UNIVERSITAT DE BARCELONA

Human Pose Analysis and Gesture Recognition from Depth Maps: Methods and Applications

Miguel Reyes Estany

UNIVERSITAT DE BARCELONA

Human Pose Analysis and Gesture Recognition from Depth Maps: Methods and Applications

Tesis realitzada pel candidat **Miguel Reyes Estany** a la Universitat de Barcelona, sota direcció de Dr. Sergio Escalera Guerrero, dins del programa de doctorat en **Ciències de la Computació**. Barcelona, October 21, 2016

Director	**Dr. Sergio Escalera** Dept. de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona & Centre de Visió per Computador
Thesis committee	**Dr. Santi Seguí** Dept. de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona Barcelona, Spain **Dr. Xavier Baró** Universitat Oberta de Catalunya & Computer Vision Center Barcelona, Spain **Dr. Jordi Gonzàlez** Dept. Ciències de la Computació & Computer Vision Center Barcelona, Spain

This document was typeset by the author using LaTeX 2_ε.

The research described in this book was carried out at the MAiA department, Universitat de Barcelona.

ISBN:

Printed by Ediciones Gráficas Rey, S.L.

Por ti Beatriz, por todo lo que me das...

Acknowledgements

Quien descubre el quién soy descubrirá el quién eres. Y el cómo, y el adónde.

Pablo Neruda

I would like to express my special appreciation and thanks to my advisor Dr. Sergio Escalera, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on both research as well as on my career have been priceless. I would also like to thank to Dr. Petia Radeva, for giving the pursue to join on research when I was finishing my undergraduate.

I would especially like to thank physicians and assistants at Instituto de Fisioterapia Global Mèzieres (IFGM) and Kinemez Medical Center. All of you have been there to support me when I recruited patients and collected data for my Ph.D. thesis. I am especially thankful with Mr. José Ramírez who has supported me throughout this work in clinical aspects. I am thankful for his aspiring guidance, invaluably constructive criticism and friendy advice during this project work. I am sincerely grateful to him for sharing their truthful and illuminating views on a number of issues related to the project.

I am really thankful to all my colleagues from HuPBA and office mates, in special to Toni, Miguel Ángel, Víctor, Xavi, Albert, Óscar and Adriana. It has been a really pleasure to share laughs and fun with you, unforgettable moments.

I express my warm thanks to MAiA department (University of Barcelona) and Computer Vision Center (Universitat Autònoma de Barcelona) for providing me with all the necessary facilities for the research.

I would also like to thank CheaLearn staff headed by Dr. Sergio Escalera and Dr. Isabelle Guyon. I am thankful and indebted to them for sharing expertise, and valuable guidance and encouragement extended to me.

Vull agrair als meus companys de l'Hospital de Sabadell Parc Taulí, en especial al Departament d'Admissions i al de Rehabilitació, tota l'empenta que em vau donar per embarcar-me en aquest projecte. Gràcies de tot cor per aquells matins, tardes i nits inolvidables.

Imposible olvidarme de los agradecimientos a mis amigos de Sabadell, en especial a los KAKTUS. Nos hemos visto crecer, podemos contar mil batallitas, nos apoyamos el uno al otro y ahí seguimos.

Finalment, vull agrair la força que he rebut de les dues persones més importants de la meva vida per dur a terme aquest treball. Bea, el teu suport ha sigut essencial, fent-me costat en els moments més tempestuosos. Amb tu he après que podem podem dur a terme qualsevol projecte que ens proposem, i els vull compartir tots amb tu. Miquel, m'he adonat que ets la meva font d'energia i, tot i que encara ets molt petit, m'has ensenyat que la vida és molt bonica.

Abstract

The visual analysis of humans is one of the most active research topics in Computer Vision. Several approaches for body pose recovery have been recently presented, allowing for better generalization of gesture recognition systems. The evaluation of human behaviour patterns in different environments has been a problem studied in social and cognitive sciences, but now it is raised as a challenging approach to computer science because of the complexity of data extraction and its analysis. The main difficulties of visual analysis in n RGB data is the discrimination of shapes, textures, background objects, changes in lighting conditions and viewpoint. In contrast to common RGB images used in Computer Vision, range images provide additional information about the 3-D world, allowing to capture the depth information of each pixel in the image. Furthermore, the use of depth maps is of increasing interest after the advent of cheap multisensor devices based on structured light, or Time of Flight (ToF) technology.

In this work we deal with the problem of analyzing human pose and motion in RGB-Depth images, and in particular: 1) human pose recovery, 2) hand pose description, and 3) gesture recognition. We will treated these three areas by using RGB-Depth data in order to take profit from visual representation and 3-D geometric information. Using both channels of information improves the efficiency of human pose and motion analysis methods. We also present efficient use of the proposed methods in real areas of application, such as eHealth and human computer interaction (HCI).

Principal objectives are establish the viability of depth map usage in human hand and body pose estimation and, in other hand, for gesture recognition. The presented research is also applied on real high impact applications.

Resumen

El análisis visual de personas es uno de los temas de investigación más activos en Visión Computacional. Varios enfoques para la recuperación de la postura corporal se han presentado recientemente, que permiten una mejor generalización de los sistemas de reconocimiento de gestos. La evaluación de los patrones de comportamiento humano en diferentes ambientes ha sido un problema de estudio en las ciencias sociales y cognitivas, pero actualmente se presenta como un reto para las ciencias informáticas, dada la complejidad de la extracción de datos y su análisis. Entre las principales dificultades del análisis visual de los datos n RGB está la discriminación de las formas, texturas, objetos de fondo, cambios en las condiciones de iluminación y puntos de vista. En contraste con las imágenes RGB comunes utilizadas en Visión Computacional, imágenes de rango aportan información adicional sobre mundo 3-D, lo que permite capturar la información de profundidad de cada pixel en la imagen. Además, el uso de mapas de profundidad es de creciente interés después de la llegada de los dispositivos multisensor baratos basados en luz estructurada, o la tecnología de Tiempo de Vuelo (TOF, por sus siglas en inglés).

En este trabajo analizaremos el problema de la postura y el movimiento humano en imágenes RGB con profundidad, y en particular: 1) la actitud humana de recuperación de la postura, 2) descripción de pocisiones de la mano, y 3) el reconocimiento de gestos. Vamos a tratar estas tres áreas mediante el uso de los datos RGB-Profundos con el fin de sacar provecho de la representación visual y la información geométrica en 3-D. El uso de los dos canales de información mejora la eficiencia de los métodos de análisis de movimiento y postura humanos. También presentamos un uso eficiente de los métodos propuestos en campos de aplicación real, como la salud y la interacción persona-ordenador (HCI).

Nuestros principales objetivos son establecer la viabilidad del uso de mapa de profundidad en la estimacin de pose de la mano y el cuerpo humano y, por otro lado, para el reconocimiento de gestos. Adicionalmente se presenta el impacto de éstas en aplicaciones reales con alto impacto social.

Contents

1 Introduction **1**
- 1.1 Background Information . . . 1
- 1.2 Research Problem . . . 3
- 1.3 Research Objectives . . . 3
- 1.4 Pose Analysis . . . 4
  - 1.4.1 Body Poses and Gesture Recognition . . . 5
  - 1.4.2 Hand Poses . . . 7
- 1.5 Thesis Achievements . . . 9
- 1.6 Conclusions . . . 11

2 Pose analysis in depth maps **13**
- 2.1 Introduction . . . 13
- 2.2 Method for Human Limb Segmentation based on RGB-D . . . 15
  - 2.2.1 Random Forest . . . 15
  - 2.2.2 Spatio-Temporal Graph-cut optimization . . . 18
  - 2.2.3 Experiments and preliminary results . . . 20
  - 2.2.4 Random forest results . . . 21
  - 2.2.5 Spatio-Temporal Graph-cuts results . . . 22
- 2.3 Material and methods . . . 23
  - 2.3.1 Sensors . . . 24
  - 2.3.2 Data and groundtruth definition . . . 24
  - 2.3.3 Methods . . . 25
- 2.4 Practical Development . . . 25
  - 2.4.1 Noise removal and surface reconstruction . . . 26
  - 2.4.2 Static posture analysis (SPA) . . . 27
  - 2.4.3 Spine curvature analysis (SCA) . . . 29
  - 2.4.4 Range of movement analysis (RMA) . . . 30
- 2.5 Results . . . 33
  - 2.5.1 Software details . . . 33
  - 2.5.2 System validation . . . 34
  - 2.5.3 Applications . . . 34
- 2.6 Conclusions . . . 36

3 Gesture Recognition in Depth Data **37**
- 3.1 Introduction . . . 37
- 3.2 Data Acquisition . . . 39
  - 3.2.1 Automatic calibration . . . 39
  - 3.2.2 Feature Vector Extraction . . . 40

3.3 Feature Weighting in DTW . . . 40
3.3.1 Begin-end of gesture detection . . . 41
3.3.2 Feature Weighting in DTW . . . 43
3.3.3 Results . . . 44
3.4 Multi-modal Gesture Recognition Challenge 2013: Dataset and Results . . . 48
3.4.1 Problem setting and data . . . 49
3.4.2 Data format and structure . . . 52
3.4.3 Protocol and evaluation . . . 54
3.4.4 Evaluation metric . . . 55
3.4.5 Results . . . 55
3.4.6 Statistics on the results . . . 56
3.4.7 Fact sheets . . . 58
3.4.8 Summary of the winner methods . . . 60
3.5 Conclusion . . . 63

4 Spherical Blurred Shape Model for 3D Object and Pose Recognition 65
4.1 Introduction . . . 65
4.2 Method . . . 67
4.2.1 Spherical Blurred Shape Model . . . 67
4.2.2 3D rotation invariant SBSM . . . 71
4.3 Quantitative analysis of SBSM . . . 72
4.3.1 Data sets . . . 72
4.3.2 Settings and evaluation metrics . . . 73
4.4 Experiments . . . 73
4.4.1 Analysis of classification performance . . . 73
4.4.2 Robustness to noise and deformations . . . 74
4.4.3 Statistical significance . . . 76
4.5 Qualitative analysis of SBSM . . . 79
4.5.1 Object spotting in 3D scenes . . . 79
4.5.2 HCI Application for medical navigation . . . 80
4.5.3 HCI Application for Intelligent retail . . . 81
4.5.4 HCI Application in Living labs . . . 82
4.6 Conclusions . . . 82

5 Conclusions 85

List of Figures

1.1 Example of the variability of low-cost and compact RGB-D devices. (a) Kinect first version launched in November 2010 by Microsoft Corp. RGB Resolution: VGA 640x480, Depth resolution: 320x240 (b) RGB-D camera launched in July 2011. RGB Resolution: SXGA 1280x1024, Depth resolution: 320x240. (c) Kinect second version launched in July 2014 by Microsoft Corp. RGB Resolution: HD 1080x1920, Depth resolution: 620x480. 2
1.2 Pose classification labels based on anatomical categories. 4
1.3 Example of signs representation based on hand pose. 8

2.1 Pipeline of the presented method, including the input depth information, Random forest, spatio-temporal Graph-cuts optimization, and the final segmentation result. 16
2.2 Interface for semi-automatic ground-truth generation. 20
2.3 Qualitative results; Ground Truth (a), RF inferred results (b), frame-by-frame GC results (c), and Temporally-coherent GC results (d). (e) Hand segmentation experiment. First row shows the ground-truth for two examples. Second row shows the RF classification results. Third row shows the final α-expansion GC segmentation results. 23
2.4 (a) Led sensor and infrared filter. (b) Validation of distances and angles. 24
2.5 Measuring spatial relations between infrared markers in order to manually label the single multi-modal frames. 25
2.6 Posture analysis system. 26
2.7 (a)-(c) Original depth map. (b)-(d) Filtered and resampled. 27
2.8 Static posture analysis example. 28
2.9 (a) 3D representation of static posture analysis example. 28
2.10 (a) Sample of analysis. (b) Automatically reconstructed 3D spinal cloud. (c) Geometric model to obtain anthropometric kyphosis and lordosis value. 30
2.11 Three-dimensional examination environment managed by the therapist. 30
2.12 Spinal curvature analysis example. 31
2.13 Skeletal model and example of selected articulations with computed dynamic range of movement (maximum opening and minimum closing values of a joint measured in degrees for a certain period of time). 33
2.14 (a) Interaction with the system while performing an anatomic spine analysis. (b) A patient is squatting in a physical rehabilitation treatment and receiving feedback from the system. 33
2.15 The system has been successfully applied in different real case scenarios: (a) posture reeducation, (b) physical rehabilitation, and (c) fitness conditioning. 35

3.1 detection and tracking in uncontrolled environments. 39
3.2 3D articulated human model consisting of 15 distinctive points. 40
3.3 Begin-end of gesture recognition of a model C in an infinite sequence Q. 43

3.4 Samples of gestures for different categories of the dataset. 45
3.5 Different calibration poses. 45
3.6 Gesture recognition example in a sequence of the dataset. 47
3.7 Data set gesture categories. 50
3.8 Skeleton joint positions.[1] 52
3.9 Different data modalities of the provided data set. 54
3.10 Best public score obtained in the validation set during the Challenge. 57
3.11 Correlation results among the validation and test results of the top ranked participants. . 57
3.12 Validation and test scores histograms. 57
3.13 Modalities considered. 58
3.14 Segmentation strategy. 58
3.15 Fusion strategy. 59
3.16 Learning strategy. 60
3.17 Programming language. 60
3.18 ExtraTreesClassifier Feature Importance. 61
3.19 Recognition of test sequence by the three challenge winners. Black bin means that the complete list of ordered gestures has been successfully recognized. 62
3.20 Deviation of the number of gesture samples for each category by the three winners in relation to the GT data. 62

4.1 Illustration of SBSM descriptor computation. (a) Sphere bins. (b) Example of neighbor bins. (c) and (d) example of the estimation of two main quaternion to rotate feature vector in the 3D space. 68
4.2 Example of point cloud voxel for an hypothetical sphere slice for $\theta = k$. Voxels of the point cloud visible on that slice are shown as red dots. An example of a voxel estimation p_z is shown in green. For this point, neighbor bins centroids are shown as black dots. For each of these relations (note that in the 3D space a total of 27 relations will be computed), equation 4.4 is computed, and the estimated value is added to descriptor position corresponding to its corresponding bin. 70
4.3 (a) Initial hand point cloud and computed center. (b) Sphere including a point cloud corresponding to a 3D hand pose. (c) The same sphere where SBSM descriptor has been computed. The density of the green dots represents the centroid bin values, and the whole descriptor has been rotated based on the quaternion codified by two main descriptor axis densities. (d) Alternative view of the computed SBSM descriptor. 70
4.4 RGB-Depth object data set category samples [49]. 72
4.5 (a) Input point cloud for a hand pose instance. (b) Example of distortion in the depth axis for (a). (c) For this distortion each voxel is randomly displaced in the z-axis with a maximum distortion of 20mm in both directions of the axis. (d) Input point cloud for a hand pose instance. (e) Example of cloud removal distortion for (d). (f) For this distortion each voxel of the original point cloud (top) is removed based on a probability value defined by the distortion (down). 75
4.6 Mean confusion matrix of the ASL data set using the SBSM descriptor $\sigma = 1$. Five most confused categories are displayed. 76
4.7 Classification performance of different classification strategies under different degrees of distortion in the depth axis on the 5 selected categories in the ASL data set. 76
4.8 Classification performance of different classification strategies under different degrees of cloud removal on the 5 selected categories in the ASL data set. 77

4.9 Object spotting in 3D scenes. (a) Example of RGB image of a multi-modal Berkeley data set. (b) Depth image of the same scene. (c) Computed point cloud from the scene. (d) Bowl spotting using SBSM (first positive 3D object prediction is shown based on minimum Euclidean distance). 78
4.10 (a) Original 3D Heart volume of http://thefree3dmodels.com. Automatic interaction with the volume with (b) translation, (c) rotation and (d) zoom manipulation. 78
4.11 HCI hand poses data set categories. 79
4.12 HCI for retail. (a) Designed 3D retail scenario using Unity engine. (b) User interaction with the scenario. (c) User manipulation by 3D rotation. 80
4.13 Hand reconstruction (c) using two Kinect^{TM} views of point clouds, right (a) and left (b). 80
4.14 82
4.15 83

List of Tables

2.1 Weights of edges in $\mathcal{E}$. 19
2.2 Average per class accuracy in % calculated over the test samples in a 5-fold cross validation. f_θ represents features of the depth comparison type from Eq.1, while g_θ - the gradient comparison feature from Eq. (2.13). O_{max} indicates the maximum absolute value for the x, y coordinates of the offsets $\mathbf{u}$ and $\mathbf{v}$. Parameter dt stands for tree depth. . . . 21
2.3 Average per class accuracy in% obtained when applying the different GC approaches –TC: Temporally coherent, Fbf: Frame-by-Frame– , and the best results from the RF approach, in the first row. 22
2.4 Pose and range of movement precision. 34
2.5 Validation of spinal analysis. 34

3.1 DTW begin-end of gesture recognition algorithm. 42
3.2 Feature Weighting in DTW cost measure. 44
3.3 Classification performance A over the gesture data set for the five gesture categories using DTW begin-end approach and including the Feature Weighting methodology. 46
3.4 Easy and challenging aspects of the data. 51
3.5 Top rank results on validation and test sets. 56
3.6 Team methods and results. Early and late refer to early and late fusion of features/classifier outputs. HMM: Hidden Markov Models. KNN: Nearest Neighbor. RF: Random Forest. Tree: Decision Trees. ADA: Adaboost variants. SVM: Support Vector Machines. Fisher: Fisher Linear Discriminant Analysis. GMM: Gaussian Mixture Models. NN: Neural Networks. DGM: Deep Boltzmann Machines. LR: Logistic Regression. DP: Dynamic Programming. ELM: Extreme Learning Machines. 61

4.1 Classification performance on the RGB-depth data set [49]. 74
4.2 Classification performance and confidence interval of the different descriptors on the novel American Sign Language data set. 74
4.3 Mean rank for the compared descriptors considering all the experiments. 77

Chapter 1

Introduction

The research in this memory is focused on human pose recovery and gesture recognition from depth maps, including its applications. These are common topics in Computer Vision and Artificial Intelligence fields, usually addressed by considering RGB images. However, there is a growing wealth of information regarding the use of depth images in place or in combination (multi-modal) with RGB images in order to deal with pose recovery and gesture recognition problems. The purpose of this section is to provide background information regarding these current depth-based approaches, current progress within the Computer Vision and Artificial Intelligence fields, our research problem and its significance, objectives, potential applications, and motivations for research.

1.1 Background Information

The visual analysis of humans is currently one of the most active research topics in Computer Vision. Several approaches for body pose recovery have been recently presented, allowing for better generalization of gesture recognition systems. The evaluation of human behaviour patterns in different environments has been a problem studied in social and cognitive sciences, but now it is raised as a challenging approach to computer science because of the complexity of data extraction and its analysis. The main difficulties of visual analysis in n RGB data is the discrimination of shapes, textures, background objects, changes in lighting conditions and viewpoint. In contrast to common RGB images used in Computer Vision, range images provide additional information about the 3-D world, allowing to capture the depth information of each pixel in the image, i.e. it is known the world coordinates of the points in the scene. Furthermore, the use of depth maps is of increasing interest after the advent of cheap multisensor devices based on structured light, or Time of Flight (ToF) technology. These RGB-Depth cameras are compact and portable, so it can be easily installed in any environment to understand 3-D scenes. This way there are multiple applications which can benefit from the analysis of 3-D objects in scenes. On the other hand, depth information is invariant to color, texture and lighting objects, making it easier to differentiate between the background and the foreground object of interest. The best-known depth-based device is Kinect. The Kinect sensor from Microsoft Corporation was introduced in the market in November 2010 as an input device for the Xbox 360 gaming console, achieving more than 10 million sells by March 2011. The Computer Vision society quickly discovered that the depth

sensing technology of Kinect could be used for a large variety of Computer Vision problems at a much lower cost than traditional 3D-cameras (such as time-of-flight based cameras). In June 2011, Microsoft released a software development kit (SDK) for Kinect, allowing it to be used as a tool for non-commercial products and spurring further interest in the product. Nowadays, it has been published several works related to this topic because of the emergence of inexpensive RGB-D devices, reliable and robust to capture the depth information along with its corresponding synchronized RGB image. In Figure 1.1 several devices using the described technology are shown.

(a) (b) (c)

Figure 1.1: Example of the variability of low-cost and compact RGB-D devices. (a) Kinect first version launched in November 2010 by Microsoft Corp. RGB Resolution: VGA 640x480, Depth resolution: 320x240 (b) RGB-D camera launched in July 2011. RGB Resolution: SXGA 1280x1024, Depth resolution: 320x240. (c) Kinect second version launched in July 2014 by Microsoft Corp. RGB Resolution: HD 1080x1920, Depth resolution: 620x480.

Examples of these low-cost depth sensors include Microsoft Kinect (aimed towards the Xbox 360) and ASUS Xtion, both of which have contributed significantly to research within this area. For example, Microsoft Kinect involve RGB-D sensors, aimed towards determining depth information through the use of structured light technology. The use of this type of sensor and technology allows for the inference of "depth values by projecting an infrared light pattern onto a scene and analysing the distortion of the projected light pattern" [36]. This type of sensor has limitations. For example, these sensors can only be used indoors, have low resolution, and have noisy depth information, all of which increases difficulties in the estimation of human poses from depth images [36]. There are other options, such as GPU, in order to increase frame rates and performance. Although this type of method has remarkable performance, difficulties arise in the operation of algorithms, especially in low-cost stems [36]. GPUs cannot be used in all methods, citing low frame rates and lack of ability to run in real time. In fact, research shows that "model-based approaches require model calibration before pose estimation" [36]. Following the high popularity of Kinect and its depth capturing abilities, there exists a strong research interest for improving the current methods for human pose and hand/body gesture recognition. While this could be achieved by inter-frame feature tracking and matching against predefined gesture models, there are scenarios where a robust segmentation of the hand/arm/body regions are needed, e.g. for observing upper limb anomalies or distinguishing between finger configurations while performing a gesture. In that respect, depth information appears quite handy by reducing ambiguities due to illumination, colour and texture diversity. Many researchers have obtained their first results in the field of human motion capture using this technology. In particular, Shotton et al 2011 [41] presented one of the greatest advances in the extraction of the human body pose from depth images, an approach that also forms the core of the Kinect human recognition framework.

This Thesis deals with the problem of analyzing human pose and motion in RGB-Depth images, and in particular: 1) human pose recovery, 2) hand pose description, and 3) gesture

recognition. These three areas will be treated by using RGB-Depth data in order to take profit from visual representation and 3-D geometric information. Using both channels of information improves the efficiency of human pose and motion analysis methods. For this reason, it has been also presented efficient use of the proposed methods in real areas of application, such as eHealth and human computer interaction (HCI).

1.2 Research Problem

There is a significant gap regarding current technologies due to rapid advancement of computer vision approaches and their abilities to relate to current population needs, especially within the medical field. This is largely due to the differences in analysis methodologies, suggesting that certain methodologies, such as RGB may result in higher costs, further suggesting that it is beneficial for newer methodologies, such as depth technology to be utilized in pose analyses and gesture recognition. It is also suggested that the gap is increasing due to the advancing possibilities of available technologies, such as through the use of human computer interactions [55]. Therefore, the research problem is established to be within the lack of current studies in relating to changing abilities of use of depth images in place of RGB techniques for pose analysis and gesture recognition. The research problem is significant for a variety of reasons. For example, through the use of depth images, it is possible to create cost effective technologies for non-invasive posture analysis and gesture recognition. It is expected that through meeting changing needs, it is possible to create new technologies that can be used in advanced situations, especially within the medical industry. Therefore, this research is significant because it can provide methodologies to establish diagnostic accuracies and verification, decreases in costs for diagnostic testing, and increases in quality of life. The study will advance current information to provide new possibilities for analysis techniques.

1.3 Research Objectives

The research objectives are developed based on the background information provided within this chapter, as well as considering the research problem and its significance. Based on this information, the research objectives are as follows:

1) To establish the viability of depth map usage in human hand and body pose estimation;

2) To establish the viability of depth map usage for gesture recognition;

3) To demonstrate how depth map use in pose analysis and gesture recognition can impact in real applications, including Human Computer Interaction and eHealth scenarios.

1.4 Pose Analysis

Pose recovery and gesture recognition approaches are very common within the computer vision and artificial intelligence field, particularly utilizing RGB images. However, there is a growing wealth of information regarding the use of depth images in place of RGB images during pose recovery and gesture recognition approaches. Video-based human activity analysis involves several research problems, such as recognition and detection/spotting, with several potential applications, such as surveillance through smart technology, the assisted living technologies, and human computer interaction [59]. Although the broad concept of video-based human activity analysis is relatively simple, the actuality is challenging [59]. Hand poses (or gestures) can be utilized in a variety of ways, prompting the use of pose analysis and gesture recognition. However, the use of hand poses suggests significant issues, especially in the design and creation of hand gesture recondition systems [55]. Recent research shows that these high costs are related to the different needs for new gesture introductions, such as the need for recording the new gestures through real subjects.

Pose analysis has shown to be useful in determining discriminative local features and their contextual information for action recognition [59]. The most of pose recovery approaches recover the human body pose in the RGB image plane. In order to update recent advances in the field of human pose recovery, it has been provided standard taxonomy to classify the State-of-the-Art of (SoA) model based approaches. The proposed taxonomy is composed of five main modules: appearance, viewpoint, spatial relations, temporal consistence, and behavior. Since this survey analyzes computer vision approaches for human pose recovery, image evidences should be interpreted and related to some previous knowledge of the body appearance. Depending on the appearance detected or due to spatio-temporal post processing, many works infer a coarse or a refined viewpoint of the body, as well as other pose estimation approaches restrict the possible viewpoints detected in the training dataset. Since the body pose recovery task implies the location of body parts in the image, spatial relations are taken into account. Figure 1.2 shows a possible human pose classification based on anatomical categories. In the same way, when a video sequence is available, the motion of body parts is also studied to refine the body pose or to analyze the behavior being performed.

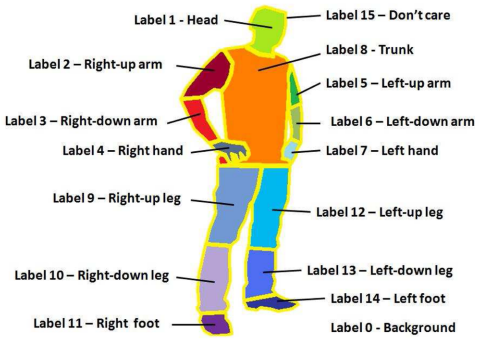


Figure 1.2: Pose classification labels based on anatomical categories.

