.7 Grid search of parameters α and W to maximize F_score in the recognition of DHGs MenuOpen and MenuClose.	54

4.4.8 Grid search of parameters β and ρ to maximize F_score in the recognition of the stationary part of simple static DHGs	54
4.4.9 FSM transitions and conforming DHGs in the execution of DHGs Take and Click.	55
5.3.1 Overview of the system.	64
5.4.1 Gestures observed from user's point of view.	65
5.4.2 Model set-ups of the arm model. $\vec{r_U}$ is a vector that goes from the shoulder to the elbow and $\vec{r_L}$ from the elbow to the wrist. The angles θ^x and θ^y are variables which define the trajectory of the arm in 5.4.2a and 5.4.2b, while ψ_0 and θ_0^z are fixed angles that define the position of the elbow at the beginning of the execution of the movement in Figure 5.4.2b and Figure 5.4.2c respectively. ψ_0 is the angle formed by $\vec{r_U}$ and $-\hat{y}$ (see 5.4.2b). θ_0^z indicates the rotation angle applied to N and S gestures, which results in the set-up shown in 5.4.2c.	67

List of Tables

2.1	Comparison of hand gestures data-sets in terms of the proposed critical factors (in bold). '-' is used for the critical factor "Pose issues" when for a certain dictionary there are no pose-based gestures.	14
2.2	Separability estimation ($\#G$ is the number of the pose-based gestures in the dictionary; $\#PI$ the number of them which present pose issues) and accuracy rates for the evaluation scheme NT (Natural content Training) detailed in Section 6.2 for the descriptors described in [Hu, 1962] and in [Molina et al., 2011].	24
2.3	Accuracy rate for the ST setups : 1, 20, 50 and 200. Training with a synthetic model captured from different Points Of View. Baseline (BL), is the accuracy rate for random gesture detection, i.e., $1/\#G$	25
3.1	Synthetic users profiles configuration parameters. The first two columns are the chosen values for the parameters. The number of users profiles are all the possible combinations of the parameters values.	31
3.2	Resulting images for set of profiles B(1). Notice that, since the number of POV is 1, only one image is rendered per user profile and gesture.	32
3.3	Accuracy rate in the detection of pose-based hand gestures on the basis of different synthetic training setups.	34
4.1	Accuracy in intra-frame SHP estimation.	45
4.2	Simple Static DHGs.	46
4.3	Simple DHGs with a specific motion pattern.	47
4.4	Compound DHGs.	49
4.5	Probabilities of correctly detecting a negative or wrongly detecting a positive. . .	50
4.6	User Independent Confusion Matrix for proposed DHGs.	57
4.7	User independent gesture recognition works comparison.	58
4.8	Non User Independent Confusion Matrix for another collection of DHGs.	59
4.9	Non User independent gesture recognition works comparison.	59
4.10	Computational Cost (msec) per frame	59

5.1	Confusion matrix for the 2.5D scenario. Gestures described in Section 5.4.2 and “U” for Unknown.	70
5.2	Confusion matrix for the 2D scenario. Gestures described in Section 5.4.2 and “U” for Unknown.	71
5.3	Computational Costs per frame and Accuracy for the two considered scenarios. .	72

Part I

Introduction

Chapter 1

Introduction

1.1 Motivation

Human Computer Interaction (HCI) technology and algorithms are evolving very rapidly. The user experience of high technological services is not always optimal and HCI might help bringing these services to the mass market. Although RGB cameras are still the most common capture technology, the trend in the last years is to use range data information. As a sign of the fast establishment of 3D user interfaces [Laviola, 2008], in the last years such kind of interfaces are becoming more important in the console gaming scenario¹²³. Besides, in desktop computers interfaces, the usage of the hand as input device provides natural HCI [Mitra and Acharya, 2007]. Traditionally, simple RGB capturing solutions were proposed for hand gesture recognition [Stenger et al., 2006, Chen and Tseng, 2007, Nickel and Stiefelhagen, 2007, Zheng et al., 2007, Teng et al., 2005, Lee and Park, 2009], but the use of 2.5D or 3D information enriches whichever human-computer interface. There are three main capture solutions for obtaining 2.5D information (i.e. two spatial coordinates plus depth information) of a scene: by the use of markers (or gloves) [Holte and Stoerring, 2004, Kelly et al., 2010, Martin Larsson, 2011, Keskin and Akarun, 2009, Usabiaga et al., 2009, Lee et al., 2009, Han, 2010, Heo et al., 2010] or accelerometers like in [Cheng et al., 2010]; using RGB stereo-vision configurations [Ho et al., 2011, Causo et al., 2009, 2008] and using TOF cameras [Soutschek et al., 2008, Kollarz et al., 2008, Molina et al., 2011]. The use of markers can result intrusive for the user of the system, while stereo-vision solutions require a complex setup and the presence of singular points in the scene in order to register the different views. Range cameras are an emerging technology which does not stop decreasing its price while increasing its capabilities (more resolution and wider viewing angle), making the confronting of possible interaction problems more affordable [Liu and Fujimura, 2004]. Addition-

¹<http://wii.com>

²<http://www.xbox.com/kinect/>

³<http://playstation.com/psmove/>

ally, the segmentation process (even in the presence of camera motion) becomes easier than with exclusively color data and with a much simpler setup than stereo-vision solutions.

The ultimate goal of this thesis is to improve user's experience when interacting with vertical surfaces, such as a TV display. Potential applications would be the navigation among maps, allowing intuitive movements of the earth surface; the control of multimedia menus [Soutschek et al., 2008] or the modification of the point of view on a virtual environment. We can find an example of TV remote control by hand gestures in [Premaratne and Nguyen, 2007] where TV commands codes are mapped on a collection of static hand gestures.

Usability constitutes a main issue in the development of HCI systems and some of the main aspects are pointed out in [Iso, 1998]; whilst a study devoted to improve user experience can be found in [Castilla et al., 2009]. The gestures selected along this work are chosen bearing in mind usability and gesture scalability criteria. As well, gesture scalability is a valuable characteristic in gesture recognition, allowing the inclusion of new detectable gestures with low cost.

1.2 Related work

When detecting hand gestures the first step is hand recognition, that is often performed by background subtraction. Depending on the application, the problem can be simplified by using a zenithal camera and an homogeneous background (the interactive surface) [Letessier and Bérard, 2004]. However, in most applications that require vertical gestures, the motion of the body of the person produces artifacts in common background subtraction techniques. Skin colour-based segmentation also presents problems in this context [Zhu et al., 2000], often due to the fact that the face of the person is also visible and corresponds to the same colour than the hand. It must be said that the use of depth information has been crucial in the last years for solving the segmentation problem. Time-of-Flight (TOF) range cameras supply depth information per pixel which makes them ideal for binary segmentation as depth information can generally separate the object from the background much better than intensity images, where colors, lighting, reflection and shadows almost always influence the performance of segmentation algorithms [Guomundsson et al., 2010a]. Some proposals for depth-based hand segmentation are proposed: in [Yang et al., 2012] the hand is expected to present wave motions; in [Molina et al., 2011] the hand is expected to be the nearest to the camera part of the body. In [Moeslund et al., 2006] some approaches to hand recognition and tracking are enumerated in the context of human motion characterization and body actions understanding, a much more generic field than the specific set-up presented in this work.

The use of depth information to improve hand gesture recognition has been recurrent in the last decade: Stereo-vision based systems such as [Grzeszczuk et al., 2000], in which background subtraction and 3D reconstruction make possible a proper hand segmentation and gesture recog-

nition within seven different static gestures, or [Nickel and Stiefelhagen, 2007], where gestures are defined by the pointing direction of hands and head and are used in remote robot-control. In [Athitsos and Sclaroff, 2003], where hand poses are detected, and in [Stenger et al., 2006], where hands are segmented and tracked, a 3D hand model is adjusted to a single view 2D input images. A recent approach is to use Time-of-Flight (TOF) range cameras that supply real-time depth information per pixel [Soutschek et al., 2008] at low cost. Some examples of the use of this technology can be found in [Guomundsson et al., 2010a] where it is used to improve people tracking in a smart room. The use of depth information results in an enrichment of the communication between user and machine by means of gestural interfaces. In this line, in [Liu and Fujimura, 2004] some advantages are remarked: robustness to illumination changes and easy segmentation even when there is camera motion. In [Breuer et al., 2007] the depth image captured by a Time-Of-Flight camera is transformed into a cloud of points to which a 3D hand model is adjusted. In [Kollorz et al., 2008] experiments are performed over a collection of twelve static hand gestures. [Soutschek et al., 2008] studies the application of this technology to the navigation among medical images using hand gestures. Another technology for obtaining range data is the one proposed in [Malassiotis and Strintzis, 2008] where the scene is illuminated with a coloured pattern, captured by a common RGB camera and later processed to infer depth information. The Kinect sensor⁴, which is based on a structured light imaging solution, is a popular and cheap solution for obtaining depth information of a scene. Recently it has been used for several gesture recognition solutions [Doliotis et al., 2011, Ren et al., 2011a, Yang et al., 2012, Antonio Hernández-Vela, 2012].

An interesting survey related with hand pose estimation was published in 2007 [Erol et al., 2007]. In [Poppe, 2007] the problem of body pose estimation, rather than hand pose, is overviewed. Several approaches have been followed for facing the problem of hand pose estimation, but it has been in the last years when the use of depth information has become really popular. When talking about 2D images (i.e. color or gray images) as input some works come up: in [Chen and Tseng, 2007], static hand postures within a dictionary are detected based on gray images. Another approach to the problem is presented in [Stenger et al., 2006] where discrimination between hand postures is performed, mainly focusing on hand features extraction starting from color images. In [Alon et al., 2009] a complete framework is presented, introducing the temporal component and detecting gestures defined by their motion pattern, such as digits drawn to the camera. The captures are still taken by a 2D camera. In [Zheng et al., 2007] a projective invariant hand feature vector is proposed and applied to person identification. In [Cheng and Trivedi, 2006, Ho et al., 2011] a hand pose estimation scheme based on 3D voxels adjusted to multicamera views of the hand is presented.

Different approaches for the description of the captured images have been proposed. In

⁴Microsoft Corp. Redmond WA. Kinect for Xbox 360.

[Chakraborty et al.] the concept of Eigenvector applied to hand gesture recognition is introduced. A hand pose angle estimation Gabor-filter based approach is described in [Huang et al., 2011]. [Li and Greenspan, 2011] treats hand gesture recognition as a secondary point. Authors of [Guomundsson et al., 2010b] present an approach for hand pose estimation with restrictions based on the adaptation of an ellipsoid 27 Degrees Of Freedom (DOF) hand model to the depth image. The variations performed in each articulation are guided by the principal components obtained from a synthetical postures data set. In [Liu and Kavakli, 2010] singular vectors are used to perform a global and local description of the capture. In [Cheng et al., 2010] a feature fusion method for 3D hand gesture recognition by learning a shared hidden space is proposed. A metric should be associated to a descriptor in order to be able to compare two descriptions. Usually this metric is the euclidean distance but some other proposals can be found in the literature [Baysal, 2010].

While most of the works which perform a training process or include a validation setup use real users data sets, there are some that make use of a synthetically generated collection [Baysal, 2010, Stenger et al., 2007].

1.3 Structure of the document

This document is divided in two main parts: one focused in the generation of a scalable hand gesture dataset, a second one which presents two examples of application of hand gesture recognition solutions:

- in Part II a dataset and associated scalability contributions are presented: in Chapter 2 a corpus, dataset and associated ground-truth, for the evaluation of hand gesture recognition approaches in Human Computer Interaction scenarios is presented. A novel collection of critical factors involved in the creation of a hand gestures dataset is proposed: capture technology, temporal coherence, nature of gestures, representativeness, complexity of gestures and scalability. Eleven users were recorded with a TOF camera. They were asked to execute hand gestures of different nature selected from several dictionaries of the State of Art. Special attention is given to the scalability of the set, proposing a method for the generation of synthetic depth images of gestures. Gestures covered in the corpus include single and multiple poses gestures (pose-based and compound), as well as gestures defined by motion and by pose and motion (motion-based and pose-motion based). Three kind of annotated data units are taken into consideration: static pose videos, gesture execution videos, both of them presenting temporal coherence (i.e. are continuous in time), and synthetically generated images. The resulting corpus, which exceeds in terms of representativity and scalability the datasets existing in the State of Art, provides a significant evaluation scenario for different kinds of hand gesture recognition solutions. In Chapter 3

the training framework for hand posture recognition systems based on a learning scheme fed with synthetically generated range images presented in Chapter 2 is extended. One of the most difficult issues when designing a hand gesture recognition system is to introduce new detectable gestures without high cost, this is known as gesture scalability. Commonly, the introduction of new gestures needs a recording session of them, involving real users in the process. Different configurations of a 3D hand model result in sets of synthetic users, which have shown good performance in the separation of gestures from several dictionaries of the State of Art. The proposed approach allows the learning of new dictionaries with no need of recording real users, so it is fully scalable in terms of gestures. The obtained accuracy rates for the dictionaries evaluated are comparable to, and for some cases better than, the ones reported for different real users training schemes.

- Part III presents two novel hand gestures recognition approaches together with the complete systems used in their validation. The dictionaries used in the evaluation of both systems are included in the dataset proposed in Chapter 2. In Chapter 4 a framework is presented, which, starting from the images captured by a TOF camera, performs hand segmentation as well as a low-level extraction of potentially relevant features which are related to the morphological representation of the hand silhouette. Classification based on these features discriminates between a set of possible Static Hand Postures (SHPs) which results, combined with the estimated motion pattern of the hand, in the recognition of Dynamic Hand Gestures (DHGs). The whole system works in real-time, allowing practical interaction between user and application. In Chapter 5 an innovative solution, also based on TOF video technology, to motion patterns recognition for real-time dynamic hand gesture recognition is presented. The resulting system is able to detect motion-based hand gestures getting as input depth images. The recognizable motion patterns are modeled on the basis of the human arm anatomy and its degrees of freedom, generating a collection of synthetic motion patterns that is compared with the captured input patterns in order to finally classify the input gesture. For the evaluation of our system one of the significant collections of gestures described in Chapter 2 is used, getting results for 2.5D pattern classification as well as a comparison with the results using only 2D information.
- Finally, the main conclusions and future work lines are compiled in Part IV.

Part II

Dataset

Chapter 2

A natural and synthetic corpus for benchmarking of hand gesture recognition systems

2.1 Introduction

2.1.1 Motivation

One of the most important limitations when designing a hand gesture recognition system is the lack of annotated datasets adapted to the particularities of the context of the application to be designed. This is critical when evaluating a system's performance. In this chapter, the design, generation and annotation of a new corpus¹ is described. It is composed of gestures of several collections of the State of Art (SoA). Furthermore, it includes some other gestures that we have considered as useful for nowadays application environments.

In a dataset design process, some critical factors should be taken into account. In Section 2.1.2, a collection of factors are identified and proposed, and a review of the SoA in this framework is included. In Section 2.1.3, the proposed dataset designed in the light of these critical factors is outlined; special attention is then given to the scalability factor, covered via synthetic content generation (see Section 2.2). In Section 2.3, the dataset content is described, indicating the considered dictionaries and the characteristics of the compiled videos and images. Some classification experiments are presented in Section 2.4 in order to estimate the separability of a dictionary (see Section 2.4.1) and to validate a posture detection approach which, by using the aforementioned synthetic content, does not require real users in the training stage (see Section 2.4.2). Finally, in Section 2.5, main conclusions are presented and future work lines are pointed out.

¹available at <http://www-vpu.eps.uam.es/DSs/HGds>

2.1.2 Critical factors involved in hand gesture datasets

The compilation of a dataset for the evaluation of hand gesture recognition should consider some factors that could be critical when evaluating a system in a specific context of application. These critical factors are identified and described below, including in the discussion the main reviewed works of the SoA (see Table 2.1).

Temporal coherence: the examined datasets just provide with either images or videos, being sets of single images per gesture sample the most common situation [Triesch and VD Malsburg, 1996, Marcel, 1999, Soutschek et al., 2008, Kollorz et al., 2008, Ren et al., 2011b]. However, video temporal continuity (as in [Marcel et al., 2000, Holte and Stoerring, 2004, Kim et al., 2007, Han and Liang, 2008, Martin Larsson, 2011]) allows temporal filtering either in the analysis or in the decision phase. Moreover, considering videos as annotation units allows a more adequate adaptation to real situations, in which gestures are performed during some consecutive frames. Besides, natural hand transitions during a gesture, which use to be the hardest poses to model, are intrinsically included in videos.

Representativeness: when designing and generating a dataset, one of the main objectives is to cover as many practical situations as possible. In this line, the representativeness of a gesture dataset increases with the number of users, with their heterogeneity and with the variations in the point of view of the captures. Besides, the more available and heterogeneous dictionaries (according to the nature of their gestures), the more scenarios could be considered when designing a recognition solution. Finally, the availability of videos instead of single images, provide transitory frames in which the performed gestures vary in appearance with respect to the iconic models of their front-side versions, which enhances representativeness.

Nature of gestures: in the SoA four different kinds of gestures can be identified:

- Pose-based, defined entirely by the pose of the hand, as in [Triesch and VD Malsburg, 1996, Marcel, 1999, Triesch and VD Malsburg, 2001, Holte and Stoerring, 2004, Dadgostar et al., 2005, Soutschek et al., 2008, Kollorz et al., 2008, Han and Liang, 2008, Ren et al., 2011b].
- Motion-based, in which the hand pose is not relevant, i.e., the hand trajectory explicitly defines the gesture, as in [Marcel et al., 2000, Holte and Stoerring, 2004, Martin Larsson, 2011].
- Pose-motion based, defined both by a pose and certain motion pattern in the execution, as in [Kim et al., 2007].
- Compound, which are gestures composed of a sequence of pose-based gestures, as in [Holte and Stoerring, 2004].

Scalability: this refers to the capacity of easily extending a dataset to include a new collection of gestures, which is a very valuable characteristic. In practice, it is hard to collect a representative group of users to perform a recording session. In [Dadgostar et al., 2005], a method for the generation of synthetic hand color images is described, but the authors suggest that their model is far from ideal. None of the analyzed datasets present a scalable solution.

Capture technology: RGB cameras are the most common technology due to its low cost. However, the trend in the last years is to use hand’s range information, either via stereo-vision [Ho et al., 2011, Causo et al., 2009, 2008] or via Time-Of-Flight (TOF) cameras [Soutschek et al., 2008, Kollorz et al., 2008, Molina et al., 2011]. TOF technology has several advantages [Liu and Fujimura, 2004]: it allows to obtain 2.5D data in a non-intrusive way, without using markers or gloves, as in [Holte and Stoerring, 2004, Martin Larsson, 2011], with a simpler set-up than stereo-vision systems, and it is robust to illumination conditions. Additionally, the hand segmentation process becomes easier than with exclusively color data, and much simpler than in stereo-vision solutions, even in the presence of camera motion.

Pose issues: some problematic factors are grouped here, either intrinsic to the gesture definition or introduced by the acquisition process, that may hinder pose detection with the existing analysis techniques. These issues can be significant when they make two or more poses more similar:

- *Finger occlusion*, owing to either crossed fingers or to a lateral point of view of the camera.
- *Hand-core occlusion*, understanding the hand-core as the part of the hand that it is not fingers. Occlusion happens when the point of view of the camera hides the palm and the opisthenar area.
- *2D silhouettes with no protuberances*: Many hand gesture detection approaches in the of the SoA, such as [Hu, 1962], use a description of the detected silhouette rather than that of the hand. When there is more than one gesture in which the fingers are not identifiable on the basis of the hand silhouette, the gesture detection task becomes more difficult. The absence of a representative 2D silhouette for more than one gesture introduces a handicap for the detection of these gestures.
- *Forearm presence*: the miss-segmentation of the forearm as part of the hand may increase the difficulty to later classify a gesture, which was trained from forearm-free samples. This only applies to videos capturing real users and depends on the acquisition process.

Content Set	Temp. Coherence				Representativeness					Scalability	Capture			Pose issues				
	pose-based videos	execution videos	real images	synthetic generation	#users	Nature					RGB data	3D data	Non intrusive	Finger occlusion	Hand-core occlusion	No protuberances	Forearm presence	
						# pose-based	# motion-based	# pose-motion based	# compound									
[Triesch and VD Malsburg, 1996]	✗	✗	✓	✗	24	10	0	0	0	✗	✓	✗	✓	✓	✓	✗	✗	✓
[Marcel, 1999]	✗	✗	✓	✗	10	6	0	0	0	✗	✓	✗	✓	✗	✓	✗	✗	✓
[Marcel et al., 2000]	✗	✓	✗	✗	10	0	4	0	0	✗	✓	✗	✓	-	-	-	-	✓
[Triesch and VD Malsburg, 2001]	✗	✗	✓	✗	19	12	0	0	0	✗	✓	✗	✓	✗	✓	✗	✓	✓
[Holte and Stoerring, 2004]	✗	✓	✗	✗	16	9	2	0	2	✗	✓	✗	✗	✓	✓	✗	✓	✓
[Kim et al., 2007]	✗	✓	✗	✗	2	0	0	9	0	✗	✓	✗	✓	✓	✗	✗	✗	✓
[Soutschek et al., 2008]	✗	✗	✓	✗	15	5	0	0	0	✗	✗	✓	✓	✓	✓	✗	✗	✗
[Kollorz et al., 2008]	✗	✗	✓	✗	34	12	0	0	0	✗	✗	✓	✓	✓	✓	✓	✓	✗
[Han and Liang, 2008]	✗	✓	✗	✗	1	0	0	9	0	✗	✓	✗	✓	✗	✗	✗	✗	✗
[Martin Larsson, 2011]	✗	✓	✗	✗	10	0	5	0	0	✗	✗	✓	✗	-	-	-	-	✓
[Ren et al., 2011b]	✗	✗	✓	✗	10	14	0	0	0	✗	✓	✓	✓	✗	✗	✓	✗	✓
Prop	✓	✓	✓	✓	11	58	8	2	2	✓	✗	✓	✓	✓	✓	✓	✓	✓

Table 2.1: Comparison of hand gestures data-sets in terms of the proposed critical factors (in **bold**). ‘-’ is used for the critical factor “Pose issues” when for a certain dictionary there are no pose-based gestures.

2.1.3 Dataset design

The last row of Table 2.1 describes the proposed dataset according to the identified critical factors. This section details the decisions taken for the creation of the proposed data-set, decisions focused on covering as much as possible these critical factors.

Temporal coherence has been considered by including three types of data in the proposed data-set:

- *Static pose videos*, obtained asking users to perform a hand pose in front of the camera for a certain amount of time.
- *Execution videos*, where users were asked to get into the capture interaction area, execute the gesture under consideration, and move outwards.
- *Images*, both synthetic images and samples from the mentioned videos.

As aforementioned, the inclusion of execution videos also affects *representativeness*, as it indirectly includes transitory frames in which the executed gestures vary in appearance with respect to the iconic models of their front-side versions. This critical factor has also been considered

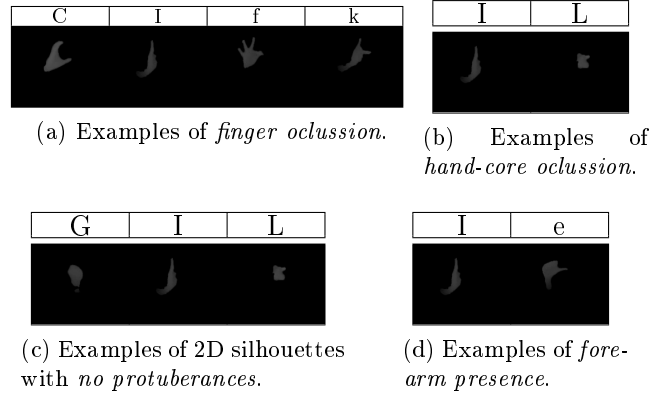


Figure 2.1.1: Incidence of *pose issues*.

by including a significant number of users (see Section 2.3.2), by developing a methodology for synthetically generating images of a hand pose from different points of view (see Section 2.2), and by including a set of dictionaries which detailed description can be found in Section 2.3.1.

Regarding the *nature of gestures*, the proposed dataset includes all the identified categories: posed-based, motion-based, pose-motion based and compound gestures.

Scalability, as already mentioned in Section 2.1.2, is a highly valuable factor, since it allows the inclusion of new gestures with no need of gathering new users. In this line, a method for the synthetic generation of depth images for new dictionaries is proposed (see Section 2.2).

The proposed dataset was recorded with a TOF camera (SR4000 developed by Mesa Imaging²), considering the advantages, mentioned in Section 2.1.2, of this *capture technology*.

The dataset was decided to include several dictionaries (see Section 2.3.1). Figure 2.1.1 illustrates some pose captures showing the aforementioned *pose issues*: *finger occlusion* in 'C' and 'I' or 'f' and 'k' from Figure 2.1.1a; *hand-core occlusion* in 'I' and 'L' from Figure 2.1.1b; 2D silhouette with *no protuberances* in 'G', 'I' and 'L' from Figure 2.1.1c; *forearm presence* in 'I' or 'e' from Figure 2.1.1d.

2.2 Generation of synthetic hand poses

This Section describes a novel method for the synthetic generation of whichever hand gesture collection; it is based on a widely used kinematic hand model [Erol et al., 2007, Ge et al., 2006] with 27 degrees of freedom (DOF). Once the model parameters have been defined for each hand pose of the desired collection (see Section 2.2.1), a volumetric hand is created via a morphological dilation process. Then, the position of the point of view is set to capture a range data image similar to the ones captured by TOF technology (see Section 2.2.2).