Finally, the block of behavior refers, on the one hand, to those methods that take into account particular activities or the information about scene to provide a feedback to the previous modules, improving the final pose recognition. In recent years, Deep Learning techniques have shown to perform well on a large variety of Computer Vision problems, reaching and often surpassing the state of the art on many tasks. The rise of deep learning is also revolutionizing the entire field of Machine Learning and Pattern Recognition pushing forward the concepts of automatic feature extraction and unsupervised learning [87]. Deep Learning, or hierarchical learning, is a branch of machine learning which comprises a set of algorithms attempting to model high-level abstractions in data through model architectures with complex structures or otherwise, composed of multiple non-linear transformations. Deep Learning is part of a broader family of machine learning methods based on learning representations of data. However, despite the strong success both in science and business, deep learning has its own limitations. It is often questioned if such techniques are only some kind of brute-force statistical approaches and if they can only work in the context of High Performance Computing with tons of data. Inside this new disruptive field, recently emerged some novel pose recovery methods using non-linear mapping based on hierarchical learning. Formulating the pose estimation as a joint regression problem using Deep Learning Networks. The location of each body joint is regressed ton using as an input the full RGB image [80]. Other approach is using depth map sequences [82], in order to effectively extract the body shape and motion information. In the following sections pose recovery problem is handled with both input signals: RGB images and depth maps. There are two categorical classifications for human pose estimation methods: model-based and learning-based. In the first classification (model-based), it is necessary to have prior knowledge of a human body model and the human pose estimation occurs through inverting kinematics or the resolution of optimization problems [36]. In the second classification (learning-based), it is not necessary to consider a human body model. Rather, human poses are directly estimated from input images with various machine learning algorithms [36]. Thus, there are two different models utilized in human pose estimation. One is a per-pixel classification method, showing classifications of each pixel located on the human body through a trained classification random forest [36]. The second is a joint position classification method through use of a regression random forest. In this case, the pixels on the human body directly votes on all of the joint positions. The classified pixels or the joint position votes are aggregated to estimate joint points by a mean shift [36].

1.4.1 Body Poses and Gesture Recognition

Video-based human behavior analysis includes different research topics, such as action/gesture recognition or detection/spotting. Its applications are countless, including surveillance through smart technology, assisted living technologies, or human computer interaction [59]. Although the broad concept of video-based human activity analysis is relatively simple, its automatic implementation is a complex task. Some of the involved challenges include individual variations of people in posture, motion and clothing, camera motion, view angle changes, illumination changes, occlusions and self-occlusions, and background clutter [59]. However, spatio-temporal interest points and dense motion trajectories have been recently proposed as spatio-temporal motion features that showed excellent performance in the context of action recognition. Its main benefit is that spatio-temporal motion features are extracted throughout the entire sequence of the video, allowing the occurrences of encoded motion features to be accumulated for action representation, a phenomenon called feature

pooling [59]. This particular study also found that for different categories of actions, motion features present distinctive spatio-temporal distribution patterns [59]. Different actions are distinguished through certain spatio-temporal locations. At the same time, the pooling scheme does not consider information regarding spatio-temporal location of motion features, presenting two significant disadvantages [59]. The first disadvantage shows that informative motion features are only responsible for a small spatio-temporal region, causing other features that may be redundant and noisy to dominate the histogram representation, causing the discriminative capability to be downgraded. The second disadvantage shows that the pooling scheme's performance is typically downgraded as a result of clutter and background motion [59]. Commonly, image feature detection is performed by using image and object classification through spatial pyramid matching, which also allows for the achievement of geometrically invariant representation. This tactic is accomplished through "aggregating statistics of local features over fixed subregions" [59]. Using spatial pyramid matching can result in misalignment due to different object locations and screen layouts in image and object classifications. This misalignment may also occur during action representation because "informative motion features may occupy a small portion of the video volume with unknown spatio-temporal positions" [59]. In order to counteract these issues, it has been proposed to use object-centric spatial partition for image classification because this tactic determines the location of the object of interest before pooling low level features. This pooling occurs in separate locations (foreground and background) utilizing object-centric spatial partition [59]. It is suggested that advanced pooling schemes are more demanding for use in action representation due to the complexity of the video background and less description of individual motion features due to the larger inter-class variation [59]. Computer vision and robotic researchers, among others, tend to research human pose estimation and gesture recognition due to their increasing number of applications. Some of these applications include human computer interaction, game control, and surveillance [36]. In fact, research has increased within this context due to "the release of low-cost depth sensors" [36]. Dynamic gestures are characterized by both the pose and the motion of the relevant body parts. Much effort has traditionally be put into detecting first body parts and then tracking their motion. In color videos, detecting hands can be quite challenging, although better performance can be achieved by placing additional constraints on the scene and the relative position of the subject and the hands with respect to the camera (Cui and Weng, 2000; Isard and Blake, 1998; Kolsch and Turk, 2004; Ong and Bowden, 2004; Stefanov et al., 2005; Stenger et al., 2003; Sudderth et al., 2004). Commonly used visual cues for hand detection such as skin color, edges, motion, and background subtraction (Chen et al., 2003; Martin et al., 1998) may also fail to unambiguously locate the hands when the face, or other "hand-like" objects are moving in the background. Dynamic gesture recognition methods can be further categorized based on whether they make the assumption that gestures have already been segmented, so that the start frame and end frame of each gesture is known. Gesture spotting is the task of recognizing gestures in unsegmented video streams, that may contain an unknown number of gestures, as well as intervals were no gesture is being performed. Gesture spotting methods can be broadly classified into two general approaches: the direct approach, where temporal segmentation precedes recognition of the gesture class, and the indirect approach, where temporal segmentation is intertwined with recognition:

- Direct methods (also called heuristic segmentation) first compute low-level motion parameters such as velocity, acceleration, and trajectory curvature (Kang et al., 2004) or mid-level motion parameters such as human body activity (Kahol et al., 2004), and then look for abrupt changes (e.g., zero-crossings) in those parameters to identify candidate gesture boundaries.

- Indirect methods (also called recognition-based segmentation) detect gesture boundaries by finding, in the input sequence, intervals that give good recognition scores when matched with one of the gesture classes. Most indirect methods (Alon et al., 2009; Lee and Kim, 1999; Oka, 1998) are based on extensions of Dynamic Programming (DP) e.g., Dynamic Time Warping (DTW) (Darrell et al., 1996; Kruskal and Liberman, 1983), Continuous Dynamic Programming (CDP) (Oka, 1998), various forms of Hidden Markov Models (HMMs) (Brand et al., 1997; Chen et al., 2003; Stefanov et al., 2005; Lee and Kim, 1999; Starner and Pentland, 1998; Vogler and Metaxas, 1999; Wilson and Bobick, 1999), and most recently, Conditional Random Fields (Lafferty et al., 2001; Quattoni et al., 2007). Also hybrid probabilistic and dynamic programming approaches have been recently published (Hernandez-Vela et al., 2013a). In those methods, the gesture endpoint is detected by comparing the recognition likelihood score to a threshold. The threshold can be fixed or adaptively computed by a non-gesture garbage model (Lee and Kim, 1999; Yang et al., 2009), equivalent to silence models in speech.

When attempting to recognize unsegmented gestures, a frequently encountered problem is the subgesture problem: false detection of gestures that are similar to parts of other longer gestures. (Lee and Kim, 1999) address this issue using heuristics to infer the users completion intentions, such as moving the hand out of camera range or freezing the hand for a while. An alternative is proposed in (Alon et al., 2009), where a learning algorithm explicitly identifies subgesture/supergesture relationships among gesture classes, from training data. Another common approach for gesture spotting is to first extract features from each frame of the observed video, and then to provide a sliding window of those features to a recognition module, which performs the classification of the gesture (Corradini, 2001; Cutler and Turk, 1998; Darrell et al., 1996; Oka et al., 2002; Starner and Pentland, 1998; Yang et al., 2002)). Oftentimes, the extracted features describe the position and appearance of the gesturing hand or hands (Cutler and Turk, 1998; Darrell et al., 1996; Starner and Pentland, 1998; Yang et al., 2002)). This approach can be integrated with recognition-based segmentation methods.

1.4.2 Hand Poses

Hand poses (key feature for gesture recognition) can be utilized in a variety of ways, prompting the use of pose analysis and gesture recognition. For example, 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 in multimedia applications [55]. However, the use of hand poses suggests significant issues, especially in the design and creation of hand gesture recondition systems. These issues are primarily related to high cost, or gesture scalability, when the hand gesture recognition system is being developed to introduce new detectable gesture patterns [55]. Recent research shows that these high costs are related to the different needs for new gesture introductions, such as the need for recording the new gestures through real subjects. Therefore, it is possible to utilize different components of a training framework for hand posture detection systems based on a learning scheme fed with synthetically generated range images [55]. Through the use of different configurations in conjunction with a 3D hand model, it is possible to find sets of synthetic subjects. This has proven to allow for the effective separation of gestures from several dictionaries. As

a result, the training framework allows for learning new dictionaries, tested through real subject recordings, resulting in gesture scalability [55]. This has created an accuracy rate that is either comparable to or better than the established dictionaries. Figure1.3 shows a representation of different signs based on hand poses.

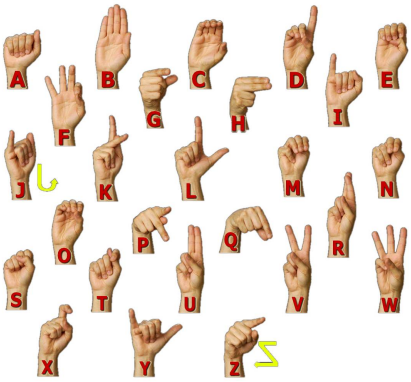


Figure 1.3: Example of signs representation based on hand pose.

The use of hand pose is to provide optimal user experiences. As a result, the use of human computer interaction can increase user experiences in a variety of ways [55]. This underlying goal is a significant cause of the rapid evolution of this type of technology. Currently, the most common capture technology is RGB cameras [55]. However, this is slowly changing to the use of range data information. These changes can largely be attributed to the fast establishment of 3D user interfaces in the last years: such kind of interfaces are becoming more important in the console gaming scenario [55]. In other cases, such as the use of desktop computers, hand usage is common through input devices, providing natural human computer interactions. Based on this study, three capture solutions exist for obtaining 3D information of a scene: by the use of markers or gloves; using RGB stereo-vision configurations; and using ToF cameras [55]. However, there are disadvantages to these solutions. For instance, marker usage can be intrusive for the user, whereas the use of stereo-vision capture solutions is complex to set up and requires singular point presences for different view registers [55]. Therefore, it is theorized that range cameras are a beneficial and increasingly effective way for pose analysis and gesture recognition to be measured because the price of range cameras does not stop decreasing even as the capabilities of these types of cameras continue increasing. Although range cameras are deemed to be cost-effective in most cases, it may be more important to recognize that the segmentation process that occurs with range cameras becomes easier through the use of a simpler set up than can be accomplished with stereo-vision solutions [55].

1.5 Thesis Achievements

1. Journals
    - A. Hernández, M. Reyes, V. Ponce, and S. Escalera. GrabCut-Based Human Segmentation in Video Sequences. Sensors. February, 2012.
    - A. Hernández, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov and S. Escalera. Human Limbs Segmentation in Depth Maps based on Spatio-Temporal Graph Cuts Optimization. Journal of Ambient Intelligence and Smart Environments. April, 2012.
    - A. Hernández, M. Reyes, S. Escalera, P. Radeva. GrabCut-based Human Segmentation in Video Sequences. International Journal of Pattern Recognition and Artificial Intelligence. October, 2012.
    - A. Clapés, M. Reyes and S. Escalera. Multi-modal user identification and Object Recognition Surveillance System. Pattern Recognition Letters, special issue in Multimodal Scene Understanding. March, 2013.
    - M. Reyes, A. Clapés, J. Ramirez, J.R. Revilla, L. Mejía, and S. Escalera. Automatic Digital Biometry Analysis Based on Depth Maps. Computers in Industry, 3D Imaging in Industry. June, 2013.
    - O. Lopes, M. Reyes, S. Escalera, J. Gonzàlez. Spherical Blurred Shape Model for 3-D Object and Pose Recognition: Quantitative Analysis and HCI Applications in Smart Environments, IEEE Transactions on System Man and Cybernetics. May, 2014.
2. Congress Participation
    - M. Reyes, S. Escalera, and P. Radeva, Real-time head pose classification in uncontrolled environments with Spatio-Temporal Active Appearance Models, CVCRD'10, Achievements and New Opportunities in Computer Vision, Barcelona, October 2010.
    - A. Hernández, M. Reyes, S. Escalera, and P. Radeva, Spatio- Temporal GrabCut Human Segmentation for Face and Pose Recovery, IEEE International Workshop on Analysis and Modeling of Faces and Gestures, Computer Vision and Pattern Recognition, San Francisco, June 2010.
    - M. Reyes, J, Ramírez, J.R. Revilla, P. Radeva, and S. Escalera, Non-Invasive Multisensor System for Automatic Data Acquisition target body, robust and reliable. Iberdiscap 2011: Ibero-American Congress of Technologies disability support, Majorca, June 2011.
    - V. Ponce, M. Reyes, X. Baró, M. Gorga, and S. Escalera, Two-level GMM Clustering of Human Poses for Automatic Human Behavior Analysis, CVCRD2011, VI CVC Workshop on the progress of Research Development, Barcelona, October 2011.
    - A. Hernández, M. Reyes, L. Igual, J. Moya, V. Violant, and S. Escalera, ADHD indicators modelling based on Dynamic Time Warping from RGBD data: a feasibility study, CVCRD2011, VI CVC Workshop on the progress of Research Development, Barcelona, October 2011.
    - M. Reyes, G. Domínguez, and S. Escalera, Feature Weighting in Dynamic Time Warping for Gesture Recognition in Depth Data, IEEE Workshop on Consumer Depth Cameras for Computer Vision, International Conference in Computer Vision , Barcelona, October 2011.

- A. Clapés, Alex Pardo, Miguel Reyes, S. Escalera, and Oriol Pujol. Intelligent Homecare Assistive Technology for People with Dementia in Smart Cities, Smart Cities Expo World Congress, Barcelona, 2012.
- A. Hernández, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov and S. Escalera. Graph Cuts Optimization for Multi-Limb Human Segmentation in Depth Maps. Computer Vision and Pattern Recognition, Providence, June 2012.
- A. Clapés, M. Reyes, and S. Escalera. User Identication and Object Recognition in Clutter Scenes based on RGB-Depth Analysis. VII Conference on Articulated Motion and Deformable Objects. Majorca, June 2012.
- A. Clapés, M. Reyes, and S. Escalera. User Identification and Object Recognition in Clutter Scenes Based on RGB-Depth Analysis, Articulated Motion of Deformable Objects, Majorca, June 2012.
- M. Reyes, A. Clapés, J. Ramírez, J.R. Revilla, and S. Escalera. Static and Dynamic Body Analysis in Physiotherapy and Rehabilitation, International Conference on Neurorehabilitation, Converging Clinical and Engineering Research on Neurorehabilitation Biosystems Biorobotics. Toledo, Spain, October 2012.
- M. Reyes, A. Clapés, J. Ramírez, J.R. Revilla, L. Mejía, and S. Escalera. Posture Analysis and Range of Movement Estimation using Depth Maps. International Congress of Pattern Recognition. Tokyo, November 2012.
- S. Escalera, X. Baró, J. Gonzàlez, M. Reyes, V. Ponce, H. J. Escalante, O. Lopes, V. Athitsos, and I. Guyon. Multi-modal Gesture Recognition Challenge 2013: Dataset and Results, Chalearn Multi-Modal Gesture Recognition Workshop, International Conference on Multimodal Interaction, ICMI, Sydney, December 2013.
- S. Escalera, X. Baró, J. Gonzàlez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce, H. J. Escalante, J. Shotton, and I. Guyon. ChaLearn Looking at People Challenge 2014: Dataset and Results, ChaLearn Looking at People, European Conference on Computer Vision, Zurich, July 2014.

3. Awards

   - M. Reyes, O. Lopes, M. Pousa, J. Gonzàlez, and S. Escalera. Third place award Chalearn Gesture Challenge, Demonstration Competition at International Congress of Pattern Recognition. Tokyo, November 2012

4. Book Chapters

   - M. Reyes, A. Clapés, J. Ramírez, J.R. Revilla, and S. Escalera. Static and Dynamic Body Analysis in Physiotherapy and Rehabilitation. Converging Clinical and Engineering Research on Neurorehabilitation. Converging Clinical and Engineering Research on Neurorehabilitation, pages 793-797. Springer Berlin Heidelberg.

5. Non Indexed Congress

   - M. Reyes, J. Vitrià, P. Radeva, and S. Escalera, Real-Time Activity Monitoring of Inpatients, Medical Image Congress of Catalonia, MICCAT, Girona, November 2011.
   - L. Igual, A. Hernández, S. Escalera, M. Reyes, J. Moya, J. Soliva, J. Faquet,O. Vilarroya, P. Radeva, Automatic Techniques for Studying Attention-Deficit/Hyperactivity Disorder, II Jornada R+D+I TIC Salut, Girona, 04/05/2011-05/05/2011, Girona, 2011.

- M. Reyes, J, Ramírez, J.R. Revilla, and S. Escalera, Posture Analysis and Range of Movement Estimation using Depth Maps. Automatic Techniques for Studying Attention-Deficit/Hyperactivity Disorder, III Jornada R+D+I TIC Salut, Girona, 2012.

6. Transfer Activity

   - M. Reyes, S. Escalera, J. Ramrez, J. R. Revilla, and P. Radeva, Registered software number B3342-11, ADiBAS Posture: Automatic Digital Biometry Analysis System, 2012.
   - Applications to the analysis of Medical Imaging and eHealth sector. Funder: Ministerio de ciencia e innovación. Reference Number: TIN2009-14404-C02-02. Headed by Petia Radeva. Associated from 2012 to 2013.
   - Mediminder: Sistema autónomo doméstico asistencial para pacientes con demencia y Alzhéimer. Funder: 6883 IMSERSO. Headed by Sergio Escalera. From 2012 to 2013.

1.6 Conclusions

Pose analysis and gesture recognition is becoming increasingly advanced as technology advances, prompting its increasing use within the medical field. Part of this is due to the expansion of capabilities, such as human computer interaction. However, part of this is due to changing needs within the human population. For example, humans are living longer, requiring more technology and quality of life enhancing devices. It has also been theorized that due to decreasing capabilities, such as low resolution, have impacted the ability to yield effective results from pose analyses and gesture recognition [36, 52]. Therefore, there is a significant gap regarding current technologies due to rapid advancement of aforementioned technologies and their abilities to relate to current population needs, especially within the medical field. This is largely due to the differences in analysis methodologies, suggesting that certain methodologies, such as RGB may result in higher costs, further suggesting that it is beneficial for newer methodologies, such as depth technology to be utilized in pose analyses and gesture recognition. It is also suggested that the gap is increasing due to the advancing possibilities of available technologies, such as through the use of human computer interactions [55, 59]. It has been recommended that a different technique be used for static body posture analysis, as well as the dynamic range of movement estimation of skeleton joints. This technique is based on 3D anthropometric information, is computed from multi-modal data, and combines RGB information obtained from a video camera and depth information obtained from an infared sensor [52]. The user is able to provide a specific set of keypoints, which allow for the alignment of RGB and depth data. Importantly, it is possible to determine range of movement for a set of joints. As a result, accurate measurements can be taken regarding posture, spinal curvature, and range of movement, resulting in a non-invasive procedure for posture analysis [52]. Through the use of this analysis, especially as it is coupled with gesture recognition, it is possible to determine physical exercise correctness through observing joint movements.

Chapter 2

Pose analysis in depth maps

Abstract

In this chapter it is presented a framework for object segmentation using depth maps based on Random Forest and Graph-cuts theory, and apply it to the segmentation of human limbs. First, from a set of random depth features, Random Forest is used to infer a set of label probabilities for each data sample. This vector of probabilities is used as unary term in $\alpha - \beta$ swap Graph-cuts algorithm. Moreover, depth values of spatio-temporal neighboring data points are used as boundary potentials. Results on a new multi-label human depth data set show high performance in terms of segmentation overlapping of the novel methodology compared to classical approaches. It is also presented a real application for human body posture assessment using this approach. It has been proposed a novel tool for static body posture analysis based on 3D anthropometric information computed from multi-modal data. These data combine RGB information from a video camera and depth from an infrared sensor. Given set of keypoints defined by the user, RGB and depth data are aligned, depth surface is reconstructed, keypoints are matching using a novel point-to-point fitting procedure, and accurate measurements about posture and spinal curvature are computed. Given a set of joints, range of movement measurements is also obtained. The system shows high precision and reliable measurements, being useful for posture reeducation purposes to prevent MSDs, such as back pain, as well as tracking the posture evolution of patients in physical rehabilitation treatments.

2.1 Introduction

Human motion capture is an essential acquisition technology with many applications in computer vision. However, detecting humans in images or videos is a challenging problem due to the high variety of possible configurations of the scenario, such as changes in the point of view, illumination conditions, and background complexity. An extensive research on this topic reveals that there are many recent methodologies addressing this problem [22, 33, 58, 83]. Most of these works focus on the extraction and analysis of visual features. These methods have made a breakthrough in the treatment of human motion capture, achieving high perfor-

mance despite the occasional similarities between the foreground and the background in the case of changes in light or viewpoint. In order to treat human pose recovery in uncontrolled scenarios, an early work used range images for object recognition or modeling [72]. This approach achieved a straighforward solution to the problem of intensity and view changes in RGB images through the representation of 3D structures. The progress and spread of this method came slowly since data acquisition devices were expensive and bulky, with cumbersome communication interfaces when conducting experiments. Recently, Microsoft has launched the Kinect, a cheap multisensor device based on structured light technology, capable of capturing visual depth information (RGBD technology, from Red, Green, Blue, and Depth, respectively).

In this chapter it is presented a framework for human-limb segmentation using depth maps based on RF and Graphcuts theory (GC) and apply it to the segmentation of human limbs. The use of GC theory has recently been applied to the problem of image segmentation, obtaining successful results [15, 38, 48]. RF is used to infer a set of probabilities for each data sample, each one indicating the probability of a pixel to belong to a particular label. Then, this vector of probabilities is used as unary term in the $\alpha - \beta$ swap GC algorithm. Moreover, depth of neighbor data points in space and time are used as boundary potentials. As a result, is obtained an accurate segmentation of depth images based on the defined energy terms. Moreover, as long as is obtained a priori likelihoods representing target appearance, the presented method is generic enough to be applicable in any other object segmentation scenario. The method is evaluated on a 3D dataset, obtaining higher segmentation accuracy compared to standard segmentation approaches. This framework illustrates that exercise recognition, based on human limbs segmentation, is reliable for physical rehabilitation purposes.

World Health Organization has categorized disorders of the muscle-skeletal system as the main cause for absence from occupational work and one of the most important causes of disability in elders in the form of rheumatoid arthritis or osteoporosis. It is estimated that 80% of the world population will suffer from musculo-skeletal disorders or dysfunctions (MSDs) during his life. As a result, MSDs lead to considerable costs for public health systems [29].

The body posture evaluation of a subject manifests, in different degrees, his level of physic-anatomical health given the behavior of bone structures, and especially of the dorsal spine. For instance, common MSDs such as scoliosis, kyphosis, lordosis, arthropathy, or spinal pain show some of their symptoms through body posture. This requires the use of reliable, noninvasive, automatic, and easy to use tools for supporting diagnostic. By the articulated nature of the human body, the development of this kind of systems is still an open issue.

Given the difficulty of finding a tool for measuring the posture of the human body at different configurations, digital biometry has become a very useful tool. Digital biometry is defined by the American Society of Anthropometric data as "the technology to obtain reliable information of the physical objects or the environment through the recording of images, its measurement or interpretation". The systems based on this technology are capable to estimate morphological or functional alterations, being a useful resource for health professionals.

The diagnostic evaluation of the anomalies follows through a careful study of musculo-skeletal structure and receptorial aspects. These diagnostic tools are based on monitoring anthropometric relationships with validated accuracy [11, 28, 30, 31, 37, 44, 63]. These kind of tools are minimally invasive and obtain good accuracy results in terms of precision but require an specific scene configuration, being necessary a camera calibration preprocessing due to use

of two-dimensional cameras in different planes. Another common handicap of these systems, is its reduced portability to perform a custom analysis for the therapist, these systems have been built highly parameterized for a specific type of analysis. Most of these systems only treat specific areas of the body, primarily the spinal deformities [3, 4, 24, 43, 44]. The solution more frequently applied to measure body posture consists of the installation and alignment of multiple cameras, applying stereo vision methodologies [45, 53]. This kind of system uses to be expensive and require specific and restricted illumination conditions. The main alternative is accelerometers [61], but these systems also use to be expensive and invasive. Furthermore, the location of accelerometers on the body of a subject is difficult and cannot obtain accurate results because of the spatial measurements of multi-axial articulations. There is another type of diagnostic tools that perform a quantitative clinical analysis, based on diagnostic predictions through analysis of anthropometric relations [5, 86]. These systems are able to predict diseases in premature state. Using these tools are highly restricted to a sector of the population, and its success rate is low for use in clinical diagnostics.
Regarding the postural and spinal analysis, compared to standard alternatives and supported by clinical specialists, the system shows high precision and reliable measurements to be included in the clinical routine.

2.2 Method for Human Limb Segmentation based on RGB-D

The depth-image based approach suggested in [75] interprets the complex pose estimation task as an object classification problem by evaluating each depth pixel affiliation with a body part label, using respective Probability Distribution Functions (PDF). The pose recognition phase is addressed by re-projecting the pixel classification results and inferring the 3D positions of several skeletal joints using the RF and mean-shift algorithms. The goal is to extend the work of [75] and combine it with a general segmentation optimization procedure to define a robust segmentation of objects in depth images. As a case study, are segmented pixels belonging to the following seven body parts[1]: LU/LW/RU/RW for arms, (from Left, Right, Upper and loWer, respectively), LH/RH for hands, and the torso. The pipeline of the segmentation framework is illustrated in Fig. 2.1.

2.2.1 Random Forest

Considering the human body a priori segmented from the background in a training set of depth images, the procedure for growing a randomized decision tree t is formulated over the same definition of a depth comparison feature as defined in [75],

$$f_\theta(I, \mathbf{x}) = \mathbf{d_I}\left(\mathbf{x} + \frac{\mathbf{u}}{\mathbf{d_I(x)}}\right) - \mathbf{d_I}\left(\mathbf{x} + \frac{\mathbf{v}}{\mathbf{d_I(x)}}\right), \tag{2.1}$$

[1]Note that the method can be applied to segment any number of labels of any object contained in a depth image.

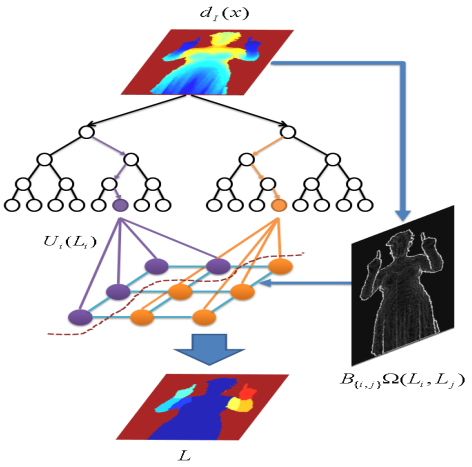


Figure 2.1: Pipeline of the presented method, including the input depth information, Random forest, spatio-temporal Graph-cuts optimization, and the final segmentation result.

where $d_I(\mathbf{x})$ is the depth at pixel $\mathbf{x}$ in image I, I is considered subspace of the Euclidean space E^2, $\theta = (\mathbf{u}, \mathbf{v})$, and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{\mathbf{2}}$ is a pair of offsets, the normalization of which ensures depth invariance. Thus, each θ determines two new pixels relative to $\mathbf{x}$, the depth difference of which accounts for the value of $f_\theta(I, \mathbf{x})$. Each tree consists of split and leaf nodes (the root is also a split node), as depicted in the upper part of Fig. 2.1. The training procedure of a given tree t over a unique set of ground truth images (avoid sharing images among trees), runs through the following steps:

1. Define a set Φ of node splitting criteria $\phi = (\theta, \tau)$, through the random selection of $\theta = (\mathbf{u}, \mathbf{v})$, and $\tau, \tau \subset \mathbb{R}$ (a set of splitting thresholds for each θ), with both θ and τ lying within some predefined range limits. After training, each split node will be assigned with its optimal ϕ value from Φ.

2. Define a set Q of training examples $Q = \{(I, \mathbf{x}) | \mathbf{x} \in \mathbf{I}\}$, over the entire set of training images for the tree, where I stands for an image $\mathbf{x}$ is a randomly selected pixel in I, and the number of pixels $\mathbf{x}$ per image is fixed. Estimate the PDF of Q over the whole set of labels C (for the data set $|C| = 7$),

$$P_Q(c) = \frac{h_Q(c)}{|Q|}, c \in C, \tag{2.2}$$

   where $h_Q(c)$ is the histogram of the examples from Q, associated with the label $c \in C$. Each example from Q enters the root node, thus ensuring optimal training of the tree t.

3. At the currently being processed node (starting from the root), split the (sub)set Q, entering this node, into two subsets Q_L and Q_R, obeying Eq. (2.1).

4. Estimate the best splitting criteria φ^* at the current node (starting from the tree root node), such that the information gain of partitioning the original set of pixels Q into

left and right subsets is maximum. The partitioning decision is taken per pixel so that

$$\begin{array}{rcl} Q_L(\phi) & = & \{(I,\mathbf{x})\,|f_\theta\,(I,\mathbf{x}) < \tau\}, \phi = (\theta,\tau) \\ Q_R(\phi) & = & Q \setminus Q_{left}, \end{array} \tag{2.3}$$

and estimate the PDF of Q_L, $P_{Q_L}(c)$, as defined in Eq. (2.2). Compute the PDF of Q_R, which may be speeded up by the following formulae,

$$P_{Q_R}(c) = \frac{|Q|}{|Q_R|}P_Q(c) - \frac{|Q_L|}{|Q_R|}P_{Q_L}(c), \qquad c \in C, \tag{2.4}$$

$$Q_R = Q_R(\phi), \quad Q_L = Q_L(\phi).$$

5. Estimate the best splitting criterion ϕ^* for the current node, so that the information gain $G_Q(\phi^*)$ of partitioning set Q entering the node, into left and right subsets to be maximum:

$$G_Q(\phi) = H(Q) - \frac{|Q_L(\phi)|}{|Q|}H(Q_L(\phi)) - \frac{|Q_R(\phi)|}{|Q|}H(Q_R(\phi)), \tag{2.5}$$

where $\phi = (\theta,\tau) \in \Phi$, and $H(Q) = -\sum_{c\in C} P_Q(c)\ln(P_Q(c))$ represents Shannon's entropy for the input (sub)set Q and its splits (Q_L and Q_R) over the set of labels C. It is more or less obvious that $G_Q(\phi) > 0$, $\phi \in \Phi$, but it is difficult to make a more analytical statement for the behaviour of $G_Q(\phi)$. That is why is also used the full search approach to evaluate $\phi^* = \arg\max_{\phi\in\Phi} G_Q(\phi)$.

6. Recursively repeat step 3 and 4 over $Q_L\,(\phi^*)$ and $Q_R\,(\phi^*)$ for the left and right node children respectively, until some preset stop conditions are met. The node where the stop condition occurred is treated as a leaf node, where, instead of ϕ^*, the respective PDF for the subset Q reaching the node, is stored (see Eq. (2.2)).

Once trained, each image pixel for recognition, i.e. an example $(I,\mathbf{x})$, is run through the tree, starting from the root and ending at a leaf node, taking a path that depends solely on the value of $f_\theta(I,\mathbf{x}) < \tau$, using the splitting criterion $\phi = (\theta,\tau)$, stored at the current tree node. The pixel acquires the PDF kept at the reached leaf node. The inferred pixel probability distribution within the forest is estimated by averaging the PDFs over all trees in the forest as follows,

$$P\,(c|I,\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} P_t\,(c|I,\mathbf{x})\,, c \in C \tag{2.6}$$

where $P\,(c|I,\mathbf{x})$ is the PDF stored at the leaf, reached by the pixel for classification $(I,\mathbf{x})$ and traced through the tree t, $t \in T$.

2.2.2 Spatio-Temporal Graph-cut optimization

GC is an energy minimization framework which has been considerably applied in image segmentation –both binary and multi-label–, with highly successful results. In this work, we extend the GC theory to be used in depth images and optimize the results obtained from the RF approach in order to deal with automatic spatio-temporal multi-label segmentation.

Given $I = \{I^1, ..., I^s, ..., I^S\}$ the set of frames of the video sequence, and $\mathcal{X} = (\mathbf{x_1}, ..., \mathbf{x_i}, ..., \mathbf{x}_{|\mathcal{P}|})$ the set of pixels of I, let us define $\mathcal{P} = (1, ..., i, ...|\mathcal{P}|)$ the set of indexes of I; $\mathcal{N}$ the set of unordered pairs $\{i, j\}$ of neighboring pixels in space and time, under a defined neighborhood system –typically 6- or 26-connectivity–, and $L = (L_1, ..., L_i, ..., L_{|\mathcal{P}|})$ a vector whose components L_i specify the labels assigned to pixels $i \in \mathcal{P}$. This framework defines an energy function $E(L)$ that combines local and contextual information, and whose minimum value corresponds to the optimal solution of the problem –in the dataset, the optimal segmentation:

$$E(L) = U(L) + \lambda B(L) \tag{2.7}$$

The first term of the energy function is called the "unary potential". This potential encodes the local likelihood of the data by assigning individual penalties to each pixel for each one of the defined labels $U(L) = \sum_{i \in \mathcal{P}} U_i(L_i)$. The second term or "boundary potential" encodes contextual information by introducing penalties to each pair of neighboring pixels as follows $B(L) = \sum_{\{i,j\} \in \mathcal{N}} B_{\{i,j\}}\ \Omega(L_i, L_j)$, where $\Omega(L_i, L_j)$ a function that introduces prior costs between each possible pair of neighboring labels. Finally, $\lambda \in \mathbb{R}^+$ is a weight that specifies the relative importance of the boundary term against the unary term.

Once the energy function is defined, a graph $\mathcal{G} =< \mathcal{V}, \mathcal{E} >$ is built following the neighborhood system used in the boundary potential $B(L)$. From a practical point of view, and considering that computer memory resources are limited, is adopted a sliding-window approach. More specifically, is defined a fixed size volume window V like the one depicted in the bottom of Fig. 2.1. The sliding-window approach starts segmenting the first $|V|$ frames, and covers all the video sequence volume, with a one-frame stride. This means that all the frames except the first and the last one are segmented at least twice, and $|V|$ times at most. In order to select the final hypothesis for each frame, is used the energy value resulting from the minimization algorithm at each execution. Therefore, the execution with the lowest energy value is the one trust as the best hypothesis. Once the graph is built with the energy function values, two main algorithms can be applied in order to find not the minimum energy, but a suboptimal approximation of it: $\alpha - \beta$ swap and α-expansion. While the first one is less restrictive and can be applied in a broader range of energy functions, the second one has been proven to obtain better results, as long as the energy function fulfills some conditions. In the case of $\alpha - \beta$ swap, the boundary term $B_{\{i,j\}}$ must be *semi-metric*, which means that the conditions in Eq. (2.8) and (2.9) must be fulfilled,

$$B(L_i, L_j) = B(L_j, L_i) \geq 0 \tag{2.8}$$

$$B(L_i, L_j) = 0 \leftrightarrow L_i = L_j \tag{2.9}$$

$$B(L_i, L_j) \leq B(L_i, L_n) + B(L_n, L_j), \tag{2.10}$$

for any $L_i, L_j, L_n \in L$, being $B(L_i, L_j) = B_{\{i,j\}}\ \Omega(L_i, L_j)$. Additionally, if is wanted to apply α-expansion, the condition in Eq. (2.10) must also be fulfilled. In that case, the boundary term $B_{\{i,j\}}$ is said to be *metric*.

In this case, Eq. (2.10) is not true for all nodes in $\mathcal{G}$, and so, is used $\alpha - \beta$ swap in the segmentation methodology for depth maps. This way, the set of nodes $\mathcal{V}$ contains a node for each pixel in I, plus two terminal nodes: α and β. Similarly, $\mathcal{E}$ is composed by two kinds of edges: terminal links t_i^{α} and t_i^{β}, and neighbor links $e_{\{i,j\}}$. The values assigned to the edges of $\mathcal{G}$ are then assigned following Table 2.1. The following subsections define the specific energy function potentials that is designed for the problem.

Unary potential. The unary potential encodes the local likelihood for each pixel to belong to each one of the labels L_i of the problem. In this case, it has been used the log-likelihood of the probabilities returned by the RF for the computation of the unary potential $U_i(L_i) = -\ln(P(c|I,x))$, obtaining a unary cost potential for each class c_i corresponding to label L_i in GC. This step is shown at the top of Figure 2.1, where the output probabilities of the leafs of the RF trees are used to compute the unary potentials $U_i(L_i)$ at the input edges of the GC graph.

Edge	Weight (cost)	For
t_i^{α}	$U_i(\alpha) + \sum_{\substack{j \in \mathcal{N}_i \\ L_j \notin \{\alpha,\beta\}}} B(\alpha, L_j)$	$L_i \in \{\alpha, \beta\}$
t_i^{β}	$U_i(\beta) + \sum_{\substack{j \in \mathcal{N}_i \\ L_j \notin \{\alpha,\beta\}}} B(\beta, L_j)$	$L_i \in \{\alpha, \beta\}$
$e_{\{i,j\}}$	$B(\alpha, \beta)$	$\substack{\{i,j\} \in \mathcal{N} \\ L_i, L_j \in \{\alpha,\beta\}}$

Table 2.1: Weights of edges in $\mathcal{E}$.

Boundary potential. In the case of the boundary potential, is used the formulation $B_{\{i,j\}} = \frac{1}{\mathsf{dist}(i,j)} e^{-\beta ||\mathbf{x}_i - \mathbf{x}_j||^2}$, where $\beta = \left(2\langle(\mathbf{x}_i - \mathbf{x}_j)^2\rangle\right)^{-1}$, and $\mathsf{dist}(i,j)$ computes the Euclidean distance between the cartesian coordinates of pixels $\mathbf{x}_i$ and $\mathbf{x}_j$. The information about the pixels $\mathbf{x}_i$ and $\mathbf{x}_j$ that is used in the exponential function is just the depth value, although in the experimental section are tested other additional approaches. Finally, are defined two different $\Omega(L_i, L_j)$ functions in order to introduce some prior costs between different labels. On one hand, is considered the trivial case where all different labels have the same cost,

$$\Omega_1(L_i, L_j) = \begin{cases} 0 & \text{for} \quad L_i = L_j \\ 1 & \text{for} \quad L_i \neq L_j \end{cases} \tag{2.11}$$

On the other hand, is introduced some spatial coherence between the different labels, taking into account the kinematic constraints of the human body limbs,

$$\Omega_2(L_i, L_j) = \begin{cases} 0 & \text{for} & L_i = L_j \\ 10 & \text{for} & L_i = \text{LU}, L_j = \text{RU} \\ & & L_i = \text{LH}, L_j = \text{RH} \\ 5 & \text{for} & L_i = \text{LW}, L_j = \text{RH} \\ & & L_i = \text{RW}, L_j = \text{LH} \\ 1 & \text{otherwise} & \end{cases} \tag{2.12}$$

With this definition of the inter-label costs, is made it difficult for the optimization algorithm to find a segmentation in which there exists a frontier between the right and left upper-arms, right and left hands, or in the lower measure, between left hand and right lower-arm, and vice-versa. Therefore, is assumed that poses in which the two hands are touching are not probable[2].

2.2.3 Experiments and preliminary results

This section starts with a brief description of the considered data and the different methods, parameters, and validation protocol of the evaluation.

Data: Is defined a new data set of several sessions where the actors are performing different gestures with their hands in front of the Kinect camera – only the upper body is considered. Each frame is composed by one 24 bit RGB image of size 640×480 pixels, one 12 bit depth buffer of the same dimension, and a skeletal graph describing important joints of the upper human body. In order to label every pixel is designed an editing tool to facilitate labelling in a semi-supervised manner. Each frame is accompanied with label buffer of the same dimension as the captured images. The label buffer is automatically initialized through rough label estimation algorithm. The pixels bounded by the cylinders between the enclosing joints of the shoulder to elbow are labelled as upper arm (LU/RU). By analogy the pixels inside the cylinder between the elbow and the joint of the hand are labelled as lower arm (LW,RW). The palm is labelled by the pixels bounded by a sphere centered in the joint of the hand (LH,RH). The RGB, depth, and skeletal data are directly obtained via the OpenNI library [1]. Finally each frame is manually edited to correct the roughly estimated labels by the initialization algorithm. The ground truth contains 2 actors in 3 sessions gathering 500 frames in total. An example of the developed interface is shown in Figure 2.14. It has been also made an extra experiment for finger segmentation defining 6 labels per hand - one label for each finger and one for the palm. In this case, are used coloured gloves and 63 frames were generated.

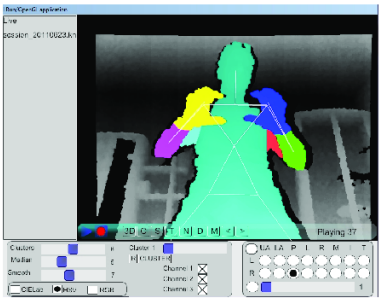


Figure 2.2: Interface for semi-automatic ground-truth generation.

[2]This label coherence cost should be estimated for each particular problem domain. In the particular data set of poses, the values of 1, 5, and 10 were experimentally computed.

Methods and validation: Inspired by the reported test parameters and accuracy results in [75], the experiments rest on the following setup: it performs a 5-fold cross-validation over the available 500 frames by training random forest of 3 trees of depth 20, 130 unique training images per tree, 1000 uniformly distributed pixels per image, 100 candidate features θ, and 20 thresholds τ per feature. Each test set consists of 100 images. Are also compare the results with another set of features: a mixture of depth Eq. (2.1) and gradient Eq. (2.13) features,

$$g_\theta(I, \mathbf{x}) = \measuredangle\left(\nabla I\left(\mathbf{x} + \frac{\mathbf{u}}{\mathbf{d_I}(\mathbf{x})}\right) - \nabla I\left(\mathbf{x} + \frac{\mathbf{v}}{\mathbf{d_I}(\mathbf{x})}\right)\right), \tag{2.13}$$

where $\nabla I(\mathbf{x})$ is the gradient vector at pixel $\mathbf{x}$, and the feature $g_\theta(I, \mathbf{x})$ represents the angle between the two gradient vectors at offsets $\mathbf{u}$ and $\mathbf{v}$ from $\mathbf{x}$. When applying GC, the λ parameter was set to 50 for all the performed experiments, the nodes of the graph are 10-connected –8 spatial neighbors + 2 temporoal neighbors–, and the size of the sliding window is set to $|V| = 5$. In order to do a richer comparison of the results, is performed an additional Graph-cuts experiment removing the temporal coherence. In this frame-by-frame approach, the α-expansion algorithm is used. Moreover, in this second experiment is also compare the use of different pixel information for the computation of the boundary term. Apart from just depth information, is also tested just RGB information, as in the standard GrabCut algorithm [38], and both RGB and depth together, resulting in a 4-D RGBD vector. Finally, is also apply the Friedman test [23] in order to look for statistical significance of the performed experiments.

Table 2.2: Average per class accuracy in % calculated over the test samples in a 5-fold cross validation. f_θ represents features of the depth comparison type from Eq.1, while g_θ - the gradient comparison feature from Eq. (2.13). O_{max} indicates the maximum absolute value for the x, y coordinates of the offsets $\mathbf{u}$ and $\mathbf{v}$. Parameter dt stands for tree depth.

	Torso	LU arm	LW arm	L hand	RU arm	RW arm	R hand	Avg. per class
100 f_θ, O_{max}=30, dt=20	92.90	73.29	71.42	57.75	74.25	76.26	59.38	72.18
100 f_θ, O_{max}=60, dt=20	94.17	79.83	77.69	**77.10**	81.04	82.65	80.17	81.81
80 f_θ, O_{max}=60, dt=20	**94.22**	79.08	76.46	74.19	81.24	83.26	79.05	81.07
60 f_θ, O_{max}=60, dt=20	94.09	78.86	75.86	73.49	79.43	82.60	78.08	80.34
100 f_θ, O_{max}=60, dt=15	94.06	79.81	78.69	76.59	81.18	83.10	**80.23**	**81.95**
100 f_θ, O_{max}=60, dt=10	91.83	**81.47**	**78.98**	72.30	**83.00**	**83.74**	76.85	81.17
60 f_θ+20 g_θ, O_{max}=60, dt=20	94.04	77.73	74.93	71.97	77.62	81.22	76.64	79.17

2.2.4 Random forest results

Table 2.2 shows the estimated average classification accuracy for each of the considered labels. Without claiming exhaustiveness of the experiments, the results from Table 2.2 allow us to make the following analysis: The maximum offset O_{max} has the greatest impact on the accuracy results at the hands regions, which are with the smallest area in the body part definition. Doubling the size of O_{max} leads to an increase in the accuracy of 20% for the

Table 2.3: Average per class accuracy in% obtained when applying the different GC approaches –TC: Temporally coherent, Fbf: Frame-by-Frame– , and the best results from the RF approach, in the first row.

	Torso	LU arm	LW arm	L hand	RU arm	RW arm	R hand	Avg. per class
RF results	94.06	79.81	78.69	76.59	81.18	83.10	80.23	81.95
TC, Depth, $\Omega_2\,(L_i, L_j)$	98.44	**78.93**	**84.38**	88.32	**82.57**	**88.85**	93.86	**87.91**
Fbf, Depth, $\Omega_1\,(L_i, L_j)$	98.86	75.05	82.87	91.45	77.57	87.35	93.96	86.73
Fbf, Depth, $\Omega_2\,(L_i, L_j)$	98.86	75.03	83.36	**92.41**	77.54	87.67	**94.20**	87.01
Fbf, RGB+Depth, $\Omega_1\,(L_i, L_j)$	**99.02**	72.02	81.86	90.29	76.56	86.84	92.14	85.53
Fbf, RGB+Depth, $\Omega_2\,(L_i, L_j)$	**99.02**	72.03	81.95	91.19	76.53	87.12	92.12	85.71

hands and 6% for the other body parts. In other words, O_{max} increases the feature diversity and the global ability to represent spatial detail. The number of candidate features Q would not have such a big impact on the accuracy as the O_{max} parameter, though a higher number helps identifying the most discriminative features. Is also tested the impact of the depth of the decision trees. Trimming the trees to depth 15 has a very little impact, showing an improvement of 0.1% on the average accuracy that may weakly be attributed to better classification at the lower arm regions. Trimming to depth 10 shows a 4% decrease in the accuracy at the hands, i.e. the tree is not trained well enough. Final test includes comparison over combination of both features f_θ and g_θ. Since the depth data provided by Kinect is noisy, is applied a Gaussian smoothing filter before calculating the image gradients and the feature from Eq. (2.13). Is chosen the gradient feature since it complements the relations of depth features with information about the orientation of local surfaces. In the test are not found significant differences in the performance of the RF approach when including this kind of features.

In order to show the generalization capability of the proposed approach, is carried out another case study, consisting of segmenting the finger regions. The results applying the same validation as in the previous case, show the best performance for the following setup: 1 tree of depth 15, 500 pixels per image, 100 candidate features Q, 20 thresholds τ per feature, and $O_{max} = 45$. The estimated average per class accuracy was 58.5% mostly due to the small number of training images. Fig. 2.3(e) displays a couple of test images comparing the ground truth and the inferred labels.

2.2.5 Spatio-Temporal Graph-cuts results

The results are obtained when applying the GC proposal over the probabilities returned by the RF are detailed in Table 2.3. Is possible to see how these results improve RF, and also the one obtained in the frame-by-frame approach. Best results are obtained when using only depth information for the computation of the boundary potential. In this case of study, adding RGB to the depth information reduce generalization of the boundary potential. Fig. 2.3(a)-(d) shows some qualitative results of the segmentations. Another interesting result is the improvement because of the influence of the prior costs given by the different $\Omega\,(L_i, L_j)$ functions. Having a look at the qualitative results in Fig. 2.3(a)-(d), one can firstly see how the spatial coherence introduced by the basic frame-by-frame Graph-cuts approach –Fig. 2.3 (c)–, allows to recover more consistent regions than the ones obtained with just the RF probabilities. Moreover, when introducing temporal coherence –Fig. 2.3 (d)–, the

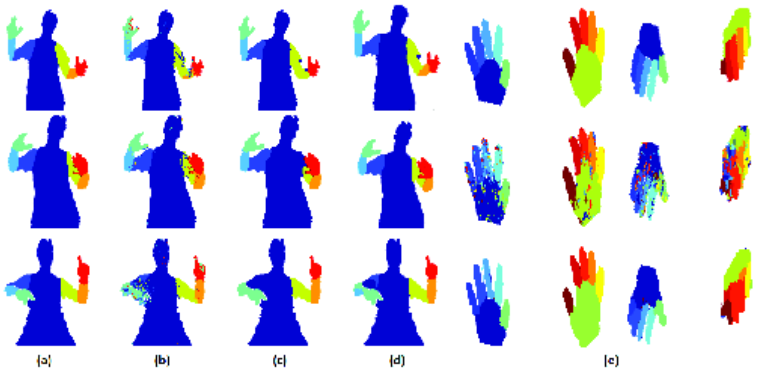


Figure 2.3: Qualitative results; Ground Truth (a), RF inferred results (b), frame-by-frame GC results (c), and Temporally-coherent GC results (d). (e) Hand segmentation experiment. First row shows the ground-truth for two examples. Second row shows the RF classification results. Third row shows the final α-expansion GC segmentation results.

classification of certain labels like the ones corresponding to the arms is improved.

Finally, in order to reject the null hypothesis that the measured ranks differ from the mean rank, and that the ranks are affected by randomness in the results, is used the Friedman test [23]. The rankings are obtained estimating each relative rank r_i^j for each test i and each segmentation strategy j, and computing the mean ranking R for each strategy as $R_j = \frac{1}{N}\sum_{i=1}^{N} r_i^j$, where N is the total number of performed tests. The Friedman statistic value is then computed as $X_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$. In this case, with $k = 6$ segmentation strategies to compare and ranks $R = [6, 2.5, 2.12, 4.12, 3.75, 2.12]$ in the order showed in Table 2.3, $X_F^2 = 20.06$. Since this value is undesirable conservative, Iman and Davenport proposed a corrected statistic $F_F = \frac{(N-1)X_F^2}{N(k+1)-X_F^2}$. Applying this correction is obtained $F_F = 7.04$. With six methods and ten experiments, F_F is distributed according to the F distribution with five and 35 degrees of freedom. The critical value of $F(5, 35)$ for 0.05 is 2.45. As the value of F_F is higher than 2.45 is possible to reject the null hypothesis, and thus, looking at the best mean performance in Table 2.3, we can conclude that the spatio-temporal GC proposal is the first choice in the presented experiments.

In the second experiment, labelling pixels from hands –in a frame-by-frame fashion–, is achieved an average per class accuracy of 70.9%, which supposes even a greater improvement than in the case of human limbs. Fig. 2.3(e) also shows some qualitative results of the GC approach.

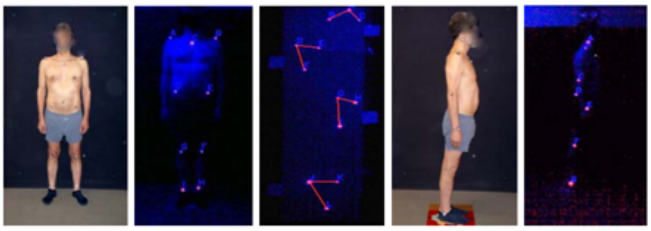
2.3 Material and methods

In this section, are provided the details about the sensors, data, and main methods involved in the development of the system for posture assessment. It uses RGB-Depth information to elaborate a semi-automatic postural and spinal analysis.

2.3.1 Sensors

The data acquisition is done using the Microsoft Kinect™ device. The device bases on a technology that combines a color camera sensor with a depth sensor. The depth sensor consists of an infrared laser projector combined with a monochrome CMOS sensor, which captures three-dimensional video data under any ambient light conditions, but it has a practical range limit of [1.2m - 3.5m] distance. The color video stream uses 8-bit VGA resolution (640×480) with a Bayer color filter, while the monochrome depth sensing video stream is in QVGA resolution (320×240 pixels) with 11-bit depth, which provides 2,048 levels of sensitivity. The device outputs video, to then be processed, at a frame rate of 30Hz.

A specific validation method to evaluate the system using infrared LEDs has been designed to assess the capabilities of Kinect™ as a measurement tool. Each infrared LED comprises a rechargeable button cell IR2032, a microswitch and a cold white color LED, 6000-8000mcd, 3V-20mW (Figure 2.4).



(a) (b)

Figure 2.4: (a) Led sensor and infrared filter. (b) Validation of distances and angles.

2.3.2 Data and groundtruth definition

In order to measure the precision of the proposed methodology in the different modules of the system, is created a novel data set consisting of two different parts.

A single multi-modal frame, comprising a still image and a depth map, is needed to perform an analysis in both static posture and spine curvature analysis. Thus, a battery of 500 single multi-modal frames has been labeled by three different observers (Figure 2.5), with an inter observer correlation superior to 99% for all planes (X, Y, Z). Each frame contains a set of angles and distances in order to simulate an analysis protocol for the study of posture, placing twelve infrared led markers on subject's skin. A total of 20 subjects participated in the validation of the method. With the aim to perform an automatic validation of the tests, infrared markers are detected by means of thresholding a HSV infrared-filtered image (filtered at 850 nanometers).

Sequences of multi-modal frames containing exercises for physical rehabilitation have been

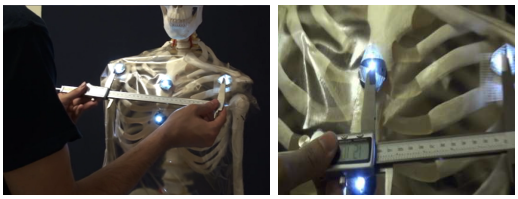


Figure 2.5: Measuring spatial relations between infrared markers in order to manually label the single multi-modal frames.

defined to validate range of motion analysis. For this purpose, a set of 25 sequences of one minute each one has been recorded, in which 5 different subjects appear performing repeatedly and correctly 5 exercises. This part of the data set contains approximately a total of 37300 semi-supervised labeled frames.