²<http://www.mesa-imaging.ch/>

2.2.1 Hand pose definition

In order to define a hand pose, it is necessary to set values to the 27 degrees of freedom which define the configuration of the aforementioned kinematic model. The nomenclature used in Figure 2.2.1 is the one followed in [Ge et al., 2006] and it is used for the definition of the hand model setups. The associated parameters are:

- Global hand parameters:
 - T_x , T_y and T_z refer to the global translation of the model.
 - θ_1^x , θ_1^y and θ_1^z refer to the global rotation.
- Finger and phalange parameters:
 - $\theta_{<PH>i}^x$, $\theta_{<PH>i}^y$ and $\theta_{<PH>i}^z$ are used for indicating the rotation of each phalange, where:
 - * $<PH>$ indicates the finger for which the rotations are performed; its possible values are: Thumb (T), Index (I), Middle (M), Ring (R) and Little (L).
 - * x , y and z indicate the relative axis about which the rotation is performed.
 - * i indicates the depth of the joint ($i = 1$ for the wrist point, $i = 3$ the last joint of the thumb and $i = 4$ for the last joints of rest of fingers) used as reference for applying the rotations.
- Fixed hand parameters, chosen to have realistic proportions in the resulting hands (see Figure 2.2.1):
 - Hand $Scale$, defines the size of the hand: in this work the scale was subjectively fixed to a value of 3.
 - The coordinates for the point in the base of the palm (using the reference axis indicated in Figure 2.2.1): $\overrightarrow{OT1} = Scale \cdot [2, 1, 0]$, $\overrightarrow{OI1} = Scale \cdot [1, 1, 0]$, $\overrightarrow{OM1} = Scale \cdot [0, 1, 0]$, $\overrightarrow{OR1} = Scale \cdot [-1, 1, 0]$, $\overrightarrow{OL1} = Scale \cdot [-2, 1, 0]$.
 - The length for the phalanges of the hand: $|\overrightarrow{T1T2}| = 4 \cdot Scale$, $|\overrightarrow{I1I2}| = |\overrightarrow{M1M2}| = |\overrightarrow{R1R2}| = 4 \cdot Scale$, $|\overrightarrow{L1L2}| = 3 \cdot Scale$, $|\overrightarrow{T2T3}| = |\overrightarrow{I2I3}| = |\overrightarrow{M2M3}| = |\overrightarrow{R2R3}| = |\overrightarrow{L2L3}| = 3 \cdot Scale$, $|\overrightarrow{I3I4}| = |\overrightarrow{M3M4}| = |\overrightarrow{R3R4}| = |\overrightarrow{L3L4}| = 2 \cdot Scale$ and also $2 \cdot Scale$ for the extreme segments of the fingers. The rotation angles about \hat{z} for the segments of the palm: for $I1I2$, $-\pi/18$; for $M1M2$, 0; for $R1R2$, $\pi/18$; for $L1L2$, $\pi/9$.

set-up parameters	hand model
$T_1^x = 0, T_1^y = 0, T_1^z = 0, \theta_1^x = 0, \theta_1^y = 0, \theta_1^z = 0$ $\theta_{T1}^x = \pi/4, \theta_{T1}^z = -\pi/8, \theta_{T2}^x = 0, \theta_{T2}^z = \pi/4, \theta_{T3}^z = \pi/4$ $\theta_{I2}^x = 0, \theta_{I2}^z = 0, \theta_{I3}^x = 0, \theta_{I4}^x = 0$ $\theta_{M2}^x = \pi/2, \theta_{M2}^z = 0, \theta_{M3}^x = \pi/2, \theta_{M4}^x = \pi/2$ $\theta_{R2}^x = \pi/2, \theta_{R2}^z = 0, \theta_{R3}^x = \pi/2, \theta_{R4}^x = \pi/2$ $\theta_{L2}^x = \pi/2, \theta_{L2}^z = 0, \theta_{L3}^x = \pi/2, \theta_{L4}^x = \pi/2$	
$T_1^x = 0, T_1^y = 0, T_1^z = 0, \theta_1^x = 0, \theta_1^y = 0, \theta_1^z = 0$ $\theta_{T1}^x = \pi/4, \theta_{T1}^z = 0, \theta_{T2}^x = 0, \theta_{T2}^z = \pi/4, \theta_{T3}^z = \pi/4$ $\theta_{I2}^x = 0, \theta_{I2}^z = 0, \theta_{I3}^x = 0, \theta_{I4}^x = 0$ $\theta_{M2}^x = 0, \theta_{M2}^z = \pi/12, \theta_{M3}^x = 0, \theta_{M4}^x = 0$ $\theta_{R2}^x = \pi/2, \theta_{R2}^z = 0, \theta_{R3}^x = \pi/2, \theta_{R4}^x = \pi/2$ $\theta_{L2}^x = \pi/2, \theta_{L2}^z = 0, \theta_{L3}^x = \pi/2, \theta_{L4}^x = \pi/2$	
$T_1^x = 0, T_1^y = 0, T_1^z = 0, \theta_1^x = 0, \theta_1^y = 0, \theta_1^z = 0$ $\theta_{T1}^x = 0, \theta_{T1}^z = -\pi/4, \theta_{T2}^x = 0, \theta_{T2}^z = 0, \theta_{T3}^z = 0$ $\theta_{I2}^x = 0, \theta_{I2}^z = 0, \theta_{I3}^x = 0, \theta_{I4}^x = 0$ $\theta_{M2}^x = 0, \theta_{M2}^z = \pi/12, \theta_{M3}^x = 0, \theta_{M4}^x = 0$ $\theta_{R2}^x = \pi/2, \theta_{R2}^z = 0, \theta_{R3}^x = \pi/2, \theta_{R4}^x = \pi/2$ $\theta_{L2}^x = \pi/2, \theta_{L2}^z = 0, \theta_{L3}^x = \pi/2, \theta_{L4}^x = \pi/2$	
$T_1^x = 0, T_1^y = 0, T_1^z = 0, \theta_1^x = 0, \theta_1^y = 0, \theta_1^z = 0$ $\theta_{T1}^x = \pi/4, \theta_{T1}^z = \pi/16, \theta_{T2}^x = 0, \theta_{T2}^z = \pi/4, \theta_{T3}^z = \pi/4$ $\theta_{I2}^x = 0, \theta_{I2}^z = 0, \theta_{I3}^x = 0, \theta_{I4}^x = 0$ $\theta_{M2}^x = 0, \theta_{M2}^z = 0, \theta_{M3}^x = 0, \theta_{M4}^x = 0$ $\theta_{R2}^x = 0, \theta_{R2}^z = \pi/32, \theta_{R3}^x = 0, \theta_{R4}^x = 0$ $\theta_{L2}^x = 0, \theta_{L2}^z = \pi/20, \theta_{L3}^x = 0, \theta_{L4}^x = 0$	
$T_1^x = 0, T_1^y = 0, T_1^z = 0, \theta_1^x = 0, \theta_1^y = 0, \theta_1^z = 0$ $\theta_{T1}^x = 0, \theta_{T1}^z = -\pi/4, \theta_{T2}^x = 0, \theta_{T2}^z = 0, \theta_{T3}^z = 0$ $\theta_{I2}^x = 0, \theta_{I2}^z = 0, \theta_{I3}^x = 0, \theta_{I4}^x = 0$ $\theta_{M2}^x = 0, \theta_{M2}^z = 0, \theta_{M3}^x = 0, \theta_{M4}^x = 0$ $\theta_{R2}^x = 0, \theta_{R2}^z = \pi/32, \theta_{R3}^x = 0, \theta_{R4}^x = 0$ $\theta_{L2}^x = 0, \theta_{L2}^z = \pi/20, \theta_{L3}^x = 0, \theta_{L4}^x = 0$	

Figure 2.2.1: Hand model set-ups.

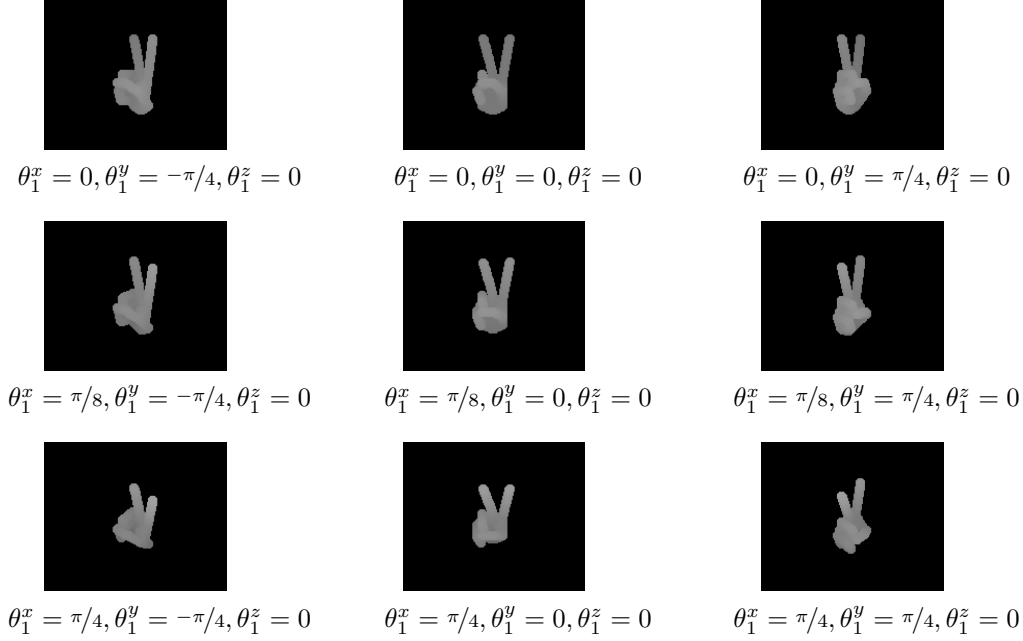


Figure 2.2.2: Synthetically generated range data images with different points of view.

Different setups of the model result in different hand poses (see Figure 2.2.1) and orientations of the hand. As mentioned, three parameters describe the global orientation of the hand: θ_1^x , θ_1^y and θ_1^z . Examples of the resulting range images for different orientations are included in Figure 2.2.2.

2.2.2 Depth image generation

Beginning from the configured hand model, the synthetic depth image is obtained by a morphological dilation process using a 3D morphological library³, using as structuring element a sphere of ratio 4 voxels. In Figure 2.2.3 the *loci* defined along this section are illustrated: *joint* points, *auxiliar* points, *filling* line and *extra* points. The resulting *seed* points, for which the 3D dilation is performed, are computed following the next steps:

1. The *joint* points (*) are defined as the set of points located between two phalanges (including also the hand origin).
2. In order to provide the synthetic hand with a natural appearance, a set of *auxiliar* points (o) are included following the next distribution:

³<http://www.mmorph.com/>

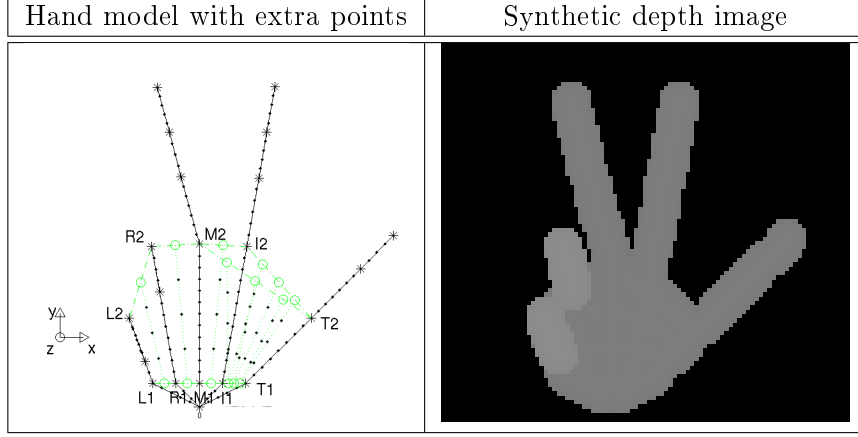


Figure 2.2.3: Synthetic depth image generation from hand model set-up. '.' *extra* points, '*' *joint* points, 'o' *auxiliar* points.

- (a) one point at the middle of the segments: $L1R1$, $R1M1$, $M1I1$, $L2R2$, $R2M2$ and $M2I2$.
 - (b) three equidistant points in the segments: $T1I1$, $T2I2$ and $T2M2$ (in a way that such segments are divided in five sub-segments of equal length).
3. The *anchor* points are defined as the union of *joint* points (*) and *auxiliar* points (o).
4. Some pairs of *anchor* points (those depicted at Figure 2.2.3) are selected in order to give volume to the palm of the hand, for the later definition of *filling* lines.
5. An interpolation process is performed over each *filling* line by introducing *extra* points (.) between the selected *anchor* points: being L the length of a *filling* line, the interpolation process consists on the division of the *filling* line in segments with a minimum length of $L_{min} = Scale/2$. Then, the number of sub-segments between two *anchor* points is $N_{seg} = trunc(L/L_{min})$, where $trunc(Q)$ returns the integer part of Q .
6. The *seed* points for the morphological 3D dilation are the union of *joint* points (*) and *extra* points (.).

Finally, the synthetic depth image is obtained by calculating the distance from the volumetric hand to a virtual camera symbolized by a 3D point and a capture direction. The relative position between such point and the volumetric hand defines the point of view and the size of the 2D projection of the syntetic hand over the view plane.

2.3 Dataset generation

This Section first describes the dictionaries selected for the proposed data-set, and then explains the procedures followed to populate the data-set, both with gestures from real users and with synthetically generated gestures.

2.3.1 Dictionary selection

The proposed data-set includes six dictionaries, chosen following *representativeness* (including several SoA dictionaries), *nature of gestures* (considering pose-based, motion-based, pose-motion based and compound) and *pose issues* (selecting dictionaries containing gestures with oclussions) criteria:

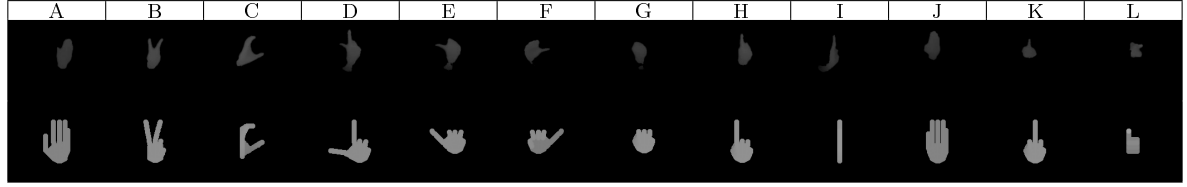
1. Dictionary proposed in [Kollarz et al., 2008] (see Figure 2.3.1a): This is a 12 pose-based gestures dictionary.
2. Dictionary proposed in [Molina et al., 2011] (see Figure 2.3.1b and 2.3.1): In this dictionary we can find 9 pose-based, 2 motion-based and 2 compound gestures.
3. Dictionary proposed in [Soutschek et al., 2008] (see Figure 2.3.1c): This is a 5 pose-based gestures dictionary.
4. Miscellaneous pose-based gestures dictionary (see Figure 2.3.1d)
5. Spanish sign language alphabet⁴ (see Figure 2.3.1e): 25 posed-based and 4 posed-motion based gestures (“n”/“ñ” and “v”/“w”, see Figure 2.3.1e).
6. Slaps dictionary (see Figure 2.3.2a): It consists of 9 motion-based gestures.

As a result, a representative dataset is provided, with dictionaries useful in different contexts of application: generic HCI interfaces for which dictionaries such as 3, 1, 2 or 4 could be of utility; interfaces for changing the point of view in a 3D virtual scene or moving a virtual object, in which 6 could apply; or some more specific, such as a sign language translator, in which 5 could be of help in first versions.

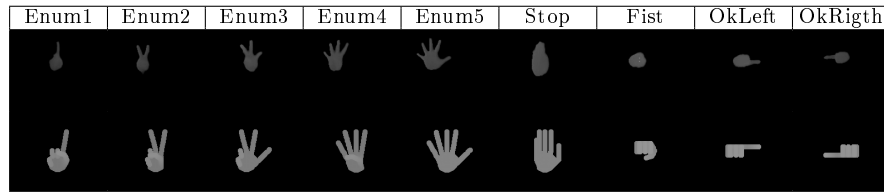
The detailed criteria followed for the selection of the gestures included in dictionaries 2 and 6 can be found respectively in Chapters 4 and 5 respectively, where two gesture detection systems are presented.

Summarizing, the dataset includes 70 gestures: 55 pose-based, 9 motion-based, 4 pose-motion based, and 2 compound.

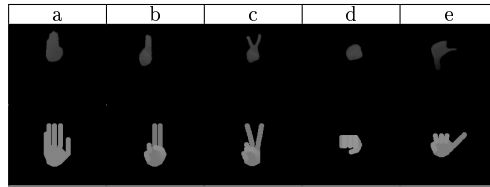
⁴<http://www.sematos.eu/lse.html>



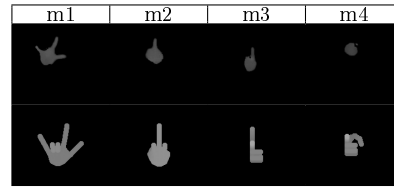
(a) [Kollorz et al., 2008]



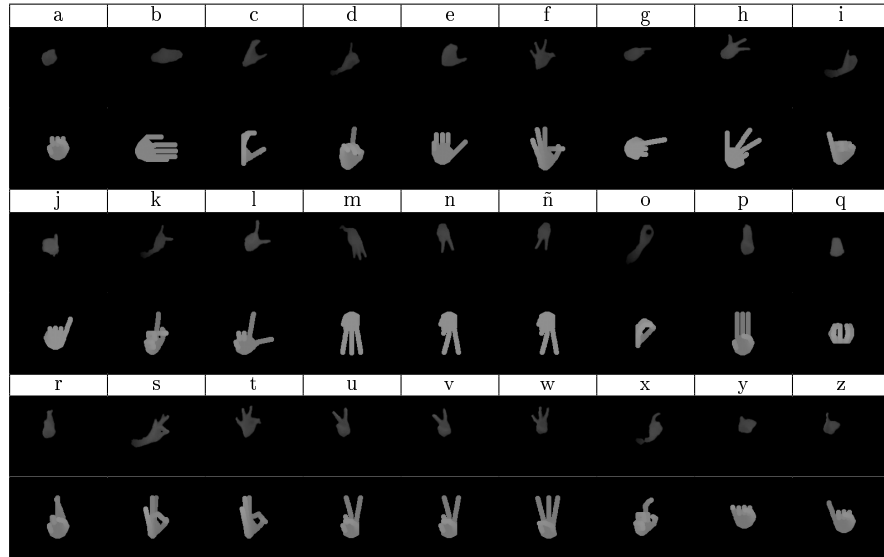
(b) [Molina et al., 2011]



(c) [Soutschek et al., 2008]

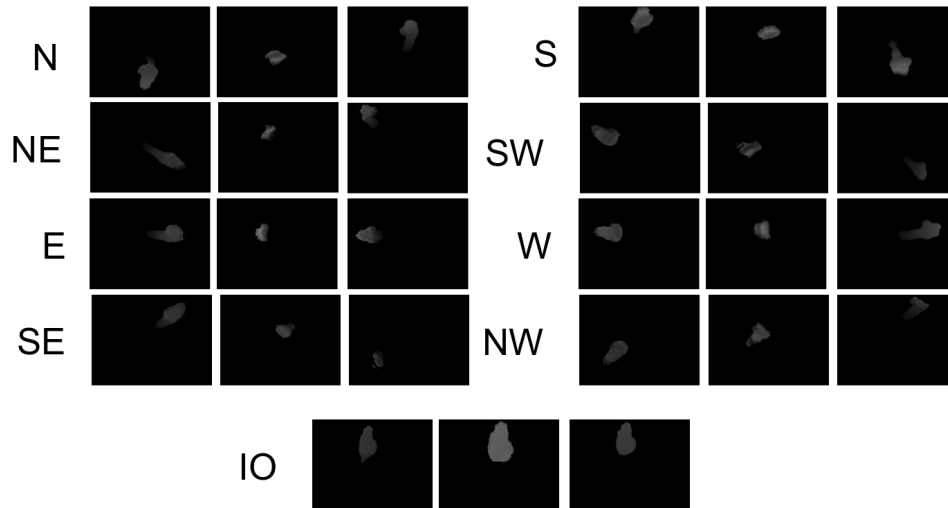


(d) Miscellaneous pose-based gestures.

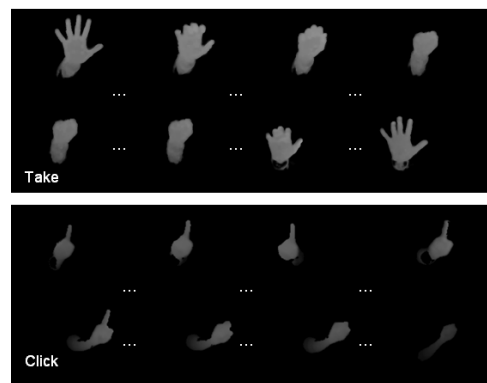


(e) Spanish sign language alphabet.

Figure 2.3.1: Captures from compiled dictionaries. First row of images of real users performing *static pose videos*. Second row of synthetic images.



(a) Captures from Slaps dictionary: N, NE, E, SE, S, SW, W and NW, the cardinal points indicating the direction of the executions. IO, meaning Inwards-Outwards.



(b) Captures of the compound gestures Take and Click, defined in [Molina et al., 2011].

Figure 2.3.2: Temporal evolution of motion-based and compound gestures.

2.3.2 Recording of natural gestures

The hardware used for recording the data-set described in this paper consisted of the Mesa Imaging SR4000 TOF camera which was placed on top of a 22" TFT display, which constituted the feedback component for HCI. This camera captures depth images with QCIF resolution (176x144 pixels) with a depth resolution of $\pm 1\text{cm}$. The camera was configured to capture 30fps and to operate in a 3m depth range (0.3m-3.3m) in order to remove background objects.

For the compilation of this data-set, 11 users of different gender, age and technical background were asked to participate in a recording session of around one hour each, with a few minutes time break between the two stages described below:

- *Static pose videos*: these were recorded just for the pose-based gestures. Users were asked to keep a static pose in front of the camera, moving through the interaction area for a short period of time (250 frames).
- *Execution videos*: these videos were recorded for all kind of gestures. They allow an off-line evaluation similar to real scenarios. Users were asked to execute each gesture 5 times. Each execution consisted in getting the hand in the interaction area, performing the gesture and moving it outwards.

2.3.3 Generation of synthetic gestures

For each pose-based gesture, several captures were compiled with randomly generated global rotation angles applied to the initial pose setup, which depends on the posture being modelled. These variation angles belong to the domain: $\Delta\theta_1^x \in [0, \pi/8]$, $\Delta\theta_1^y \in [-\pi/8, \pi/8]$, $\Delta\theta_1^z \in [-\pi/8, \pi/8]$. Different setups are generated, corresponding to the number of points of view (POV) captured per gesture: 1, 20, 50 and 200 POV. So, for each pose-based gesture, we have 1, 20, 50 and 200 samples, depending on the setup.

2.4 Discussion

As a result of experimentation with the proposed data-set, this section focuses on two relevant aspects: the first is to describe the observation that the ability of a detector to identify pose-based gestures on a specific dictionary is highly correlated with the number of gestures with *pose issues*; the second is to validate the proposed scalability approach.

These experiments have been conducted using two different hand descriptors, fully described in [Molina et al., 2011, Hu, 1962]. While the descriptor presented in [Hu, 1962] can be used for the description of whichever 2D visual shape, the one presented in [Molina et al., 2011] (see Appendix V) was specifically conceived for the modelling of range data associated to hand gestures. Both descriptors provide with a fixed length description of the contour of the detected hand.

Dictionary ↓	Sep. estimation			(NT) Accuracy rates	
	# G	# PI	# $PI/\#G$	[Hu, 1962]	[Molina et al., 2011]
[Soutschek et al., 2008]	5	2	0.400	0.856	0.913
[Kollorz et al., 2008]	12	7	0.583	0.602	0.846
[Molina et al., 2011]	9	2	0.222	0.706	0.971
Misc. pose-based	4	3	0.750	0.791	0.861
Spanish SL alphabet	25	11	0.440	0.354	0.624

Table 2.2: Separability estimation ($\#G$ is the number of the pose-based gestures in the dictionary; $\#PI$ the number of them which present pose issues) and accuracy rates for the evaluation scheme NT (Natural content Training) detailed in Section 6.2 for the descriptors described in [Hu, 1962] and in [Molina et al., 2011].

2.4.1 Estimating the separability of pose-based gestures

There are two factors which might have *a priori* influence in the detection accuracy in a collection of pose-based gestures: *the size of the dictionary*, the higher the number of gestures, the more difficult might be to separate among them; and the *pose issues*, which can make some gestures of a dictionary very similar between them.

In the proposed dictionaries, it is quite straightforward to identify subsets of hand postures that, due to the occurrence of *pose issues*, present similar visual appearance (see Figure 2.3.1): a-d for [Soutschek et al., 2008]; A-J, G-J-L and H-I-K for [Kollorz et al., 2008] ; Stop-Fist for [Molina et al., 2011]; m2-m3-m4 for the Misc pose-based dictionary; and d-r, f-k-s-t and a-o-p-q-y for the Spanish SL alphabet .

Table 2.2 summarizes three basic indicators for each dictionary: the number of gestures ($\#G$), the number of gestures with *pose issues* ($\#PI$), and the ratio between them ($\#PI/\#G$). It can be observed that it exists a significant correlation (p-value<5%) between the computed values of $\#PI$ and the accuracy rates (detailed in the next section) obtained using descriptors [Hu, 1962] and [Molina et al., 2011]: these correlations are 0.949 and 0.963 respectively. The correlations between $\#PI/\#G$ and the accuracies are not significant, while for $\#G$ the correlation with the accuracy rate obtained with [Hu, 1962] is 0.981, and with [Molina et al., 2011] it is not significant. It can be concluded that the number of gestures with *pose issues* ($\#PI$) in a pose-based gestures dictionary is significant for the estimation of its separability; and that this indicator is more relevant than the size of the dictionary ($\#G$) or the ratio $\#PI/\#G$.

Dict. ↓ Desc.⇒	Baseline (BL)	[Hu, 1962] ST [1 20 50 200]	[Molina et al., 2011] ST [1 20 50 200]
[Soutschek et al., 2008]	0.200	0.619 0.672 0.664 0.674	0.498 0.667 0.770 0.810
[Kollorz et al., 2008]	0.083	0.240 0.262 0.286 0.283	0.483 0.562 0.553 0.569
[Molina et al., 2011]	0.111	0.460 0.551 0.562 0.568	0.652 0.834 0.851 0.882
Misc.	0.250	0.413 0.481 0.423 0.518	0.708 0.731 0.725 0.729
Spanish SL	0.040	0.130 0.168 0.174 0.179	0.219 0.298 0.317 0.309

Table 2.3: Accuracy rate for the ST setups : 1, 20, 50 and 200. Training with a synthetic model captured from different Points Of View. Baseline (BL), is the accuracy rate for random gesture detection, i.e., $1/\#G$

2.4.2 Validation of the scalability approach

The tests that have been run to validate the scalability approach for each dictionary of the proposed data-set are described here. As synthetic content is only available for pose-based gestures, the evaluation only targets their detection. Two evaluation schemes have been set out in order to obtain and compare the accuracy rates in the detection of gestures for each dictionary, and also compare them with the *a priori* probabilities of each posture in each dictionary, named as Baseline (BL):

1. Training with natural content (NT): Leave-One-Out for a user independent cross-validation using Nearest Neighbour as classifier, which is the scheme described in [Kollorz et al., 2008].
2. Training with syntethic content (ST): in this scheme, the system is trained with the syntethic compiled data, while the evaluation is performed over recordings with real users.