2.3.3 Methods

The postural analysis is done through the examination of 3D anthropometric values. Given a set of markers/keypoints defined by the user, the method performs the following steps: a) RGB and depth data are aligned, b) noise is removed and depth surface is reconstructed, c) user keypoints and predefined protocols are matched using a novel point-to-point fitting procedure, d) static measurements about posture and spinal curvature are accurately computed. The static measurements about posture consist of spatial relations between those keypoints: pairwise distances, distances relative to a vertical or horizontal axis, and angles among triplets of keypoints or angles between pairs relative to an axis. Regarding the spinal measurements, it is interpolated a curve representing the spine and also clinical spatial relations (distances and angles) among vertebrae are computed.

In order to evaluate exercise recognition for rehabilitation purposes, is presented a generic framework for object segmentation using depth maps based on RF and Graph-cuts theory (GC) and apply it to the segmentation of human limbs. The use of GC theory has recently been applied to the problem of image segmentation, obtaining successful results RF is used to infer a set of probabilities for each data sample, each one indicating the probability of a pixel to belong to a particular label. Then, this vector of probabilities is used as unary term in the $\alpha - \beta$ swap GC algorithm. Moreover, depth of neighbor data points in space and time are used as boundary potentials. As a result, is obtained a robust segmentation of depth images based on the defined energy terms.

2.4 Practical Development

Is designed a full functional system devoted to help in the posture reeducation task with the aim of preventing and correcting musculo-skeletal disorders, as well as tracking the posture evolution of patients in physical rehabilitation treatments. The system is composed by three main functionalities: a) static posture analysis (SPA), b) spine curvature analysis (SCA), and c) range of movement analysis, including automatic gesture recognition (RMA). The

architecture of the system is shown in Figure 2.6. First, a pre-processing step to remove noise and reconstruct surfaces is performed. Next, is described each of these stages.

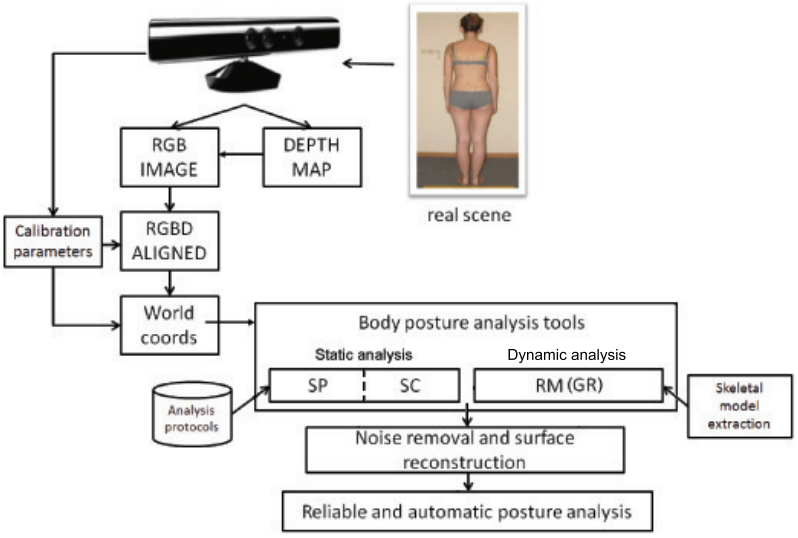


Figure 2.6: Posture analysis system.

2.4.1 Noise removal and surface reconstruction

After aligning RGB and depth data [64, 90], and even though the used depth information is compelling it is still inherently noisy. Depth measurements often fluctuate and depth maps contain numerous holes where no readings are obtained. In order to obtain a valid and accurate depth map, is performed a depth preprocessing step to eliminate erroneous information caused by noise and to reconstruct surfaces not well defined. Performing the following methodology:

Noise removal: For each point, is computed the mean distance from it to all its neighbors. By assuming that the resulted distribution is Gaussian with a mean and a standard deviation, all points whose mean distances are outside an interval defined by the global distances mean and standard deviation are considered as outliers.

Surface reconstruction: The surface reconstruction process rests on an inpainting image processing adaptation for 3D point cloud, based on compactly supported radial basis functions [84]. Radial basis function (RBF) interpolation is an important method for surface reconstruction [17] from 3D scatter points. The 3D point cloud extracted from the scene is encoded as a depth map. Then, the algorithm converts 2D image inpainting problem into implicit surface reconstruction problem from 3D points set. By performing this reconstruction small holes are reconstructed with high precision. Using compactly supported radial basis functions decreases the computational complexity in comparison to other common resampling approaches [88], which attempt to recreate the missing parts of the surface with higher order

polynomial interpolation among the surrounding data points. Figure 2.7 shows an example of this process [3].

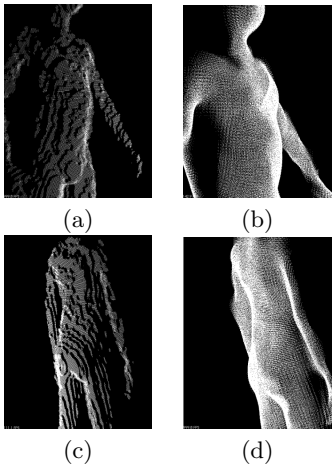


Figure 2.7: (a)-(c) Original depth map. (b)-(d) Filtered and resampled.

Once the data is aligned and depth maps are filtered, the user can access to the main functionalities of the system described below.

2.4.2 Static posture analysis (SPA)

This module computes and associates a set of three-dimensional angles and distances to keypoints defined by the user. These keypoints correspond to the dermal markers placed on the patient's skin. These dermal markers have to be physically placed by a therapist and, then, manually selected interacting with the RGB data displayed in the screen (which internally is aligned with the corresponding depth data, providing real 3D information) by the application to define the set of virtual markers (or keypoints). The keypoints could not be directly placed in the virtual scenario, because they have to correspond to specific body bone structures that need to be located by palpation.

The module also allows the therapist the possibility of designing a protocol of analysis. That is, a predefined set of angular-distance measurements among a set of body keypoints, all of them defined and saved by the user for a posterior automatic matching. The application of these protocols makes possible static posture analysis performed quickly, being highly customized, and done in an automatic way. Thus, when keypoints have been defined, the user can choose the most appropriate protocol and the system will finally provide the measurements automatically. Figure 2.8 shows an example of a predefined protocol (the set of manual annotated keypoints together with the list of distance and angle relations to

[3]Experimentally found that our approach for noise removal and background reconstruction obtained better results than standard approaches based on accumulating temporal images (e.g. 30 frames of a stationary subject) for noise reduction and hole filling.

be computed). Figure 2.9 shows an static posture analysis on 3D human body representation.

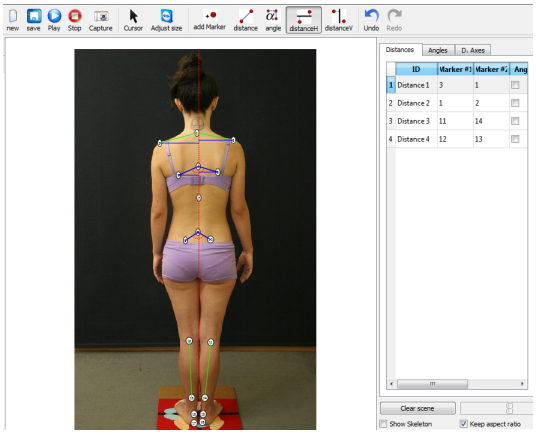
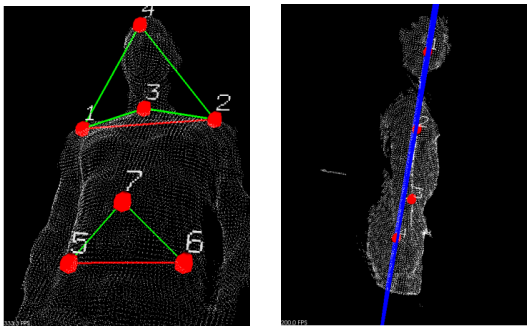


Figure 2.8: Static posture analysis example.



(a) (b)

Figure 2.9: (a) 3D representation of static posture analysis example.

In order to obtain an intelligent and automatic estimation of posture measurements, is defined a correspondence procedure among manually placed virtual markers and protocol markers. Formulating markers matching as an optimization problem. Suppose a protocol analysis (template) T composed by L markers, $T = \{T_1, T_2, ..., T_L\}$, $T_i = (x_i, y_i, z_i)$, and the current analysis C composed by the same number of markers, $C = \{C_1, C_2, ..., C_L\}$ (predefined template and current set of keypoints defined by the user, respectively). Our goal is to make a one-to-one correspondence so that is minimized the sum of least square distances among assignments as follows:

$$\text{argmin}_{C'} \sum_{i=1}^{L} \left\| C'_i - T_i \right\|^2, \tag{2.14}$$

where C' is evaluated as each of the possible permutations of the elements of C. For this task, first, is performed a soft pre-alignment between C and T using Iterative Closest Point (ICP) [26], and then, is proposed a sub-optimal approximation to the least-squares minimization problem. ICP is based on the application of rigid transformations (translation and rotation) in order to align both sequences C and T. This attempts to minimize the error of alignment $E(.)$ between the two marker sequences as follows:

$$E(\mathcal{R},\mathcal{T}) = \sum_{i=1}^{L}\sum_{j=1}^{L} w_{i,j}\|T_i - \mathcal{R}(C_j) - \mathcal{T}\|^2, \tag{2.15}$$

being $\mathcal{R}$ and $\mathcal{T}$ the rotation matrix and translation vector, respectively. $w_{i,j}$ is assigned 1 if the i-th point of T described the same point in space as the j-th point of C. Otherwise $w_{i,j} = 0$. Two things have to be calculated: First, the corresponding points, and second, the transformation ($\mathcal{R}$, $\mathcal{T}$) that minimizes $E(\mathcal{R},\mathcal{T})$ on the base of the corresponding points. For this task, is applied Singular Value decomposition (SVD). At the end of the optimization, the new projection of the elements of C is considered for final correspondence. Then, Eq. 2.14 is approximated as follows: Given the symmetric matrix of distances $\mathcal{M}$ of size $L \times L$ which codifies the set of $L\cdot(L-1)/2$ possible distances among all assignments between the elements of C and T, is set a distance threshold $\theta_{\mathcal{M}}$ to define the adjacency matrix A:

$$A(i,j) = \begin{cases} 1 & \text{if} \quad \mathcal{M}(i,j) < \theta_{\mathcal{M}} \\ 0, & \text{otherwise.} \end{cases} \tag{2.16}$$

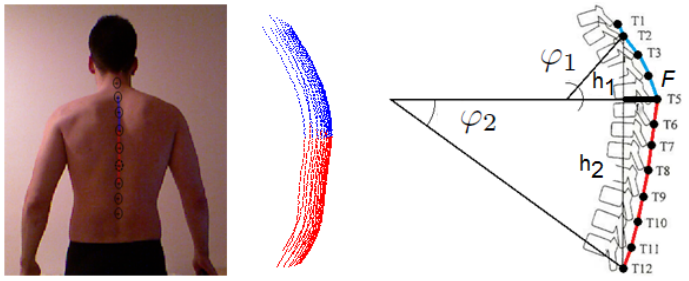
Then, instead of looking for the set of $L!$ possible assignments of elements of C and T that minimizes Eq. 2.14, only the possible assignments (C_i, T_j) that satisfies $A(i,j) = 1$ are considered, dramatically reducing the complexity of the correspondence procedure[4].

2.4.3 Spine curvature analysis (SCA)

The objective of this task is to evaluate sagittal spine curvatures (curves of the spine projected on the sagittal plane) by noninvasive graphic estimations in kyphotic and lordotic patients. Kyphosis and lordosis are, respectively, conditions of over-curvature of the thoracic spine (upper back) and the lumbar spine (lower back). The methodology proposed by Leroux et al [51] offers a three-dimensional analysis valid for clinical examinations of those conditions. In order to perform this analysis is proceeded as follows. First, the therapist places the virtual markers on the spine (an example of spine interaction and computation are shown in Figure 2.10(a)). Then, a few markers are selected and the 3D curve that represents the spine is reconstructed by linear interpolation (Figure 2.10(b)). Finally, the anthropometric kyphosis $\mathcal{K}_a$ and lordosis $\mathcal{L}_a$ are obtained.

The geometric model to compute $\mathcal{K}_a$ is represented in the Figure 2.10(c). F divides the curve representing the thoracic spine in two asymmetric arcs with different radius. hat the F component begins at the farthest marker (apex, corresponding to T5) and it ends at the intersection with the T2-to-T12 line. *h1* and *h2* are the distances from T2 to the intersection

[4]Experimentally found that high values of θ_M obtain optimal results and reduces then computational cost in comparison to other approaches, such as Shape Context [56].



(a) (b) (c)

Figure 2.10: (a) Sample of analysis. (b) Automatically reconstructed 3D spinal cloud. (c) Geometric model to obtain anthropometric kyphosis and lordosis value.

and the distance from the intersection to T12, respectively. Then, the summation of two angles, φ_1 and φ_2, represents the kyphosis curve value, where:

$$\begin{aligned} \varphi_1 &= 180 - 2 \cdot \arctan\left(\tfrac{h_1}{F}\right), \\ \varphi_2 &= 180 - 2 \cdot \arctan\left(\tfrac{h_2}{F}\right). \end{aligned} \tag{2.17}$$

$\mathcal{L}_a$ is calculated in a similar way, though the therapist should note the markers in the lumbar spine region.

The capacity analysis of the spine is reinforced by a three-dimensional environment that can be managed by the therapist for a thorough examination (Figure 2.11 and Figure 2.12).

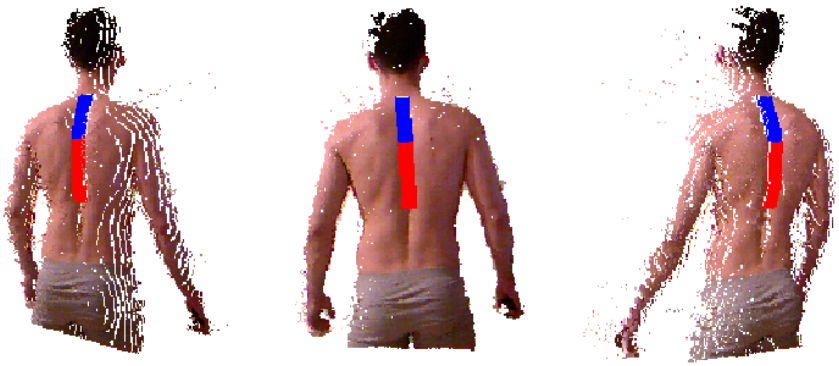


Figure 2.11: Three-dimensional examination environment managed by the therapist.

2.4.4 Range of movement analysis (RMA)

Computer vision has many applications in relation to human motion capture. However, since there are many different possible configuration possibilities, such as view point, conditions

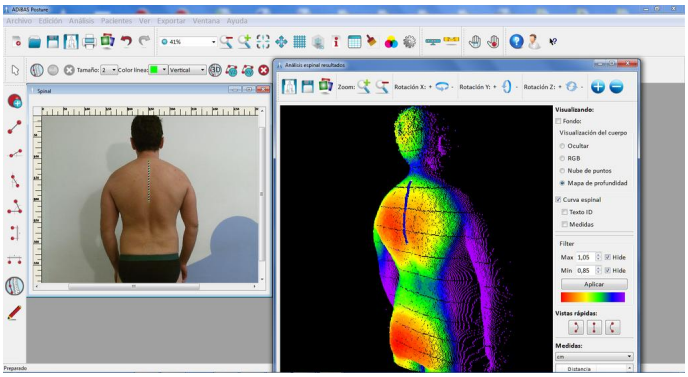


Figure 2.12: Spinal curvature analysis example.

of illumination, and complexity of background. Early studies utilized range image in order to accommodate for uncontrolled scenarios within human pose. This allowed for resolving intensity problems and view changes through RGB images. RGB images, in this context, are developed through 3D structure representation. Data acquisition devices during this time period, were exceedingly expensive and bulky. As a result, these data acquisition devices had communication interfaces that were cumbersome.

RGB-D technology has been used for obtaining first results in capturing human motions. The core of the Kinect human recognition framework allows for human body pose extraction through depth images [41]. This is conducted through Random Forest (RF), allowing for inferences to occur regarding pixel label probabilities. As a result, estimations of human joints are made utilizing the mean shift. This allows the human body to be represented in skeletal form. In fact, recent work conducted by researchers has used the skeletal model, as well as computer vision techniques, so that complex poses can be detected in cases where multiple actors are involved.

Human limb segmentation can effectively be done utilizing depth maps. These depth maps may be based on both RF and Graph cuts theory (GC). These depth maps can be applied to human limb segmentation. At the same time, GC theory has been applied to a host of situations, including image segmentation problems. However, RF has been successfully used in order to infer probability sets for unique data samples. As a result, each data sample has a unique probability in consideration of individual pixels and labels. Following this, probabilities are analyzed in the establishment of the GC algorithm. As a result of using boundary potentials based on neighbor data points depth, especially in consideration of space and time, it is possible to determine a robust segmentation of these depth images, significant because of defined energy terms. This method is compared to RF and GC approaches for accuracy, allowing for higher segmentation (Hernández-Vela et al., 2012[38]).

Utilizing the depth image approach allows for complex pose estimation to be interpreted through object classification, resolved through the evaluation of depth pixel affiliation in conjunction with a body part label, commonly using respective Probability Distribution Functions (PDF) (Hernández-Vela et al., 2012[38]). At the same time, pixel classification results are addressed by re-projection. This also allows for opportunities for the inferences of 3D positions using RF and mean-shift algorithms for several skeletal joints. The purpose of

utilizing depth maps is to create a segmentation optimization procedure in a general manner in order to allow for robust segmentation.

The depth difference accounts for the development of two new pixels. A given tree has split and lead nodes, creating images of a set of ground truth.

The process involves the definition of node splitting criterion. This is conducted through determining the random splitting thresholds. At the same time, range limits are predefined.

GC theory is applied in image segmentation and consists of a framework aimed at minimizing energy. Successful applications have occurred in binary and multi-label segmentation. However, the GC theory can be utilized in conjunction with depth images for RF approach results optimization (Hernndez-Vela et al., 2012[38]). This allows for automatic spatio-temporal multi-label segmentation. This specific framework defines energy functions that focus on the combination of local and contextual information. Furthermore, it is important to determine the minimum value that corresponds to the problem's optimal solution. The energy function has different terms. The first of which is the unary potential, which assigns penalties to pixels at defined labels, allowing for encoding of local likelihood (Hernndez-Vela et al., 2012[38]). This is followed by the second term, known as the boundary potential, which allows for encoding of contextual information for pairs of neighboring pixels (Hernndez-Vela et al., 2012[38]). The sliding-window approach is used in order to adapt to limited computer memory resources. The fixed size volume window is defined through segmenting different frames, covering all video sequence volumes through a one-frame stride. As a result, all frames, with the exception of the first and last frames, are segmented in two different instances. The lowest energy value is the one accepted for the best hypothesis. This is determined from the minimization algorithm that occurs at each execution occurrence. The use of two main algorithms, the minimum energy suboptimal approximation (swap and expansion) is found. Swap is noted to be less restrictive, especially in context of edge weight, allowing for energy functions to be analyzed in broader ranges. Expansion has better results, provided the function meets certain conditions (Hernndez-Vela et al., 2012[38]).

In order to complement the posture analysis procedure, is computed the range of movement of the different body articulations. This is a facility aimed to assist in diagnoses and physical rehabilitation treatments. For this purpose, is performed human limb segmentation using the RF - GC, approach described above [38] and, then, compute the skeletal model.

Computing the intersection borders among mean shift clusters estimated over the obtained limb labels of the RF - GC, is obtained a three-dimensional skeletal model composed by nineteen joints. The physician then selects joint articulations and automatically obtains their maximum opening and minimum closing values measured in degrees for a certain period of time (Figure 2.13).

However, to provide a fully-automatic functional range of movement estimation, it has been considered those cases in which a therapist is not present to operate with the system, for instance, when the patient is performing the rehabilitation exercises at home. Obviously, it would be useless obtaining range of movement measurements when an exercise is not well performed. This problem was solved by including an action/gesture recognition framework to the software.

Figure 2.13: Skeletal model and example of selected articulations with computed dynamic range of movement (maximum opening and minimum closing values of a joint measured in degrees for a certain period of time).

2.5 Results

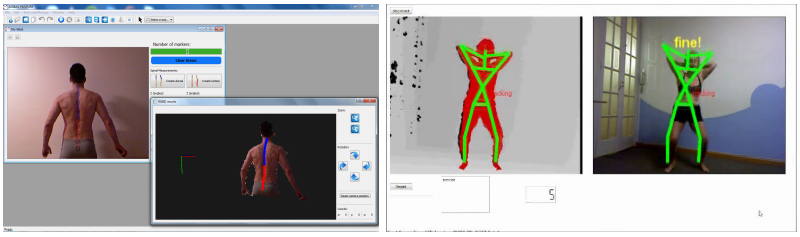
In order to present the results, are discussed the software details and validation procedures.

2.5.1 Software details

As a result of this work, it has been obtained a system that combines hardware - the Kinect™ device - and software.

Regarding the software implementation, it has been used the Kinect™ SDK framework. Also used the Point Cloud Library (PCL) to treat cloud points, and to support a free and three-dimensional visualization it has been used the Visualization Toolkit library (VTK). The user interface has been developed in the multi-platform Digia Qt technology. The programming language used was C++. The system runs fluently on a standard 2-CORE PC with 4GB of RAM.

In Figure 2.14, it is shown an example of interaction with the system through its graphical user interface.



(a) (b)

Figure 2.14: (a) Interaction with the system while performing an anatomic spine analysis. (b) A patient is squatting in a physical rehabilitation treatment and receiving feedback from the system.

2.5.2 System validation

The data described previously, in material and methods section, has been defined and used for the validation of the system.

Results for different distance of the device to the scene are shown in Table 2.4. AAV and '∘' correspond to the average absolute value and degree, respectively. This analysis validates the accuracy of the SPA and RMA in millimeters and degrees, respectively. Note the high precision in both tests. In addition, in order to validate the curvature analysis of the spine (SCA), it has been used a group of 10 patients and performed the Leroux protocol [51], placing nine markers over the spine. The relationship between lateral radiographic and anthropometric measures was assessed with the mean difference. It has used Cobb technique on the lateral radiograph in order to obtain the coefficients of kyphosis and lordosis. The results of the SPA validation are shown in Table 2.5. Moreover, after discussing with specialists in physiotherapy they agreed that the accuracy of the results is more than sufficient for diagnostic purposes.

Distance subject-device (m)	1,3	1,9	2,2
AAV (∘ movement)	2,2	3,8	5,2
AAV (mm)	0,98	1,42	2,1
AAV (∘ angles)	0,51	1,04	1,24
AAV (%)	0,46	0,77	1,3
Standard Error (%)	1,01	1,18	1,71

Table 2.4: Pose and range of movement precision.

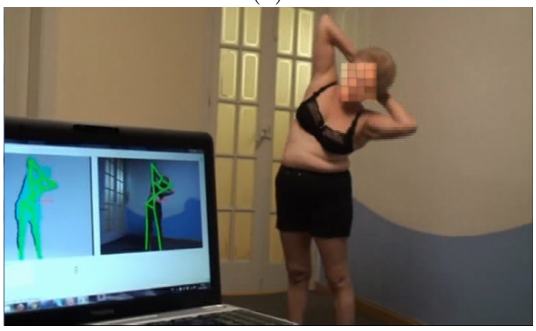
	Khyposis range	Lordosis range
AAV (∘)	5	6

Table 2.5: Validation of spinal analysis.

2.5.3 Applications

The main scenarios of application of the system are posture analysis, physical rehabilitation, and fitness conditioning. In Figure 2.15, some real examples where are tested the system are shown.

(a)



(b)

(c)

Figure 2.15: The system has been successfully applied in different real case scenarios: (a) posture reeducation, (b) physical rehabilitation, and (c) fitness conditioning.

2.6 Conclusions

It has been proposed a generic framework for object segmentation using depth maps based on Random Forest and Graph-cuts theory in order to benefit from the use of spatial and temporal coherence, and applied it to the segmentation of human limbs. Random Forest estimated the probability of each depth sample point to belong to a set of possible object labels, while Graph-cuts was used to optimize, both locally, spatially and temporally, the RF probabilities. Results on two novel data sets showed high performance segmenting several body parts in depth images compared to classical approaches.

It has been designed and built a system for semi-automatic static posture and spine curvature analysis from 3D anthropometric data acquired by a low cost RGB-Depth camera. The aim of the system is to assist in the prevention and treatment of musculo-skeletal dysfunctions. It has been integrated several frontier computer vision and artificial intelligence techniques including three-dimensional visual data processing, statistical learning, and time series analysis. Furthermore, the system is meant to be highly adaptable and customizable to the needs of the therapist. The validation study shows high precision and reliable measurements in terms of distance, degree and range of movement estimation. Supported by clinical specialists, it is suitable to be included in the clinical routine, including posture reeducation, rehabilitation, and fitness conditioning scenarios.

It is necessary to have effective body posture analysis and evaluation due to different physical conditions, including “common MSDs such as scoliosis, kyphosis, lordosis, arthropathy, or spinal pain”. Body posture analysis is needed to show the symptoms related to some of these disorders, prompting the need for “reliable, noninvasive, automatic, and easy to use tools for supporting diagnostic”. The development of these systems is still difficult due to continuous changes in the human body. However, it is suggested that digital biometry is an effective way to measure human body posture. Digital biometry is considered to be technology that is used to obtain reliable information regarding physical objects through recording images, measuring images, or interpreting images. The novel system uses RGB-Depth information to elaborate a semi-automatic postural and spinal analysis, and to automatically estimate the range of movement of the different limbs. Posture assessment and range of movement estimation are also reliable for physical rehabilitation purposes.

Chapter 3

Gesture Recognition in Depth Data

Abstract

In this chapter it is presented a gesture recognition approach for depth video data based on a novel Feature Weighting approach within the Dynamic Time Warping framework. Depth features from human joints are compared through video sequences using Dynamic Time Warping, and weights are assigned to features based on inter-intra class gesture variability. Feature Weighting in Dynamic Time Warping is then applied for recognizing begin-end of gestures in data sequences. The obtained results recognizing several gestures in depth data show high performance compared with classical Dynamic Time Warping approach. In order to promote the research advance on this field, it has been also organized in collaboration with "ChaLearn Look at People", a challenge on multi-modal gesture recognition. It has been made available a large video database of 13,858 gestures from a lexicon of 20 Italian gesture categories recorded with a Kinect^{TM} camera, providing the audio, skeletal model, user mask, RGB and depth images. The focus of the challenge was on *user independent multiple gesture learning*. There are no resting positions and the gestures are performed in continuous sequences lasting 1-2 minutes, containing between 8 and 20 gesture instances in each sequence. As a result, the dataset contains around 1.720.800 frames; 54 international teams participated in the challenge, and outstanding results were obtained by the first ranked participants.

3.1 Introduction

Visual analysis of human motion is currently one of the most active research topics in Computer Vision. Several segmentation techniques for body pose recovery have been recently presented, allowing for better generalization of gesture recognition systems. The evaluation of human behavior patterns in different environments has been a problem studied in social and cognitive sciences, but now it is raised as a challenging approach to computer science due to the complexity of data extraction and its analysis.

In this chapter, is presented a system for gesture recognition using depth data. From the point of view of data acquisition, many methodologies treat images captured by visible-

light cameras. Computer Vision are then used to detect, describe, and learn visual features [22, 57]. The main difficulties of visual descriptors on RGB data is the discrimination of shapes, textures, background objects, changing in lighting conditions and viewpoint. On the other hand, depth information is invariant to color, texture and lighting objects, making it easier to differentiate between the background and the foreground object. The first systems for depth estimation were expensive and difficult to manage in practice. Earlier research used stereo cameras to estimate human poses or perform human tracking [72]. In the past few years, some research has focused on the use of time-of-flight range cameras (TOF) [42, 67, 81]. Nowadays, it has been published several works related to this topic because of the emergence of inexpensive structured light technology, reliable and robust to capture the depth information along with their corresponding synchronized RGB image. This technology has been developed by the PrimeSense [65] company and marketed by Microsoft XBox under the name of Kinect. Using this sensor, Shotton et al. [76] present one of the greatest advances in the extraction of the human body pose from depth images, representing the body as a skeletal form comprised by a set of joints.

Once features are extracted from video data, the second step is the classification of gestures with the aim of describing human behavior. This step is extremely challenging because of the huge number of possible configurations of the human body that defines human motion. In this case, is based on the fifteen joints extracted by the approach of [76] as the set of features that will define the different gestures to recognize. A common approach for gesture recognition to model sequential data is Hidden Markov Model (HMM) [21], which is based on learning the transition probabilities among different human state configurations. Recently, there has been an emergent interest in Conditional Random Field (CRF) [77] for the learning of sequences. However, all these methods assume that is known the number of states for every motion. Other approaches make use of templates or global trajectories of motion [18], being highly dependent of the environment where the system is built. In order to avoid all these situations, the proposal is focused within the Dynamic Time Warping framework (DTW) [91]. Dynamic Time Warping allows to align two temporal sequences taking into account that sequences may vary in time based on the subject that performs the gesture. The alignment cost can be then used as a gesture appearance indicator.

The main contribution here is the introduction of a new method based on DTW for gesture recognition using depth data. It is proposed a Feature Weighting approach within the DTW framework to improve gesture/action recognition. First, it is estimated a temporal feature vector of subjects based on the 3D spatial coordinates of fifteen skeletal human joints. From a set of different ground truth behaviors of different length, DTW is used to compute the inter-class and intra-class gesture joint variability. These weights are used in the DTW cost function in order to improve gesture recognition performance. It has been tested the approach on several human behavior sequences captured by the Kinect sensor. It is showed the robustness of the novel approach recognizing multiple gestures, identifying beginning and end of gestures in long term sequences, and showing performance improvements compared with classical DTW framework.

The rest of the chapter describes the Challenge organized, called “Multi-modal Gesture Recognition Challenge”, in order to promote the research advance on gesture recognition based on multi-modal data.

3.2 Data Acquisition

This section describes the processing of depth data in order to perform the segmentation of the human body, obtaining its skeletal model, and computing its feature vector.

For the acquisition of depth maps is use the public API OpenNI software[2]. This middleware is able to provide sequences of images at a rate of 30 frames per second. The depth images obtained are 340×280 pixels resolution. These features are able to detect and track people to a maximum distance of six meters from multi-sensor device, as shown in Figure 3.1.

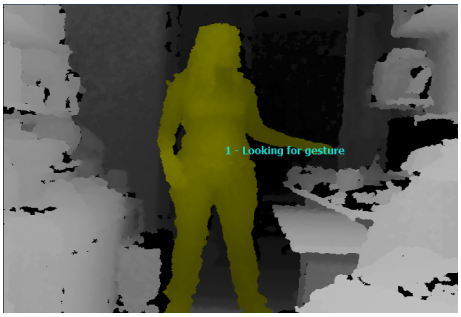


Figure 3.1: detection and tracking in uncontrolled environments.

It is used the method of [76] to detect the human body and its skeletal model. The approach of [76] uses a huge set of human samples to infer pixel labels through Random Forest estimation, and skeletal model is defined as the centroid of mass of the different dense regions using mean shift algorithm. Experimental results demonstrated that it is efficient and effective for reconstructing 3D human body poses, even against partial occlusions, different points of view or no light conditions. The main problem of the skeletal representation is that it requires from a reference pose for initialization. In this sense, is performed an automatic calibration to fit human model and automatically obtain skeletal representation.

3.2.1 Automatic calibration

In order to define an automatic fitting of the human body without the need of detecting a specific pose for calibration, is defined a set of silhouettes associated to plausible normalized models of the skeletal form $\mathbf{C}$. These models make possible initialization of the skeletal model for a set of silhouettes. In order to find the model that best fits the subject of the scene is used a similarity function of structure θ between consecutive frames. This similarity function is based on the alignment of the consecutive skeletal joints to obtain their respective Euclidean distance in two-dimensional space, using ζ as a similarity threshold value. Thus, the initialization of silhouette is given by,

$$\mathsf{argmin}_i \theta(\mathbf{C}, \mathbf{C_i}), \quad \theta(\mathbf{C}, \mathbf{C_i}) < \zeta \tag{3.1}$$

being $\mathbf{C}$ the skeletal model at the current frame and $\mathbf{C_i}$ the i-th trained model.

3.2.2 Feature Vector Extraction

The articulated human model is defined by the set of 15 reference points shown in Figure 3.2. This model has the advantage of being highly deformable, and thus, able to fit to complex human poses.

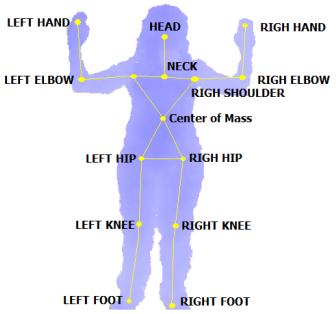


Figure 3.2: 3D articulated human model consisting of 15 distinctive points.

In order to subsequently make comparisons and analyze the different extracted skeletal models, is needed to normalize them. In this sense, it is used the neck joint of the skeletal model as the origin or coordinates (OC). Then, the neck is not used in the frame descriptor, and the remaining 14 joints are using in the frame descriptor computing their 3D coordinates with respect to the OC. This transformation allows us to relate pose models that are at different depths, being invariant to translation, scale, and tolerant to corporal differences of subjects. Thus, the final feature vector $\mathbf{V_j}$ at frame j that defines the human pose is described by 42 elements (14 joints $\times$ three spatial coordinates),

$$V_j = \{\{v^1_{j,x}, v^1_{j,y}, v^1_{j,z}\}, ..., \{v^{14}_{j,x}, v^{14}_{j,y}, v^{14}_{j,z}\}\}$$

3.3 Feature Weighting in DTW

The original DTW algorithm [60] was defined to match temporal distortions between two models, finding an alignment warping path between the two time series $Q = \{q_1, .., q_n\}$ and $C = \{c_1, .., c_m\}$. In order to align these two sequences, a $M_{m\times n}$ matrix is designed, where the position (i, j) of the matrix contains the distance between c_i and q_j. The Euclidean distance is the most frequently applied. Then, a warping path,

$$W = \{w_1, .., w_T\}, \max(m, n) \leq T < m + n + 1$$

is defined as a set of "contiguous" matrix elements that defines a mapping between C and Q. This warping path is typically subjected to several constraints:

Boundary conditions: $w_1 = (1, 1)$ and $w_T = (m, n)$.

Continuity: Given $w_{t-1} = (a', b')$, then $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$.

Monotonicity: Given $w_{t-1} = (a', b')$, $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$, this forces the points in W to be monotically spaced in time.

The focus of interest is at the final warping path that satisfying these conditions minimizes the warping cost,

$$DTW(Q, C) = \min \left\{ \frac{1}{T} \sqrt{\sum_{t=1}^{T} w_t} \right\} \tag{3.2}$$

where T compensates the different lengths of the warping paths. This path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance $\gamma(i, j)$ as the distance $d(i, j)$ found in the current cell and the minimum of the cumulative distance of the adjacent elements,

$$\gamma(i, j) = d(i, j) + \min\{\gamma(i - 1, j - 1), \gamma(i - 1, j), \gamma(i, j - 1)\} \tag{3.3}$$

Given the nature of the system to work in uncontrolled environments, it is continuously reviewing the stage for possible actions or gestures. In this case, the input feature vector Q is of "infinite" length, and may contain segments related to gesture C at any part.

Next, it is described the algorithm for begin-end of gesture recognition and the Feature Weighting proposal within the DTW framework.

3.3.1 Begin-end of gesture detection

In order to detect a begin-end of gesture $C = \{c_1, .., c_m\}$ in a maybe infinite sequence $Q = \{q_1, .., q_\infty\}$, a $M_{m \times \infty}$ matrix is designed, where the position (i, j) of the matrix contains the distance between c_i and q_j, quantifying its value by the Euclidean distance, as commented before. Finally, the warping path is defined by $W = \{w_1, .., w_\infty\}$ as in the standard DTW approach. The aim is focused on finding segments of Q sufficiently similar to the sequence C. The system considers that there is correspondence between the current block k in Q and a gesture if satisfying the following condition,

$$M(m, k) < \mu, k \in [1, .., \infty].$$

for a given cost threshold μ. This threshold value is estimated in advance for each of the categories of actions or gestures using leave-one-out cross-validation strategy. This involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. At each iteration, it is evaluated the similarity value between the candidate and the rest of the training set. Finally it is choosen the threshold value which is associated with the largest number of hits within a category.

Once detected a possible end of pattern of gesture or action, the working path W can be found through backtracking of the minimum path from $M(m,k)$ to $M(0,z)$, being z the instant of time in Q where the gesture begins. The algorithm for begin-end of gesture detection for a particular gesture C in a large sequence Q using DTW is summarized in Table 3.1. Note that $d(i,j)$ is the cost function which measures the difference among the descriptors V_i and V_j. An example of a begin-end gesture recognition for a model and infinite sequence together with the working path estimation is shown in Figure 3.3.

Input: A gesture model $C = \{c_1, .., c_m\}$, its similarity threshold value μ, and the testing sequence $Q = \{q_1, .., q_\infty\}$. Cost matrix $M_{m \times \infty}$ is defined, where $N(w), w = (i,t)$ is the set of three upper-left neighbor locations of w in M.

Output: Working path W of the detected gesture, if any

// Initialization
for $i = 1 : m$ **do**
  for $j = 1 : \infty$ **do**
    $M(i,j) = \infty$
  end
end
for $j = 1 : \infty$ **do**
  $M(0,j) = 0$
end
for $t = 0 : \infty$ **do**
  for $i = 1 : m$ **do**
    $x = (i,t)$
    $M(w) = d(w) + \min_{w' \in N(w)} M(w')$
  end
  if $M(m,t) < \mu$ **then**
    $W = \{\text{argmin}_{w' \in N(w)} M(w')\}$
    return
  end
end

Table 3.1: DTW begin-end of gesture recognition algorithm.

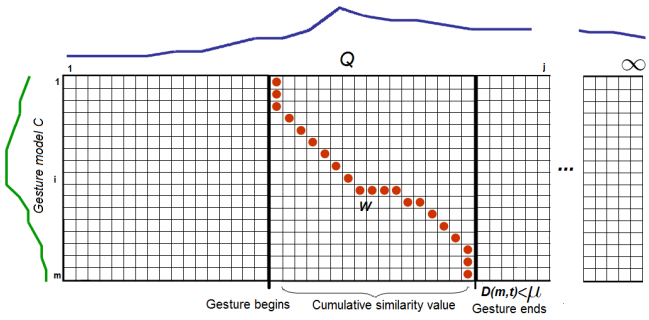


Figure 3.3: Begin-end of gesture recognition of a model C in an infinite sequence Q.

3.3.2 Feature Weighting in DTW

Is proposed a Feature Weighting approach to improve the cost distance computation $d(w)$ of previous begin-end DTW algorithm.

In standard DTW algorithm, cost distances among feature vectors c_i and q_j (3D coordinates of the skeletal models in this case) are computed equally for each feature of the descriptors.

However, it is intuitive that not all skeletal elements of the model participate equally for discriminating the performed gesture. For instance, the movement of the legs when performing hand shaking should not have influence, and thus, computing their deviation to a correspondence model of the gesture adds noise to the cost similarity function. In this sense, the proposal is based on associating a discriminatory weight to each joint of the skeletal model depending on its participation in a particular gesture. In order to automatically compute this weight per each joint, it is proposed an inter-intra gesture similarity algorithm.

First, it is performed a weight training algorithm based on a ground truth data of gestures. Given the data composed by $\{n_1, .., n_N\}$ gesture categories described using skeletal descriptors, the objective is to obtain the inter-intra coefficient of the joints for the data set. This estimation is performed per each joint using a symmetric cost matrix $D_{N\times N}$. Each matrix element $D^p(i,j)$ for the matrix of joint p contains the mean DTW cost between all pairs of samples $C_i, C_j, \forall C_i \in n_i, \forall C_j \in n_j$ only considering the features of the descriptor related to the p-th joint, where n_i and n_j represent the set of samples for gesture categories i and j of the data set.

The mean DTW value at each position of the matrix D^p represents the variability of joint p between a pair of gestures. Note that the diagonal of D represents the intra-gesture variability per joint for all the gesture categories, meanwhile the rest of the elements compare the variability of joint p for two different gesture categories, codifying the inter-gesture variability. Since gestures, as any other object recognition system, will be more discriminative when increasing inter-distance and reducing intra-distance, a discriminative weight is defined as shown in Algorithm 3.2, which assigns high cost to joints high high intra-inter difference values and low cost otherwise. Moreover, the assigned weight is normalized in the same range to be comparable for all joints. Note that at the end of this procedure it is obtained a final global weight vector $\nu = \{\nu^1, .., \nu^z\}$, with a weight value ν^p for the p-th joint, which is

included in the re-definition of the begin-end DTW algorithm cost function $d(w)$ to improve gesture recognition performance as follows,

$$d(c_i, c_j) = \sqrt{\sum_{p=1}^{|c_i|}((c_i^p - c_j^p) \cdot \nu^p)^2}, \tag{3.4}$$

where $|c_i|$ is the length of the feature vector c_i. The Feature Weighting algorithm for computing the weight vector $\nu = \{\nu^1, .., \nu^z\}$ is summarized in 3.2.

Input: Ground-truth data formed by N sets of gestures $\{n_1, .., n_N\}$.
Output: Weight vector $\nu = \{\nu^1, .., \nu^z\}$ associated with skeletal joints so that $\sum_{i=1}^{z} \nu^i = 1$.
$\nu = \emptyset$
for $p = 1 : z$ **do** // Number of joints
  for $i = 1 : N$ **do**
    for $j = i : N$ **do**
      $D^p(i,j) = \mathsf{mean}(DTW(C_v^i, C_w^j)), \forall v, w$ gesture samples of categories i and j.
    end
  end
  $\nu_{\mathsf{intra}} = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} D^p(i,j)}{\frac{N\times(N-1)}{2}}$ // Computer intra-class variability
  $\nu_{\mathsf{inter}} = \frac{\mathsf{Trace}(D^p)}{m}$ // Computer inter-class variability
  $\nu^p = max(0, \frac{\nu_{\mathsf{intra}} - \nu_{\mathsf{inter}}}{\nu_{\mathsf{intra}}})$ // Compute global weight for joint p
  $\nu = \nu \cup \nu^p$
end
Normalize ν so that $\sum_{i=1}^{z} \nu^i = 1$

Table 3.2: Feature Weighting in DTW cost measure.

3.3.3 Results

Before the presentation of the results, first, are discussed the data, methods and parameters, and validation protocol of the experiments.

Data: It is designed a new data set of gestures using the Kinect device consisting of five different categories: jumping, bendding, clapping, greeting, and noting with the hand. It has been considered 10 different actors, 10 different backgrounds, and 100 sequences per subject for recording the data set. Thus, the data set contains the high variability from uncontrolled environments. The resolution of the video depth sequences is 340×280 at 30 FPS. The data set contains a total of 1000 gesture samples considering all the categories. The ground-truth of each sequence is performed manually by examining and noting the position in the video when some actor begin-ends a gesture.

Some samples of the captured gestures for different categories are shown in Figure 3.4.

Methods and parameters: For the implementation of the system is used C/C++, efficiently using dynamic programming to evaluate the recurrence which defines the cumulative distance between vectors of features on each frame. The people detection system used is provided by the public library OpenNI. This library has a high accuracy in people detection, allowing multiple detection even in cases of partial occlusions. The detection is accurate as people remain at a minimum of 60cm from the camera and up to 4m, but can reach up to 6m but with less robust and reliable detection. For automatically initialization of the system are used 20 calibration poses. These calibration models are also obtained through the library OpenNI. This calibration set has been built with high variability in order to automatically obtain the feature vector in different human pose configurations. During the calibration process it is used a structural coherence function θ from the fifth consecutive frame. This assures stabilization to obtain a reliable ζ for a better fit of the skeletal model. In Figure 3.5 are shown different initialization models.

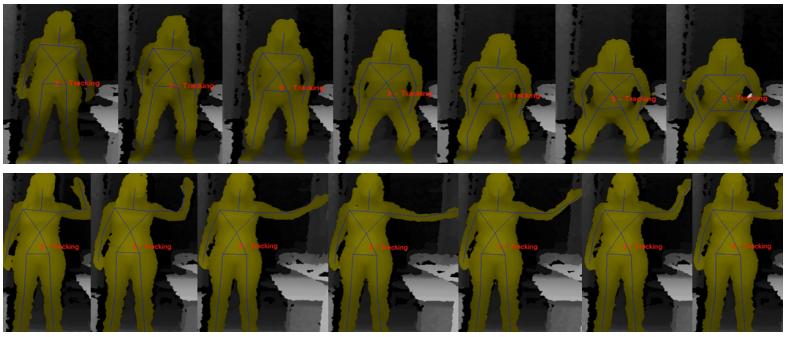


Figure 3.4: Samples of gestures for different categories of the dataset.

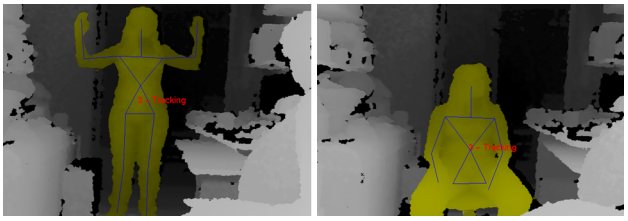


Figure 3.5: Different calibration poses.

Validation protocol: For the validation of the approach and classical DTW algorithm, is computed the Feature Weighting vector ν and gesture cost threshold μ over a leave-one-out validation. The validation sequences may have different length size since they can be aligned using DTW algorithm and trained for the different estimated values of μ. Is validated the begin-end of gesture DTW approach and compare with the Feature Weighting methodology within the same framework. As a validation measurement is computed the confusion matrix for each test sample of the leave-out-out strategy. This methodology allows us to perform an exhaustive analysis of the methods and data set. Adding all test confusion matrices in a performance matrix C_m, final accuracy A is computed using the following formula,

$$A = 100 \cdot \frac{\mathsf{Trace}(C_m)}{NC + \sum_{i=1}^{m} \sum_{j=1}^{m} C_m(i,j)} \tag{3.5}$$

Where NC contains the number of samples of the data set that has not been classified by any gesture since the classification threshold μ has not been satisfied. This evaluation is pessimistic and realistic since both a sample which is not classified or is classified more than once penalizes the final evaluation measurement.

The obtained results applying DTW begin-end gesture recognition and including the Feature Weighting approach on the new data set are shown in Table 3.3. The results show the final performance per gesture over the whole data set using both classification strategies. The best performance per category is marked in bold. Note that for all gesture categories, the begin-end DTW technique with Feature Weighting improves the accuracy of standard DTW. Only in the case of the jump category the performance is maintained. An example of gesture recognition in a sequence of the data set is shown in Figure 3.6.

Classification Results Feature Weighting DTW		
Gesture	Begin-end DTW	Feature Weighting
Jump	**68**	**68**
Bend	63.4	**68**
Clap	42	**55**
Greet	64.2	**73**
Note	68	**76**

Table 3.3: Classification performance A over the gesture data set for the five gesture categories using DTW begin-end approach and including the Feature Weighting methodology.

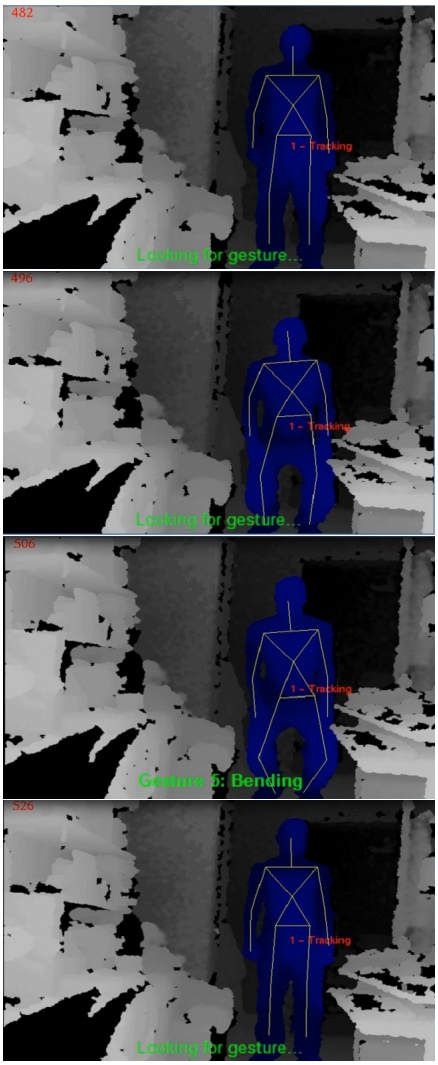


Figure 3.6: Gesture recognition example in a sequence of the dataset.

3.4 Multi-modal Gesture Recognition Challenge 2013: Dataset and Results

Predictive modelling competitions or *Challenges* have been fostering progress in Computer Vision in recent years. One of the most important challenges so far has been undoubtedly the PASCAL Visual Object Classes Challenges[1], organized by Everingham et al. [27], which have contributed to push the state-of-the-art on image classification, detection and segmentation.

These initial efforts concentrating on image understanding are now moving towards the analysis of video data, in which recognizing human activities in visual data has received much attention. Improving automatic recognition of human actions in visual data will allow the development of novel applications useful in surveillance and security, new generation of eHealth applications such as assisted living, the development of natural interfaces for Human Computer Interaction, and improved control in leisure scenarios.

Several contests have been organized in activity recognition with applications in video surveillance, for example the VIRAT Action Recognition Challenge[2], the 3D human reconstruction and action recognition Grand Challenge [3], the Human Activities Recognition and Localization Competition [4] and the Contest on Semantic Description of Human Activities[5].

Following this trend, it has been organized a challenge called "Multi-modal Gesture Recognition Challenge"[6] focusing on recognizing multiple gestures from a novel data set of videos recorded with a Microsoft Kinect™ camera. Kinect™ has revolutionized computer vision in recent years by providing an affordable 3D camera. The applications, initially driven by the game industry [75], have been rapidly diversifying and include video surveillance, computer interfaces, robot vision and control, and education [39].

Our previous one-shot learning challenge [35] was devoted to learning a gesture category from a single example of gesture coming from a limited vocabulary, using RGB and depth data. Our novel data set offers several gesture categories labeled from a dictionary of 20 Italian sign gesture categories. Several features make this new competition extremely challenging, including the recording of continuous sequences, the presence of distracter gestures (not included in the dictionary), the relatively large number of categories, the length of the gesture sequences, and the variety of users. To attack such a difficult problem, several modalities are provided in the data set, including audio, RGB, depth maps, user masks, and user skeletal model. The presentation of the data set and the results obtained in the Mutimodal Gesture Recognition Challenge are explained in the next sections.

[1] http://pascallin.ecs.soton.ac.uk/challenges/VOC
[2] http://www.umiacs.umd.edu/conferences/cvpr2011/ARC
[3] http://mmv.eecs.qmul.ac.uk/mmgc2013/
[4] http://liris.cnrs.fr/harl2012/
[5] http://cvrc.ece.utexas.edu/SDHA2010/
[6] http://gesture.chalearn.org

3.4.1 Problem setting and data

The focus of the challenge is on *multiple instance, user independent learning of gestures from multi-modal data*, which means learning to recognize gestures from several instances for each category performed by different users, drawn from a vocabulary of 20 gesture categories. A gesture vocabulary is a set of unique gestures, generally related to a particular task. In this challenge it has been focused on the recognition of a vocabulary of 20 Italian cultural/anthropological signs, see Figure 3.7 for one example of each Italian gesture.

(1) Vattene (2) Viene qui (3) Perfetto (4) E un furbo (5) Che due palle

(6) Che vuoi (7) Vanno d'accordo (8) Sei pazzo (9) Cos hai combinato (10) Nonme me friega niente

(11) Ok (12) Cosa ti farei (13) Basta (14) Le vuoi prendere (15) Non ce ne piu

(16) Ho fame (17)Tanto tempo fa (18) Buonissimo (19) Si sono messi d'accordo (20) Sono stufo

Figure 3.7: Data set gesture categories.

In all the sequences, a single user is recorded in front of a Kinect^{TM}, performing natural communicative gestures and speaking in fluent Italian. The main characteristics of the dataset of gestures are:

• 13.858 gesture samples recorded with the Kinect^{TM} camera, including audio, skeletal model, user mask, RGB, and depth images.

• RGB video stream, 8-bit VGA resolution (640×480) with a Bayer color filter, and depth sensing video stream in VGA resolution (640×480) with 11-bit. Both are acquired in 20 fps on average.

• Audio data is captured using Kinect^{TM} 20 multi-array microphone.

• A total number of 27 users appear in the data set.

• The data set contains the following number of sequences, development: 393 (7.754 gestures), validation: 287 (3.362 gestures), and test: 276 (2.742 gestures), each sequence lasts between 1 and 2 minutes and contains between 8 and 20 gesture samples, around 1.800 frames. The total number of frames of the data set is 1.720.800.

• All the gesture samples belonging to 20 main gesture categories from an Italian gesture dictionary are annotated at frame level indicating the gesture label.

• 81% of the participants were Italian native speakers, while the remaining 19% of the users were not Italian, but Italian-speakers.

• All the audio that appears in the data is from the Italian dictionary. In addition, sequences may contain distracter words and gestures, which are not annotated since they do not belong to the main dictionary of 20 gestures.

This dataset, available at `http://sunai.uoc.edu/chalearn`, presents various features of interest as listed in Table 3.4.