In Table 2.3, the accuracies for the proposed synthetic trained setups are presented. Notice that in most of the cases, the accuracy rate increases with the number of points of view considered. In Figure 2.4.1, a grapichal comparison of the results for real users training and synthetic training is presented. The best results for synthetic training are promising: achieving more than 80% correct detections for two of the dictionaries and going beyond twice the baseline accuracy for all of them.

Deepening into the confussion matrixes of each dictionary applying the ST200 evaluation scheme (see A(200) matrixes in Appendix A), it is possible to observe that the loss in accuracy for the ST evaluation scheme is mainly due to misclassification in some specific situations: d is detected as b,a in [Soutschek et al., 2008]; A as J, G as J and E as D in [Kollorz et al., 2008]; Stop as Fist in [Molina et al., 2011]; m3 as m2 in the Misc pose-based dictionary; and a as o, c as l, g as b, o as p, q as c, a and m, in the Spanish SL alphabet.

From these results it can be concluded that pose-based gestures with no protuberances are specially problematic when there are more than one in the dictionary. In the Spanish SL dictionary there are five of these (i.e. a, b, o, p and q), which, for the used descriptors, produce a

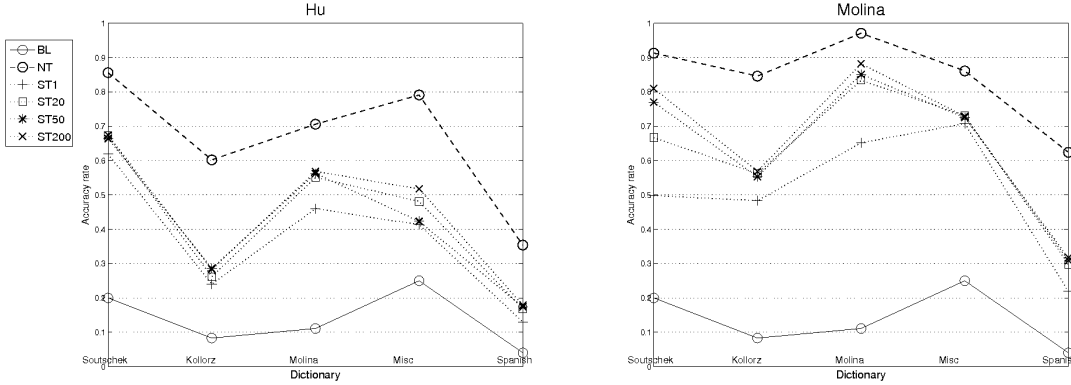


Figure 2.4.1: Accuracy comparison for different descriptors and synthetic train setups.

lot of errors. The dictionary presented in [Kollorz et al., 2008], apart of three gestures with no protuberances (i.e. A, G and J), presents two gestures, E and D, only differing in a protuberance and, as the descriptors have shown, are therefore not easy to detect. The information associated to the contour of the hand is very valuable when detecting poses. The proposed model does not properly represent poses with similar protuberances. This could be solved varying fixed hand parameters (i.e. scale and intra-hand proportions) in order to create different synthetic users profiles, increasing this way, the representativity of the synthetic collection.

2.5 Conclusions

In this chapter a corpus of hand depth images for benchmarking of hand detection systems has been presented. It has been compared in terms of a set of novel critical factors with several datasets of the SoA. It has real users recordings of several dictionaries described in other papers, as well as synthetically generated depth images associated to the hand poses of those gestures. The used capture technology provides 2.5D data in a non intrusive way. The compiled collection includes posed-based, motion-based, pose-motion based and compound gestures. In terms of representativity, 11 different users participated in the compilation of the collection. Moreover, point of view variations can be introduced in the synthetic data, increasing the significance of the collection. In relation with the temporal coherence of the compiled corpus, for the real data, the annotation units are videos, allowing the introduction of a temporal preprocessing module in the design of a hand gesture detection solution. A separability study for pose-based gestures is proposed and validated, providing with a method for the estimation of the difficulty to separate a set of gestures, with no need of applying classification techniques. The proposed method for the generation of synthetic images makes posible the generation of images for new gestures with a simple design process and with no need of training users, what constitutes an

important advantage in terms of scalability. The synthetic generation method has been validated with synthetic content training schemes presenting promising results, which are close (for some dictionaries) to those obtained by the the real users training scheme.

Chapter 3

A Method for enhancing gesture scalability

3.1 Introduction¹

As mentioned in Chapter 2, one of the most important limitations when designing a hand gesture recognition system is the lack of annotated datasets (which are critical for evaluating the system's performance and are commonly used in the learning stage) adapted to the particularities of the context of the application to be designed. Gesture scalability is a highly desirable characteristic for a gesture recognition system as pointed out in Chapter 2. A solution for achieving this scalability is the use of synthetic data in the training stage. Some works use synthetically generated images: for example, in [Stenger et al., 2007] 2D hand images are generated for evaluating color and shape features in hand pose classifications; in [Baysal, 2010], configurations of a 3D hand model are used for modelling a hand manipulating objects; in Chapter 2 a hand model and the process to obtain range data images from it are described in detail.

The main contribution presented in this chapter is the introduction of synthetic users profiles to extend the synthetic collection presented in Chapter 2, improving the scalability of the solution proposed there. As the obtained results show, some synthetic training configurations obtain results comparable to the ones with real users, and saving the complexity of such approach. Differently to [Baysal, 2010], the proposed solution evaluates the system with real users records, whilst with respect to [Stenger et al., 2007] their input images contain color information not depth as the ones considered in this work.

The chapter is structured as follows: in Section 3.2, the concept of synthetic user is introduced and the parameters for generating a set of them is explained; in Section 3.3 the real users and

¹This chapter is based on : *J. Molina, and J. M. Martínez, "A Synthetic Training Framework for providing gesture scalability to 2.5D pose-based hand gesture recognition systems", Machine Vision and Applications (under review).*

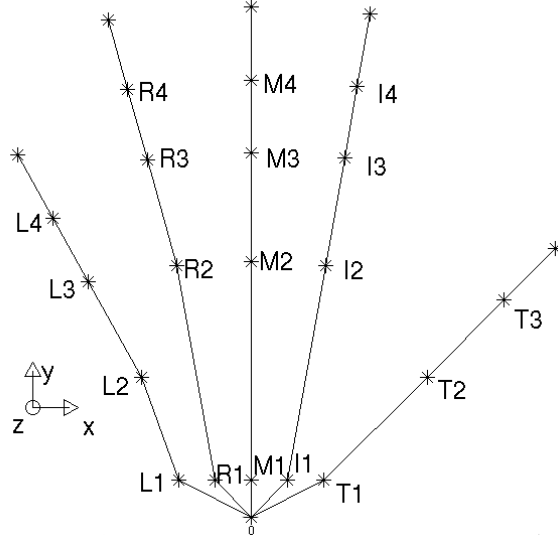


Figure 3.2.1: Hand Model, the joints are denoted with *.

the generated synthetic data set are described, the used evaluation schemes are enumerated, and the results for the different natural and synthetic configurations are presented; the conclusions and future work lines are included in Section 3.4.

3.2 Generation of synthetic users profiles

3.2.1 Previous work

In Chapter 2, a method for the generation of synthetically generated range images is presented. It is based on a 27 Degrees of Freedom (DOF) kinematic model widely used [Erol et al., 2007, Ge et al., 2006]. As already described, for each configuration of the hand model a volumetric hand is created via a morphological dilation process. Variations in the point of view (POV) are introduced to increase the representativity of the collection: the range of variation of the angles are $\Delta\theta_1^x \in [0, \pi/8]$, $\Delta\theta_1^y \in [-\pi/8, \pi/8]$, and $\Delta\theta_1^z \in [-\pi/8, \pi/8]$, being θ_1^x , θ_1^y and θ_1^z the global rotation angles about x , y and z axes (see Figure 3.2.1).

3.2.2 Hand parametrization: user profiles

In order to increase the significance of the synthetic collection, new degrees of freedom are now introduced in the generation stage. Synthetic users profiles are created on the basis of the combinations of values for the following parameters, which definition are formulated using the reference points indicated in Figure 3.2.1:

	FP-AR	PW	# user profiles
A(#POV)	{1}	{3}	$1 \times 1 = 1$
B(#POV)	{1, 1.1, 1.2, 1.3}	{2, 3, 4}	$4 \times 3 = 12$
C(#POV)	{1, 1.05, 1.1, 1.15, 1.2}	{2, 2.5, 3, 3.5, 4}	$5 \times 5 = 25$
D(#POV)	{0.9, 0.95, 1, 1.05, 1.1}	{2, 2.5, 3, 3.5, 4}	$5 \times 5 = 25$

Table 3.1: Synthetic users profiles configuration parameters. The first two columns are the chosen values for the parameters. The number of users profiles are all the possible combinations of the parameters values.

- Fingers-Palm Aspect Relation (FP-AR): the ratio between the length of the phalanges of fingers and hand.

$$FP - AR = \frac{|\overrightarrow{M1M2}|}{2 \star |\overrightarrow{M2M3}|}$$

Notice that certain restrictions are applied to the hand model (see Chapter 2), among them: $|\overrightarrow{M1M2}| = |\overrightarrow{I1I2}| = |\overrightarrow{R1R2}| = |\overrightarrow{L1L2}|$ and $|\overrightarrow{M2M3}| = |\overrightarrow{I2I3}| = |\overrightarrow{R2R3}| = |\overrightarrow{L2L3}|$

- Palm Width (PW): the width of the palm.

$$PW = |\overrightarrow{T1L1}|$$

3.2.3 Sets of user profiles

Different sets of synthetic users profiles (from now on will be referred as sets) are generated applying variations in the parameters described above. In Table 3.1, the possible values of the two parameters are indicated (notice that the set A(#POV) is the one evaluated in Chapter 2), while in Figure 3.2 the 12 range images for one of the postures and for the set B(1) (see Table 3.1) are shown.

3.3 System validation

3.3.1 Dataset

For the validation of the proposed solution the *static poses videos* compiled from real users (see Section 2.3.2) are used.

The dictionaries under consideration are the ones with posed-based gestures (see Section 2.3.1): [Soutschek et al., 2008], [Kollorz et al., 2008], [Molina et al., 2011], Miscellaneous pose-based gestures dictionary and Spanish sign language alphabet.

The synthetic dataset is compiled generating each of the sets of user profiles (see Section 3.2.3) for the above listed dictionaries. Notice that the gestures synthetically generated do not













		PW		
		2	3	4
FP-AR	1			
	1.1			
	1.2			
	1.3			

Table 3.2: Resulting images for set of profiles B(1). Notice that, since the number of POV is 1, only one image is rendered per user profile and gesture.

include motion-based ones, since these are out of the scope of this work. Each real user, for each of the gestures, is asked to record up to 250 frames, using in the evaluation 200 of them.

3.3.2 Experimental setups

The used descriptor is the one described in [Molina et al., 2011], which showed good results for real users training. This descriptor is a feature characterization scheme based on a geodesic description of the 3D surface of the hand. In order to evaluate the proposed solution for hand gesture detection, two evaluation schemes are proposed:

- Natural Training, NT(N): the *static pose videos* of N of the 11 users (see Section 2.3.2) are used for training while the rest for evaluation. Notice that the number of combinations to be evaluated for each value of N is $\binom{11}{N}$. For each value of N, the mean accuracy rate for the resulting combinations is computed. This scheme can be understood as a measure of the performance of the system trained with N users, the more training users, the more representative the system is. Three values of N have been used in the evaluation stage: $N = 1, 2$ and 10. Notice that when $N = 10$, this scheme is named Leave-One-Out [Kollorz et al., 2008, Molina et al., 2011].
- Synthetic Training, A(#POV), B(#POV), C(#POV) and D(#POV), in the learning stage one of these synthetic sets is used (see Table 3.1), while for the evaluation the *static pose videos* of the eleven users in the real users content are used.

3.3.3 Results

The accuracy rates for the evaluation schemes enumerated in Section 3.3.2 can be found in Table 3.3. The used descriptor in these evaluations is described in [Molina et al., 2011]. In Figure 3.3.1, the most significant evaluation schemes are compared in terms of their accuracy rates, for later discussing the inferred conclusions.

As expected, the accuracy rates for NT(N) increase with N: the higher number of users used in the training stage, the more significant the resulting model will be. Notice that, since the used descriptor was designed for working with the dictionary proposed in [Molina et al., 2011], this is the one for which the descriptor works best, independently of the number of gestures of each dictionary. The low accuracy rates obtained for dictionaries [Kollorz et al., 2008] and Spanish sign language, in natural and synthetic evaluation schemes, make think that the used descriptor [Molina et al., 2011] does not perform adequately for the particularities of these dictionaries. Anyway the main contribution of this work is to present a training framework that does not need real users participation, and that is open to new descriptors more adapted to the particularities of the gestures to classify.

Dict.\setup->	NT(1 2 10)	A(1 20 50 200)	B(1 20 50 200)	C(1 20 50 200)	D(1 20 50 200)
[Soutschek et al., 2008]	0.834 0.876 0.913	0.498 0.667 0.770 0.810	0.631 0.860 0.849 0.843	0.688 0.846 0.857 0.858	0.592 0.858 0.833 0.859
[Kollorz et al., 2008]	0.734 0.775 0.846	0.483 0.562 0.553 0.569	0.544 0.541 0.582 0.558	0.543 0.567 0.566 0.551	0.492 0.542 0.525 0.595
[Molina et al., 2011]	0.875 0.92 0.971	0.652 0.834 0.851 0.882	0.703 0.876 0.875 0.879	0.620 0.865 0.879 0.872	0.578 0.845 0.882 0.913
Misc.	0.764 0.799 0.861	0.708 0.731 0.725 0.729	0.749 0.682 0.672 0.663	0.757 0.662 0.714 0.694	0.693 0.709 0.712 0.758
Spanish SL*	0.529 0.596 0.624	0.219 0.298 0.317 0.309	0.308 0.361 0.359 0.357	0.328 0.359 0.345 0.339	0.271 0.305 0.316 0.307

Table 3.3: Accuracy rate in the detection of pose-based hand gestures on the basis of different synthetic training setups.

As shown in Table 3.3, the best synthetic training configuration (for four of the five dictionaries) is D(200), this is, the set of profiles D, described in Table 3.1, with 200 randomly generated variations of the POV in the ranges mentioned in Section 3.2.1. The worst synthetic configuration is A(1), the simplest one, with only one synthetic user profile and with no variations in the POV.

As can be checked in Figure 3.3.1, the setup D(200) rose above the accuracy rates obtained with NT(1) and is near to the ones for NT(2) for two dictionaries [Soutschek et al., 2008, Molina et al., 2011], being comparable for another one, the Miscellaneous dictionary. No synthetic sets present better results than the ones reported for NT(10). These results show that the proposed synthetic training solution is not able to perform as good as a system trained with ten real users, but it is if the system is trained with only one user. As well, it can be said that the accuracy rate obtained in Section 2.4.2 for its best setup (A(200)) is overcome by, among others, the setup D(200). The confusion matrixes associated to the synthetic training schemes A(200), B(200), C(200) and D(200) are compiled in Appendix A.

3.4 Conclusions

In this Chapter a method for the generation of synthetic hands range data is proposed, introducing the concept of synthetic user profile. This approach is evaluated in terms of accuracy rate, with a training stage performed using synthetic data and an evaluation with real users. This solution offers a gesture scalable approach, which allows the learning of new gestures with no need of recording real users. As well, the use of synthetic users with different hand particularities makes the collection more representative, as the evolution of the results for the Synthetic Training schemes shows. The proposed training framework is able to work for some dictionaries

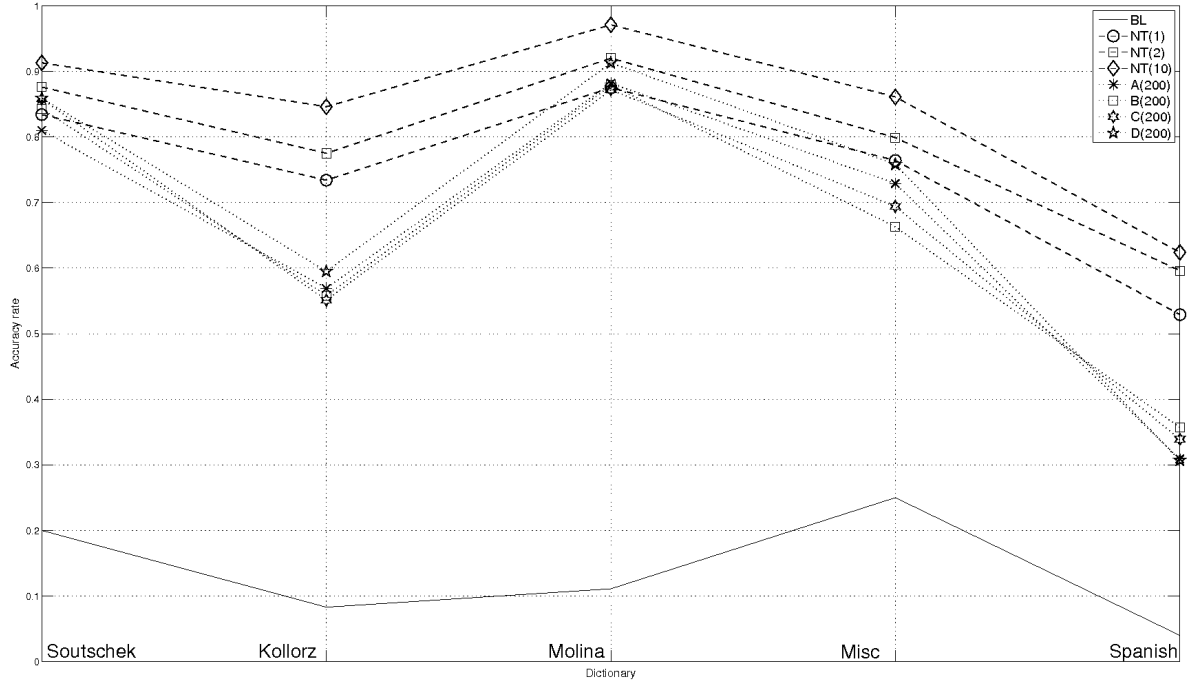


Figure 3.3.1: Accuracy comparison for different evaluation schemes. BL refers to Base Line, these accuracy rates are obtained assuming the gestures of each dictionary equiprobables.

as good as a system trained with one real user.

The particularities of the gestures for some of the dictionaries under consideration make difficult their correct separation. Some specially problematic pose-based gestures are pointed out in Section 2.4.2, describing possible causes for the missclassification in their detection. The experiments carried out in this chapter corroborates these proposed causes. The future use of different visual descriptors more adapted to the particularities of the collection of gestures to be detected would improve the results as already pointed out in the previous chapter.

Part III

Recognition systems

Chapter 4

Simple, compound and motion-based hand gesture recognition using static and dynamic models

4.1 Introduction¹

The framework presented in this chapter works with a set of gestures [Molina et al., 2011] that have been shown to be both user friendly and descriptive enough to cover most common human-computer interactions [Castilla et al., 2009], such as controlling multimedia menus or interacting with virtual environments. This set of gestures is described in Section 2.3.1 and some captions of real users executions can be found in Figure 2.3.1b.

Starting from the captured range data, precise hand segmentation is performed, to introduce depth information in its low-level description and to determine the start and end times of the gestures. A hand feature characterization scheme based on a geodesic description of the 3D surface of the hand is used [Molina et al., 2011]. This characterization feeds two middle-level classification stages: Static Hand Postures (SHP) and Dynamic Hand Gestures (DHG) recognition. First, the SHP recognition stage determines the posture of the hand for each frame with a set of classification machines, one per posture in the proposed dictionary; then, the probability of presence of at least a positive for each SHP within a temporal window is calculated. Additionally, the motion of the hand within the considered temporal window is stored and fitted to a set of possible motion patterns. Finally, the restrictions introduced by a Finite State Machine (FSM) and the estimated motion pattern result in the final recognition of simple, compound and motion based gestures, this is, DHGs.

¹This chapter is based on: *J. Molina, M. Escudero-Viñolo, A. Signoriello, M. Pardás, C- Ferrán, J. Bescós, F. Marqués, and J. M. Martínez, “Real-time user independent hand gesture recognition from time-of-flight camera video using static and dynamic models”, Machine Vision and Applications, pp. 1-18, 2011 (on-line first).*

The remainder of the chapter is structured as follows. Firstly the State Of Art is presented in Section 4.2, for later including a system overview in Section 4.3. In Section 4.4.2 the SHP is defined as a concept, describing the chosen approach to its recognition. In Section 4.4.3, starting from the intra-frame recognition of SHPs, the methodology for the recognition of the DHGs included in the proposed dictionary is introduced. In Section 4.5.2 results for the proposed gestures, as well as for the ones described in [Soutschek et al., 2008], are presented for finally pointing out in Section 4.6 conclusions and future work lines.

4.2 Related work

Regarding the hand gesture recognition, in [Chen and Tseng, 2007], static hand postures within a dictionary are detected based on 2D gray images. Another approach to the problem is presented in [Stenger et al., 2006], where discrimination between hand postures is performed, mainly focusing on hand features extraction starting from 2D images. In [Alon et al., 2009], a complete framework is presented, introducing the temporal component and detecting gestures defined by their motion pattern, such as digits drawn to the camera. The captures are still taken by a 2D camera. In [Zheng et al., 2007], a projective invariant hand feature vector is proposed and applied to person identification. In [Teng et al., 2005], static hand posture detection is applied to the recognition of some gestures of the Chinese sign language. In [Zaki and Shaheen, 2011, Holden et al., 2005], American and Australian Sign Languages gestures are recognized basing on hand shape, place of articulation, hand orientation and movement, while [Kelly et al., 2010] do so only focusing in static postures.

There are several approaches for processing a temporal sequence of observations: [Mitra and Acharya, 2007] presents the use of Finite State Machines (FSM) and Hidden Markov Models (HMM) for gesture recognition in a general way for then pointing out different contexts of application. In the proposed approach, a gesture is assumed to start when the hand is detected in the scene, and the end of the gesture is declared when the hand is not detected anymore.

In comparison with the SoA, a segmentation technique based on range data captured by a TOF camera is proposed, thus performing a non intrusive hand segmentation (unlikely to glove dependent approaches: [Kelly et al., 2010, Keskin and Akarun, 2009, Usabiaga et al., 2009]) robust to low illumination conditions (not as in color camera based systems: [Stenger et al., 2006, Chen and Tseng, 2007, Nickel and Stiefelhagen, 2007, Zheng et al., 2007, Teng et al., 2005]) and face to hand occlusion (differently than skin color based systems like: [Zhu et al., 2000, Athitsos and Sclaroff, 2003, Grzeszczuk et al., 2000, Alon et al., 2009, Zaki and Shaheen, 2011]). Regarding the referred range data based systems [Soutschek et al., 2008, Breuer et al., 2007, Kollorz et al., 2008, Malassiotis and Srinivas, 2008], none of them faces the problem of detecting compound or motion based gestures.

Summarizing, the key differentiating feature of the proposed framework from existing work in gesture recognition is the recognition of simple, compound and motion based gestures in a unified real-time framework and in a non intrusive fashion. For this novel recognition framework a novel abstraction of gestures, depending on their temporal evolution is proposed: Static Hand Poses (SHP) and Dynamic Hand Gestures (DHG). The detection of SHPs is performed by a novel Support Vector Machines (SVM) scheme, while a configuration of a Finite State Machine (FSM) is presented for DHG recognition.

4.3 System overview

The proposed system allows the user to control applications in a vertical display using hand gestures for the interaction. Analogously to the recording sessions described in Section 2.3.2, a TOF camera (SR4000 developed by Mesa Imaging²) is placed above a display and the system analyzes the performed gestures. This camera captures depth images with QCIF resolution (176x144 pixels) and a depth precision of $\pm 1\text{cm}$. The camera has been configured to capture 30fps and to operate in a 3m depth range (0.3m-3.3m), in order to remove background objects. From now on this range will be named the interaction area.

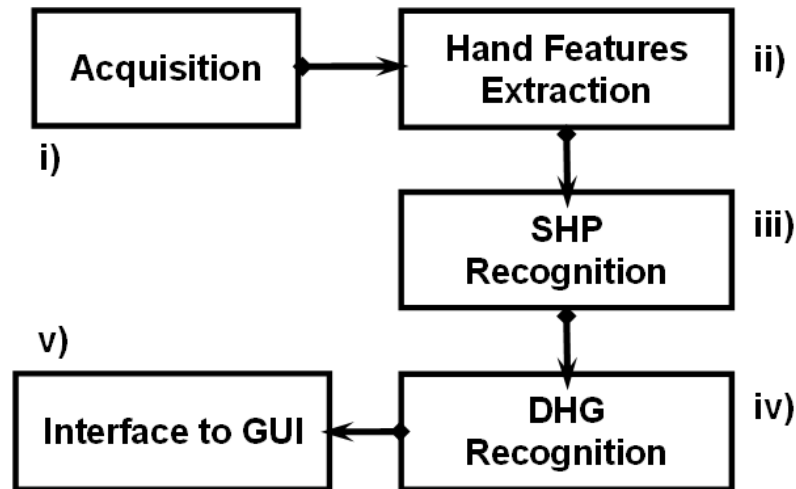


Figure 4.3.1: System Overview.

The recognition system consists of the following modules:

- (i) acquisition of depth images: this initial stage performs the acquisition of the depth image performed by a TOF camera, which results in a volumetric surface of the captured hand.