Table 3.4: Easy and challenging aspects of the data.

Easy
Fixed camera Near frontal view acquisition Within a sequence the same user Gestures performed mostly by arms and hands Camera framing upper body Several available modalities: audio, skeletal model, user mask, depth, and RGB Several instances of each gesture for training Single person present in the visual field
Challenging
Within each sequence: Continuous gestures without a resting pose Many gesture instances are present Distracter gestures out of the vocabulary may be present in terms of both gesture and audio *Between sequences*: High inter and intra-class variabilities of gestures in terms of both gesture and audio Variations in background, clothing, skin color, lighting, temperature, resolution Some parts of the body may be occluded Different Italian dialects

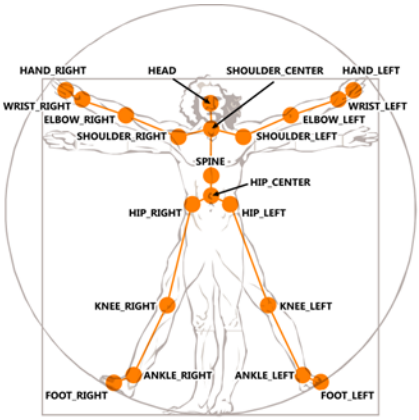


Figure 3.8: Skeleton joint positions.[7]

3.4.2 Data format and structure

Is provided the X_audio.ogg, X_color.mp4, X_depth.mp4, and X_user.mp4 files containing the audio, RGB, depth, and user mask videos for a sequence X, respectively, see Figure 3.9. Is also provided a script in order to export the data in Matlab, which contains the following Matlab structures:

- **NumFrames:** Total number of frames.
- **FrameRate:** Frame rate of the video in fps.
- **Audio:** Structure that contains WAV audio data.
  - **y:** Audio Data.
  - **fs:** Sample rate for the data.
- **Labels:** Structure that contains the data about labels contained in the sequence, sorted in order of appearance. The labels considered to the 20 gesture categories as shown in Figure 3.7.

[7] Image from
http://msdn.microsoft.com/en~us/library/microsoft.kinect.jointtype.aspx

- **Name:** The name given to this gesture.

  1. vattene
  2. vieniqui
  3. perfetto
  4. furbo
  5. cheduepalle
  6. chevuoi
  7. daccordo
  8. seipazzo
  9. combinato
  10. freganiente
  11. ok
  12. cosatifarei
  13. basta
  14. prendere
  15. noncenepiu
  16. fame
  17. tantotempo
  18. buonissimo
  19. messidaccordo
  20. sonostufo

- **RGB:** This matrix represents the RGB color image, expressed in 8-bit VGA resolution (640×480) with a Bayer color filter.

- **Depth:** The Depth matrix contains the pixel-wise z component, VGA resolution (640×480) represented with 11 bits. The value of depth is expressed in millimeters.

- **UserIndex:** The user index matrix represents the player index of each depth pixel. A non-zero pixel value means that a tracked subject occupies the pixel, and a value of 0 denotes that no tracked subject occupies the pixel..

- **Skeleton:** An array of Skeleton structures is contained within a Skeletons array. It contains the joint positions, and bone orientations comprising a skeleton. The format of a Skeleton structure is:.

  - **JointType:** Skeleton joints that make up a tracked skeleton. Figure 3.8 visualizes these joint types.

    1. HipCenter
    2. Spine
    3. ShoulderCenter
    4. Head
    5. ShoulderLeft
    6. ElbowLeft
    7. WristLeft
    8. HandRight
    9. HipLeft
    10. KneeLeft
    11. AnkleLeft
    12. FootLeft
    13. HipRight
    14. KneeRight
    15. AnkleRight
    16. FootRight

  - **JointPosition:** It contains the joint positions in the next three coordinates:

    * **WorldPosition:** The world coordinates position structure represents the global position of a tracked joint. The format is **X, Y, Z** which represents the x, y, and z components of the subject´s global position (in millimeters).

    * **PixelPosition:** The pixel coordinates position structure represents the position of a tracked joint. The format of the Position structure is **X, Y** which represent the x and y components of the joint location over the RGB map (in pixels coordinates).

    * **WorldRotation:** The world rotation structure contains the orientations of skeletal bones in terms of absolute transformations and is formed by a 20×4

matrix, where each row contains the W, X, Y, Z values of the quaternion related to the rotation. The world rotation structure provides the orientation of a bone in the 3D camera space. The orientation of a bone is relative to the child joint and the Hip Center joint still contains the orientation of the player/subject.

Figure 3.9: Different data modalities of the provided data set.

3.4.3 Protocol and evaluation

The timeline of the competition was as follows:

• **April 30th, 2013**: Beginning of the challenge competition, release of first data examples.

• **May 20th, 2013**: Full release of training and validation data. Training data with ground truth labels.

• **August 1st, 2013**: Encrypted Final evaluation data and ground truth labels for the validation data are made available.

• **August 15th, 2013**: End of the challenge competition. Deadline for code submission. The organizers start the code verification by running it on the final evaluation data and obtaining the team scores.

• **August 25th, 2013**: Deadline for fact sheets.

• **September 1st, 2013**: Release of the verification results to the participants for review.

The challenge consisted of two main components: a development phase (April 30th to Aug 1st) and a final evaluation phase (Aug 2nd to Aug 15th). The submission and evaluation of the challenge entries was via the *Kaggle* platform [8]. The official participation rules were provided on the website of the challenge. In addition, publicity and news on the ChaLearn Multi-modal Gesture Recognition Challenge were published in well-known online platforms, such as LinkedIn, Facebook, Google Groups and the ChaLearn website.

During the development phase, the participants were asked to build a system capable of learning from several gesture samples a vocabulary of 20 Italian sign gesture categories. To that end, the teams received the development data to train and self-evaluate their systems.

[8] https://www.kaggle.com/c/multi-modal-gesture-recognition

In order to monitor their progress they could use the validation data for which the labels were not provided. The prediction results on validation data could be submitted online to get immediate feed-back. A real-time leaderboard showed to the participants their current standing based on their validation set predictions.

During the final phase, labels for validation data are published and the participants performed similar tasks as those performed in previous phase, using the validation data and training data sets in order to train their system with more gesture instances. The participants had only few days to train their systems and upload them. The organizers used the final evaluation data in order to generate the predictions and obtain the final score and rank for each team. At the end, the final evaluation data was revealed, and authors submitted their own predictions and fact sheets to the platform.

3.4.4 Evaluation metric

For each unlabeled video, the participants were instructed to provide an ordered list of labels R corresponding to the recognized gestures. It has been compared this list with the truth labels T i.e. the prescribed list of gestures that the user had to play during data collection. It has been computed the Levenshtein distance $L(R,T)$, that is the minimum number of edit operations (substitution, insertion, or deletion) that one has to perform to go from R to T (or vice versa). The Levenhstein distance is also known as "edit distance". For example: $L([124],[32]) = 2$, $L([1],[2]) = 1$, $L([222],[2]) = 2$. The overall score is the sum of the Levenshtein distances for all the lines of the result file compared to the corresponding lines in the truth value file, divided by the total number of gestures in the truth value file. This score is analogous to an error rate. For simplicity, in what follows, called it Error Rate, although it can exceed 1.0. A public score appeared on the leaderboard during the development period and was based on the validation data. Subsequently, a private score for each team was computed on the final evaluation data released at the end of the development period, which was not revealed until the challenge was over. The private score was used to rank the participants and determine the prizes.

3.4.5 Results

The challenge attracted high level of participation, with a total of 54 teams and near 300 total number of entries. This is a good level of participation for a computer vision challenge requiring very specialized skills. Finally, 17 teams successfully submitted their prediction in final test set, while providing also their code for verification and summarizing their method by means of a fact sheet questionnaire.

After verifying the codes and results of the participants, the final scores of the top rank participants on both validation and test sets were made public: these results are shown in Table 3.5, where winner results on the final test set are printed in bold. In the end, the final error rate on the test data set was around 12%.

Table 3.5: Top rank results on validation and test sets.

TEAM	Validation score	Test score
IVA MM	0.20137	**0.12756**
WWEIGHT	0.46163	**0.15387**
ET	0.33611	**0.16813**
MmM	0.25996	0.17215
PPTK	0.15199	0.17325
LRS	0.18114	0.17727
MMDL	0.43992	0.24452
TELEPOINTS	0.48543	0.25841
CSI MM	0.32124	0.28911
SUMO	0.49137	0.31652
GURU	0.51844	0.37281
AURINKO	0.31529	0.63304
STEVENWUDI	1.43427	0.74415
JACKSPARROW	0.86050	0.79313
JOEWAN	0.13653	0.83772
MILAN KOVAC	0.87835	0.87463
IAMKHADER	0.93397	0.92069

3.4.6 Statistics on the results

Figure 3.11 shows the correlation of the validation and test error scores obtained by the top ranked participants of the challenge. One can see that most of them obtain similar results in both sets. However, there exist a few outliers that show non-correlated results among validation and test scores. Most of the participants that achieved top positions in test scores also achieved high recognition rates on the validation set. However, some participants that achieved low error on validation, considerably increased their error on the test set. It could be mainly related to overfitting of method parameters on the validation data, which could not be able to generalize to the variability of new users present on the test set. On the other hand, the final scores on the test set are in general lower than in the validation set. This may be produced because more data is trained by the participants when testing for final evaluation set, and thus, if propertly trained, this small final improvement is expected. The best public score on the validation set achieved during the period of the challenge taking into account all team submissions over time is summarized in Figure 3.10.

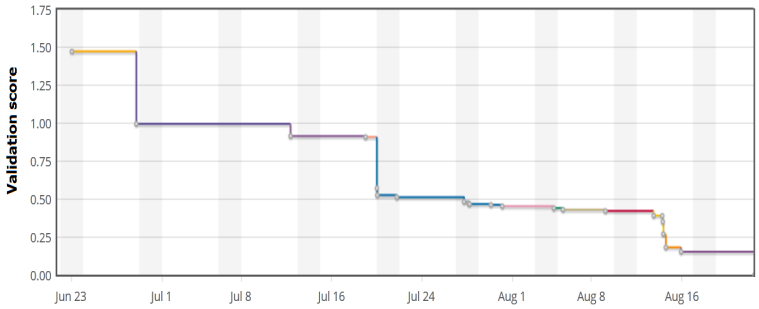


Figure 3.10: Best public score obtained in the validation set during the Challenge.

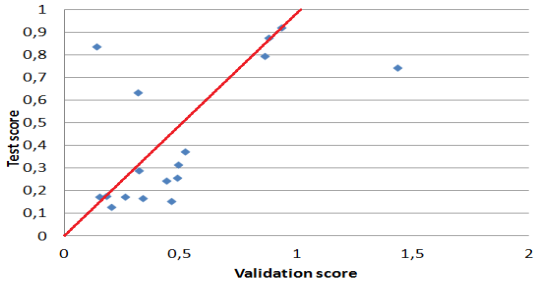


Figure 3.11: Correlation results among the validation and test results of the top ranked participants.

In Figure 3.12, is showed the histograms of validation and test scores based on the best score achieved by each team on each of the two sets. One can see that the scores on the validation set become more sparse, and the teams that finally submitted their predictions to the test set, except for two cases, achieved scores inferior to 1 Levenstein score error.

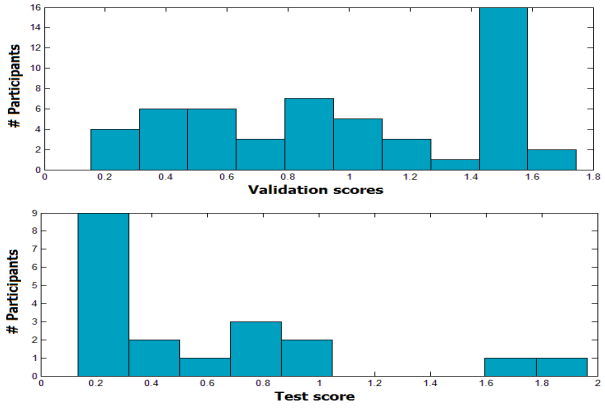


Figure 3.12: Validation and test scores histograms.

3.4.7 Fact sheets

It has been asked the participants to fill out a survey about the methods employed. The top ranked 17 test participants filled out this survey. It has been briefly summarize the results next.

From the questions within the survey, the most relevant aspects for the challenge where: the modalities considered for the methods, the temporal segmentation methodology applied, the considered fusion strategy, as well as the recognition techniques considered. The information about modalities, segmentation strategy, fusion, classifier, and programming language are summarized in Figures 3.13, 3.14, 3.15, 3.16, and 3.17, respectively. The details of each particular team strategy are shown in Table 3.6.

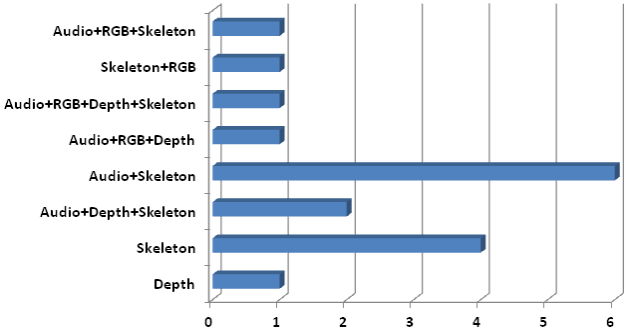


Figure 3.13: Modalities considered.

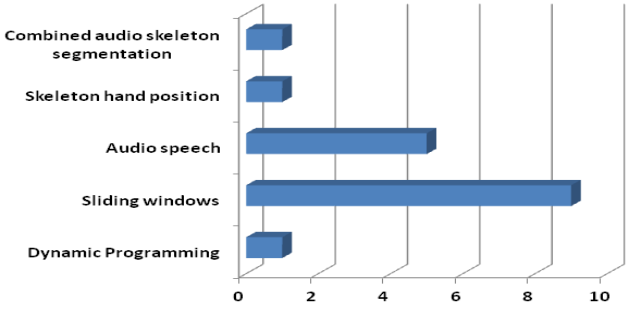


Figure 3.14: Segmentation strategy.

Looking at the considered modalities in Figure 3.13, one can see that none of the teams considered only audio information, and most of them used multiple modalities for describing the data. In particular, skeleton was the most considered feature when no multiple modalities were used. In general, combining audio plus skeleton information was the predominant strategy among the participants, and the one considered by the first three ranked teams on the test set.

The considered temporal segmentation strategies are shown in Figure 3.14. One can see that two strategies were mainly applied: based on features and based on temporal windows or classifiers (such as Dynamic Time Warping). For the first case, audio information and joint positions were the most considered information to split the continuous sequence into gesture candidates. In the second case, Sliding-Windows technique was the preferred choice of the participants.

Regarding the fusion strategy (Figure 3.15), several authors did not apply any, since only one cue was used in their approach or different cues were independently used in different stages, such as one for temporal segmentation and the other one for describing segmented candidate gestures and final classification. Regarding the participants that fused different modalities, few of them combined feature vectors in an early fusion fashion before training a classifier. The preferred strategy was to train classifiers on different feature sets from different modalities and fuse the weighted outputs of classifiers in a late fusion fashion.

Regarding the classifiers (Figure 3.16), it is interesting to see the broad variety of strategies, covering most of the state-of-the-art Machine Learning strategies. Both discriminative and generative classifiers were considered, in all cases applying supervised learning. The preferred strategy was Hidden Markov Models. On the other hand dynamic programming, and in particular Dynamic Time Warping, which of often one of the preferred methods for gesture recognition, was not widely applied in our challenge (fourth choice). In particular, Random Forest and Neural Network variants were the second and third choice of the participants, respectively.

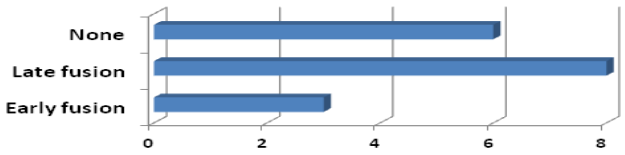


Figure 3.15: Fusion strategy.

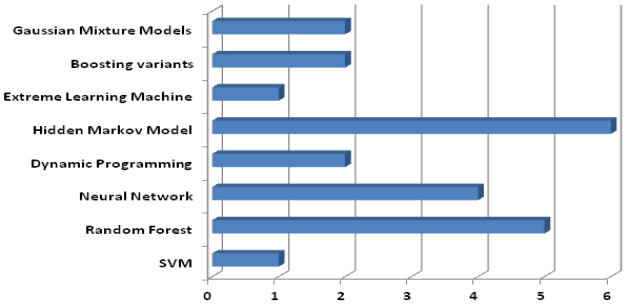


Figure 3.16: Learning strategy.

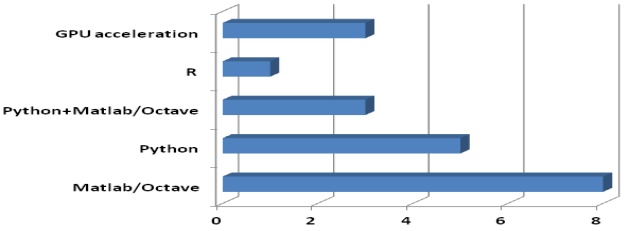


Figure 3.17: Programming language.

Finally, Figure 3.17 summarizes the programming languages: mainly Matlab/Octave and Python languages were used, and in few cases the codes were speeded up by means of GPU programming.

Table 3.6 shows the particular strategy from these statistics applied for each team of the top ranked positions of the challenge. Interestingly, the three top ranked participants agree in the modalities and segmentation strategy considered, although they differ in the final applied classifier.

3.4.8 Summary of the winner methods

Table 3.6 shows the summary of the strategies considered by each of the top ranked participants on the test set. Next, it has been briefly described in more detail the approach designed by the three winners of the challenge.

The first ranked team *IVAMM* on the test set used a feature vector based on audio and skeletal information, and applied late fusion to obtain final recognition results. A simple time-domain end-point detection algorithm based on joint coordinates is applied to segment continuous data sequences into candidate gesture intervals. A Gaussian Hidden Markov Model is trained with 39-dimension MFCC features and generates confidence scores for each gesture category. A Dynamic Time Warping based skeletal feature classifier is applied to provide complementary information. The confidence scores generated by the two classifiers are firstly normalized and then combined to produce a weighted sum. A single threshold approach is employed to classify meaningful gesture intervals from meaningless intervals caused by false detection of speech intervals.

Table 3.6: Team methods and results. Early and late refer to early and late fusion of features/classifier outputs. HMM: Hidden Markov Models. KNN: Nearest Neighbor. RF: Random Forest. Tree: Decision Trees. ADA: Adaboost variants. SVM: Support Vector Machines. Fisher: Fisher Linear Discriminant Analysis. GMM: Gaussian Mixture Models. NN: Neural Networks. DGM: Deep Boltzmann Machines. LR: Logistic Regression. DP: Dynamic Programming. ELM: Extreme Learning Machines.

TEAM	Test score	Rank	Modalities	Segmentation	Fusion	Classifier
IVA MM	0.12756	1	Audio,Skeleton	Audio	None	HMM,DP,KNN
WWEIGHT	0.15387	2	Audio,Skeleton	Audio	Late	RF,KNN
ET	0.16813	3	Audio,Skeleton	Audio	Late	Tree,RF,ADA
MmM	0.17215	4	Audio,RGB+Depth	Audio	Late	SVM,Fisher, GMM,KNN
PPTK	0.17325	5	Skeleton,RGB,Depth	Sliding w.	Late	GMM,HMM
LRS	0.17727	6	Audio,Skeleton,Depth	Sliding w.	Early	NN
MMDL	0.24452	7	Audio,Skeleton	Sliding w.	Late	DGM+LR
TELEPOINTS	0.25841	8	Audio,Skeleton,RGB	Audio,Skeleton	Late	HMM,SVM
CSI MM	0.28911	9	Audio,Skeleton	Audio	Early	HMM
SUMO	0.31652	10	Skeleton	Sliding w.	None	RF
GURU	0.37281	11	Audio,Skeleton,Depth	DP	Late	DP,RF,HMM
AURINKO	0.63304	12	Skeleton,RGB	Skeleton	Late	ELM
STEVENWUDI	0.74415	13	Audio,Skeleton	Sliding w.	Early	DNN,HMM
JACKSPARROW	0.79313	14	Skeleton	Sliding w.	None	NN
JOEWAN	0.83772	15	Skeleton	Sliding w.	None	KNN
MILAN KOVAC	0.87463	16	Skeleton	Sliding w.	None	NN
IAMKHADER	0.92069	17	Depth	Sliding w.	None	RF

The second ranked team *WWEIGHT* combined audio and skeletal information, using both joint spatial distribution and joint orientation. The method first searches for regions of time with high audio-energy to define 1.8-second-long windows of time that potentially contained a gesture. This had the effect that the development, validation, and test data were treated uniformly. Feature vectors are then defined using a log-spaced audio spectrogram and the joint positions and orientations above the hips. At each time sample the method subtracts the average 3D position of the left and right shoulders from each 3D joint position. Data is down-sampled onto a 5 Hz grid considering 1.8 seconds. There were 1593 features total (9 time samples $\times$ 177 features per time sample). Since some of the detected windows can contain distracter gestures, an extra 21st label is introduced, defining the 'not in the dictionary' gesture category. Python's scikit-learn was used to train two models: an ensemble of randomized decision trees (ExtraTreesClassifier, 100 trees, 40% of features) and a K-Nearest Neighbor model (7 neighbors, L1 distance). The posteriors from these models are averaged with equal weight. Finally, a heuristic is used (12 gestures maximum, no repeats) to convert posteriors to a prediction for the sequence of gestures.

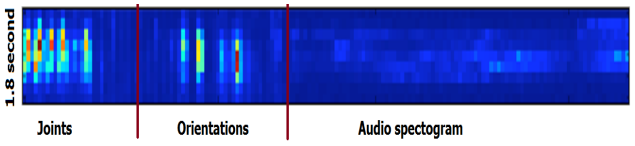


Figure 3.18: ExtraTreesClassifier Feature Importance.

Figure 3.18 shows the mean feature importance for the windows size of 1.8 seconds for the three sets of features: joint coordinates, joint orientations, and audio spectogram. One can note that features from the three sets are selected as discriminative by the classifier, although skeletal features becomes more useful for the ExtraTreesClassifier. Additionally, the most discriminative features are those in the middle of the windows size, since begin-end features are shared among different gestures (transitions) and thus are less discriminative for the classifier.

Figure 3.19: Recognition of test sequence by the three challenge winners. Black bin means that the complete list of ordered gestures has been successfully recognized.

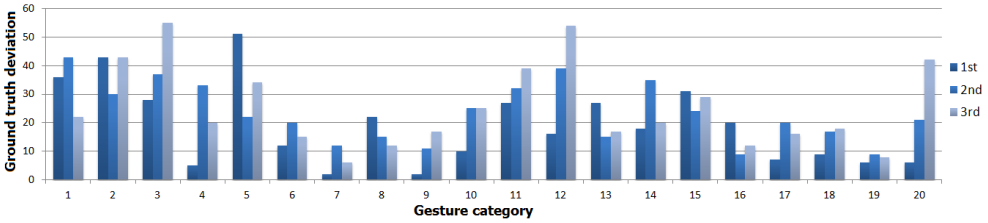


Figure 3.20: Deviation of the number of gesture samples for each category by the three winners in relation to the GT data.

The third ranked team *ET* combined the output decisions of two designed approaches. In the first approach, they look for gesture intervals (unsupervised) using the audio files and extracts features from this intervals (MFCC). Using these features, authors train a random forest and gradient boosting classifier. The second approch uses simple statistics (median, var, min, max) on the first 40 frames for each gesture to build the training samples. The prediction phase uses a sliding window. The authors create a weighted average of the output of these two models [62]. The features considered were skeleton information and audio signal.

Finally, it has been extracted some statistics from the results of the three challenge winners in order to analyze common points and difficult aspects of the challenge. Figure 3.19 shows the recognition of the 276 test sequences by the winners. Black bin means that the complete list of ordered gestures was successfully recognized for those sequences. Once can see that there exists some kind of correlation among methods. Taking into account that consecutive sequences belong to the same user performing gestures, it means that some some gestures are easier to recognize than others. Since different users appears in the training and test sequences, it is sometimes difficult for the models to generalize to the style of new users, based on the gesture instances used for training.

It has been also investigated the difficulty of the problem by gesture category, within each of the 20 Italian gesture categories. Figure 3.20 shows for each winner method the deviation between the number of gesture instances recognized and the total number of gestures, for each category. This was computed for each sequence independently, and adding the deviation for all the sequences. In that case, a zero value means that the participant method

recognized the same number of gesture instances for a category that was recorded in the ground truth data. Although is not possible to guarantee with this measure that the order of recognized gesture matches with the ground truth, it gives us an idea of how difficult the gesture sequences were to segment into individual gestures. Additionally, the sum of total deviation for all the gestures for all the teams was 378, 469, and 504, which correlates with the final rank of the winners. The figure suggests a correlation between the performance of the three winners. In particular, categories 6, 7, 8, 9, 16, 17, 18, and 19 were the ones that achieved most accuracy for all the participants, meanwhile 1, 2, 3, 5, and 12 were the ones that introduced the highest recognition error. Note that the public data set provides accurate label annotations of end-begin of gestures, and thus, a more detailed recognition analysis could be performed applying a different recognition measurement to Leveinstein, such as Jaccard overlapping or sensitivity score estimation, which will also allow for confusion matrix estimation based on both inter and intra user and gesture category variability.

3.5 Conclusion

In this chapter, it was proposed a fully-automatic general framework for real time action/gesture recognition in uncontrolled environments using depth data. The system analyzes data sequences based on the assignment of weights to gesture descriptors so that DTW cost measure improves discrimination. The feature vectors are extracted automatically through a calibration set, obtaining 3D coordinates of skeletal models with respect an origin of coordinates, making description invariant to translation, scale, and tolerant to corporal differences among subjects. The final gesture is recognized by means of a novel Feature Weighting approach, which enhance recognition performance based on the analysis on inter-intra class variability of vector features among gesture descriptors. The evaluation of the method has been performed on a novel depth data set of gestures, automatically detecting begin-end of gesture and obtaining performance improvements compared to classical DTW algorithm.

The second part of this chapter describes the "ChaLearn Multi-modal Gesture Recognition Challenge". For this purpose, it has been designed a large data set, including several people that perform gestures from a vocabulary 20 Italian sign gesture categories. Data also include distracter gestures to make the recognition task challenging. The modalities provided included audio, RGB, depth maps, user masks, and skeletal model. The datasets has been manually annotated to provide ground truth of temporal segmentation of the signal into individual gestures. The dataset has been made publicly available.

Different classifiers for gesture recognition were used by the participants. The preferred one was Hidden Markov Models (used by the first ranked team of the challenge), followed by Random Forest (used by the second and third winners). Although several state of the art learning and testing gesture techniques were applied at the last stage of the methods of the participants, still the feature vector descriptions are mainly based on MFCC audio features and skeleton joint information. This supports the use of complementary source of information, but it is expected that the use of more sophisticated features in a near feature will be useful to reduce the current error rate achieved in the data set. For instance, is concluded that structural hand information around hand joint could be useful to discriminate among gesture categories that may share similar trajectories of hand/arms.