²<http://www.mesa-imaging.ch/>

- (ii) hand features extraction: with an appropriate post-processing and thresholding the segmentation of the hand is performed, removing the forearm. The hand is therefore assumed to be the nearest scene object. If no object is clearly detected at this point, it is assumed that it is because the hand is not in the defined interaction area. Later, the descriptor described in Appendix V is extracted, modeling the hand by a feature vector.
- (iii) SHP recognition: on the basis of the mentioned description, a Support Vector Solution is proposed (Section 4.4.2).
- (iv) DHG recognition: starting from the recognized SHPs and the estimated hand motion and by means of a Finite State Machine, the temporal sequence is processed (Section 4.4.3)
- (v) communication with the GUI: the recognized DHGs and the 3D hand coordinates are sent to the GUI allowing the control of the application under consideration.

4.4 Hand gesture recognition approach

4.4.1 Introduction

The gesture recognition approach consists on three stages:

- preprocessing and hand description: The accurate detection of the hand silhouette, even using a TOF camera, is not a solved problem. When the gestures are executed in the central part of the image, a simple thresholding provides good segmentation results which do not depend on the posture of the hand. However, when the gestures are executed far from the central part of the image it is common to find the forearm as part of the mask. In the left image of Figure 4.4.1, the forearm is clearly visible becoming a problem when applying the descriptor extraction detailed in Appendix V. In order to eliminate the forearm before the feature extraction, the brightest pixel (the nearest point to the camera) is identified. A mask is generated including in it at least 20 gray levels below the brightest one (twenty centimeters from the nearest point). In this way good segmentation results are obtained, since for most of the cases the forearm is removed without losing hand pixels. When the gestures are executed in the outermost area of the screen, it is harder to eliminate the forearm, which results in a problem in the recognition of certain gestures, as it will be explained in Section 4.4.3.2. The used descriptor is fully described in Appendix V; it is a model of the silhouette built using the following parameters: Geodesic Center (3D coordinates), ellipse parameters, Minimum Depth Point, number of maxima, and for each maximum: intensity, amplitude and angle.



Figure 4.4.1: Depth Segmentation: the captured depth image of a hand and its segmentation

- SHP recognition for a hand captured in an instant of time (see Section 4.4.2). On the basis of the contour description already introduced, a classification among a dictionary of postures is performed using Support Vector Machines (SVM).
- DHG recognition results of processing the temporal evolution of the detected SHPs and of the estimated hand coordinates (see Section 4.4.3). For this purpose, a Finite State Machine (FSM) is configured.

4.4.2 Static hand posture recognition

The concept of SHP is introduced as an intermediate level to achieve the detection of DHGs. A SHP is understood as a posture of the hand captured in an instant of time. Relative position to the camera is not significant when separating SHPs, that is, a SHP can be performed anywhere in the interaction area.

4.4.2.1 Dictionary of static hand postures

In [Castilla et al., 2009], an experiment with real users was conducted to define a gestural dictionary allowing a user to interact with a system in a natural way. The interaction can be done with real/non real and horizontal/vertical metaphors. Examples of interactions with real metaphors are rotation, grabbing or catching and with non-real metaphors ‘cancel’. The amount of interaction possibilities with the metaphors is very high, but the experiment identifies the most frequent ones and its associated hand gesture.

Those gestures and interactions that can improve the user experience without fatiguing him have been selected for recognition. Therefore, among the complete set of gestures obtained from the experiment a sub-set has been selected to define the dictionary used in this work, based on a trade-off between usability and recognition. This dictionary was included in the dataset

proposed in Section 2.3.1 (see Figure 2.3.1b). *Static pose videos* (see Section 2.3.2) from just three of the users (users 1, 2 and 3) were selected for training purposes, selecting 200 frames per video. This resulted in 600 positives samples per SHP.

4.4.2.2 Learning static hand postures

The recognition of a SHP is performed at frame level using SVMs [Chang and Lin, 2001], which requires a representative data collection (see Section 4.4.2.1), a selection of an adequate feature vector and a design of the training schema.

Feature vector The feature vector designed to describe each frame is compound of a combination of the global parameters and contour characteristics mentioned described in Appendix V. Summarizing, the chosen features are:

- Number of maxima.
- For each maximum: intensity, amplitude and angle.
- A description of the fitted ellipse: major axis angle and major and minor axis length.

These values constitute the feature vector of intra-hand characteristics, let us call it, v^i , where i indicates that the feature vector is extracted from a frame with the SHP- i , where i is the id of the SHP (see Table 4.1). Considering always 5 maxima (if there are less their coordinates are filled with ‘-1’), a vector with 19 coordinates is obtained.

Training scheme A SVM [Chang and Lin, 2001] has been trained for each SHP using the samples of a specific SHP as positives and the samples of the rest of SHPs as negatives. Each coordinate of the input pattern has been normalized with mean 0 and standard deviation 1. The Radial Basis Function (RBF) kernel has been used:

$$K(x_i, x_j) = C * e^{-\gamma \|v^i - v^j\|^2} \quad (4.4.1)$$

, where C and γ are the parameters of the kernel and v^i, v^j two feature vectors. In order to identify the optimal configuration of this kernel, it has been evaluated with the following grid: $C = 2^{-3, -2, \dots, 3}$, combined with $\gamma = 2^{-3, -2, \dots, 0}$. For this purpose a 5-fold cross-validation has been performed, optimizing the F-score [Goutte and Gaussier, 2005] fixing $\beta = 0.5$:

$$F_\beta = \frac{(1 + \beta^2) * (precision * recall)}{\beta^2 * precision + recall} \quad (4.4.2)$$

This value of β weights the precision over the recall. Experimentally, this measure shows a reasonable incidence of positive recognition without deteriorating excessively the precision. Table 4.1 compiles the achieved values for F-0.5 as well as the number of true positives (tp), false positives (fp), true negatives (tn) and false negatives (fn).

SHP	Id	F-0.5	tp	fp	tn	fn
EnumOne	1	0.983	577	7	4793	23
EnumTwo	2	0.989	600	9	4791	0
EnumThree	3	0.969	595	25	4775	5
EnumFour	4	0.972	566	11	4789	34
EnumFive	5	0.993	585	1	4799	15
Stop	6	0.984	572	4	4796	28
Fist	7	0.960	586	29	4771	14
OkLeft	8	0.983	575	6	4794	25
OkRight	9	0.995	600	4	4796	0

Table 4.1: Accuracy in intra-frame SHP estimation.

As shown in Table 4.1, the predictions for each SHP indicate different reliabilities (e.g., predictions for EnumOne or EnumTwo are more reliable than predictions for Fist). In order to achieve better results, the use of the temporal context of each frame is proposed (see Section 4.4.3.2).

4.4.3 Dynamic hand gesture recognition

SHPs can be combined with motion in order to obtain a semantically richer dictionary of gestures, the DHGs. Moreover, this approach allows for a more robust recognition of the performed gesture, assuming that a user cannot change the gesture frame-to-frame along a gesture execution.

Notice that in SHP recognition more than a SVM can return a positive output at each frame thus, more than a single SHP might be detected per frame. Some temporal coherence evaluation of the SVM output would help to extract the right sequence of SHPs and, consequently, the performed DHG.

Figure 4.4.2 illustrates this situation for a particular DHG. At each time instant and hand position (the y coordinate corresponds to the center of the ellipse as explained in Section 4.4.3.2) the SHP recognition can yield several results (SHP Ids). From this information the performed gesture has to be identified, corresponding this to a hand performing SHP-Fist and moving slightly down and up (i.e. oscillating y) along twenty five frames.

The recognition of a DHG should then involve an evaluation of its corresponding SHP along time and of the hand motion pattern.

4.4.3.1 Dictionary of dynamic hand gestures

Most DHGs can be understood as a sequence of a single SHP; these are denoted as simple DHGs in which the hand is placed statically or pseudo-statically in front of the camera. Recognizing these gestures entails testing SHP coherence inside the analysis window and hand stillness. In

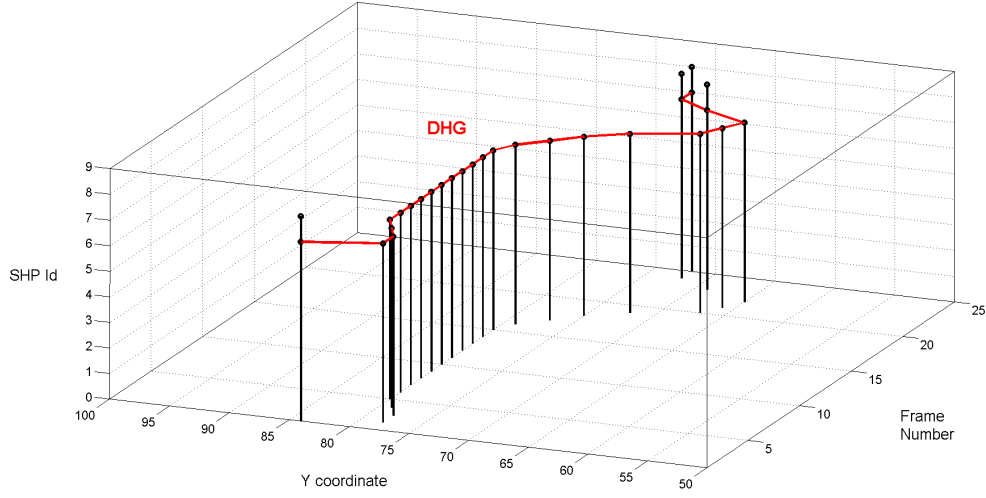


Figure 4.4.2: Execution of DHG Fist performed by training user 3

Figure 4.4.3 some DHGs belonging to this category are shown, including all static gestures. Table 4.2 shows the dictionary of these type of simple DHGs, which allow the user to interact with complex applications in an easy and straight way as explained in [Castilla et al., 2009]. This dictionary could be easily configured for remote controlling a TV menu, in the same line as in [Premaratne and Nguyen, 2007].

DHG	SHPs Id sequence	Motion Pattern
EnumerateOne	1	Partial or totally static
EnumerateTwo	2	Partial or totally static
EnumerateThree	3	Partial or totally static
EnumerateFour	4	Partial or totally static
EnumerateFive	5	Partial or totally static
Cancel	6	Partial or totally static
Fist	7	Partial or totally static
MenuRight	8	Partial or totally static
MenuLeft	9	Partial or totally static

Table 4.2: Simple Static DHGs.

One of the requirements of the targeted applications is to allow users interaction with spreading out menus. In this sense, and also considering the gesture definition studies developed in [Castilla et al., 2009], two additional simple DHGs which additionally involve hand motion are proposed: MenuOpen and MenuClose (see Table 4.3). Figure 4.4.4 depicts the hand evolution for realizations of these two DHGs. SHPs conforming these DHGs widely change along the

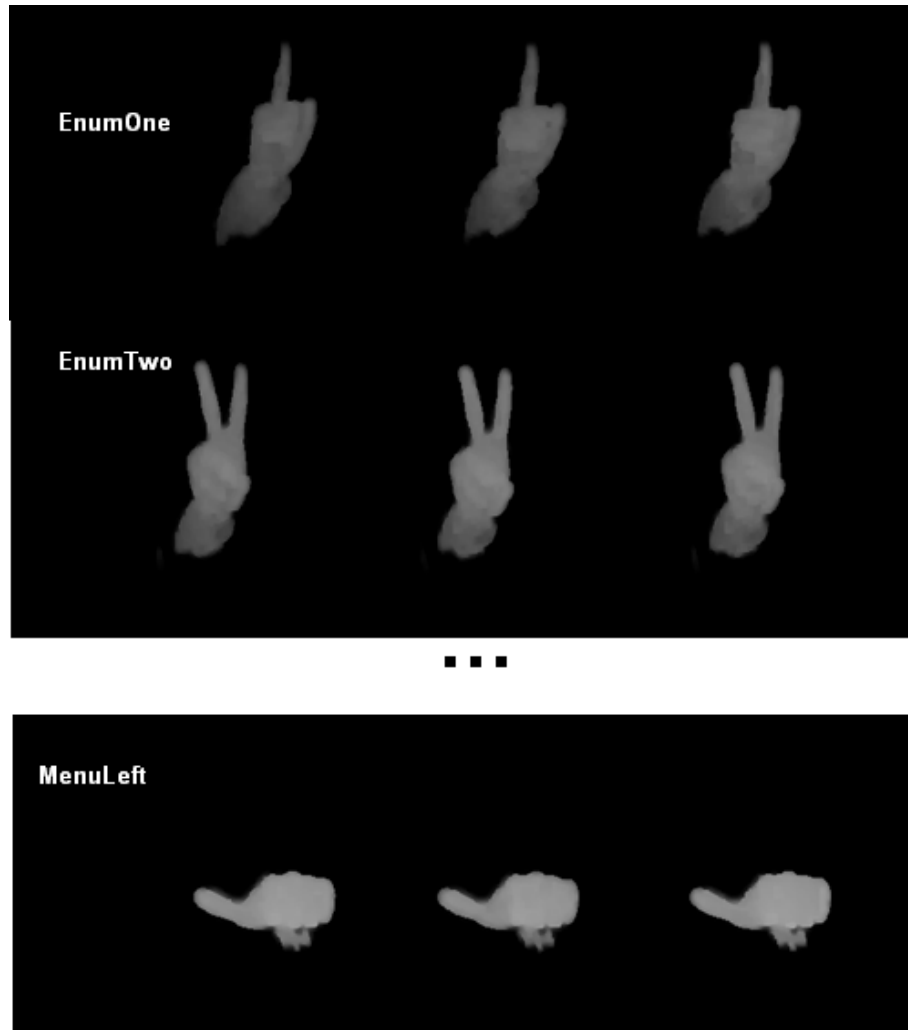


Figure 4.4.3: Examples of static DHGs: its recognition just relies on SHP recognition and hand stillness.

execution. Consequently, SHP coherence is not required to detect these gestures.

DHG	SHPs Id sequence	Motion Pattern
MenuOpen	any	move up
MenuClose	any	move down

Table 4.3: Simple DHGs with a specific motion pattern.

In response to the requirement of catching, grasping and releasing items in an application, the compound DHGs Catch and Release were included, both defined as combinations of simple DHGs: first the interface item is searched; when located, it is caught and dragged to desired position; finally, the item is released. This is modeled with a combination of two DHGs (see

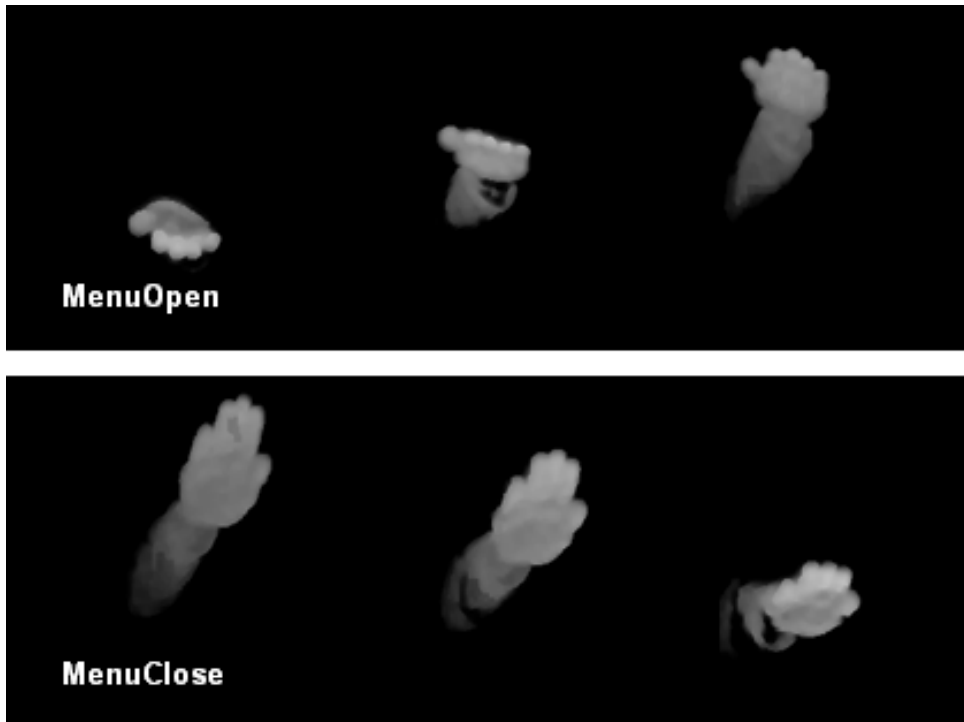


Figure 4.4.4: Examples of simple DHGs involving motion, which requires SHP recognition and a estimation of the hand motion pattern. Notice that these two gestures are the same as N and W in Figure 2.3.2a.

Figure 2.3.2b in Chapter 2): first the user moves with EnumerateFive until the item to take is found; then the user closes his/her hand over it executing DHG Fist. At this point the item is selected and the Catch is detected. The user can make any displacement of the item and finally the system detects Release when the user returns to EnumerateFive. The whole process results in an intuitive and natural gesture. An illustration of this DHG is shown in Figure 2.3.2b (top) (in Chapter 2) and described in Table 4.4.

Following the same line, in response to the requirement of selecting an item, DHG Click was introduced consisting of two simple DHGs: EnumerateOne and Fist (see Figure 2.3.2b- bottom). Its description can be found in Table 4.4.

Compound DHGs are the most challenging to detect. Users are allowed to perform any kind of motion and displacement over the interaction area, obviously including motion patterns like those associated to DHGs MenuOpen and MenuClose. Furthermore, SHPs conforming these DHGs are subjected to transitions between postures and are suitable to be captured as shapes different to those modeled. These situations are considered and controlled by the system as explained in Section 4.4.3.2.

Even though the definition of the DHGs has been application driven, it is important to remember that the interpretation of those DHGs must rely on a higher semantic level of the

DHG	DHGs sequence	Motion Pattern
Catch	EnumerateFive Fist	Any
Release	Fist EnumerateFive	Any
Take	Catch Release	Any
Click	EnumerateOne Fist	Any

Table 4.4: Compound DHGs.

system, as clicks and movements of a mouse are translated to actions by particular applications or operating systems.

The *execution videos*³ (see Section 2.3.2) records of the users not considered in training (i.e. users 4, 5, 6, 7, 8, 9, 10 and 11) were used for the final evaluation (see 4.5.2) while the records of training users (i.e. 1, 2 and 3) are used for motion pattern learning (see Section 4.4.3.2). So, 40 videos per DHG, executed by users are not used in training.

4.4.3.2 Detecting the dynamic hand gestures⁴

In contrast to the learning approach for SHPs (see Section 4.4.2.2), no training is performed in the case of DHGs. The recognition is based on three fundamental information sources:

- the SHPs predictions detailed in Section 4.4.3.2.
- the observed motion behaviors in the DHG collection data, defined in Section 4.4.3.2.
- the definition of the DHG themselves, with their associated restrictions and transitions modeled by the Finite State Machine (FSM) described in Section 4.4.3.2.

Combining these sources of information a gesture recognition at the output of the FSM is obtained.

SHP recognition in a temporal window When a user performs a SHP, it is reasonable to expect him/her to keep it for some frames. In order to take advantage of this temporal redundancy, a first approach could consist on just counting the incidences of each detected SHP within a temporal window, and decide by majority vote on the performed DHG. However, there are two main limitations in the intra-frame SHP recognition process (see Section 4.4.2.2) that can be summarized as follows:

³Available at <http://www-vpu.eps.uam.es/~vision/paper/indexpaper.html>

⁴Thanks to Marcos Escudero-Viñolo for his contributions in the design of the Finite State Machine and the motion-based gestures recognition system.

1. Discrimination capabilities of extracted low-level features may not be enough to separate among the SHPs of the dictionary, specially in compound DHGs, as mentioned in Section 4.4.3.1, and in transitions frames.
2. For every single SHP, the variety derived from different users and scenarios prevents any data collection from being representative enough to model it.

These two limitations might seriously affect SHPs recognition. However, SHPs have been selected to be as different as possible among them, and the obtained results (see Section 4.4.2.2) demonstrate that the indicated limitations can be overcome.

Considering intra-frame statistics compiled in Table 4.1, which indicate that there are SVMs more reliable than others, the probability of having a negative output when introducing a low level feature vector v^j (describing a frame with SHP-j) into a SVM trained with data describing SHP-i ($i \neq j$) can be computed as

$$p^i(0/pred = 0) = \frac{tn^i}{tn^i + fn^i} \quad (4.4.3)$$

and the probability of wrongly detecting it as positive:

$$p^i(0/pred = 1) = \frac{fp^i}{tp^i + fp^i} \quad (4.4.4)$$

The values of these probabilities, calculated for the training *static pose videos* of the gestures described in Section 4.4.2.1, are compiled in Table 4.5.

SHP/id	$p^i(0/pred = 0)$	$p^i(0/pred = 1)$
EnumOne/1	0.995	0.012
EnumTwo/2	1	0.015
EnumThree/3	0.999	0.040
EnumFour/4	0.993	0.019
EnumFive/5	0.997	0.002
Stop/6	0.994	0.007
Fist/7	0.997	0.047
OkLeft/8	0.995	0.010
OkRight/9	1	0.007

Table 4.5: Probabilities of correctly detecting a negative or wrongly detecting a positive.

Notice that reliability of predictions for a SVM trained with SHP-i patterns is better for low values of $p^i(0/pred = 1)$ and for high values of $p^i(0/pred = 0)$. In the light of the differences among SHPs probabilities, it makes sense to treat predictions for each SHP differently. The prediction for this SVM-i (SVM trained with SHP-i patterns) for the frame n is modeled as a

function named $pred^i(n)$. The possible values for this function are ‘0’ and ‘1’ (i.e. negatives and positives).

A temporal window is considered, defined by

$$\Delta T_{n_0} \equiv \{n : n - n_0 < N\} \quad (4.4.5)$$

where n_0 is the frame number in which the temporal window begins and N its duration ($N = 13$ was experimentally adopted)

The probability of occurrence of no positives for within ΔT_{n_0} , is:

$$\begin{aligned} p_{\Delta T_{n_0}}^i(\#pos = 0) &= p_{\Delta T_{n_0}}^i(\#neg = |\Delta T_{n_0}|) \\ &= \prod_{\nabla n \in \Delta T_{n_0}/pred^i(n)=1} p^i(0/pred = 1) \star \prod_{\nabla n \in \Delta T_{n_0}/pred^i(n)=0} p^i(0/pred = 0), \end{aligned} \quad (4.4.6)$$

The above expression corresponds to a product with two differentiated groups of factors: first, the probabilities of being wrong when having detected positives; second, the probabilities of being right when having detected negatives. The whole product, consequently, is the probability of having all negatives within ΔT_{n_0} .

So, the probability of occurrence of, at least, one positive stands:

$$p_{\Delta T_{n_0}}^i(\#pos \geq 1) = 1 - p_{\Delta T_{n_0}}^i(\#pos = 0) \quad (4.4.7)$$

In conclusion, starting from the binary predictions of SHP-i in a temporal window, it is possible to estimate the probability of the incidence of, at least, one positive SHP-i. Computing probabilities for each of the considered SHPs and comparing them, it is possible to estimate the one with higher probability of having been performed within the temporal window.

Motion pattern analysis The trajectory described by the hand at each DHG execution is the main parameter to characterize the MenuOpen and MenuClose DHGs. Furthermore, it can be used as a control parameter to improve the recognition of the simple static DHGs described in Table 4.2. In any case, the definition of a process to extract hand’s trajectory, its velocity and the boundaries among the different trajectories is needed.

Complex processes to extract hand trajectory may result in unfeasible analysis in a real time environment. Fortunately, the evolution of the position of the characteristic points (see Appendix V) extracted for each DHG execution offers a rough description of the hand trajectory.

Selection of point to track Figure 4.4.5 shows the evolution of the Y component of three of these points (the Geodesic Center of the hand, the Ellipse Center, and the Minimum Depth

Point) for different DHGs *execution videos* (see Section 2.3.2) performed by the training users. Each row includes fifteen executions (three training users, five repetitions each).

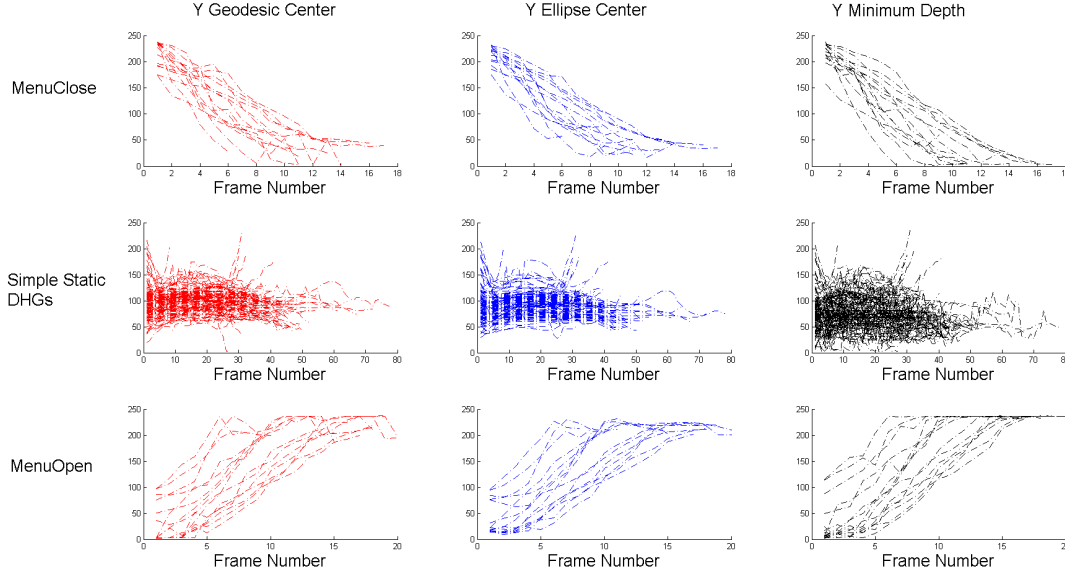


Figure 4.4.5: Stability of the Y coordinate evolution for each of the characteristic points. Three types of DHGs, performed by training users, are evaluated: MenuClose (1st row), Simple Static Gestures(2nd row) and MenuOpen (3rd row).

Although all the points describe correctly the Y motion pattern, the Ellipse Center seems to be the most stable in terms of transition softness, as can be observed in Figure 4.4.5. The standard deviation of the first derivative for the Y evolution of these three points at each DHG execution is computed. Computing the mean of these standard deviations for the training users and for each of the three groups of DHGs indicated in Figure 4.4.5, the following values are obtained: 11.9816 (Geodesic Center), 10.0692 (Ellipse Center) and 11.6709 (Minimum Depth Point). In the light of these results, which supported the observation, the Ellipse Center was selected as the most suitable point to characterize the Y evolution of the hand for the considered DHGs. The same process could be performed to extract coordinate X's evolution and coordinate Z's evolution.

Motion pattern clasification Nevertheless, as DHGs have been defined, coordinate Y's evolution is enough to discriminate between MenuUp and MenuClose and between these and the simple static DHGs (Table 4.2). Furthermore, coordinate Y's evolution can also be used to separate non-stationary parts at the beginning and end of the simple static gestures executions (see second row of Figure 4.4.5), parts modeled as a time percentage, ρ , of the gesture length. In order to discriminate motion patterns (see Section 4.4.3.1), a point trajectory is characterized by

its slope during a temporal window of length W . Figure 4.4.6 shows six slope areas, associated to different motion patterns. The intervals for this slope angle are $[0, \Pi/2)$ and $(-\Pi/2, 0]$ radians, where 0 means staticity and $\Pi/2$ and $-\Pi/2$ instantaneous hand movements, upwards and downwards respectively. The β parameter establishes a margin to categorize a trajectory as static, and the α defines a margin for an upward trajectory and a downward one. The area between the boundaries is tagged as unknown and is destined to some of the unpredictable trajectories of the compound DHGs and to the non-stationary parts of the simple static gestures.

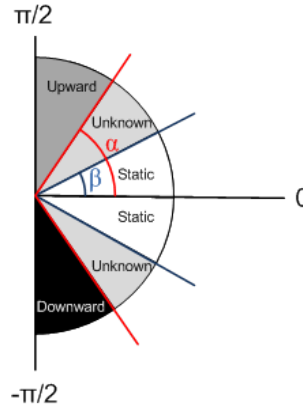


Figure 4.4.6: Y coordinate trajectory characterization

So far two couples of parameters which have to be set were identified: α and W , targeted to control the recognition of upward and downward motion; and β and ρ , targeted to model the non-stationary parts of simple static DHGs. In order to set those parameters different optimization processes were followed for each couple.

First, α and W have been set by optimizing the F-score measure in the recognition of MenuOpen and MenuClose when analyzing trajectories from every training DHG and user (including compound DHGs). A grid search is performed with grid parameters α and W uniformly distributed between the intervals $[0, \Pi/2]$ and $[2, 13]$ and with, respectively, parameter steps $\Pi/120$ and 1. Results are included in Figure 4.4.7 and show the α and W combination that maximizes both DHGs F-score in recognition: $\alpha = 1,388$ rad and $W = 13$ frames.

A similar optimization has been performed to set optimal values for β and ρ . The aim is to discriminate between stationary and non-stationary parts of a simple static gesture execution. For this, it is assumed that there is a percentage of transitory frames at the beginning and the end of each DHG execution. A grid search is performed with grid parameters β and ρ uniformly distributed between the intervals $[0, \Pi/2]$ and $[5\%, 20\%]$ and with, respectively, parameter steps $\Pi/120$ and 1%. The resulting graph is included in Figure 4.4.8 and shows its maximum value for $\beta = 0.739$ rad and $\rho = 8\%$.

Although the speed of gestures can vary from user to user, the parameters chosen have proved

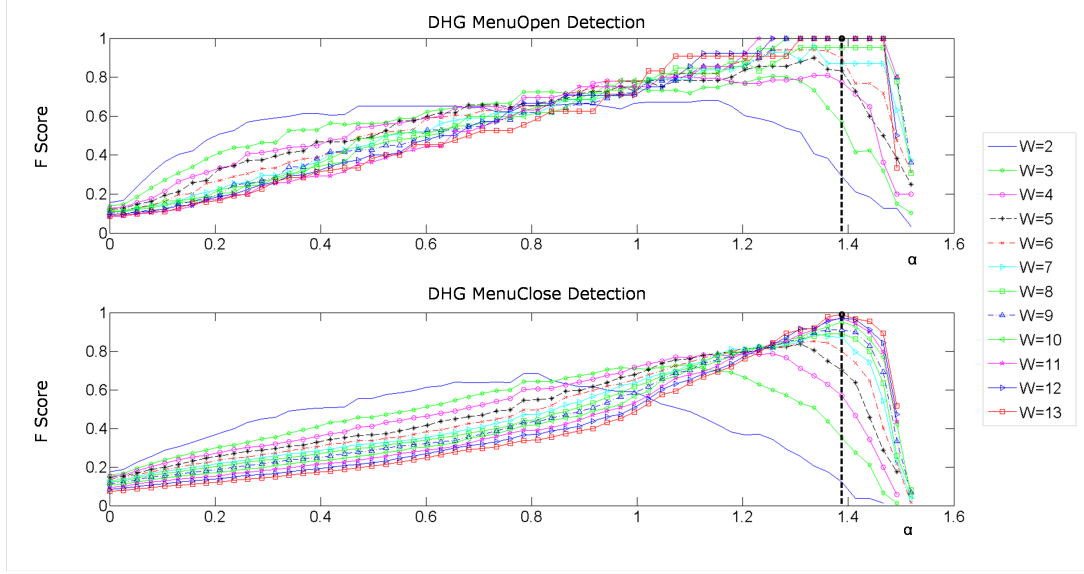


Figure 4.4.7: Grid search of parameters α and W to maximize F_score in the recognition of DHGs MenuOpen and MenuClose.

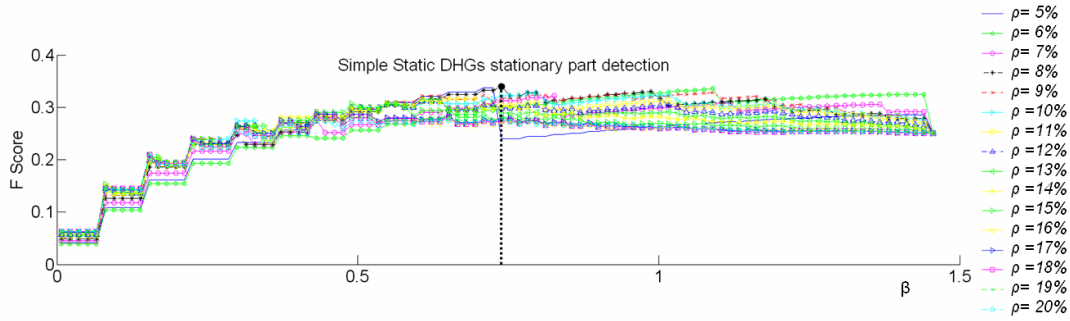


Figure 4.4.8: Grid search of parameters β and ρ to maximize F_score in the recognition of the stationary part of simple static DHGs

to be adequate for the compiled dataset, which contains a wide range of execution speeds, as users were not asked to perform the motion-based gestures with a certain rhythm.

Gesture recognition by means of a FSM The decision on the performed SHP over a frame sequence and the estimated motion pattern are the inputs to the FSM developed for this system. The FSM establishes priorities in the DHG recognition and also avoids forbidden transitions (whichever different from the ones described in Section 4.4.3.1).

As opposed to the FSM used by [Hong et al., 2000], where it constitutes the main strategy to recognize hand gestures, the proposed FSM works as a supervisor module; its specific functions are:

- To control that just one DHG is declared at each gesture execution.
- To apply restrictions according to the estimated motion pattern (see Section 4.4.3.2).
- To model transitions in the execution of compound DHGs.
- To discard unconsidered DHGs.

It is important to notice that except for compound DHGs, just one DHG is returned by the system, immediately after recognition and at each gesture execution. No gesture might be detected when the system is deactivated, i.e., when the hand is out of the interaction area. The FSM fulfills this restriction by remaining in a deactivation state. FSM deactivation, which can occur at any time, forces the system to decide on a DHG. If the system is unable to detect one, it returns Unknown.

The recognition of the compound DHGs Take and Click is also a task of the FSM (see Figure 4.4.9), which in this case needs to keep track of previously detected DHGs. It is important to remember that the estimated motion of these gestures does not need to fit any specific pattern.

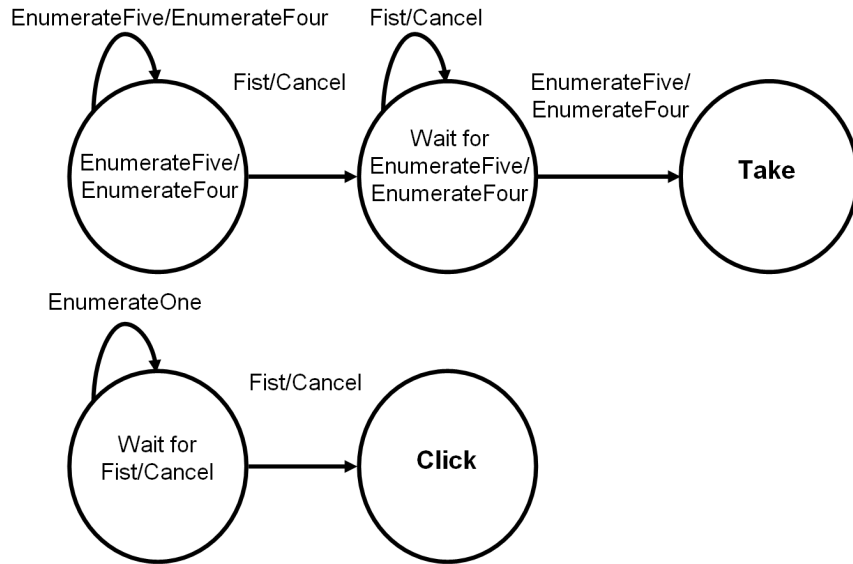


Figure 4.4.9: FSM transitions and conforming DHGs in the execution of DHGs Take and Click.

Let us define a conforming DHG as a DHG that has meaning by itself but it is also part of a compound DHG. For instance, DHGs EnumerateFive and EnumerateFour are gestures (see Section 4.4.3.1) and also conforming DHGs of DHG Take. The same applies to EnumerateOne, to Fist and to Cancel, which are considered as conforming DHGs just if any of the other conforming DHGs has been detected before. If any of the aforementioned conforming DHGs is detected, the system delays its declaration waiting for a subsequent recognition of a compatible conforming DHG. If this occurs, then the FSM advances in the state and either finally declares the compound DHG or starts waiting for the next conforming DHG. Otherwise, the system returns the initial

conforming DHG when one of these situations occurs:

- i. A deactivation takes place.
- ii. The last detected conforming DHG is repeatedly detected: a time-out counter was set in order not to force the user to deactivate the system to recognize isolated conforming DHGs. Experimentally this counter was set to a time equivalent to three non-overlapping temporal windows ($TO = 3N$), which resulted to be an adequate time for the users asked to perform the evaluation.

Regarding to the estimation of the hand trajectory, if it is categorized as Unknown Trajectory, no gesture will be declared; this aims to avoid wrong recognitions caused by unconsidered hand movements where SHP may be not modeled. Exceptions to this rule are compound DHGs in which any motion pattern is valid (see Table 4.4). If the trajectory is tagged as Static, the system could return any of the DHGs compiled in Table 4.2 but also any of the compound DHGs. If the angle that describes the trajectory is over α (Upward) or under $-\alpha$ (Downward), the DHGs MenuOpen and MenuClose are declared, independently of the SHP sequence, if and only if none of the conforming DHGs have been previously detected (see Section 4.4.3.2). Finally, the FSM configuration forces the system to return Unknown when user executes forbidden actions as trying to execute a new gesture without deactivating the system.

FSM additional concerns FSM transitions for taking and clicking (see Figure 4.4.9) are specially problematic since the executions of DHGs Take and Click occur indistinctly in all the screen area. This produces the loose of the thumb for several cases, producing for example the recognition of SHP EnumFour instead of EnumFive (this is due to the fact that the fingers are expected to point up to be correctly detected [Molina et al., 2011]) and Stop instead of Fist (the segmentation process described in Section 4.4.1 does not completely eliminate the forearm when execution takes place near the corners of the display, making these two SHPs more similar). This is the reason for equating the associated DHGs to the mentioned SHPs: EnumerateFour (as EnumerateFive) and Cancel (as Fist).

4.5 Experiments

4.5.1 Experimental setup

This section presents two different evaluation scenarios for DHG recognition: user independent and non user independent. For this purpose the real users *static pose videos* and the *execution videos* described in Section 2.3.2 are used. The hand is modeled with descriptor described in Appendix V.

1. User independent: in this evaluation scheme, the *static pose videos* of three users are used in the training stage while the *execution videos* of the other eight users are used in the

	DHG predictions													
DHG	U	1	2	3	4	5	6	7	8	9	10	11	12	13
EnumerateOne (1)	1	39	0	0	0	0	0	0	0	0	0	0	0	0
EnumerateTwo (2)	0	0	35	0	0	0	0	0	1	4	0	0	0	0
EnumerateThree (3)	0	0	4	31	0	3	0	0	0	2	0	0	0	0
EnumerateFour (4)	5	0	0	0	31	2	0	0	1	0	0	0	1	0
EnumerateFive (5)	0	0	0	0	1	37	0	0	0	0	0	0	2	0
Cancel (6)	0	0	0	0	0	0	35	4	0	0	0	1	0	0
Fist (7)	0	0	0	0	0	0	1	38	1	0	0	0	0	0
MenuOpen (8)	0	0	0	0	0	0	0	0	40	0	0	0	0	0
MenuClose (9)	0	0	0	0	0	0	2	1	0	35	1	0	0	1
MenuLeft (10)	0	0	0	0	0	0	0	0	1	0	39	0	0	0
MenuRight (11)	0	0	0	0	0	0	0	0	0	0	0	40	0	0
Take (12)	7	0	0	0	0	0	0	1	0	0	0	0	32	0
Click (13)	0	1	0	0	0	0	0	2	0	0	1	0	0	36

Table 4.6: User Independent Confusion Matrix for proposed DHGs.

evaluation.

2. Non user independent: the *static pose videos* of all the users are used for training the system, the *execution videos* for these same users are used in the evaluation.

The resulting accuracy rates for several dictionaries are compared with the State of Art for user and non user independent works.

4.5.2 Results

In this section the results for DHG evaluation are presented. Notice that the used data set is not the same one as the ones used in the papers with which the comparisons are performed, new training (*static pose videos*) and evaluation (*execution videos*) video collections were recorded (see Chapter 2).

The evaluation results for the DHG data collection are compiled in the confusion matrix of Table 4.6 where identifiers of the DHGs are listed in rows while final predictions of the system are included in columns (column labeled as “U” correspond to Unknown). From the confusion matrix the achieved accuracy rate can be calculated, 0.900, an encouraging value taking into account the number of DHGs separated, 13, the number of evaluation executions compiled, 40 per gesture, and the quality of separations described in some similar works. Considering only static gestures the obtained accuracy is 0.939.

From the results compiled in Table 4.6 there are several aspects subject to improvement:

1. Some executions of EnumerateThree are wrongly classified as EnumerateTwo. This is

because sometimes the thumb is not detected as a prominence since the system expects fingers to point up, due to the limitations of the used descriptor (see Appendix V).

2. Sometimes the Static motion pattern is not detected for DHGs for which it is mandatory. This reverts in some misclassifications: EnumerateFour detected as Unknown.
3. The recognition of non considered DHGs in some DHG Take executions (see Section 4.4.3.2) produces various Unknown recognitions.

System→	Proposed	Proposed (static)	[Teng et al., 2005]	[Zaki and Shaheen, 2011]	[Kelly et al., 2010]	[Malassiotis and Strintzis, 2008]	[Kollorz et al., 2008]
Accuracy	0.900	0.939	0.906	0.891	0.918	0.872	0.946
# gestures	13	7	30	30	10	9	12

Table 4.7: User independent gesture recognition works comparison.

A strict comparison with the State Of Art systems is difficult, because each system uses a different set of gestures and a different database to compute the results, with different number of users and executions per gesture. However, for the sake of completeness, in Table 4.7 the results achieved in the literature for some user independent systems are compiled. In [Teng et al., 2005], the obtained accuracy rate for 30 static gestures is 0.906. In [Zaki and Shaheen, 2011], separating 30 different gestures they obtain an accuracy of 0.891 using 90 repetitions for training and 30 for test. [Kelly et al., 2010] manages to get an accuracy of 0.918 separating 10 static hand postures without facing the segmentation problem (i.e. using a color glove or a fixed and static background). In [Malassiotis and Strintzis, 2008], assuming a scenery similar to the proposed (referred as 'Session B' in that paper) and separating 9 static gestures, the best obtained result is 0.872. In [Kollorz et al., 2008], 12 gestures are separated and evaluated with 34 executions per gesture, achieving an accuracy rate of 0.946; but it is important to point out that none of the considered gestures are compound or present a specific motion pattern.

A comparison with non user independent system has also been performed. Using the whole set of gestures in non user independent context, the systems are trained with the SHPs records (i.e. *static pose videos*) of all users and evaluated with the DHGs executions (i.e. *execution*

videos) for all the users, as well. The obtained accuracy for the proposed gestures raises from the 0.900 obtained in Table 4.6 to 0.933. In order to compare the proposed system with the State of Art, it has also been evaluated in a non user independent context using the gestures proposed in [Soutschek et al., 2008]. The resulting confusion matrix can be found in Table 4.8. The overall accuracy is 0.971, slightly better than the one reported in [Soutschek et al., 2008] for the best set-up: 0.943, obtained in a non user independent approach. A compilation of the results obtained for different non user independent systems is provided in Table 4.9.

DHG [Soutschek et al., 2008]	DHG predictions					
	U	1	2	3	4	5
Translation (1)	0	49	1	0	5	0
Cursor (2)	0	0	53	1	0	1
Click (3)	0	0	0	55	0	0
Rotation (4)	0	0	0	0	55	0
Reset (5)	0	0	0	0	0	55

Table 4.8: Non User Independent Confusion Matrix for another collection of DHGs.

System→	Proposed	Proposed	[Soutschek et al., 2008]	[Keskin and Akarun, 2009]
Accuracy	0.933	0.971	0.943	0.941
Dataset	Proposed	[Soutschek et al., 2008]	[Soutschek et al., 2008]	[Keskin and Akarun, 2009]

Table 4.9: Non User independent gesture recognition works comparison.

4.5.3 Computational cost

The computational cost of the different stages of the system has been measured, resulting in the execution times compiled in Table 4.10, which have been measured on an Intel(R) Core(TM)2 Duo CPU E7500 @ 2.93Ghz with 2.98GB RAM. Notice that these processing times allow the system to work up to 33.9 fps, enabling real-time human-computer interaction.

	Segmentation	Desc. (Molina et al. [2011])	Classification
msec	16.114	4.311	9.054

Table 4.10: Computational Cost (msec) per frame

4.6 Conclusions

A non intrusive system for the recognition of hand gestures based on a TOF camera has been presented in this chapter. It is able to work in realtime as it has been measured (see Section 5.5.3). Gesture classification is based on features which are based on extraction of crucial points

of the silhouette using the geodesic distance to the center of the segmented hand. Three different types of hand gestures are considered: simple, compound and based on motion pattern. The system has been evaluated with a significant number of users, obtaining user independent results that improve the ones reported in the State Of Art for simple gestures. The proposed system shows remarkable performance even when comparing to non user independent systems. In terms of usability, the system properly works in real-time with a low response time, allowing the interaction with application interfaces.

In the light of the results described in Section 4.5.2 three main future work lines are considered: the improvement of the hand segmentation, which would be useful for solving the forearm elimination in the outermost areas of the screen; the improvement of the hand descriptor, to avoid the need of having fingers pointing up to obtain a proper description; and the integration of a Hidden Markov Model based solution for making the system more robust to noisy SHPs recognitions.

Chapter 5

Motion-based hand gesture recognition using synthetic trajectories

5.1 Introduction¹

The ultimate goal of this work is to provide the user with a natural interaction and a good experience when interacting with a computer in contexts of application such as the interaction with maps², allowing intuitive movements of the earth surface. Other contexts of application of this approach can be the control of multimedia menus [Soutschek et al., 2008] or the point of view on a virtual environment. Other motion based gestures recognition could allow the interpretation of sign languages [Holden et al., 2005, Kelly et al., 2010].

Starting from the captured range data, a Point Of Interest (POI) is defined and computed in a temporal window in order to describe the trajectory of the user's hand. This captures are then compared to synthetically generated trajectories covered at different speeds in order to recognize the proposed motion-based gestures collection.

Two POIs are taken under consideration along this Chapter: the nearest point to the camera and the geodesic center of the binary mask resulting of limiting the range of depth information. The comparison of the modelled and captured trajectories is performed using Dynamic Time Warping (DTW), a distance that offers more robustness than euclidean to phase lag between compared temporal signals. The final detected gesture for a input sequence depends on the nearest synthetic motion pattern to the one captured.

The chapter is structured as follows: In Section 5.2 the State Of Art is exposed and the innovations of the proposed system are pointed out before giving an overview of it in Section 5.3. In Section 5.4 the proposed dictionary of gestures and the compilation of users executions is

¹This chapter is based on : *J. Molina, J. A. Pajuelo and J. M. Martínez, "Real-time Motion-based Hand Gestures Recognition from Depth Sensor video," IEEE Transactions on Consumer Electronics (under review).*

²Atlas Gloves: A DIY Hand Gesture Interface for Google Earth, <http://atlasgloves.org/about>

described for later presenting the approach followed for its recognition is explained. In Section 5.5, the significant user-independent evaluation figures are presented for later enumerating the achieved conclusions in Section 5.6.

5.2 Related work

The evolution of the capturing technologies is important when talking about HCI, a main difference should be done: RGB based solutions [Stenger et al., 2006, Chen and Tseng, 2007, Nickel and Stiefelhagen, 2007, Zheng et al., 2007, Teng et al., 2005, Lee and Park, 2009, Kelly et al., 2010, Wenjun et al., 2010]; 2.5D [Molina et al., 2011, Doliotis et al., 2011, Yang et al., 2012] and 3D solutions [Keskin and Akarun, 2009, Usabiaga et al., 2009, Cheng et al., 2010]. In relation with the intrusiveness of the capturing solutions, as mentioned in Chapter 1, there are several intrusive solutions for obtaining 2.5D or 3D information [Keskin and Akarun, 2009, Usabiaga et al., 2009, Cheng et al., 2010], while TOF technology is getting more importance in the last years [Soutschek et al., 2008, Kollorz et al., 2008, Molina et al., 2011].

There are several works which focus on the detection of motion pattern based gestures. In [Wenjun et al., 2010], a system for the detection of shape and motion based gestures is presented, using 2D images as input. It is evaluated for four different gestures, but only two different trajectories. [Yoon et al., 2001] recognizes 26 alphabetical gestures on the basis of features of location, angle and velocity. In [Cheng et al., 2010], based on 3D motion captures obtained with an accelerometer, digits 0 to 9 drawn to the air are recognized. [Kim et al., 2008] presents a solution based on neural networks fed with spatiotemporal information. In Chapter 2 several dictionaries are proposed, one of them is the one used in this study (see Section 5.4.2). Notice that two of the gestures under account (i.e. N and S) are also used in [Molina et al., 2011] (see Chapter 4). Some recognition solutions based on the Kinect sensor³ have been proposed in the last years: in [Doliotis et al., 2011] numbers drawn in the air are recognized with a Nearest Neighbour scheme; in [Yang et al., 2012] eight motion based gestures are separated using a HMM solution. In [Chai et al., 2010] a kinematics chain model for upper body is proposed for defining synthetic human body gestures based on body parts position.

In this chapter, a novel non intrusive (i.e. there is no need of gloves or markers like in [Kelly et al., 2010, Keskin and Akarun, 2009, Usabiaga et al., 2009] or accelerometers like in [Cheng et al., 2010]) real-time approach to the detection of intuitive motion based gestures usable in different application contexts is presented. The learning phase of the proposed approach does not need the capture of ground-truth real data, since the patterns are defined synthetically by using a human arm model (see Section 5.4.3) making it user independent (differently to [Wenjun et al., 2010, Yoon et al., 2001, Cheng et al., 2010, Kim et al., 2008, Yang et al., 2012]). This

³Microsoft Corp. Redmond WA. Kinect for Xbox 360.

kinematic model is similar to part of the one proposed in [Chai et al., 2010], but in our case it is used to define hand trajectories rather than position based configurations. In [Doliotis et al., 2011] only two test users are considered while in the proposed work, eleven users were asked to record the gestures collection. During evaluation, performed with the collaboration of several users (see Section 2.3.2), the system worked properly, as the results confirm (see Section 5.5). Thanks to the proposed normalization (see Section 5.4.6) and the representativity of the chosen arm model (see Section 5.4.3), the system is robust to variations in the distance to the camera, in the height of the user and in the size of arm and hand. The use of TOF technology, apart from providing an accurate segmentation robust to low illumination conditions (not as in color camera based systems [Stenger et al., 2006, Chen and Tseng, 2007, Nickel and Stiefelhausen, 2007, Zheng et al., 2007, Teng et al., 2005]), offers a representative point of the hand motion, the closest one to the camera, with no need of application of traditional segmentation techniques.

5.3 System overview

In Figure 5.3.1, an overview of the system is presented:

- (i) acquisition of depth images, first of all, the depth data range is limited to a maximum distance of 3 meters, as explained in Section 2.3.2.
- (ii) feature extraction, the Point Of Interest (POI) to be tracked is computed, storing its coordinates from frame to frame (*i.e.* each p_i represents the 3 coordinates of the POI at frame i) which are an estimation of the hand trajectory. More concretely, the proposed POI is the point detected closest to the camera. An alternative POI is also proposed for evaluating purposes, this is the geodesic center of the segmented hand mask (see section 5.4.5).
- (iii) motion patternmodelling, synthetically generated motion patterns (*i.e.* each ξ_{ai} represents the coordinates of pattern associated to gesture a at sample i) are generated on the basis of a proposed arm model (See Section 5.4.3).
- (iv) motion patterns recognition, five samples trajectory segments (*i.e.* four translation segments) are compared with the synthetically generated motion patterns. using the Dynamic Time Warping (DTW) distance as explained in Section 5.4.6. So, each translation segment will be locally labeled with the closest synthetic pattern. This results, along a gesture execution, in a collection of assigned labels to several translation segments. The final label of the gesture will be the most common assigned label.
- (v) communication with the GUI, the recognized motion patterns are sent to the GUI allowing the control of the application under consideration.