Although the current error rate on the data set is about 12% using de Levenshtein edit distance among order of predicted gestures, it still offers range for improvement and test for other metrics given the provided ground truth, such are Jaccard overlapping index or sensitivity.

Chapter 4

Spherical Blurred Shape Model for 3D Object and Pose Recognition

Abstract

The use of depth maps is of increasing interest after the advent of cheap multisensor devices based on structured light, such as Kinect^{TM}. In this context, there is a strong need of powerful 3D shape descriptors able to generate rich object representations. Although several 3D descriptors have been already proposed in the literature, the research of discriminative and computationally efficient descriptors is still an open issue. We propose a novel point cloud descriptor, called *Spherical Blurred Shape Model* (SBSM), which successfully encodes the structure density and local variabilities of an object based on shape voxel distances and a neighborhood propagation strategy. The proposed SBSM is proven to be rotation and scale invariant, robust to noise and occlusions, highly discriminative for multiple categories of complex objects like the human hand, and computationally efficient since the SBSM complexity is linear to the number of object voxels. Experimental evaluation in public depth multi-class object and a novel hand poses data sets show significant performance improvements in relation to state-of-the-art approaches. Moreover, the effectiveness of the proposal is also proved for object spotting in 3D scenes and for real-time automatic hand pose recognition in Human Computer Interaction scenarios.

4.1 Introduction

Computer vision research on 3D point cloud analysis has recently received a lot of attention because of the availability of cheap multisensor devices based on structured light, such as Kinect^{TM}. This RGB-Depth camera is compact and portable, so it can be easily installed in any environment to understand 3D scenes. This way there are multiple applications which can benefit from the analysis of 3D objects in scenes [34]. However, recognition of 3D objects is still a challenging problem: in addition to the typical issues tackled by 2D object recognition approaches (such as robustness to noise and occlusions, discriminate power and computational complexity), the captured sequences are usually sampled at discrete points,

so the finer details of the 3D object are usually lost.

Under these assumptions, there exists a strong interest for designing new 3D object descriptors [7, 40, 68, 78]. Is revisited the literature by dividing the existing approaches into those descriptors based on pure 3D geometric properties and those extended from already existing 2D object descriptors.

Describing 3D geometric information has been proven to be useful when classifying everyday objects like cans, glasses or doors, and for 3D scene analysis. For example, some approaches take into account the set of normals of the surface defined by a given point and its neighbors [6, 54]. As an example, the SHOT descriptor proposed in [79] defines a surface representation based on point normals. It is based on counting the points that fall into bins according to a function of the angle between the normal at each point within the corresponding part of the grid and the normal at the feature point. However, in this case the descriptor is local and uses to require from a previous keypoint detection step, which difficult its adaptation to recognize non-rigid shapes. The use of normals are useful to recognize 3D objects since they encode the implicit surface that neighboring points define, although they depend on the density of the underlying points and the smoothness of 3D object surfaces to give accurate results. Also, spherical harmonics [16] have been used to design 3D descriptors invariant to rotation [46] or have been considered directly as features [73]. Conformal factors have also been considered [10], measuring the relative curvature of a vertex given the total curvature. The result can be viewed as a vector which is not only invariant to rigid body transformations, but also to changes in the pose. The Point Feature Histogram (PFH) local descriptor proposed in [71] is used to recognize points conforming planes, cylinders and other geometric primitives. As an extension, the Fast PFH (FPFH) descriptor [69] is based on codifying angle relations among 3D points. FPFH optimizes the PFH computation to make it usable in real time 3D registration applications. The Viewpoint Feature Histogram (VFH) [70] combines an extended version of FPFH with statistics between the viewpoint and the surface normals on the 3D object. Recently, authors in [85] have presented an Ensemble of Shape Functions (ESF) approach to describe 3D objects, which benefits from several combinations of histograms for codifying 3D object relations of angles, areas and distances among points.

Unfortunately, point clouds captured from Kinect^{TM}-like devices usually contain "holes", since data are sampled at discrete points. Consequently, in all the aforementioned approaches (which rely on an accurate computation of 3D geometric primitives) their performance is usually downgraded. Alternatively, some recent 3D regional descriptors have been defined as an extension of classical derivative-based 2D features, such as HOG, SIFT, and SURF [47, 66]. For example, as a generalization of the 2D shape context descriptor presented in [9], the authors of [32] propose a 3D shape context descriptor which is compared with a classical spin-image representation and a novel Harmonic Shape Context (HSC) descriptor for 3D car model classification. Despite the excellent results reported, these methods require the computation of a large number of shape points relations.

In this chapter, it has been proposed a novel 3D object descriptor, called Spherical Blurred Shape Model (SBSM). SBSM is inspired in the Blurred Shape Model (BSM) descriptor presented in [25]. The novel SBSM descriptor codifies the object structure density and local variabilities in the 3D space. Similar to the Zoning descriptor and 3D shape histograms of [8], SBSM bases on a linear computation of spatial relation of shape points to 3D bin centroid, but including a propagation *blurring* degree to define a compact and discrimina-

tive 3D object descriptor. In this sense, Zoning and [8] can be seen as an instance of the proposed descriptor when the defined blurring degree is null. As it is reported in the results, when increasing the blurring factor the overall classification rate of the system is improved. In addition to provide a 3D generalization of BSM, SBSM introduces the following enhancements:

(i) A 3D spherical grid which partitions the 3D space into 3D shape bins,

(ii) A 3D Gaussian-based weight propagation schema controlling the blurring level based on shape voxel distances, and

(iii) A quaternion-based rotation strategy based on sphere axis densities to define a 3D rotation invariant descriptor.

As a result, the proposed SBSM is a global descriptor that encodes the shape of an object, being rotation and scale invariant, computationally efficient, and highly discriminative.

it has been evaluated the descriptor on public and novel 3D object and hand poses data set, showing significant performance improvements in comparison to state-of-the-art approaches. It has been tested the descriptor in front of deformation in depth coordinates and noise point removal. As a result it is shown that SBSM copes with the noise and occlusions typically present in the point clouds acquired by range scanner sensors. Additionally, are shown four real applications where is applied the descriptor. In the first, we perform object spotting in public 3D scenes. The last three applications correspond to Human Computer Interaction scenarios (HCI). In the second, is presented a real-time fully-automatic HCI system for medical image volume navigation, segmenting human hands and classifying multiple hand poses using the proposed SBSM descriptor. In the third, the same approach is applied in a multi-camera setup to perform intelligent retail. Finally, we present a prototype for intelligent navigation through a repository of books in a living lab which may represent the library of the future. Although pointing recognition and gaze tracking in multi-camera setups for HCI has been previously addressed in literature [20, 89], here is shown how complex multi-camera setups can be avoided and recognition for interaction can be improved within different HCI application contexts using the proposed approach.

4.2 Method

In this section, is presented the novel Spherical Blurred Shape Model to describe 3D objects.

4.2.1 Spherical Blurred Shape Model

The Spherical Blurred Shape Model is inspired in 3D grid approaches and in the discriminative power of SIFT and HOG descriptors to codify object information based on the distribution of object gradients and orientations. However, instead of performing computation of 3D object derivatives, SBSM just requires the computation of object shape voxel distances between neighbors in order to codify the object structure density and local variabilities in the 3D space. As a result, SBSM is a computationally efficient descriptor, with a complexity

linear to the number of object voxels $O(|P|)$, with an upper bound of $27 \cdot |P|$ simple operations for a point cloud P of $|P|$ shape points (defined based on a 26-connectivity of regions in the 3D space of bins).

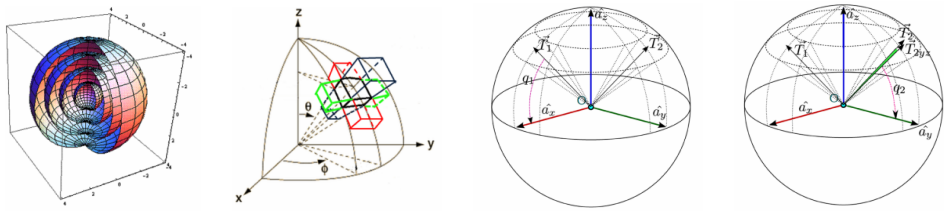


Figure 4.1: Illustration of SBSM descriptor computation. (a) Sphere bins. (b) Example of neighbor bins. (c) and (d) example of the estimation of two main quaternion to rotate feature vector in the 3D space.

As in the case of 2D and 3D object descriptors, an initial grid is fitted to contain the region of interest to describe. In this case, to describe 3D regions, an spherical grid containing a set of 3D bins is defined, which contain the set of voxels P of the point cloud to be described. The description methodology computes for each voxel P contained in the grid a set of voxel-bin spatial relations, which are include in a global region descriptor. Next, is described in detail each step of the description procedure.

In the first step, a discrete spherical grid partitions the 3D space in a set of bins, as shown in Fig. 4.1(a). Let $P = \{p_i | p_i \in \mathbb{R}^3\}$, C, N_L, N_θ, N_ϕ, R, and σ define the set of voxels of the point cloud, point cloud centroid, number of layers, number of angular divisions for θ, number of angular divisions for ϕ, radius length, and sigma value for the gaussian distance metric, respectively. Some of these parameters are illustrated in Fig. 4.1(b). Then, $d_R = R/N_L$, $d_\theta = 2\pi/N_\theta$, and $d_\phi = 2\pi/N_\phi$ are computed as the distance between consecutive layers and the degrees in θ and ϕ polar coordinates between consecutive sectors, respectively. Using this division, next is defined the set B of sphere bins $b_{\{i,j,k\}}$ as follows:

$$B = \{b_{\{0,0,0\}}, ..., b_{\{i,j,k\}}, ..., b_{\{N_L-1,N_\theta-1,N_\phi-1\}}\},$$
$$\forall i \in \{0, 1, .., N_L - 1\}, \forall j \in \{0, 1, .., N_\theta - 1\},$$
$$\forall k \in \{0, 1, .., N_\phi - 1\}, \tag{4.1}$$

where bin $b_{\{i,j,k\}}$ is the 3D bin defined as the cartesian product of intervals $[i \cdot d_R, (i+1) \cdot d_R)$, $[j \cdot d_\theta, (j+1) \cdot d_\theta)$, and $[k \cdot d_\phi, (k+1) \cdot d_\phi)$ for de distance between consecutive layers in relation to the center of the spherical grid, θ, and ϕ, respectively. This way B defines a partition in $\mathbb{R}^3$ of the object of interest. Then, the centroid coordinates for all each bin $b^*_{\{i,j,k\}} \in B^*$ are computed as follows:

$$b^*_{\{i,j,k\}} = (i \cdot d_R + \frac{d_R}{2}, j \cdot d_\theta + \frac{d_\theta}{2}, k \cdot d_\phi + \frac{d_\phi}{2}),$$
$$\forall i \in \{0, 1, .., N_L - 1\}, \forall j \in \{0, 1, .., N_\theta - 1\},$$
$$\forall k \in \{0, 1, .., N_\phi - 1\}. \tag{4.2}$$

An example of some 3D bin neighbors of the spherical descriptor is shown in Fig. 4.1(b). Once the 3D spatial bins are defined, the SBSM feature vector is initialized as:

$$W_i = 0, \forall i \in \{1, 2, .., N_L \cdot N_\theta \cdot N_\phi\}. \tag{4.3}$$

Subsequently, for each voxel in the point cloud $p_z \in P | P \subset B$, the distance of that voxel to its neighbor bins is estimated based on a Gaussian distance metric, and the normalized weights are added to the corresponding descriptor bin locations. For this task, let $b_z = b_{\{i,j,k\}} | p_z \subset b_{\{i,j,k\}}$ be the bin containing voxel p_z. First, the lists containing bin weights and index bins for p_z are initialized to $W^* = \{1\}$ and $I^* = \{\{i, j, k\}\}$, respectively. Then, the iterative procedure updates W^* and I^* for each $b_{\{i,j,k\}} \in N(b_z)$, where $N(b_z)$ is the set of neighbors bins of b_z in a 27-neighborhood for inner sphere bins and 18-neighborhood for external sphere surface bins (including the reference bin). So the list of weights is updated as:

$$W^* = W^* \cup \left\{ e^{-\frac{||p_z - b^*_{\{i,j,k\}}||}{R \cdot \sigma}} \right\}, \tag{4.4}$$

and the list of indexes as:

$$I^* = I^* \cup \{\{i, j, k\}\}. \tag{4.5}$$

As a result, the normalized weights for p_z are added to its corresponding positions of W as follows:

$$W_{I^*_i} = W_{I^*_i} + \frac{W^*_i}{\sum_{j=1}^{|W^*|} W^*_j}, \forall i \in \{1, 2, .., |I^*|\}. \tag{4.6}$$

In this way, a new shape point of the point cloud will include a weight to its belonging bin centroid and neighbor centroid based on a Gaussian function of the distance and a blurring level defined by σ. This values defines the degree of influence of each neighbor bin for each point cloud voxel. Note that when σ parameter is set to zero, the descriptor is equivalent to a dense sampling of the point cloud as in the classical state-of-the-art Zoning descriptor, but defined in the 3D space [25]. It is important to remark that the voxels that are not contained within the spherical grid bins are not considered in the descriptor computation. On the other hand, the voxels that intersect with the spherical surface are considered as inner voxels, and thus, considered in the descriptor estimation. Figure 4.2 shows an example of an hypothetical sphere slice for $\theta = k$ and the analysis of a point cloud voxel to update the SBSM descriptor.

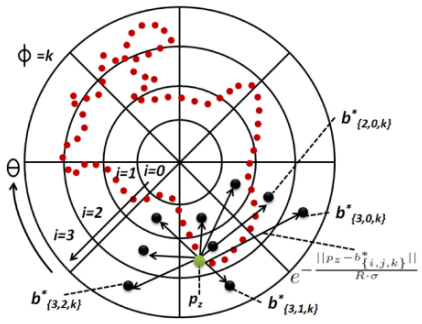


Figure 4.2: Example of point cloud voxel for an hypothetical sphere slice for $\theta = k$. Voxels of the point cloud visible on that slice are shown as red dots. An example of a voxel estimation p_z is shown in green. For this point, neighbor bins centroids are shown as black dots. For each of these relations (note that in the 3D space a total of 27 relations will be computed), equation 4.4 is computed, and the estimated value is added to descriptor position corresponding to its corresponding bin.

Once the procedure is repeated for all points $p_z \in P$, the final feature vector W is normalized as follows:

$$W_i = \frac{W_i}{\sum_{j=1}^{N_L \cdot N_\theta \cdot N_\phi} W_j}, \forall i \in \{1, 2, .., N_L \cdot N_\theta \cdot N_\phi\}. \tag{4.7}$$

Given that all the voxels within the point cloud where the descriptor is computed contribute with the same cost and that the final vector is normalized, it becomes scale invariant. Thus, if different instances of a 3D object category are fitted with the spherical descriptor, even with different sizes, all the descriptors are comparable and can be trained with the same classifier.

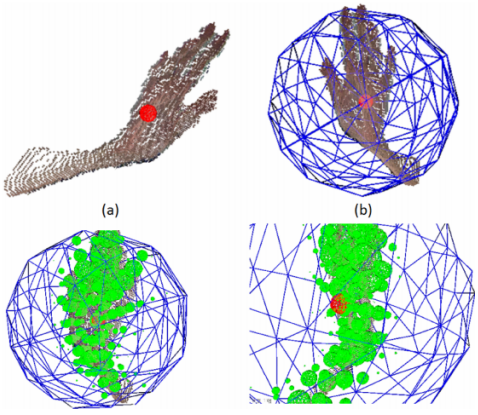


Figure 4.3: (a) Initial hand point cloud and computed center. (b) Sphere including a point cloud corresponding to a 3D hand pose. (c) The same sphere where SBSM descriptor has been computed. The density of the green dots represents the centroid bin values, and the whole descriptor has been rotated based on the quaternion codified by two main descriptor axis densities. (d) Alternative view of the computed SBSM descriptor.

4.2.2 3D rotation invariant SBSM

Once SBSM is computed based on the predefined number of layers, bin orientations, and σ value for the Gaussian function, the descriptor is able to encode the local density and spatial relations of 3D shape points for a particular granularity degree. Rotation invariance is achieved by considering the main spherical axis densities to compute the main vector orientations in quaternion coordinates as a reference axis that rotates feature vector bins. As a result, this feature vector reordering step makes the descriptor rotation invariant for similar 3D objects.

The use of unit quaternion instead of rotation matrices provides a fast computation for rotation invariance and, at the same time, it is simpler to enforce that quaternions have unit magnitude than constrain rotation matrices to be orthogonal [74]. The procedure is detailed next.

First, it is computed the density of the descriptor for each axis defined by the angles θ and ϕ as follows:

$$f(\theta,\phi)=\sum_{r=1}^{N_L} W_{\{r,\theta,\phi\}}, \tag{4.8}$$

and the two maximum axis densities are found:

$$\overrightarrow{T}_1=\arg\max_{\theta,\phi} f(\theta,\phi),\ \overrightarrow{T}_2=\arg\max_{\theta,\phi\setminus\overrightarrow{T}_1} f(\theta,\phi). \tag{4.9}$$

Back to Cartesian coordinates, it is computed the component of $\overrightarrow{T}_2$ vertical to $\overrightarrow{T}_1$ by projecting $\overrightarrow{T}_2$ onto the plane perpendicular to $\overrightarrow{T}_1$ as follows:

$$\overrightarrow{T}_{2yz}=\overrightarrow{T}_2-\overrightarrow{T}_1^T\overrightarrow{T}_2\overrightarrow{T}_1 \tag{4.10}$$

Subsequently, it is computed the rotation that aligns the axis $\overrightarrow{T}_1$, $\overrightarrow{T}_2$ with $\hat{a}_x$ and $\hat{a}_y$, respectively. Where $\hat{a}_x=[100]^T$ and $\hat{a}_y=[010]^T$. The rotation quaternion q can be computed as the combination of two quaternions $q=q_2q_1$, where q_1 rotates $\overrightarrow{T}_1$ to $\hat{a}_x$ and q_2 aligns $\overrightarrow{T}_2$ with $\hat{a}_y$ (see Fig. 4.1(d)).

Finally, values of the bin locations are rotated based on the quaternion q, such that each bin $b^{*r}_{i,j,k}\in B^*$ is computed as:

$$b^{*r}_{i,j,k}=qb^*_{i,j,k}q^*, \tag{4.11}$$

using the Hamilton product, where q^* is the conjugate of the quaternion q. Abusing of notation, b^* and b^{*r} also denote the corresponding pure quaternion to each bin. So it has been taken advantage of this rotation order to obtain the rotation invariant feature vector $W^r_{\{i,j,k\}}=W_{\{i,j,k\}}$.

An example of the two main quaternions for an hypothetical spherical toy problem is shown in Fig. 4.1(c) and (d), respectively. In Fig. 4.3(a) a real example of a 3D hand pose is shown. Figure 4.3(b) shows the centered point cloud within the 3D correlogram containing SBSM bins. The result after computing the SBSM descriptor from hand point cloud and performing rotation invariance is shown in Fig. 4.3(c) and (d) for two different points of view.

4.3 Quantitative analysis of SBSM

In order to present the results, first it is described the training data and settings of the experiments.

4.3.1 Data sets

It has been tested the methodology on two data sets: a public 3D object category data set and a new 3D hand pose data set.

The RGB-D Object Data set

The RGB-D Object data set is a large collection of 300 common household objects [49]. All these objects are organized into 51 categories arranged using WordNet hypernym-hyponym relationships (similar to ImageNet). This data set was recorded using a Kinect^{TM} style 3D camera that recorded a set of synchronized and aligned 640×480 RGB-D images at 30 Hz. Each object was placed on a turntable and the video sequences were captured for a single, whole rotation. For each object, 3 video sequences were recorded with the camera mounted at different heights so that the object can be viewed from different angles w.r.t. the horizon. Example of segmented objects are shown in Fig. 4.4.

Figure 4.4: RGB-Depth object data set category samples [49].

The American Signal Language Data set

It has been recorded a novel 3D hand poses data set based on the American Sign Language vocabulary. The data set is composed of 23 categories with around 47K instances of both hands. The Kinect^{TM} device was used to extract the hands and their point clouds data using standard segmentation and detection algorithms. Also, both hands were captured not only considering a frontal view but also including variabilities in terms of scale, hand orientation,

and finger joint articulations. This way, the complexity and variability of the overall data set was enriched.

4.3.2 Settings and evaluation metrics

A multi-class classifier is trained using the proposed SBSM descriptor. Specifically, feature vectors are trained in a one-versus-one SVM classifier using a RBF kernel, and optimizing the parameters C and γ by means of cross-validation using LibSVM [19]. SBSM descriptor size was experimentally set to $N_L = 8, N_\theta = 8, N_\phi = 8$ for all the experiments, with a total descriptor length of 512. [1] It has been compared the SBSM descriptor with different state-of-the-art methods on the different experiments [12, 13, 49, 50]. It is also included in the comparative the VFH [70] and ESF [85] descriptors by also training the feature vectors with one-versus-one SVM classifier using a RBF kernel and optimizing the parameters as in the case of SBSM. VFH and ESF have been selected since they are recent, representative, and robust well-known descriptors for shape estimation and codification of normal vectors distribution.

It has been validated the object classification experiments by means of recognition rate applying stratified ten-fold cross-validation and estimating the confidence interval with a two-tailed t-test. We also test and compare the descriptors against depth distortions and noisy data to compute their statistical significance based on Friedman and Nemenyi statistics [23].

4.4 Experiments

Next it is presented the multi-class 3D object categorization performance using SBSM in the RGB-D Object and Sign Language data sets. Once it has been demonstrated the performance of the proposed descriptor, it has been tested against different 3D object distortions.

4.4.1 Analysis of classification performance

For the RGB-D Object Data set, it has been used the turntable data for both training and evaluation, thus classifying 51 different 3D object categories using depth information only. For the object recognition experiments on cropped images, it is applied a leave-one-out strategy as described in [49]. For comparison with the state-of-the-art, it has been compared SBSM performance with the previous results provided on the same data set using the same data partitions for evaluation [12, 13, 49, 50] and ESF [85] and VFH [70] descriptors, as shown in Table 4.1.

Subsequently it is shown the importance of the weight propagation strategy in the SBSM descriptor by setting $\sigma = 1$ and $\sigma = 0$. These two values define the presence or absence of the

[1] The SBSM descriptor code is included as supplemental material.

propagation step, respectively. [2] Based on the mean data set samples volume radius length, it has been set $R = 0.15$. Results reported in Table 4.1 show that the SBSM descriptor clearly outperforms previous state-of-the-art results on this data set. In particular, it is concluded that using neighbor propagation, the performance improves by more than 16% the best result reported in [13] for this data set. This experiment shows that when a neighboring measure of the shape point is taken into account to update neighbor bins, the local variations of shape objects are better learnt by the classifier. Consequently, intra-class variability is reduced without the need of increasing the computational complexity of the descriptor.

Table 4.1: Classification performance on the RGB-depth data set [49].

Method	Depth performance
SIFT + Texton + Color + Spin [49]	64.7%
Sparse Distance Learning [50]	70.2%
VFH + SVM	77.5%
RGB-D Kernel Descriptors [12]	80.3%
Hierarchical Matching Pursuit [13]	81.2%
ESF + SVM	84.9%
SBSM, $\sigma = 0$ + SVM	96.7%
SBSM, $\sigma = 1$ + SVM	**97.9%**

It has been also showm the performance of the VFH, ESF and SBSM on the novel ASL data set. The final performance is obtained by applying a stratified ten-fold cross-validation and testing the confidence interval with a two-tailed t-test. In this case, the spherical grid size is fitted to the minimum spherical grid size containing all the voxels for each data sample. Results are shown in Table 4.2. One can see how the descriptor obtains better performances than using VFH or ESF, and that the best performance is achieved when weight propagation is taken into account.

Method	Depth performance
VFH + SVM	88.7%±0.2
ESF + SVM	92.3%±0.2
SBSM, $\sigma = 0$ + SVM	97.1%±0.6
SBSM, $\sigma = 1$ + SVM	**99.3%±0.4**

Table 4.2: Classification performance and confidence interval of the different descriptors on the novel American Sign Language data set.

4.4.2 Robustness to noise and deformations

In this section it is demonstrated the robustness of the SBSM when describing and classifying 3D objects that suffer from noisy captures and deformations due to different ambient

[2] It has been also tested for different values of σ and experimentally found $\sigma = 1$ to obtain the best results.

conditions or deviations captured by the sensor, as well as partial occlusions. To achieve this goal, are designed two different settings. In the first one it is analyzed the robustness of the descriptor when objects suffer from deviations in the depth dimension in a range from 0 up to 20 millimeters in both directions of the z-axis (Fig. 4.5(a)-(c)). This distortion simulates non-accurate "reading" errors of the sensors because of distance precision and ambient conditions. In the second test we progressively remove shape points from the point cloud from 0 up to 50% of the voxels for each object sample (Fig. 4.5(d)-(f)). This distortion simulated local occlusions and "reading" errors that may produce the removal of some voxel points of the region of interest. Thus, in the depth distortion, the resulting point cloud has the same number of available voxels, though they are distorted in the z-axis, meanwhile in the removing distortion, the resulting point cloud contains less voxel points based on the percentage defined by the distortion percentage. It has been fixed this range of distortions to be representative of the maximum distortion that is possible to find on recorded samples in real scenarios under different conditions and errors produced by different kind of sensors.

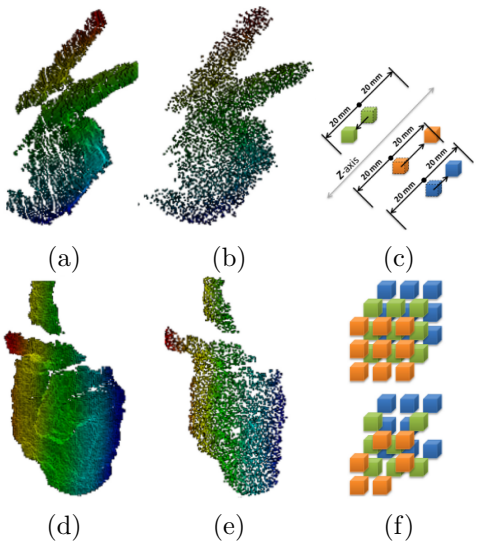


Figure 4.5: (a) Input point cloud for a hand pose instance. (b) Example of distortion in the depth axis for (a). (c) For this distortion each voxel is randomly displaced in the z-axis with a maximum distortion of 20mm in both directions of the axis. (d) Input point cloud for a hand pose instance. (e) Example of cloud removal distortion for (d). (f) For this distortion each voxel of the original point cloud (top) is removed based on a probability value defined by the distortion (down).

In order to perform these analysis, are selected those five categories from the novel 3D human poses data set that achieved the highest confusion in the previous section. The chosen hand categories are displayed in the confusion matrix of Fig. 4.6. In this image the confusion matrix is the mean computed for SBSM on that particular data set.