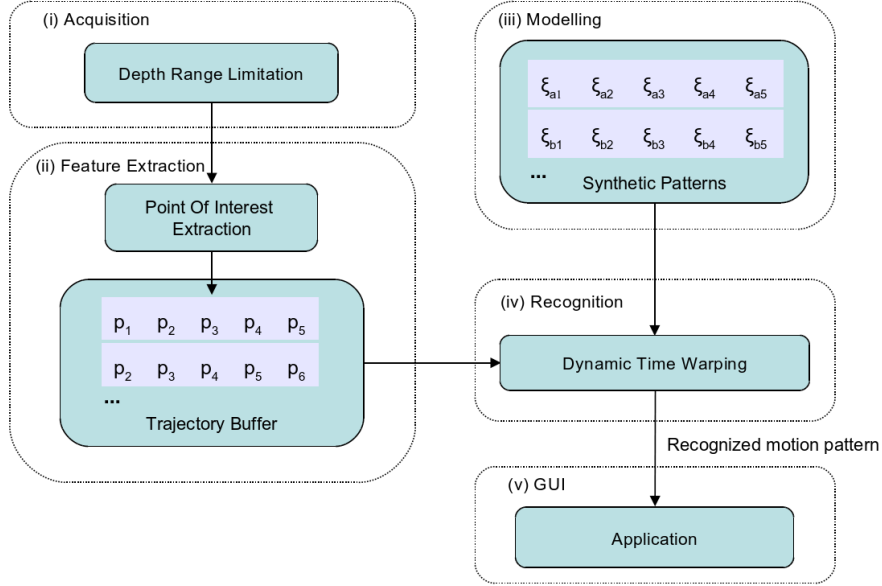


Figure 5.3.1: Overview of the system.

5.4 Hand gesture recognition approach

5.4.1 Introduction

The proposed approach consists of the definition of synthetic motion patterns (see Section 5.4.4) on the basis of a human arm model (see Section 5.4.3). This collection of synthetic motion patterns will be compared with the hand motion estimations (see Section 5.4.5) computed from the natural data *execution videos* (see Section 5.4.2).

5.4.2 Dataset

The dictionary of gestures is proposed following usability criteria: slaps executed in different directions are an intuitive way of interacting with a virtual environment. Two usability objectives [ISO9241-11, 1998] have been taken into account in the gestures selection process: learnability and minimization of support requirements. In terms of learnability, it can be said that none of the users showed difficulties in learning the dictionary and that they only required of a brief introduction: they were asked to perform the indicated gestures as if they were interacting with a menu environment. In terms of minimization of support requirements, it can be said that no user presented doubts about how to execute the gestures.

Nine gestures with clear motion patterns independently from the hand pose were selected (see Figure 5.4.1): slaps in 8 directions (named as the cardinal directions: N, NE, E, SE, S, SW, W and NW) and one slap getting closer and further to the camera (named IO, Inwards-Outwards).

The *execution videos* described in 2.3.2 make a total of 495 videos⁴ (11 users, 9 gestures and 5 repetitions per user and gesture). This collection is entirely used for evaluation purposes, since the knowledge used by the detection system is expressed by the motion patterns defined via the arm model described in Section 5.4.3. The recorded users were not asked to keep a certain distance to the camera neither to perform the gestures with any speed restriction. As well, the users had different heights, what makes the collection certainly representative of the potential users of the system. Some captures of this data set can be found in Figure 2.3.2a in Chapter 2..

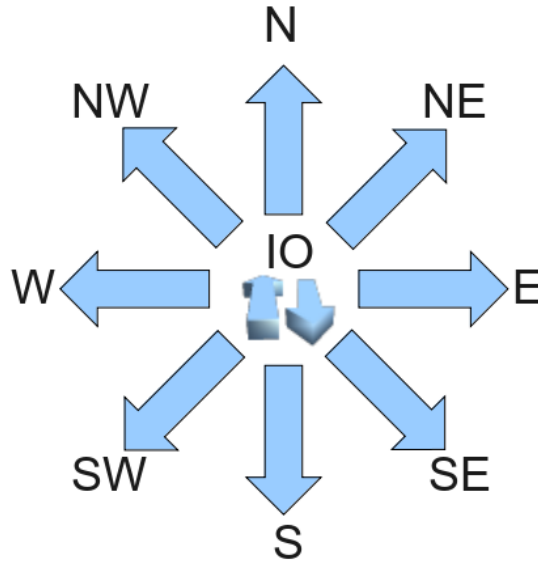


Figure 5.4.1: Gestures observed from user's point of view.

5.4.3 Motion pattern modelling

An arm model, responding to human anatomy, has been proposed for the definition of the considered motion patterns. Two arm segments are considered (see Figure 5.4.2): the upper arm represented by the vector \vec{r}_U which goes from the shoulder to the elbow and the lower arm represented by \vec{r}_L , from the elbow and to the wrist. The hand is not considered explicitly in this model, since the variation that could introduce is non significant in comparison with the ones shown by the arm movements. The lengths for these upper and lower segments were defined with fixed length: $|\vec{r}_U| = |\vec{r}_L| = 1$. Finally, the vector that describes the trajectory of the wrist to be analyzed is $\vec{r} = \vec{r}_U + \vec{r}_L$. In Figure 5.4.2 some set-ups of the arm model are shown. Notice that for a variation of $\Delta\theta$ in angles θ^x and θ^y for the upper segment, the lower segment presents

⁴<http://www-vpu.eps.uam.es/publications/papermotion/indexpaper.html>,
(user: vision, password: visionpaper)

a variation of $2\Delta\theta$, accumulating this way the variation of the upper segment. The expression of the vectors $\vec{r_U}$ and $\vec{r_L}$ are the following:

- For gestures N and S (see Figure 5.4.2a):

$$\vec{r_U} = [0, -\sin(\theta^x), \cos(\theta^x)]$$

$$\vec{r_L} = [0, -\sin(2\theta^x), \cos(2\theta^x)]$$

where $\theta^x \in [0, \pi/2]$. For gesture N θ^x goes from $\pi/2$ to 0, while for gesture S from 0 to $\pi/2$. Notice that these two motion patterns are contained in plane yz and that the two gestures only differ in the direction of execution: N begins with the hand in front of the chest, while S with the arm pointing down.

- For gestures E and W (see Figure 5.4.2b):

$$\vec{r_U} = -\sin(\psi_0) \left[\cos(\theta^y), \frac{\cos(\psi_0)}{\sin(\psi_0)}, -\sin(\theta^y) \right]$$

$$\vec{r_L} = [-\cos(2\theta^y - \pi/2), 0, \sin(2\theta^y - \pi/2)]$$

where $\theta^y \in [\pi/4, 3\pi/4]$ and $\psi_0 = \frac{25^\circ \times \pi \text{rad}}{180^\circ}$. ψ_0 is the angle formed by the upper segment of the arm and $-\hat{y}$. For gesture E θ^y goes from $3\pi/4$ to $\pi/4$, while for gesture W from $\pi/4$ to $3\pi/4$. Notice that these two motion patterns are contained in plane xz, only differing in the direction of execution: E, from left to right; W, from right to left.

- For NE, SE, SW and NW : a rotation about the z axis is performed over the gestures N and S (see Figure 5.4.2c). This rotation matrix, R , is:

$$R = \begin{bmatrix} \sin(\theta_0^z) & \cos(\theta_0^z) & 0 & 0 \\ -\cos(\theta_0^z) & \sin(\theta_0^z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and so, the homogenous coordinates for vectors $\vec{r_U}$ and $\vec{r_L}$ are:

$$\vec{r_U^{hom}} = R \times [0, -\sin(\theta^x), \cos(\theta^x), 0]'$$

$$\vec{r_L^{hom}} = R \times [0, -\sin(2\theta^x), \cos(2\theta^x), 0]'$$

where $\theta^x \in [0, \pi/2]$, as for gestures N and S, $\theta_0^z = \pi/4$ for gestures NW and SE and $\theta_0^z = 3\pi/4$ for gestures NE and SW. The application of these rotation matrixes implies that

the modelled patterns are contained in the plane xz rotated about the z axis.

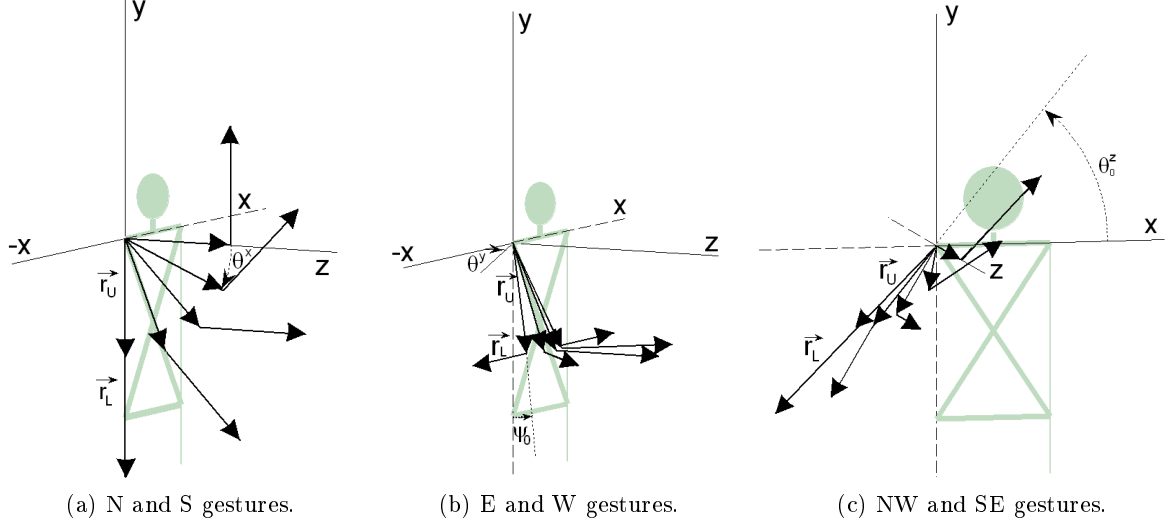


Figure 5.4.2: Model set-ups of the arm model. \vec{r}_U is a vector that goes from the shoulder to the elbow and \vec{r}_L from the elbow to the wrist. The angles θ^x and θ^y are variables which define the trajectory of the arm in 5.4.2a and 5.4.2b, while ψ_0 and θ_0^z are fixed angles that define the position of the elbow at the beginning of the execution of the movement in Figure 5.4.2b and Figure 5.4.2c respectively. ψ_0 is the angle formed by \vec{r}_U and $-\hat{y}$ (see 5.4.2b). θ_0^z indicates the rotation angle applied to N and S gestures, which results in the set-up shown in 5.4.2c.

5.4.4 Motion pattern definition

The direction in which the defined intervals are covered depends on the direction of execution of the specific gesture, for example, in the case of gesture N θ^x for \vec{r}_U begins in 0 and ends in $\pi/2$, while for gesture S is the other way around. In order to consider different speeds in the execution of the gestures 6 different patterns per gesture are presented: 1 for the whole arc, 1 for each half and 1 for each third. This makes 6 synthetic patterns per gesture. The selected length for these patterns was 5 samples (i.e. 4 translation segments) what defines the temporal window used for the comparison of synthetic and real patterns (see Figure 5.3.1).

For the definition of the IO synthetic pattern no angles or arm model were considered, just a simpler approach was followed: the pattern was defined as a sequence of movements in the z axis. Three kinds of translations segments (i.e., an homogeneous motion interval) were considered: I, translation getting closer to the camera; O, moving away from the camera; S, staticity between two frames (applying the normalization described in Section 5.4.6 spurious

translations are considered as staticity). Following the line of considering different execution speeds, various motion patterns (composed by 4 translation segments) were defined: IIII, IIS, IISS, SSOO, SOOO, OOOO, IIIO, IIOO and IOOO. For example, if the execution of the gesture is very fast and only 5 samples are captured during it, the expected segments pattern would be IISS or SSOO. While, if the execution is slower sequences such as IIII or OOOO could be detected.

5.4.5 Motion pattern capturing

In order to capture a representative trajectory of the hand motion it is important to choose an easily traceable point. An inestable point would present noisy translations that could produce wrong estimations of the hand motion. The use of range information provides us with a robust to illumination and easy to detect POI, the closest to the camera. For the detection of this point it is not even necessary to previously segment the image.

With the intention of showing the advantages of using depth information, an approach that makes no use of depth information (except for the depth range limitation) is also presented: it extracts the tracking point considering the segmentation mask image resulting from the depth range limitation as binary (considering foreground all the pixels of the depth image with value over zero). In this case, the chosen tracking POI is the geodesic center of the binary mask, which is estimated by performing the ultimate erosion [Lantuejoul and Maisonneuve, 1984] up to a point.

5.4.6 Patterns comparison

For calculating the distance between two patterns a previous normalization is performed, consisting of setting to one the length of each displacement between two sucesives samples frames of the POI. This solution has been used in problems such as hand writing recognition [Hu et al., 1996] or motion hand based gestures detection, like in [Wenjun et al., 2010] where the length of the translations is not used as a feature, something equivalent to fixing their length. In order to filter spurious errors in the detection of the tracked point when it is static (for gesture IO), this normalization is only applied when the magnitude of the translation of the POI between consecutive frames is over the third of the maximum one within the gesture execution. This defines an enough wide range of speeds for the proposed gestures which are intuitively executed in an homogenous way. The presented normalization makes the system independent to variations in the distance to the camera, in the angle of view, in the heigth of the user and in the size of the arm.

Once the synthetic (see Section 5.4.2) and captured motion patterns (see Section 5.4.5) are normalized, they are compared. The Dynamic Time Warping (DTW) distance has shown good performance when comparing temporal patterns executed at different speeds, concretely it has

been widely applied to speech recognition problems [Sakoe and Chiba, 1978]. An example of its application to hand gesture recognition can be found in [Wenjun et al., 2010]. Notice that each new captured motion pattern has four translation vectors, which describe the hand trajectory for five frames. It is then compared using DTW with each of the synthetic motion patterns present in the collection described in Section 5.4.3. This way we obtain a histogram of incidence of the closest synthetic patterns to this new captured motion pattern. The most common one gives us the label to assign to the gesture capture. If there is a tie between labels, the label 'Unknow' is the one assigned.

5.5 Experiments

5.5.1 Experimental setup

This section presents two different evaluation scenarios, both of them user independent since the learning process is performed using synthetic data and the evaluation is done with 11 different users (see Section 5.4.2):

1. 2.5D scenario: the tracked POI is the closest point to the camera and its depth coordinate (apart from x and y coordinates) is used for modelling the trajectory.
2. 2D information scenario: this second scenario was set-up considering the input images as binary masks as explained in Section 5.4.5. The depth information is implicitly used in the set-up of the camera (see Section 5.4.2), resulting in a segmentation mask, but this information is not used in the estimation of the hand trajectory. In this case, the tracked POI is the geodesic center of the binary mask, obtained with an iterative algorithm process [Molina et al., 2011]. Although the depth information is used for the calculation of this mask, the z coordinate is not used in the comparison of the patterns.

The comparison of the results obtained for these two set-ups will permit to obtain conclusions about the utility of using depth information in hand gesture recognition.

5.5.2 Results

This section compiles the results obtained for the two evaluation scenarios introduced in section 5.5.1:

1. 2.5D scenario: the resulting confusion matrix can be found in Table 5.1. The obtained accuracy rate is 0.951.
2. 2D information scenario: The obtained accuracy rate is 0.780 (see Table 5.2).

From the results compiled in Table 5.1 there are several aspects to point out:

- The label IO is the one assigned more times erroneously. It introduces 10 false negatives for executions of other gestures. This is due to the fact that users tend to introduce the hand in the interaction area (and move it away) with upward and downward trajectories. These patterns are present in the definition of other gestures, apart from IO, producing misclassifications.
- When the assigned labels within an execution results on the same score for 2 or more gestures the assigned label is *Unknown* (U). This situation produces 7 misclassifications.
- Without taking into account the misclassifications produced by the inclusion of the IO gesture (i.e. the only one which translation is fundamentally takes place in the depth coordinate), the obtained accuracy rates are, 0.873 for the 2D scenario and 0.977 for the 2.5D one. So, the use of depth information improves the results even when the gestures are apparently detectable using only 2D information.

Table 5.2 presents not such good results, mainly due to the instability of the geodesic center. Since no depth information is considered, the representative point to be tracked needs to be estimated on the basis of a segmentation which is noisy due to variation in its shape and size. So, noisy translations are added to the real translations of the hand.

	U	N	S	W	E	SW	NW	SE	NE	IO
N	0	52	0	0	0	0	0	0	0	3
S	1	0	50	0	0	0	0	0	0	4
W	1	0	0	53	0	0	0	0	0	1
E	0	0	0	0	55	0	0	0	0	0
SW	0	0	0	0	0	55	0	0	0	0
NW	2	2	0	0	0	0	51	0	0	0
SE	2	0	0	0	0	0	0	51	0	2
NE	1	0	0	0	1	0	0	0	53	0
IO	0	1	0	2	0	0	0	0	1	51

Table 5.1: Confusion matrix for the 2.5D scenario. Gestures described in Section 5.4.2 and “U” for Unknown.

No user-indepent evaluations performed for motion based gestures detection were found in the State Of Art, consequently the evaluation figures of some works in which the absence of overlap between train and evaluation corpora is not ensured are enumerated. In [Wenjun et al., 2010], a 0.97 accuracy rate is obtained in separating only two motion patterns. [Cheng et al.,

	U	N	S	W	E	SW	NW	SE	NE	IO
N	0	51	0	0	0	0	0	0	0	4
S	1	0	26	0	0	1	0	0	0	27
W	1	0	0	37	0	0	16	0	0	1
E	0	0	0	0	38	0	0	6	9	2
SW	0	0	0	1	0	47	1	0	0	6
NW	2	4	0	2	0	0	44	1	0	2
SE	2	0	0	0	0	0	0	49	0	4
NE	1	2	0	0	0	0	0	0	51	1
IO	0	1	0	1	0	0	7	1	2	43

Table 5.2: Confusion matrix for the 2D scenario. Gestures described in Section 5.4.2 and “U” for Unknown.

2010] presents results for an intrusive approach based on the use of an accelerometer: obtaining 0.93 for 5-fold cross validation and 0.98 for 10-fold cross validation, in the detection of 0 to 9 digits. [Kim et al., 2008] separates 6 gestures on the basis of the posture and motion of the hand, obtaining an accuracy of 0.975 for the best setup. In [Yoon et al., 2001], the highest accuracy rate in the detection of 26 gestures drawn to the air is 0.932. In [Molina et al., 2011] (i.e. Chapter 4), two of the considered gestures were N and S, obtaining a mean recall of 0.938 in their detection. So it can be said that the proposed approach achieves results comparable to the ones of the State Of Art, even when they do not present user-independent evaluations.

5.5.3 Computational cost

The computational cost can be expressed as a function depending on the number of translation segments for each motion pattern, N , and the number of synthetical patterns, N_{SynPat} , contained in the collection described in section 5.4.3. It has been considered, as significant, the periods necessary for performing a sum, T_S , a product, T_P , and a square root T_{sqrt} . The different stages considered on this work will present the following computational times per frame:

1. POI sampling: In the case of the 2.5D scenario, this is the time needed to compute the position of the closest pixel, for what is necessary to perform $width \times height - 1$ comparisons, so $T_{A-2.5D} = (width \times height - 1) \times (N + 1) \times T_S$. In the 2D scenario the time for extracting the geodesic center of the binary mask as described in [Molina et al., 2011], $T_{A-2D} = 4.311msec$, has been taken into account
2. Trajectory computation: This is the time needed for calculating the trajectory vector on the basis of the point coordinates, $T_B = 3 \times N \times T_S$.
3. Trajectory Normalization: as described in section 5.4.6, $T_C = N \times (5 \times T_S + 6 \times T_P + T_{sqrt})$.

4. DTW computation: $T_D = N^2 \times N_{SynPat} \times (5 \times T_S + 3 \times T_P + T_{sqr})$.

Current Float Point Units offer a solution for the computation of arithmetic operations with dedicated hardware, achieving computational times in the same order of magnitude for sum, product and squared root. On the basis of Pentium speed tests⁵ the following relation between T_S , T_P and T_{sqr} can be established, defining T_0 as the reference computational time: $T_S \simeq T_P = T_0$ and $T_{sqr} = 2 \times T_0$. Doing so, and on the basis of the presented expressions, a total computational time of $T = T_A + T_B + T_C + T_D = T_A + T_0 \times N \times (16 + 10 \times N \times N_{SynPat})$ is obtained. With $N = 4$ and $N_{SynPat} = 54$ $T = T_A + 8704 \times T_0$. A CPU performance test was run on an Intel(R) Core(TM)2 Duo CPU E7500 @ 2.93Ghz with 2.98GB RAM, as in Molina et al. [2011], being the obtained T_0 below $1nsec$. So $T_{2.5D} = T_{A-2.5D} + 8704 \times T_0 = 135419 \times T_0$ ($T_{2.5D} < 0.136msecs$) and $T_{2D} = T_{A-2D} + 8704 \times T_0$ ($T_{2D} < 4.321msecs$).

Scenario→	2.5D	2D
Comp. Cost($msec/frame$)	< 0.136	< 4.321
Accuracy	0.951	0.780

Table 5.3: Computational Costs per frame and Accuracy for the two considered scenarios.

As shown in Table 5.3, the described approaches require much less than $1/25sec$ per frame, enabling real-time HCI.

5.6 Conclusions

In this chapter a non intrusive motion-based hand gesture detection system using range data is presented. It is able to work in real-time allowing the interaction between a user and a virtual environment or computer menu. It is robust to the relative camera position and to the speed of execution of the gestures. It is, as well, user-independent, being able to work with a collection of gestures executed by users of different heights and arm's sizes. A novel definition of the motion patterns, based on human anatomy, is presented: the obtained results bear witness to its remarkable representation capacity.

From the results, it can be concluded that the use of depth information for the hand trajectory estimation implies a significant increase in gesture recognition accuracy rate, even with no need of segmentation algorithms apart from limiting the depth range of the capture (2.5D scenario). As well, it can also be asserted that the use of the closest point as POI performs better than the geodesic center of the hand mask, which is more computationally costly. The achieved accuracy rate for the proposed dictionary, performing a user-independent evaluation, is 0.951, a very promising value, as already mentioned, comparable to the results of the State Of Art.

⁵<http://www.obliquity.com/computer/speedtest.html>

The experiments performed in this work also show that the 2.5D approach performs better than the 2D, even without considering the only gesture with a clear translation just in the depth coordinate, the IO gesture.

Part IV

Conclusions

Chapter 6

Conclusions and future work

6.1 Summary of achievements

The aim of this thesis has been to produce contributions in the hand gesture recognition scope. As already mentioned in the different chapters of this document, the main contributions are related with hand gesture usability, scalability and representativity.

Firstly, in Chapter 2 a corpus of hand depth images for benchmarking of hand gesture recognition systems was presented. As well a set of novel critical factors was proposed and several datasets of the SoA compared in terms of them. The dataset has real users recordings and synthetically generated depth images (of the hand pose-based gestures) of several dictionaries described in other papers as well as other novel ones. The compiled collection responds to a taxonomy consisting of posed-based, motion-based, pose-motion based and compound gestures. In terms of representativity, 11 different users participated in the compilation of the collection. Moreover, point of view variations can be introduced in the synthetic data, increasing the representativeness of the collection. It is significant that the proposed method for the generation of synthetic range data images makes possible the recognition of new gestures with a simple design process and with no need of training users. This constitutes an important advantage in terms of scalability. The synthetic generation method has been validated with synthetic content training schemes presenting promising results, which are close (for some dictionaries) to those obtained by the real users training scheme. In Chapter 3 the method for the generation of synthetic hands range data is extended, introducing the concept of synthetic user. This approach is evaluated in terms of accuracy rate, with a training stage performed using synthetic data and an evaluation with real users. This solution improves the previous gesture scalable approach as the results show: this framework is able to work for some dictionaries as good as a system trained with one real user.

Making use of the records associated to some of the dictionaries described in Chapter 2 two recognition systems were developed and evaluated. A framework was proposed (see Chapter 4),

with two main objectives, to be usable and to have a system able to work in a scenario with real users. Gesture classification is based on a descriptor which consists of crucial points of the silhouette using the geodesic distance to the center of the segmented hand. These features have proven to be more robust than features based on moments of the contour [Molina et al., 2011]. Three different types of hand gestures are considered: simple, compound and based on motion pattern. The system has been evaluated with a significant number of users, obtaining user independent results that improve the ones reported in the State Of Art for simple gestures. The proposed system shows remarkable performance even when comparing to non user independent systems. In terms of usability, the system properly works in real-time with a low response time, allowing the interaction with application interfaces. In Chapter 5 a motion-based hand gesture recognition system is presented. It is able to work in real-time allowing the interaction between a user and a virtual environment or computer menu. It is robust to the relative camera position and to the speed of execution of the gestures. It is, as well, user independent, being able to work with a collection of gestures executed by users of different heights and arm's sizes. A novel definition of the motion patterns, based on human anatomy, is presented: the obtained results bear witness to its remarkable representation capacity. From the results it can be confirmed that the use of depth information implies a significant increase in gesture recognition accuracy rate. The proposed approach (2.5D scenario) works without the need of applying any segmentation algorithm or calculating the geodesic center of the hand mask, as in the 2D scenario, which means a lower computation time. The achieved accuracy rate for the proposed dictionary, performing a user-independent evaluation, is a very promising value comparable to the results of the State Of Art.

6.2 Future work

In the light of the obtained results some future work lines are under consideration:

- In relation with the hand gesture scalability (see Chapter 3): Future work lines include the variation of fixed hand parameters (i.e. scale and intra-hand proportions) and/or the resolution of their grid in order to create different synthetic user profiles, increasing, this way, the representativity of the synthetic collection. Future research lines could also include the design of gesture detection approaches adapted to the nature of the gestures of each dictionary. Another research line could be the testing of the proposed solution with different visual descriptors, more adapted to the particularities of each gesture collection under consideration.
- In relation with the simple, compound and motion based gestures recognition framework presented in Chapter 4: the improvement of the hand segmentation, which would also be useful for solving the forearm elimination in the outermost areas of the screen; the

improvement of the hand descriptor, to avoid the need of having fingers pointing up to obtain a proper description; and the integration of a Hidden Markov Model based solution for making the system more robust to noisy SHPs recognition. As well, it could be of utility the use of not only depth information but this registered with color information.

- In relation with the motion based gestures recognition framework presented in Chapter 5: In the light of the obtained results, two main future work lines are considered: the use of a Hidden Markov Model in order to manage the temporal sequence of detected labels, trying to solve the misclassification situations in which the order of the translation detections is relevant; the use of color-depth registration approaches could improve the quality of the hand motion estimation and make feasible the detection of more complex gestures.

Bibliography

- J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff. A unified framework for gesture recognition and spatiotemporal gesture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1685–1699, 2009.
- X. P.-S. V. P. L. X. B. O. P. C. A. S. E. Antonio Hernández-Vela, Miguel Ángel Bautista Martín. Bovdw: Bag-of-visual-and-depth-words for gesture recognition. In *21st International Conference on Pattern Recognition*, Tsukuba International Congress Center, Tsukuba, Japan, 2012.
- V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:432, 2003.
- C. Baysal. Implementation of fuzzy similarity methods for manipulative hand posture evaluation. In *Systems Man and Cybernetics. IEEE International Conference on*, pages 1320 –1324, 2010.
- P. Breuer, C. Eckes, and S. Muller. Hand gesture recognition with a novel ir time-of-flight range camera: A pilot study. In *Computer Vision/Computer Graphics Collaboration Techniques Third International Conference, MIRAGE*, pages 247–260, 2007.
- D. Castilla, I. Miralles, M. Jorquera, C. Botella, R. Baños, J. Montesa, and C. Ferran. Analysis and testing of metaphors for the definition of a gestual language based on real users interaction: vision project. In *13th International Conference on Human-Computer Interaction*, San Diego, CA, USA, 2009.
- A. Causo, E. Ueda, Y. Kurita, Y. Matsumoto, and T. Ogasawara. Model-based hand pose estimation using multiple viewpoint silhouette images and unscented kalman filter. In *Robot and Human Interactive Communication. The 17th IEEE International Symposium on*, pages 291 –296, 2008.
- A. Causo, M. Matsuo, E. Ueda, K. Takemura, Y. Matsumoto, J. Takamatsu, and T. Ogasawara. Hand pose estimation using voxel-based individualized hand model. In *Advanced Intelligent Mechatronics. IEEE/ASME International Conference on*, pages 451 –456, 2009.

- Y. Chai, S. Shin, K. Chang, and T. Kim. Real-time user interface using particle filter with integral histogram. *Consumer Electronics, IEEE Transactions on*, 56(2):510–515, 2010.
- P. Chakraborty, P. Sarawgi, A. Mehrotra, G. Agarwal, and R. Pradhan. Hand gesture recognition: A comparative study.
- C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines, 2001.
- Y. T. Chen and K. T. Tseng. Developing a multiple-angle hand gesture recognition system for human machine interactions. In *Industrial Electronics Society, 2007. IECON 2007. 33rd Annual Conference of the IEEE*, pages 489–492, 2007.
- J. Cheng, C. Xie, W. Bian, and D. Tao. Feature fusion for 3d hand gesture recognition by learning a shared hidden space. *Pattern Recognition Letters*, (In Press), 2010.
- S. Y. Cheng and M. Trivedi. Multimodal voxelization and kinematically constrained gaussian mixture models for full hand pose estimation: An integrated systems approach. page 34, 2006.
- F. Dadgostar, A. L. C. Barczak, and A. Sarrafzadeh. A color hand gesture database for evaluating and improving algorithms on hand gesture and posture recognition. *Res. Lett. Inf. Math. Sci.*, 7:127–134, 2005.
- P. Doliotis, A. Stefan, C. McMurrough, D. Eckhard, and V. Athitsos. Comparing gesture recognition accuracy using color and depth information. In *Proceedings of the 4th International Conference on Pervasive Technologies Related to Assistive Environments*, PETRA '11, pages 20:1–20:7, New York, NY, USA, 2011. ACM.
- A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1-2):52 – 73, 2007.
- S. Ge, Y. Yang, and T. Lee. Hand gesture recognition and tracking based on distributed locally linear embedding. In *Robotics, Automation and Mechatronics. IEEE Conference on*, pages 1–6, 2006.
- C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In D. E. Losada and J. M. Fernández-Luna, editors, *Advances in Information Retrieval*, volume 3408 of *Lecture Notes in Computer Science*, pages 345–359. Springer Berlin / Heidelberg, 2005.
- R. Grzeszczuk, G. Bradski, M. Chu, and J. Bouguet. Stereo based gesture recognition invariant to 3d pose and lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages I: 826–833, 2000.

- S. Guomundsson, M. Pardás, R. Larsen, H. Aanaes, and J. R. Casas. TOF imaging in smart room environments towards improved people tracking. *Computer Vision and Image Understanding*, 114 (12):1376–1384, jun 2010a.
- S. A. Guomundsson, J. R. Sveinsson, M. Pardas, H. Aanaes, and R. Larsen. Model-based hand gesture tracking in tof image sequences. In *AMDO*, volume 6169 of *Lecture Notes in Computer Science*, pages 118–127. Springer, 2010b.
- L. Han and W. Liang. Continuous hand gesture recognition in the learned hierarchical latent variable space. In *Proceedings of the 5th international conference on Articulated Motion and Deformable Objects*, AMDO '08, pages 32–41, Berlin, Heidelberg, 2008. Springer-Verlag.
- Y. Han. A low-cost visual motion data glove as an input device to interpret human hand gestures. *Consumer Electronics, IEEE Transactions on*, 56(2):501–509, 2010.
- H. Heo, E. C. Lee, K. R. Park, C. J. Kim, and M. Whang. A realistic game system using multi-modal user interfaces. *Consumer Electronics, IEEE Transactions on*, 56(3):1364–1372, aug. 2010.
- P. C. C. Hernandez, J. Czyz, F. Marqués, T. Umeda, X. Marichal, and B. M. Macq. Bayesian approach for morphology-based 2-d human motion capture. *IEEE Transactions on Multimedia*, 9(4):754–765, 2007.
- M.-F. Ho, C.-Y. Tseng, C.-C. Lien, and C.-L. Huang. A multi-view vision-based hand motion capturing system. *Pattern Recognition*, 44:443–453, February 2011.
- E.-J. Holden, G. Lee, and R. Owens. Australian sign language recognition. *Machine Vision and Applications*, 16:312–320, 2005.
- M. B. Holte and M. Stoerring. Pointing and command gestures under mixed illumination conditions: video sequence dataset, <http://www-prima.inrialpes.fr/fgnet/data/03-pointing/index.html>, 2004.
- P. Hong, M. Turk, and T. S. Huang. Constructing finite state machines for fast gesture recognition. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 3, pages 691–694 vol.3, 2000.
- J. Hu, M. K. Brown, and W. Turin. Hmm based on-line handwriting recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18:1039–1045, October 1996.
- M.-K. Hu. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179–187, February 1962.

- D.-Y. Huang, W.-C. Hu, and S.-H. Chang. Gabor filter-based hand-pose angle estimation for hand gesture recognition under varying illumination. *Expert Systems with Applications*, 38(5): 6031 – 6042, 2011.
- I. O. F. S. Iso. Iso 9241-11: Guidance on usability. *Ergonomic requirements for office work with visual display terminals*, 1998.
- ISO9241-11. Ergonomic requirements for office work with visual display terminals (vdts) - part 11: Guidance on usability, 1998.
- D. Kelly, J. McDonald, and C. Markham. A person independent system for recognition of hand postures used in sign language. *Pattern Recognition Letters*, 31(11):1359 – 1368, 2010.
- C. Keskin and L. Akarun. Stars: Sign tracking and recognition system using input-output hmms. *Pattern Recognition Letters*, 30(12):1086 – 1095, 2009. ISSN 0167-8655. Image/video-based Pattern Analysis and HCI Applications.
- H.-J. Kim, J. Lee, and J.-H. Park. Dynamic hand gesture recognition using a cnn model with 3d receptive fields. In *Neural Networks and Signal Processing, 2008 International Conference on*, pages 14 –19, 2008.
- T.-K. Kim, S.-F. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, june 2007.
- E. Kollorz, J. Penne, J. Hornegger, and A. Barke. Gesture recognition with a time-of-flight camera. *International Journal of Intelligent Systems Technologies and Applications*, 5(3/4): 334–343, 2008.
- F. Kuhl and C. Giardina. Elliptic Fourier Features of a Closed Contour. *Computer Graphics and Image Processing*, 18:236–258, 1982.
- C. Lantuejoul and F. Maisonneuve. Geodesic methods in quantitative image analysis. *Pattern Recognition*, 17(2):177–187, 1984.
- J. J. Laviola. Bringing vr and spatial 3d interaction to the masses through video games. *IEEE Computer Graphics and Applications*, 28(5):10–15, 2008.
- D. Lee and Y. Park. Vision-based remote control system by motion detection and open finger counting. *Consumer Electronics, IEEE Transactions on*, 55(4):2308 –2313, 2009.
- D.-W. Lee, J.-M. Lim, J. Sunwoo, I.-Y. Cho, and C.-H. Lee. Actual remote control: a universal remote control using hand motions on a virtual menu. *Consumer Electronics, IEEE Transactions on*, 55(3):1439 –1446, 2009.

- J. Letessier and F. Bérard. Visual tracking of bare fingers for interactive surfaces. In *UIST '04: Proceedings of the 17th annual ACM symposium on User interface software and technology*, pages 119–122, New York, NY, USA, 2004. ACM. ISBN 1-58113-957-8.
- H. Li and M. Greenspan. Model-based segmentation and recognition of dynamic gestures in continuous video streams. *Pattern Recognition*, In Press, Corrected Proof:–, 2011. ISSN 0031-3203.
- J. Liu and M. Kavakli. Hand gesture recognition based on segmented singular value decomposition. 6277:214–223, 2010.
- X. Liu and K. Fujimura. Hand gesture recognition using depth data. In *Automatic Face and Gesture Recognition. Sixth IEEE International Conference on*, pages 529 – 534, 2004.
- S. Malassiotis and M. Srinivas. Real-time hand posture recognition using range data. *Image and Vision Computing*, 26(7):1027–1037, 2008.
- S. Marcel. Hand posture recognition in a body-face centered space. In *CHI '99 extended abstracts on Human factors in computing systems*, CHI EA '99, pages 302–303, New York, NY, USA, 1999. ACM.
- S. Marcel, O. Bernier, J.-E. Viallet, and D. Collobert. Hand gesture recognition using input-output hidden markov models. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 456 –461, 2000.
- D. K. Martin Larsson, Isabel Serrano Vicente. Cvap arm/hand activity database, http://www.nada.kth.se/~danik/gesture_database/, 2011.
- S. Mitra and T. Acharya. Gesture recognition: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37(3):311–324, 2007.
- T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, November 2006.
- J. Molina, M. Escudero-Viñolo, A. Signoriello, M. Pardás, C. Ferrán, J. Bescós, F. Marqués, and J. M. Martínez. Real-time user independent hand gesture recognition from time-of-flight camera video using static and dynamic models. *Machine Vision and Applications (online first)*, pages 1–18, 2011.
- K. Nickel and R. Stiefelhagen. Visual recognition of pointing gestures for human-robot interaction. *Image and Vision Computing*, 25(12):1875–1884, 2007.

- R. Poppe. Vision-based human motion analysis: An overview. *Comput. Vis. Image Underst.*, 108:4–18, 2007.
- P. Premaratne and Q. Nguyen. Consumer electronics control system based on hand gesture moment invariants. *Iet Computer Vision*, 1(1):35–41, 2007. ISSN 1751-9632.
- Z. Ren, J. Meng, and J. Yuan. Depth camera based hand gesture recognition and its applications in human-computer-interaction. In *Information, Communications and Signal Processing (ICICS) 2011 8th International Conference on*, pages 1–5, dec. 2011a.
- Z. Ren, J. Meng, J. Yuan, and Z. Zhang. Robust hand gesture recognition with kinect sensor. In *Proceedings of the 19th ACM international conference on Multimedia*, ACM MM '11, pages 759–760. ACM, 2011b.
- H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.
- P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003. ISBN 3540429883.
- S. Soutschek, J. Penne, J. Hornegger, and J. Kornhuber. 3-d gesture-based scene navigation in medical imaging applications using time-of-flight cameras. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6, 2008.
- B. Stenger, A. Thayananthan, P. H. S. Torr, and R. Cipolla. Model-based hand tracking using a hierarchical bayesian filter. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(9):1372–1384, 2006.
- B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Estimating 3d hand pose using hierarchical multi-label classification. *Image and Vision Computing*, 25(12):1885 – 1894, 2007. The age of human computer interaction.
- C. H. Teh and R. T. Chin. On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):496–513, 1988.
- X. Teng, B. Wu, W. Yu, and C. Liu. A hand gesture recognition system based on local linear embedding. *J. Vis. Lang. Comput.*, 16:442–454, October 2005.
- J. Triesch and C. VD Malsburg. Robust classification of hand postures against complex backgrounds. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 170 –175, oct 1996.
- J. Triesch and C. VD Malsburg. A system for person-independent hand posture recognition against complex backgrounds. 23(12):1449–1453, December 2001.

- J. Usabiaga, A. Erol, G. Bebis, R. Boyle, and X. Twombly. Global hand pose estimation by multiple camera ellipse tracking. *Machine Vision and Applications*, 21:1–15, 2009. ISSN 0932-8092.
- T. Wenjun, W. Chengdong, Z. Shuying, and J. Li. Dynamic hand gesture recognition using motion trajectories and key frames. In *Advanced Computer Control (ICACC), 2010 2nd International Conference on*, volume 3, pages 163 –167, 2010.
- C. Yang, Y. Jang, J. Beh, D. Han, and H. Ko. Gesture recognition using depth-based hand tracking for contactless controller application. In *Consumer Electronics (ICCE), 2012 IEEE International Conference on*, pages 297 –298, jan. 2012.
- H.-S. Yoon, J. Soh, Y. J. Bae, and H. S. Yang. Hand gesture recognition using combined features of location, angle and velocity. *Pattern Recognition*, 34(7):1491 – 1501, 2001.
- M. M. Zaki and S. I. Shaheen. Sign language recognition using a combination of new vision based features. *Pattern Recognition Letters*, 32(4):572 – 577, 2011.
- G. Zheng, C. J. Wang, and T. E. Boulton. Application of projective invariants in hand geometry biometrics. *IEEE Transactions on Information Forensics and Security*, 2(4):758–768, 2007.
- X. Zhu, J. Yang, and A. Waibel. Segmenting hands of arbitrary color. In *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, page 446, Washington, DC, USA, 2000. IEEE Computer Society.

Part V

Appendixes

Appendix A: Confussion matrixes for synthetic training schemes

Confussion Matrixes for the Synthetic Training Scheme A with 200 POV

	A	B	C	D	E	F	G	H	I	J	K	L
A	761	0	1	0	0	0	1	0	0	1437	0	0
B	5	2194	0	0	0	0	0	0	0	0	1	0
C	8	1561	388	3	0	32	0	0	9	0	199	0
D	204	648	0	1348	0	0	0	0	0	0	0	0
E	617	0	0	1177	406	0	0	0	0	0	0	0
F	3	0	14	0	1	2122	0	0	0	0	0	60
G	71	0	417	0	0	0	303	0	58	1195	0	156
H	437	0	0	0	0	0	0	961	0	24	778	0
I	1	0	0	0	0	0	0	16	2022	13	148	0
J	6	0	26	0	0	0	16	0	60	2089	3	0
K	233	0	0	0	0	0	0	701	0	17	1088	161
L	10	0	100	0	1	0	509	19	5	118	97	1341

Table 6.1: Confussion matrix for dictionary described in [Kollorz et al., 2008] applying evaluation scheme A(200).

	Enum1	Enum2	Enum3	Enum4	Enum5	Stop	Fist	OkLeft	OkRigth
Enum1	2150	0	0	0	0	27	5	11	7
Enum2	21	2128	3	0	0	39	3	2	4
Enum3	1	112	1982	29	4	52	11	3	6
Enum4	0	2	205	1940	41	6	3	0	3
Enum5	0	0	10	482	1705	0	2	0	1
Stop	19	0	3	0	0	2157	18	0	3
Fist	1	0	0	0	0	610	1186	3	400
OkLeft	2	17	0	0	0	23	12	2045	101
OkRigth	24	0	0	0	0	0	7	2	2167

Table 6.2: Confussion matrix for dictionary described in [Molina et al., 2011] applying evaluation scheme A(200).

	a	b	c	d	e
a	1619	556	0	24	1
b	7	2132	0	2	59
c	5	85	2094	0	16
d	308	425	0	1418	49
e	348	191	0	12	1649

Table 6.3: Confussion matrix for dictionary described in [Soutschek et al., 2008] applying evaluation scheme A(200).

	m1	m2	m3	m4
m1	1730	378	0	92
m2	0	1941	3	256
m3	4	1296	570	330
m4	0	21	2	2177

Table 6.4: Confussion matrix for Miscellaneous pose-based dictionary applying evaluation scheme A(200).

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	x	y	z	
a	33	218	11	0	0	0	0	0	22	53	0	0	462	2	763	33	566	3	0	3	0	0	0	0	31	0
b	0	1836	0	0	20	102	32	0	0	9	0	0	0	0	0	153	15	0	28	1	0	0	0	4	0	
c	0	17	8	0	10	13	0	62	0	0	269	1763	0	0	0	0	0	0	11	0	47	0	0	0	0	
d	0	1129	0	239	0	0	0	0	0	0	511	26	0	0	0	284	0	0	0	0	0	0	11	0	0	
e	0	78	0	0	1957	3	0	2	1	1	0	155	0	0	1	0	0	0	2	0	0	0	0	0	0	
f	0	7	3	4	0	1145	0	0	0	5	349	0	0	0	0	7	0	20	246	165	0	236	13	0	0	
g	0	1413	5	0	382	14	0	5	0	226	43	16	0	0	2	0	8	18	1	0	66	1	0	0	0	
h	0	89	1	0	2	698	0	1372	0	0	29	3	0	0	0	0	0	0	1	1	3	1	0	0	0	
i	0	231	5	66	0	0	0	0	14	253	104	0	1	5	0	1397	0	0	98	0	0	0	26	0	0	
j	0	87	25	385	168	44	0	0	15	672	72	0	0	1	0	385	1	0	177	156	0	0	9	0	3	
k	0	8	45	0	0	9	0	0	0	0	1250	442	0	0	0	2	0	0	3	16	423	0	2	0	0	
l	0	4	10	0	0	1	1	29	1	2	64	2011	0	0	0	0	0	0	0	0	77	0	0	0	0	
m	0	5	0	1	0	0	4	0	0	0	0	0	764	17	0	610	0	37	59	699	0	0	0	1	3	
n	0	0	0	0	0	0	1	0	0	0	7	0	722	211	0	1045	0	54	4	152	0	0	0	4	0	
o	0	192	12	0	0	0	0	0	0	0	22	1	8	22	0	1922	0	0	0	0	19	2	0	0	0	
p	0	0	7	0	0	0	0	0	0	0	0	0	0	54	0	1510	0	442	107	79	0	0	1	0	0	
q	252	0	554	0	0	8	0	0	18	20	0	0	304	46	255	225	1	26	40	329	0	0	22	100	0	
r	0	1	5	144	0	0	0	0	0	202	210	0	39	14	0	86	0	1185	72	108	0	0	102	0	32	
s	0	284	0	1	0	131	0	102	0	0	332	0	0	0	0	370	0	101	147	368	188	176	0	0	0	
t	0	0	42	0	0	1309	0	0	0	46	227	0	0	0	0	15	0	3	444	65	0	30	19	0	0	
u	0	0	0	0	0	0	0	0	0	0	87	223	0	0	0	0	0	0	0	0	1890	0	0	0	0	
v	0	0	3	0	0	582	1	523	0	0	34	2	0	0	0	0	0	0	20	262	233	540	0	0	0	
x	0	104	16	43	0	0	0	0	0	251	1349	0	0	0	0	296	0	63	0	0	0	0	75	0	3	
y	90	4	15	70	120	0	0	0	179	200	3	0	80	0	72	443	158	0	638	100	0	0	0	27	1	
z	42	145	0	244	80	5	20	0	284	122	91	0	4	0	155	68	227	1	471	47	0	0	45	93	56	

Table 6.5: Confusion matrix for Spanish sign language alphabet dictionary applying evaluation scheme A(200).

Confussion Matrixes for the Synthetic Training Scheme B with 200 POV

	A	B	C	D	E	F	G	H	I	J	K	L
A	1512	0	0	0	0	0	1	0	4	682	0	1
B	238	1913	0	0	0	0	0	10	0	39	0	0
C	42	261	1477	0	0	2	0	97	9	6	306	0
D	985	644	0	549	0	0	0	10	0	3	9	0
E	756	0	0	743	685	0	0	1	0	1	0	14
F	4	0	30	0	1	2093	0	0	0	0	0	72
G	249	0	185	0	0	0	930	0	32	799	0	5
H	283	0	0	0	0	0	0	1008	0	58	832	19
I	14	0	0	0	0	0	0	358	1223	3	602	0
J	477	0	13	0	0	0	40	0	21	1649	0	0
K	3	0	0	0	0	0	7	664	0	458	826	242
L	72	0	0	0	1	1	1156	19	0	0	76	875

Table 6.6: Confussion matrix for dictionary described in [Kollorz et al., 2008] applying evaluation scheme B(200).

	Enum1	Enum2	Enum3	Enum4	Enum5	Stop	Fist	OkLeft	OkRigth
Enum1	2172	0	1	0	0	12	3	8	4
Enum2	23	2164	0	0	0	4	3	3	3
Enum3	0	72	1955	29	4	103	11	23	3
Enum4	1	3	20	1696	39	436	2	1	2
Enum5	0	0	23	365	1705	104	2	0	1
Stop	0	7	0	0	0	2157	36	0	0
Fist	0	0	0	0	0	285	1233	8	674
OkLeft	3	0	0	0	0	1	11	2164	21
OkRigth	13	0	0	0	0	0	30	3	2154

Table 6.7: Confussion matrix for dictionary described in [Molina et al., 2011] applying evaluation scheme B(200).

	a	b	c	d	e
a	2128	15	0	57	0
b	136	1993	0	1	70
c	0	611	1589	0	0
d	52	293	0	1815	40
e	240	137	0	72	1751

Table 6.8: Confussion matrix for dictionary described in [Soutschek et al., 2008] applying evaluation scheme B(200).

	m1	m2	m3	m4
m1	1731	428	2	39
m2	1	1283	193	723
m3	5	1246	715	234
m4	0	6	90	2104

Table 6.9: Confussion matrix for Miscellaneous pose-based dictionary applying evaluation scheme B(200).

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	x	y	z
a	66	28	295	0	0	0	0	0	8	12	0	0	72	19	634	38	1013	0	0	0	0	0	0	15	0
b	0	2156	0	0	0	0	15	0	0	1	0	0	0	4	8	0	14	0	0	0	0	0	0	2	0
c	0	231	77	0	43	1	0	39	0	0	1128	674	0	0	2	0	2	0	0	0	2	0	1	0	0
d	0	350	0	245	0	0	0	0	0	127	490	0	1	175	0	766	0	4	0	0	0	0	42	0	0
e	0	55	1	0	2118	2	0	1	0	0	0	0	0	0	6	0	4	0	0	0	0	0	0	13	0
f	0	16	24	0	0	723	0	31	0	13	207	0	0	5	17	345	22	26	422	346	0	3	0	0	0
g	0	915	2	76	334	0	9	0	23	818	0	0	0	0	3	0	4	15	0	0	1	0	0	0	0
h	0	205	3	0	93	292	0	1455	1	0	6	0	0	0	0	10	6	1	4	87	0	0	0	37	0
i	0	5	1	165	0	0	0	0	21	595	21	0	27	127	1	1175	0	16	17	0	0	0	29	0	0
j	0	2	99	338	48	3	0	0	40	616	53	0	0	37	23	620	151	0	34	117	0	0	14	1	4
k	0	0	3	0	0	1	0	0	0	0	2075	36	0	0	0	37	0	13	17	16	0	0	2	0	0
l	0	66	2	0	12	0	1	24	1	2	204	1805	0	0	26	0	42	0	0	0	15	0	0	0	0
m	0	0	0	0	0	0	0	0	20	0	0	0	1041	292	0	399	0	59	29	356	0	0	0	3	1
n	0	0	0	0	0	0	0	0	1	0	0	0	736	752	0	572	0	74	2	58	0	0	0	5	0
o	0	433	14	0	0	0	55	0	0	1	35	0	11	49	1	1599	0	2	0	0	0	0	0	0	0
p	0	0	0	0	0	0	0	0	0	1	0	0	323	94	7	1469	0	239	64	2	0	0	1	0	0
q	344	0	324	0	0	2	0	0	15	15	0	0	422	133	132	105	417	54	3	48	0	0	15	171	0
r	0	0	19	137	0	0	0	0	0	250	145	0	34	15	1	44	3	1194	39	2	0	0	317	0	0
s	0	119	0	30	0	61	0	112	0	129	242	0	37	201	0	843	0	5	142	26	127	126	0	0	0
t	0	0	11	1	0	548	0	59	0	51	163	0	0	0	27	237	0	2	722	379	0	0	0	0	0
u	0	0	0	0	0	0	0	0	0	0	17	9	0	0	0	0	0	0	0	0	2174	0	0	0	0
v	0	14	22	0	2	728	6	496	0	14	27	0	0	0	8	55	10	7	1	69	206	535	0	0	0
x	0	1	0	98	0	0	0	0	0	1024	457	0	4	223	0	347	0	9	0	0	0	0	37	0	0
y	73	0	38	59	16	0	0	0	191	151	5	0	56	52	79	490	582	1	302	19	0	0	2	84	0
z	74	80	1	193	52	0	21	0	341	196	84	0	2	39	215	110	334	2	249	20	0	0	76	69	42

Table 6.10: Confussion matrix for Spanish sign language alphabet dictionary applying evaluation scheme B(200).

Confussion Matrixes for the Synthetic Training Scheme C with 200 POV

	A	B	C	D	E	F	G	H	I	J	K	L
A	1462	0	0	0	0	0	1	0	6	730	0	1
B	48	1927	2	0	0	0	0	5	0	212	6	0
C	31	385	1286	0	0	2	0	0	9	9	478	0
D	836	834	0	518	0	0	0	6	0	5	1	0
E	744	0	0	779	659	0	0	0	0	0	0	18
F	4	0	28	0	1	2093	0	0	0	1	0	73
G	157	0	218	0	0	0	880	0	42	902	0	1
H	373	0	0	0	0	0	0	727	0	92	1002	6
I	20	0	0	0	0	1	0	291	1349	17	522	0
J	448	0	11	0	0	0	24	0	39	1678	0	0
K	0	0	0	0	0	0	0	433	0	511	1030	226
L	82	0	1	0	1	1	1087	15	0	7	80	926

Table 6.11: Confussion matrix for dictionary described in [Kollorz et al., 2008] applying evaluation scheme C(200).

	Enum1	Enum2	Enum3	Enum4	Enum5	Stop	Fist	OkLeft	OkRigth
Enum1	2166	0	1	0	0	14	3	6	10
Enum2	27	2158	0	0	0	6	3	3	3
Enum3	0	85	1958	25	4	99	11	14	4
Enum4	0	3	33	1688	41	430	3	1	1
Enum5	0	0	20	396	1705	77	2	0	0
Stop	0	0	0	0	0	2172	28	0	0
Fist	0	0	0	0	0	323	1203	8	666
OkLeft	3	1	0	0	0	24	11	2047	114
OkRigth	12	0	0	0	0	0	26	2	2160

Table 6.12: Confussion matrix for dictionary described in [Molina et al., 2011] applying evaluation scheme C(200).

	a	b	c	d	e
a	2150	10	0	40	0
b	138	1984	0	1	77
c	0	539	1661	0	0
d	69	291	0	1798	42
e	253	134	0	73	1740

Table 6.13: Confussion matrix for dictionary described in [Soutschek et al., 2008] applying evaluation scheme C(200).

	m1	m2	m3	m4
m1	1730	417	6	47
m2	1	1458	168	573
m3	5	1170	769	256
m4	0	6	47	2147

Table 6.14: Confussion matrix for Miscellaneous pose-based dictionary applying evaluation scheme C(200).

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	x	y	z
a	68	37	389	0	0	0	0	0	2	16	0	0	36	3	395	39	1174	0	0	0	0	0	0	41	0
b	0	2175	0	0	3	0	4	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0
c	0	350	84	0	46	1	0	31	0	0	623	1060	0	0	1	0	0	0	0	0	4	0	0	0	0
d	0	566	0	278	0	0	0	0	0	33	487	0	15	83	0	719	0	1	0	0	0	0	18	0	0
e	0	47	0	0	2146	2	0	0	1	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
f	0	37	18	0	0	618	0	33	0	10	207	0	0	1	10	348	10	25	637	232	0	11	3	0	0
g	0	983	1	17	317	0	12	1	53	743	0	7	0	0	3	0	1	54	0	3	5	0	0	0	0
h	0	211	0	0	164	291	0	1384	3	1	7	0	0	0	0	9	6	0	2	82	0	5	0	35	0
i	0	25	0	124	0	0	0	0	19	547	36	0	39	69	4	1238	1	38	26	0	0	0	33	1	0
j	0	0	80	295	56	8	0	0	48	601	63	0	21	15	70	668	99	0	61	98	0	0	14	1	2
k	0	17	10	0	0	1	0	0	0	0	1963	85	0	0	0	63	0	11	32	14	0	0	4	0	0
l	0	16	24	0	53	0	1	26	0	3	263	1735	0	0	23	0	6	0	0	0	50	0	0	0	0
m	0	2	0	0	0	0	12	0	5	0	0	0	668	172	0	535	0	196	293	313	0	0	1	2	1
n	0	0	0	0	0	0	0	0	1	0	0	0	461	414	0	1108	0	133	73	6	0	0	0	4	0
o	0	484	10	0	0	0	0	0	0	0	30	0	22	58	0	1595	0	0	0	0	1	0	0	0	0
p	0	0	0	0	0	0	0	0	0	1	0	0	138	103	0	1540	0	309	88	20	0	0	1	0	0
q	352	0	355	0	0	0	0	0	5	13	0	0	408	81	116	96	427	65	68	68	0	0	12	134	0
r	0	0	11	93	0	0	0	0	0	221	207	0	39	31	7	54	1	1248	44	0	0	0	241	1	2
s	0	168	0	13	0	92	0	98	0	68	262	0	0	187	0	868	0	17	167	25	134	101	0	0	0
t	0	0	19	0	0	451	0	80	0	50	167	0	0	0	18	241	0	3	947	224	0	0	0	0	0
u	0	0	18	0	0	0	0	0	0	0	2	3	0	0	0	0	0	0	0	0	2177	0	0	0	0
v	0	16	11	0	0	770	4	531	0	6	49	1	0	0	0	38	3	2	5	105	208	451	0	0	0
x	0	44	0	29	0	0	0	0	0	648	816	0	0	206	0	332	0	37	0	0	0	0	88	0	0
y	54	0	47	40	10	0	0	0	171	149	3	0	23	77	73	494	609	2	348	14	0	0	3	83	0
z	79	91	1	148	50	0	13	0	278	168	105	0	12	36	212	121	392	2	287	11	0	0	55	95	44

Table 6.15: Confusion matrix for Spanish sign language alphabet dictionary applying evaluation scheme C(200).

Confussion Matrixes for the Synthetic Training Scheme D with 200 POV

	A	B	C	D	E	F	G	H	I	J	K	L
A	1133	0	0	0	0	0	0	0	0	1066	0	1
B	7	2192	0	0	0	0	0	0	0	0	1	0
C	3	1662	352	0	0	100	0	0	9	6	68	0
D	939	288	0	957	0	0	0	3	0	13	0	0
E	838	0	0	995	366	0	0	1	0	0	0	0
F	0	0	7	0	2	2112	44	0	0	15	0	20
G	551	0	286	0	0	0	143	0	53	1052	0	115
H	41	0	0	0	0	0	0	959	0	573	627	0
I	1	0	0	0	0	0	0	8	1931	32	228	0
J	123	0	15	0	0	0	4	0	22	2036	0	0
K	0	0	0	0	0	0	0	505	0	632	994	69
L	214	0	92	0	2	1	649	30	11	41	82	1078

Table 6.16: Confussion matrix for dictionary described in [Kollorz et al., 2008] applying evaluation scheme D(200).

	Enum1	Enum2	Enum3	Enum4	Enum5	Stop	Fist	OkLeft	OkRigth
Enum1	2187	0	0	0	1	1	4	4	3
Enum2	2	2186	0	0	0	4	4	2	2
Enum3	1	119	1922	22	33	83	10	6	4
Enum4	0	10	48	1920	55	162	4	0	1
Enum5	0	0	33	370	1760	35	2	0	0
Stop	0	0	0	1	0	2185	4	0	10
Fist	0	0	0	0	0	664	1215	9	312
OkLeft	4	0	0	0	0	1	15	2165	15
OkRigth	27	0	0	0	0	0	6	2	2165

Table 6.17: Confussion matrix for dictionary described in [Molina et al., 2011] applying evaluation scheme D(200).

	a	b	c	d	e
a	2149	38	1	12	0
b	60	2115	0	3	22
c	0	175	2025	0	0
d	242	755	0	1166	37
e	278	31	0	155	1736

Table 6.18: Confussion matrix for dictionary described in [Soutschek et al., 2008] applying evaluation scheme D(200).

	m1	m2	m3	m4
m1	1730	423	1	46
m2	0	2030	1	169
m3	2	1523	386	289
m4	0	38	1	2161

Table 6.19: Confussion matrix for Miscellaneous pose-based dictionary applying evaluation scheme D(200).

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	x	y	z
a	92	208	5	0	0	0	0	0	5	59	0	0	404	20	686	180	500	13	0	0	0	0	0	28	0
b	0	2169	0	0	0	0	5	0	0	0	0	0	0	0	0	2	24	0	0	0	0	0	0	0	
c	0	10	14	0	33	0	0	11	0	0	215	1911	0	0	0	0	0	0	0	0	6	0	0	0	
d	0	702	0	138	0	0	0	0	0	0	518	0	0	0	0	821	0	0	0	0	0	18	3	0	
e	0	68	0	0	2118	0	0	2	1	0	0	1	0	0	0	0	2	0	0	0	0	0	8	0	
f	0	23	5	0	0	830	0	273	0	0	163	0	0	0	0	395	0	26	358	110	0	13	4	0	
g	0	1518	1	0	429	0	0	0	19	159	49	9	0	0	0	0	0	0	0	0	16	0	0	0	
h	0	387	0	0	62	153	0	1456	9	0	0	95	0	0	0	1	0	0	0	1	7	0	0	13	
i	0	102	0	34	0	0	0	0	12	232	154	0	2	6	0	1605	0	0	21	4	0	22	6	0	
j	0	2	24	480	89	42	0	0	5	349	111	8	0	1	0	933	0	0	75	72	0	0	9	0	
k	0	16	2	0	0	34	0	0	0	0	1171	641	0	0	0	60	0	0	0	12	263	0	1	0	
l	0	28	0	0	8	0	2	17	0	2	28	2104	0	0	0	0	1	0	0	0	9	0	1	0	
m	0	0	0	5	0	0	9	0	0	0	0	0	47	34	0	1494	0	74	58	477	0	0	0	2	
n	0	0	0	0	0	0	1	0	1	0	0	0	2	120	0	1906	0	159	0	8	0	0	3	0	
o	0	347	7	0	0	0	0	0	0	0	11	2	6	8	0	1808	0	0	0	0	11	0	0	0	
p	0	0	2	0	0	0	0	0	0	0	0	0	10	30	0	1871	0	243	28	16	0	0	0	0	
q	184	0	359	0	0	3	0	0	10	10	0	0	472	200	211	310	12	122	70	165	0	0	17	55	
r	0	0	4	61	0	0	0	0	0	12	324	0	6	8	0	144	0	1476	39	24	0	3	99	0	
s	0	250	0	26	1	80	0	191	0	0	213	0	0	0	0	988	0	84	90	4	123	150	0	0	
t	0	0	18	0	0	983	0	137	0	40	151	0	0	0	0	258	0	7	490	107	1	7	1	0	
u	0	0	0	0	0	0	0	0	0	0	0	191	0	0	0	0	0	0	0	2009	0	0	0	0	
v	0	15	2	0	2	713	3	598	0	0	6	7	0	0	0	11	0	0	1	36	276	530	0	0	
x	0	0	0	24	0	0	0	0	0	117	1374	0	0	0	0	480	0	97	0	0	0	0	108	0	
y	121	4	12	141	62	0	0	0	273	244	11	0	85	0	69	595	56	0	454	42	0	4	7	20	
z	68	164	0	189	71	5	22	0	266	235	93	0	1	0	81	265	215	0	269	37	0	20	54	45	
																								100	

Table 6.20: Confussion matrix for Spanish sign language alphabet dictionary applying evaluation scheme D(200).

Appendix B: Hand Descriptor¹

Some global parameters that will be used both for extracting additional silhouette features and for the gesture classification are first computed. These parameters are the Geodesic Center (C) of the hand, the length and orientation of the axes of the ellipse fitted to the hand silhouette and the Minimum Depth Point. C is estimated by performing the ultimate erosion [Lantuejoul and Maisonneuve, 1984] up to a point (see Figure 6.2.1). This is an approximation to the center of gravity of the mask which is guaranteed to be within the mask. The ultimate erosion is used to reduce the bias introduced by the finger shapes in the estimation of this central point. The depth of this Center is directly obtained from the TOF camera.



Figure 6.2.1: Computation to Geodesic Center of the hand silhouette.

The ellipse fitted to the contour (see Figure 6.2.2) provides global information about the silhouette shape (the hand size and its orientation). The axis length and orientation of the ellipse are computed as the eigenvalues and eigenvectors of the covariance matrix of the coordinates of the silhouette.

The Minimum Depth Point corresponds to the nearest point to the camera, that is, the brightest pixel.

¹This appendix is based on: *J. Molina, M. Escudero-Viñolo, A. Signoriello, M. Pardás, C- Ferrán, J. Bescós, F. Marqués, and J. M. Martínez, “Real-time user independent hand gesture recognition from time-of-flight camera video using static and dynamic models”, Machine Vision and Applications, pp. 1-18, 2011 (on-line first).* The described descriptor was developed by The Image and Video Processing Group, Polytechnic University of Catalonia, Barcelona, Spain.

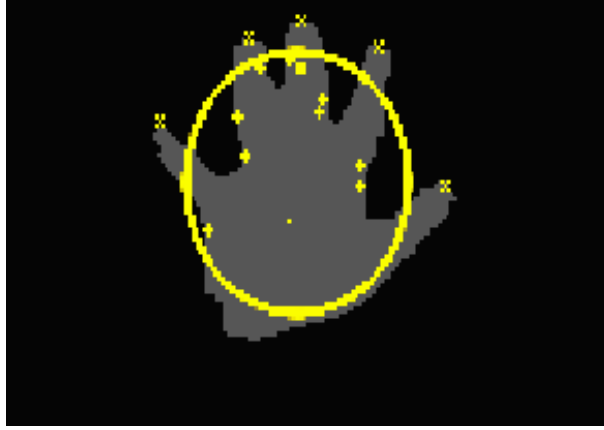


Figure 6.2.2: Description superimposed to the hand mask

Once the silhouette mask has been obtained and its global parameters computed, the shape has to be analyzed in order to extract the additional features needed for its classification. Different features can be used for this aim. For instance, shape descriptors such as Fourier descriptors [Kuhl and Giardina, 1982], Zernike [Teh and Chin, 1988] or Hu moments [Hu, 1962]. Another traditional shape analysis technique consists in modeling the skeleton of the silhouette. We have chosen to extend the method used in [Hernandez et al., 2007] for searching the crucial points of a 2D human body for pose estimation. One of the advantages of this approach is the possibility to include semantic information in the feature extraction process, thus making the system more robust to noise or spurious detections (corresponding for instance to the arm). In our case, the presence of extended fingers must be detected. This can be achieved analyzing the distance from the Geodesic Center (C) to the silhouette border. We use a robust method to search for crucial points based on the extraction of the points of the silhouette that represent the local geodesic distance maxima with respect to C. Semantic information is included by selecting only those maxima that accomplish certain conditions in the distance function related to the width and length of the fingers.

The geodesic distance from C to any point x in the silhouette S is defined as:

$$d_S(C, x) = n \Leftrightarrow x \in \delta_S^n(C) \text{ and } x \notin \delta_S^{n-1}(C) \quad (6.2.1)$$

where $\delta_S^n(C)$ is the geodesic dilation of size n of C within S, which can be expressed as:

$$\delta_S^n(C) = \overbrace{\delta \left(\dots \left(\delta \left(\delta(C) \cap S \right) \cap S \right) \dots \right) \cap S}^{n \text{ times}} \quad (6.2.2)$$

being δ the morphological dilation of minimal size.

A one-dimensional function $f(p)$ linking each pixel position p in the silhouette border with its

geodesic distance with respect to C is then computed. This function $f(p)$ yields the local maxima associated with the crucial points as well as other minor local maxima due to segmentation noise. Noise peaks present lower intensities than peaks associated with crucial points. Therefore, they are removed by applying a H-maxima operator [Soille, 2003].

This filtering is obtained by applying an opening by reconstruction (γ_{rec}) to the function $f(p)$ that provides us with a function $g(p)$:

$$g(p) = \gamma_f^{rec}(f(p) - H) = \delta_f^\infty(f(p) - H) \quad (6.2.3)$$

where $\gamma_f^{rec}(f - H)$ is the unidimensional geodesic dilation of the function $f - H$ within f iterated until idempotence, and H is a constant value related to the estimated noise intensities. In this way, local maxima that have a smaller intensity than H are eliminated. In Figure 6.2.3, the function representing the geodesic distance evaluated on the silhouette is shown before ($f(p)$) and after ($g(p)$) the H-maxima. The removed local maxima are marked by circles and the selected local maxima are selected by vertical lines.

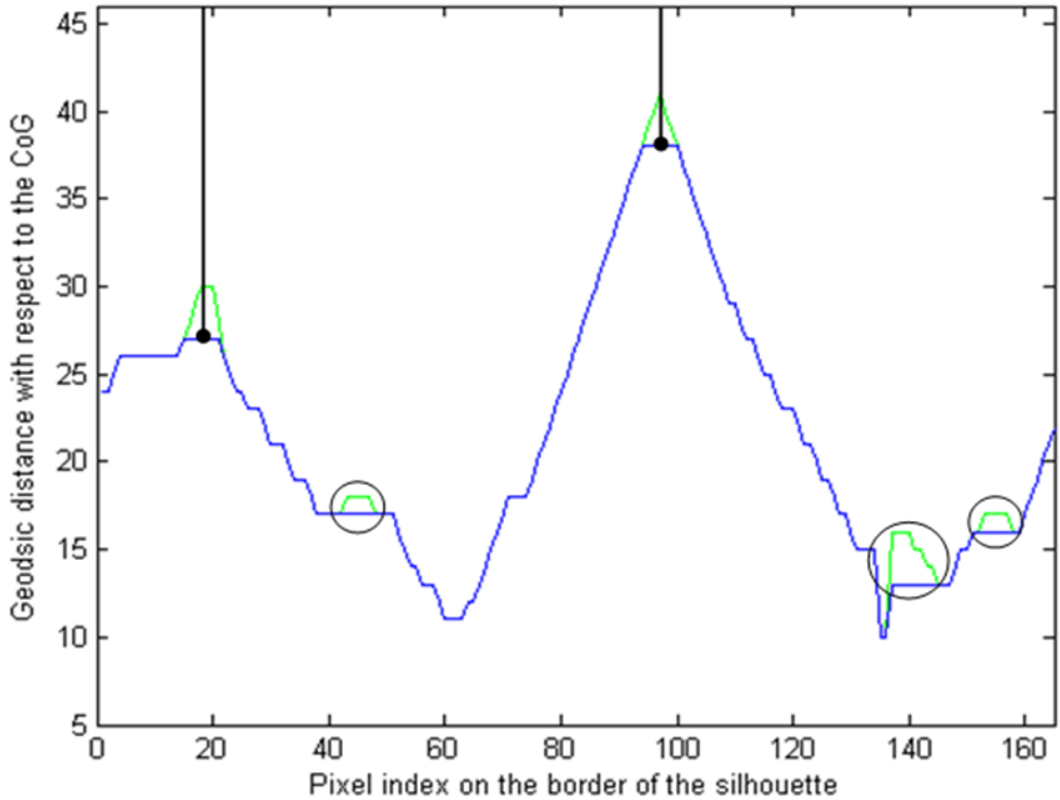


Figure 6.2.3: Geodesic distance from Geodesic Center to silhouette border points before (in green) and after (in blue) applying the H-maxima. Local maxima correspond to prominent hand features (the fingers).

Although the geodesic distance using (6.2.1) has a high computational cost, and due to the fact that we only need the geodesic distance on the silhouette contour, in a limited computational capacity context a simplified calculation can be performed, as described in [Hernandez et al., 2007].

Appendix C: Glossary

HCI *Human-Computer Interaction*

TOF *Time-Of-Flight*

DOF *Degrees Of Freedom*

NT *Natural Training*

ST *Synthetic Training*

SHP *Static Hand Posture*

DHG *Dynamic Hand Gesture*

SVM *Support Vectors Machine*

FSM *Finite State Machine*

DTW *Dynamic Time Warping*

Appendix D: Publications

The following publications have been produced in association with this thesis:

- Related with visual descriptors and classification techniques:

J. Molina, E. Spyrou, N. Sofou, and J. M. Martínez, “On the selection of mpeg-7 visual descriptors and their level of detail for nature disaster video sequences classification,” in *Proceedings of the semantic and digital media technologies 2nd international conference on Semantic Multimedia*, SAMT’07, (Berlin, Heidelberg), pp. 70-73, Springer-Verlag, 2007.

J. Molina, J. M. Martínez, V. Mezaris, G. T. Papadopoulos, S. Nikolopoulos, I. Kompatsiaris, A. Dimou, P. Villegas, J. Rodríguez-Benito, E. Bru, T. Adamek, G. Toliás, E. Spyrou, N. Sofou, P. Kapsalas, and Y. S. Avrithis, “Mesh participation to trecvid2008 hlfe,” in *TRECVID* (P. Over, G. Awad, R. T. Rose, J. G. Fiscus, W. Kraaij, and A. F. Smeaton, eds.), National Institute of Standards and Technology (NIST), 2008.

- Associated with the dataset proposed in Chapter 2:

J. Molina, J. A. Pajuelo, M. Escudero-Viñolo, J. Bescós, and J. M. Martínez, “A real and synthetic corpus for benchmarking of hand gesture recognition systems,” *Pattern Recognition Letters* (under review).

- In relation with the proposed approach for gesture scalability (Chapter 3):

J. Molina, and J. M. Martínez, “A Synthetic Training Framework for providing gesture scalability to 2.5D hand gesture recognition systems,” *Machine Vision and Applications* (under review).

- In relation with the proposed simple, compound and motion-based gestures detection framework (Chapter 4):

J. Molina, M. Escudero-Viñolo, A. Signoriello, M. Pardás, C- Ferrán, J. Bescós, F. Marqués, and J. M. Martínez, “Real-time user independent hand gesture recognition from time-of-flight camera video using static and dynamic models,” *Machine Vision and Applications*, pp. 1-18, 2011.

- In relation with the proposed motion-based gestures detection solution (Chapter 5):

J. Molina, J. A. Pajuelo, and J. M. Martínez, “Real-time Motion-based Hand Gestures Recognition from Depth Sensor video,” *IEEE Transactions on Consumer Electronics* (under review).

Appendix D: Conclusiones y trabajo futuro

El objetivo de esta tesis es el de ofrecer contribuciones en el ambito de reconocimiento de gestos manuales. Las principales contribuciones atienden a mejoras en usabilidad, escalabilidad y representatividad.

En primer lugar, en el Capítulo 2 se presenta una colección de videos capturados y de imágenes generadas sintéticamente, con información de profundidad. A su vez, se propone una novedosa colección de factores críticos a tener en cuenta en el proceso de elaboración de una colección de contenido asociado a gestos manuales. Distintas colecciones del Estado del Arte son analizadas en relación a la cobertura de estos factores críticos. La colección recopilada presenta una taxonomía en el tipo de gestos: basados en postura, basados en trayectoria, basados en postura y trayectoria y por último, compuestos. En terminos de representatividad, 11 usuarios reales participaron en la grabación de la colección. Además, variaciones del punto de vista son introducidas en el contenido sintético, aumentando así tambien la representatividad de la colección. Es importante destacar que el uso del metodo de generación sintética propuesto supone una importante mejora en terminos de escalabilidad, permitiendo la introducción de nuevas colecciones de gestos sin necesidad de grabar a usuarios reales. El sistema de generación de posturas sintéticas ha sido validado mediante un esquema de evaluación en el que el sistema es entrenado con contenido sintético mientras que la evaluación se realiza con usuarios reales. Los resultados son prometedores, alcanzando cifras comparables, para algunos diccionarios, a esquemas de entrenamiento con usuarios reales. En el Capítulo 3 el método para la generación de contenido sintético es extendido mediante el concepto de usuario sintético. Esta aproximación es evaluada en terminos de tasa de acierto, entrenando con la colección sintética y evaluado con usuarios reales. Los resultados mejoran en relación a la primera aproximación de generación sintética, alcanzando resultados, para algunos de los diccionarios, mejores que con esquemas de entrenamiento con un solo usuario real.

Haciendo uso de algunos de los diccionarios presentados en el Capítulo 2 dos sistemas de reconocimiento de gestos manuales son presentados y evaluados. En el Capítulo 4 se presenta

un sistema que premia la usabilidad, contemplando para ello gestos de distinta naturaleza. El descriptor utilizado para llevar a cabo la clasificación está basado en la identificación de puntos característicos dentro de la silueta de la mano, que es calculada usando el centro geodésico de la misma. Este descriptor, como se puede comprobar en [Molina et al., 2011] mejora los resultados obtenidos usando otros descriptores basados en información de contorno. Tres tipos de gestos son considerados: basados en postura, basados en trayectoria y compuestos (i.e. secuencia de posturas). El sistema ha sido evaluado con un número significativo de usuarios, obteniendo resultados independientes de usuario y comparables a los recogidos en el Estado del Arte. En cuanto a la usabilidad el sistema es capaz de funcionar en tiempo real, permitiendo una correcta interacción hombre-máquina. En el Capítulo 5 se presenta un segundo sistema de reconocimiento de gestos, en este caso basados en trayectoria. Además de ser capaz de funcionar en tiempo real, es robusto a variaciones en el punto de vista de la cámara y a la velocidad de ejecución de los gestos. A su vez, ha demostrado un correcto funcionamiento independientemente del usuario, habiendo sido realizada una evaluación con usuarios de distinta altura y distinto tamaño de brazo. Se propone una novedosa definición de los patrones de trayectoria basados en un modelo de brazo inspirado en la anatomía humana, los resultados tras una evaluación con usuarios reales acreditan la capacidad de representación del modelo propuesto. Se concluye que el uso de información de profundidad, además de mejorar computacionalmente la estimación de la trayectoria de la mano, mejora los resultados de clasificación sobre el uso de solo información 2D. Los resultados son comparables a los presentados en otros trabajos del Estado del Arte.

Como consecuencia de los resultados y conclusiones alcanzados en el desarrollo de esta tesis se proponen tres líneas de trabajo futuro:

- En cuanto a la escalabilidad (ver Capítulo 3) parece interesante la creación de nuevos usuarios sintéticos, con distintas características en sus manos artificiales. Estas modificaciones podrían introducirse variando el rango y la resolución con las que se muestrean los parámetros de configuración de la mano (ver Sección 3.2.2). Otra opción consiste en la variación de más parámetros de la mano, por ejemplo, el grosor de los dedos.
- En cuanto al sistema de reconocimiento de gestos manuales genéricos descrito en el Capítulo 4 se pueden plantear varias líneas de mejora: el refinamiento de la segmentación de la mano, por ejemplo, mediante la utilización de información de color; la mejora del descriptor utilizado; la integración de un modelo de cadenas ocultas de Markov para procesar la secuencia temporal de detecciones de posturas, para así hacer el sistema más robusto ante detecciones erróneas.
- En relación al sistema de reconocimiento de gestos manuales basados en trayectoria presentado en el Capítulo 5: la integración de un modelo de cadenas ocultas de Markov para procesar la secuencia temporal de detecciones de translaciones; la utilización de informa-

ción de color y profundidad registrada que podría permitir el reconocimiento de gestos más complejos.