Recognition rate results are shown when applying distortion in the depth axis in Fig. 4.7: for each degree of distortion, the mean recognition rate and confidence interval for 10 runs of ESF, VFH, and SBSM $\sigma = 1$ descriptors are computed. At each iteration, the percentage of distortion is randomly computed for each object voxel within different depth ranges in

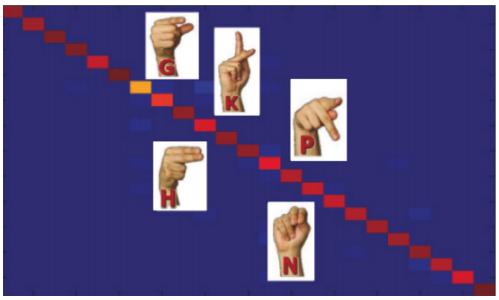


Figure 4.6: Mean confusion matrix of the ASL data set using the SBSM descriptor $\sigma = 1$. Five most confused categories are displayed.

millimeters. The maximum value was set to 20 mm since it is hard to obtain higher deformations produced by the sensor. As expected, the recognition rate for all three descriptors decrease w.r.t. the depth distortion. One can see that for all the different tests of this experiments and descriptors, SBSM still obtains the best performance and VFH suffers the worst decrease in recognition (around 4%).

In Fig. 4.8 the results are shown when applying cloud removal. For each percentage of removed voxels, the mean recognition rate and confidence internal are shown. At each iteration, a percentage of distortion is randomly generated, and different voxels are removed at each time satisfying the percentage of information to be removed. One can see that the general performance ranking in recognition is SBSM, VFH and finally ESF descriptor. Moreover, independently of the percentage of removed number of shape points, the recognition rate for the three methods is maintained in a small range of performance.

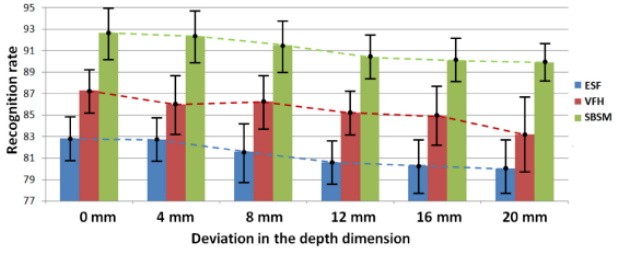


Figure 4.7: Classification performance of different classification strategies under different degrees of distortion in the depth axis on the 5 selected categories in the ASL data set.

4.4.3 Statistical significance

In order to compare the performances computed by the different experiments considered, Table 4.3 shows the mean rank for each descriptor considering 14 different experiments (2 data sets and 6×2 distortion experiments). The rankings are obtained by estimating each particular ranking r_i^j for each data set and experiment i and each descriptor strategy j, and

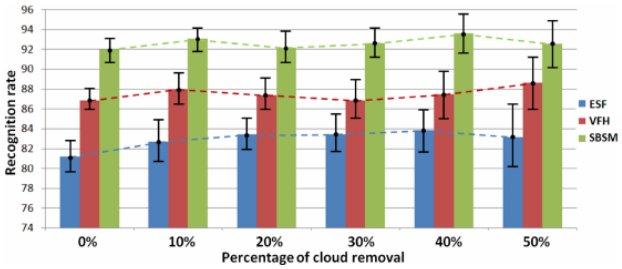


Figure 4.8: Classification performance of different classification strategies under different degrees of cloud removal on the 5 selected categories in the ASL data set.

computing the mean ranking R for each configuration as $R_j = \frac{1}{J}\sum_i r_i^j$, where J is the total number of tests.

Table 4.3: Mean rank for the compared descriptors considering all the experiments.

	VFH	ESF	**SBSM**
Mean rank	2.14	2.85	**1.00**

In order to reject the *null hypothesis*, i.e. measured ranks may differ from the mean rank and these may be also affected by randomness in the results, it has been used the Friedman test [23]. The Friedman statistic value is computed as follows:

$$X_F^2 = \frac{12J}{K(K+1)}\left[\sum_j R_j^2 - \frac{K(K+1)^2}{4}\right]. \tag{4.12}$$

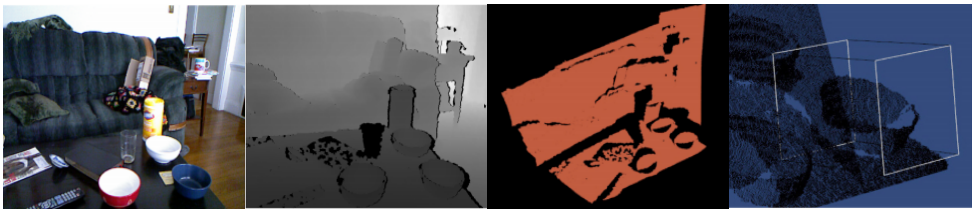
In this case, since $K = 3$ descriptors are compared, $X_F^2 = 24.57$. This value is undesirable conservative, so Iman and Davenport proposed a corrected statistic instead:

$$F_F = \frac{(J-1)X_F^2}{J(K-1) - X_F^2}. \tag{4.13}$$

Applying this correction it is obtained $F_F = 93.17$. With 3 methods and 14 experiments, F_F is distributed according to the F distribution with 2 and 26 degrees of freedom. The critical value of $F(2, 26)$ for 0.05 is 3.37. As the value of $F_F = 93.17$ is clearly higher than 3.37, it is possible to reject the null hypothesis.

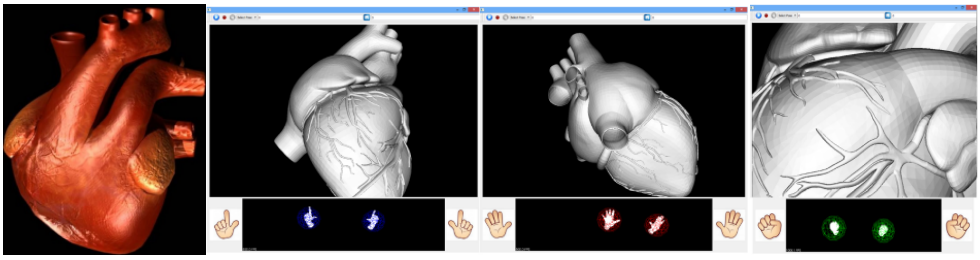
Once it has been checked for the non-randomness of the results, it is performed a post ad-hoc test to check if one of the configurations could be statistically singled out. For this purpose it is used the Nemenyi test, in which the Nemenyi statistic is obtained as follows:

$$CD = q_\alpha \sqrt{\frac{K(K+1)}{6J}}. \tag{4.14}$$



(a) (b) (c) (d)

Figure 4.9: Object spotting in 3D scenes. (a) Example of RGB image of a multi-modal Berkeley data set. (b) Depth image of the same scene. (c) Computed point cloud from the scene. (d) Bowl spotting using SBSM (first positive 3D object prediction is shown based on minimum Euclidean distance).



(a) (b) (c) (d)

Figure 4.10: (a) Original 3D Heart volume of http://thefree3dmodels.com. Automatic interaction with the volume with (b) translation, (c) rotation and (d) zoom manipulation.

In this case with $K = 3$ description strategies to compare and $J = 14$ tests, the critical value for a 95% of confidence ($q_\alpha = 2.35$) is $CD = 0.88$. As the ranking of the proposed SBSM approach does not intersect with any rank for that value of the CD, it is possible to state that for the reported experiments the results are statistically significant w.r.t. VFH and ESF results. In the case of VFH and ESF descriptors, since their rank intersect with the CD value is not possible to state that there exists statistical differences between both strategies.

4.5 Qualitative analysis of SBSM

In this section it is presented four real applications that use the proposed SBSM descriptor, object spotting in 3D scenes and three fully-functional applications for a real-time HCI: 3D medical volume navigation, intelligent retail, and living labs: the library of the future. For this task, an additional novel set of hand poses for the Human Computer Interaction navigation applications was designed.

Figure 4.11: HCI hand poses data set categories.

4.5.1 Object spotting in 3D scenes

After showing the discriminative power and robustness of our proposed descriptor, next it is illustrated the generality of SBSM when applied in real scenarios. To achieve this end, it has been considered a single scene obtained in the public 3D Berkeley data set. [3] The public RGB and corresponding depth map for the selected 3D scene are shown in Fig. 4.9(a) and (b), respectively. Using these data, it is computed the point cloud scene shown in Fig. 4.9(c). In this particular case, are selected one *bowl* object to describe it and perform object spotting within the whole scene. It has been manually selected one bowl from a training image, computed its SBSM descriptor, and performed sliding window search over the three dimensions of the point cloud shown Fig. 4.9(c) for different scale hypotheses of the target object. SBSM descriptor size was set to $N_L = 8, N_\theta = 8, N_\phi = 8$, and $\sigma = 1$, with a total descriptor length of 512. The radius of the sphere is set in the range 5 to 20 cm, with an increment of 1 centimeter, and 2 voxel displacement increments in the three axis among iterations of the sliding windows approach. The first matched region of interest based on the best score obtained by the minimum Euclidean distance among the computed descriptors in the scene is shown in Fig. 4.9(d). Note the accurate fitting of the captured 3D bowl object in the test 3D scene. Moreover, the system spends 13 seconds in a conventional 2.7GHz 2Core 4Gb RAM computer to run this experiment for the tested scene. Given the performed exhaustive search and that it has been ran the experiment iteratively without any

[3] http://kinectdata.com/

kind of parallelism, this experiment show the generality of the descriptor to be applied for scene analysis purposes.

4.5.2 HCI Application for medical navigation

Given the high discriminative power and fast computation of the proposed descriptor in comparison to state-of-the-art approaches, it has been also designed different fully-functional applications to take benefit from it. The first application is an automatic HCI system for medical volume navigation, which is able to detect user, hands, poses, and gestures, manipulating a medical volume of interest.

(a) (b) (c)

Figure 4.12: HCI for retail. (a) Designed 3D retail scenario using Unity engine. (b) User interaction with the scenario. (c) User manipulation by 3D rotation.

(a) (b) (c)

Figure 4.13: Hand reconstruction (c) using two Kinect^{TM} views of point clouds, right (a) and left (b).

The application was developed based upon the MS Kinect^{TM} SDK to capture the RGB-D data stream from the depth sensor. The depth maps were converted into world coordinates projecting the registered 3D image by means of pin-hole model and intrinsic camera parameters. Then, Point Cloud Library (PCL) is used to work with the point cloud, and VTK framework is used for medical image visualization purposes. The hand detection algorithm takes advantage of the Body Pose Skeleton to find the hand wrist joints. Using this information, a fixed radius of interest of 15cm centered in each joint is defined to segment hand point clouds. The usage of the skeleton also enables us to define a heuristic to discard false positive hand poses in the cases that the hands are near the body, or below the waist line. In a later step, the detected hand point clouds within the sphere are used to refine sphere center

by computing center of mass and relocating sphere center, and points are classified using the proposed SBSM methodology. A reduced data set of 18K samples for the six classes shown in Fig. 4.11 were recorded and trained with SBSM and SVM for this purpose.

The detected pose label is then combined with the hand movement (3D object displacements in real coordinates), and used as input of a Hidden Markov Model to obtain a hand gesture classification. The system is able to recognize and control zoom, rotation, and translation operations. In the user interface, the user can visualize its detected hand point clouds in real-time, having a feed-back with the displayed prototypes of both recognized hand poses. Also, the user can visualize the volumetric medical model and the real-time interactions caused by the hand gestures. The overall application enables a powerful and automatic volumetric medical model interaction and visualization. Figure 4.10 shows some examples of detected poses, gestures, interactions, and visualizations on a public 3D heart volume. [4]

4.5.3 HCI Application for Intelligent retail

Based on the same procedure for hand pose recognition than in the medical volume navigation approach, it has been designed a fully-functional application for intelligent retail. In this scenario, hands are tracked in a multi-device setup and two main poses (open and close hand) are recognized. The user can automatically interact with a set of products in an interface designed with the Unity engine [5] so that information about the product and manipulation by 3D rotation based on hand trajectories on two main object axis can be performed. In this scenario, given that it works with bigger displays and and only one Kinect may not cover most part of the pointcloud related to one hand, is included an extra Kinects in order to recover voxels of the same hand from "near complementary" views and reconstruct a new hand with more voxel information. Thanks to the fusion of two views, more information about the tracked hand is available, and thus it has been made the descriptor more discriminative to classify multiple hand poses useful for interaction purposes.

In order to achieve registration of two Kinect^{TM} cameras, we have mounted two cameras in a rigid setup with a baseline of 1.30 meters and an angle of 18 degrees. In this way it is possible to acquire depth images and convert each pixel to real-world coordinates by applying the intrinsic parameters of the cameras. In order to complete the registration, are necessary the extrinsic parameters between both depth cameras, represented by a rotation matrix and a translation vector that allows us to convert $Kinect_2$ depth points to $Kinect_1$ depth coordinate system. Intrinsic and extrinsic parameters can be obtained by doing a stereo calibration. In this case, it has been used the *Camera Calibration Toolbox* [14], that gives us a rotation vector related through the *rodrigues* formula and a translation vector. For this procedure are only necessary few images from both depth cameras looking at the same planar checker-board pattern. Once the calibration is performed are related $Kinect_1$ and $Kinect_2$ depth points as follows:

$$\mathcal{P}_{reference} = \mathcal{R}\mathcal{P} + \mathcal{T},$$

being $Kinect_1$ be the camera reference, $\mathcal{P}_{reference} = (\mathcal{X}_{ref}, \mathcal{Y}_{ref}, \mathcal{Z}_{ref})$ a point in the camera reference coordinate system, $\mathcal{R}$ the rotation matrix obtained in calibration step, $\mathcal{P}$ a point of the $Kinect_2$, and $\mathcal{T} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$ the translation vector obtained in calibration step.

[4] See the supplemental material video for a system demonstration.
[5] http://unity3d.com/

Figure 4.13 shows an example of a hand reconstruction by two different views. From the 3D reconstructed hand, SBSM is computed, and the automatic interactive process is performed. The designed 3D scenario and real-time use case scenarios are shown in Fig. 4.12. In this case is found that increasing the point of view of the hand because of the use of two cameras allowed for a more natural movements of the user while keeping the recognition rate of the system [6].

4.5.4 HCI Application in Living labs

For this last scenario, it is designed a first prototype for intelligent interaction in a virtual repository of books within a Smart Environment or Libing Lab corresponding to a new generation of libraries (VL^3 'Volpelleres Library Living Labs' project). For this scenario, a one Kinect^{TM} camera setup was designed, and the same recognition procedure than in the retail scenario was performed. The 3D scenario was designed with the Unity engine. The real environment to implement the prototype is shown in Fig. 4.14, which is located in the region of Volpelleres in Barcelona. In this prototype the user is able to navigate through a catalogue of books and read the selected ones. The designed 3D scenario and real-time use case scenarios are shown in Fig.4.15.

Finally, it is important to remark that the Kinect acquired a variable frame rate between the range 20-30 FPS. The procedure works real time for all the presented HCI applications, and thus, it has been able to process (segment, describe, and classify segmented point clouds from the region of interest) for all the frames acquired by the Kinect device.

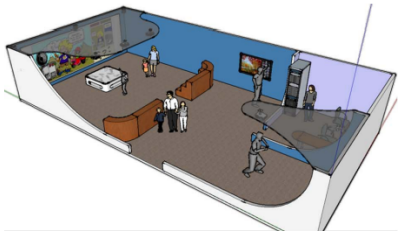


Figure 4.14:

4.6 Conclusions

We presented the Spherical Blurred Shape Model descriptor. SBSM is computationally efficient and highly discriminative. The computed descriptor codifies the spatial relations among object voxels and spherical bins given a granularity degree and a blurring factor defined by a Gaussian-based weight propagation function. The descriptor is rotation invariant by re-locating descriptor bins using the two main quaternion based on the two major 3D

[6]See the supplemental material video for a system demonstration.

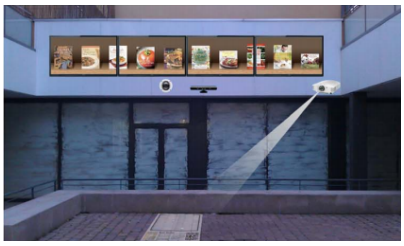


Figure 4.15:

descriptor axis densities. In our experimental evaluation, we found that our methodology outperformed state-of-the-art results up to 16.7% on public depth multi-class object recognition data. Moreover, we tested the descriptor against different 3D data distortions, obtaining high recognition rates and significant performance improvements in relation to standard approaches.

We also tested the descriptors in four real scenarios: object spotting in 3D scenes, within a probabilistic gesture recognition pipeline for real-time Human Computer Interaction in medical volume navigation scenarios, HCI for intelligent retail in a multi-camera setup, and within a prototype for catalogue navigation in a living lab library, showing the high discriminative power, efficiency and generality of the proposed descriptor to be applied in new generation of HCI applications and smart environments.

Chapter 5

Conclusions

Pose Analysis and Gesture recognition are becoming increasingly advanced as technology advances, and with them widen horizons for the analysis of humans. Part of this is due to the expansion of capabilities, such as human computer interaction and the changing needs within the human population. In this thesis it has been studied on the problem of analyzing human pose and motion in RGB-Depth images. In this sense are obtained interesting results, which is not idle modes review to conclusions.

It has been proposed a generic framework for object segmentation using depth maps based on Random Forest and Graph-cuts theory in order to benefit from the use of spatial and temporal coherence, and applied it to the segmentation of human limbs. Our results with the data sets in this point showed high performance segmenting several body parts in depth images compared to classical approaches.

It has been designed and built a system for semi-automatic static posture and spine curvature analysis from 3D anthropometric data (with a low cost RGB-Depth camera) with the aim of providing assistance in the prevention and treatment of musculoskeletal dysfunctions. It has been integrated several frontier computer vision and artificial intelligence techniques including three-dimensional visual data processing, statistical learning, and time series analysis. The system is meant to be highly adaptable and customizable to the needs of the therapist. The validation study shows high precision and reliable measurements in terms of distance, degree and range of movement estimation. Supported by clinical specialists, it is suitable to be included in the clinical routine, including posture reeducation, rehabilitation, and fitness conditioning scenarios.

It has been proposed a fully-automatic general framework for real time action/gesture recognition in uncontrolled environments using depth data. The system analyzes data sequences based on the assignment of weights to gesture descriptors so that DTW cost measure improves discrimination. The feature vectors are extracted automatically through a calibration set, obtaining 3D coordinates of skeletal models with respect an origin of coordinates, making description invariant to translation, scale, and tolerant to corporal differences among subjects. The final gesture is recognized by means of a novel Feature Weighting approach, which enhance recognition performance based on the analysis on inter-intra class variability of vector features among gesture descriptors. The evaluation of the method has been performed on a novel depth data set of gestures, automatically detecting begin-end of gesture

and obtaining performance improvements compared to classical DTW algorithm.

On the other hand, it has been presented the Spherical Blurred Shape Model descriptor. SBSM is computationally efficient and highly discriminative. Experimental evaluation, it has been found that our methodology outperformed state-of-the-art results up to 16.7% on public depth multi-class object recognition data. Moreover, it has been tested the descriptor against different 3D data distortions, being able to deal with multiple deformations in data, obtaining high recognition rates and significant performance improvements in relation to standard approaches.

It has been also tested the descriptors in four real scenarios: object spotting in 3D scenes, within a probabilistic gesture recognition pipeline for real-time Human Computer Interaction in medical volume navigation scenarios, HCI for intelligent retail in a multi-camera setup, and within a prototype for catalogue navigation in a living lab library, showing the high discriminative power, efficiency and generality of the proposed descriptor to be applied in new generation of HCI applications and smart environments.

From the review of the Pose Analysis and Gesture recognition topic, one can observe that still it is possible that progress will also be made in feature extraction by making better use of the multimodal development data for robust transfer learning. Also recent approaches have shown that Random Forest and Deep Learning, such as considering Convolutional Neural Networks, are powerful alternatives to classical Posture Analysis and Gesture recognition approaches, which still open the door for future design of new posture descriptors and gesture recognition classifiers.

In the case of using Convolutional Neural Networks for image classification, most of the methods are still based on sliding windows approaches, which makes recognition a time-consuming task. Thus, the research on methods that can generate multimodal human body posture descriptors candidates, or gestures/actions, from data in a different fashion are still an open issue.

Bibliography

[1] Openni. http://www.openni.org.

[2] Open natural interface. Last viewed 14-07-2011 13:00.

[3] Ajemba, P., Durdle, N., and Raso, V. Characterizing torso shape deformity in scoliosis using structured splines models. In *Biomedical Engineering, IEEE Transactions on* (Jun-2009), vol. 56.

[4] Ajemba, P., Durdle, N., and Raso, V. J. Clinical monitoring of torso deformities in scoliosis using structured splines models. In *Medical and Biological Engineering and Computing* (Dec-2008), vol. 46.

[5] Ajemba, P., Ramirez, L., Durdle, N., Hill, D., and Raso, V. A support vectors classifier approach to predicting the risk of progression of adolescent idiopathic scoliosis. In *IEEE Transactions on Information Technology in Biomedicine* (2005), vol. 9.

[6] Alexa, M., and Adamson, A. On normals and projection operators for surfaces defined by point sets. *Eurographics Symposium on Point-Based Graphics* (2004), 149–155.

[7] Alexandre, L. 3D descriptors for object and category recognition: a comparative evaluation. In *International Conference on Intelligent Robots and Systems (IROS)* (2012).

[8] Ankerst, M., Kastenmüller, G., Kriegel, H.-P., and Seidl, T. 3d shape histograms for similarity search and classification in spatial databases. *Proceedings of the 6th International Symposium on Advances in Spatial Databases 28* (1999), 207–226.

[9] Belongie, S., Malik, J., and Puzicha, J. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 24* (2004), 509–522.

[10] Ben-Chen, M., and Gotsman, C. Characterizing shape using conformal factors. *Eurographics conference on 3D object retrieval (3DOR)* (2008), 1–8.

[11] Betsch, M., Wild, M., Jungbluth, P., Hakimi, M., Windolf, J., Horstmann, B. H. T., and Rapp, W. Reliability and validity of 4d rasterstereography under dynamic conditions. In *Computers in biology and medicine* (2011), vol. 41.

[12] Bo, L., Ren, X., and Fox, D. Depth kernel descriptors for object recognition. *International Conference on Intelligent Robots and Systems (IROS)* (2011).

[13] Bo, L., Ren, X., and Fox, D. Unsupervised feature learning for rgb-d based object recognition. *ISER* (2012).

[14] Bouguet, J.-Y. Camera calibration toolbox for matlab.

[15] Boykov, Y., and Funka-Lea, G. Graph cuts and efficient n-d image segmentation. *International Journal on Computer Vision, 70:109131. ISSN 0920-5691* (2006).

[16] Burel, G., and Henoco, H. Determination of the orientation of 3d objects using spherical harmonics. *Graph Models Image Process 57*, 5 (1995), 400–408.

[17] Carr, J. C., Beatson, R. K., and Cherrie, J. B. Reconstruction and representation of 3d objects with radial basis functions. proceedings of "SIGGRAPH".

[18] Chan, C. S., Liu, H., and Brown, D. J. Recognition of human motion from qualitative normalised templates. *J. Intell. Robotics Syst. 48* (January 2007), 79–95.

[19] Chang, C., and Lin, C. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2* (2011), 1–27.

[20] Colombo, C., Bimbo, A. D., and Valli, A. Visual capture and understanding of hand pointing actions in a 3-d environment. *IEEE Transactions on Systems, Man, and Cybernetics-Part B 33*, 4 (2003), 677–686.

[21] D. Gehrig, H. Kuehne, A. W.-T. S. Hmm-based human motion recognition with optical flow data. In *IEEE International Conference on Humanoid Robots (Humanoids 2009)* (Paris, France, 2009).

[22] Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. *CVPR 1* (2005), 886–893.

[23] Demsar, J. Statistical comparisons of classifiers over multiple data sets. *JMLR 7* (2006), 1–30.

[24] Drerup, B., and Hierholzer, E. Back shape measurement using video rasterstereography and three-dimensional reconstruction of spinal shape. In *Clinical Biomechanics* (1994), vol. 9.

[25] Escalera, S., Fornes, A., Pujol, O., Llados, J., and Radeva, P. Circular blurred shape model for multiclass symbol recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 41*, 2 (2011), 497–506.

[26] Estepar, R. S. J., Brun, A., and Westin, C.-F. Robust generalized total least squares iterative closest point registration. In *In Seventh International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI04), Lecture Notes in Computer Science* (Rennes - Saint Malo, France, 2004).

[27] Everingham, M., Gool, L. V., Williams, C., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *IJCV 88*, 2 (2010), 303–338.

[28] Ferreira, E. A., Duarte, M., Maldonado, E. P., Bersanetti, A. A., and Marques, A. P. Quantitative assessment of postural alignment in young adults based on photographs of anterior, posterior, and lateral views. In *J Manipulative Physiol Ther* (2011), vol. 34.

[29] for Safety, E. A., and at Work, H. Osh in figures: Work-related musculoskeletal disorders in the eu - facts and figures, 2010.

[30] Fortin, C., Feldman, D., Cheriet, F., , and Labelle, H. Validity of a quantitative clinical measurement tool of trunk posture in idiopathic scoliosis. In *Spine* (Sept-2010), vol. 35.

[31] Fortin, C., Feldman, D., Cheriet, F., Gravel, D., Gauthier, F., and Labelle, H. Reliability of a quantitative clinical posture assessment tool among persons with idiopathic scoliosis. In *Physiotherapy* (2011).

[32] Frome, A., Huber, D., Kolluri, R., Bulow, T., and Malik, J. Recognizing objects in range data using regional point descriptors. *European Conference on Computer Vision (ECCV)* (2004).

[33] G. Cheung, T. Kanade, J.-Y. B., and A, M. H. Real time system for robust 3d voxel reconstruction of human motions. 714–720. In Proceedings of CVPR2000, Hilton Head Island,(USA),.

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

[35] Guyon, I., Athitsos, V., Jangyodsuk, P., Escalante, H., and Hamner, B. Results and analysis of the chalearn gesture challenge 2012. *ICPR* (2012).

[36] Hanguen, K., Sangwon, L., Dongsung, L., Soonmin, C., Jinsun, J., and Hyun, M. Real-time human pose estimation and gesture recognition from depth images using superpixels and svm classifier. *Pattern Recognition* (2015).

[37] Harrison, D., Janik, T., Cailliet, R., Harrison, D., Normand, M., Perron, D., and Oakley, P. Upright static pelvic posture as rotations and translations in 3-dimensional from three 2-dimensional digital images: validation of a computerized analysis. vol. 31.

[38] Hernandez, A., Reyes, M., Escalera, S., and Radeva, P. Spatio-temporal grabcut human segmentation for face and pose recovery. In *AMFG10* (2010), pp. 33–40. IEEE Conference on Computer Vision and Pattern Recognition.

[39] Hernandez-Vela, A., Zlateva, N., Marinov, A., Reyes, M., Radeva, P., Dimov, D., and Escalera, S. Human limb segmentation in depth maps based on spatio-temporal graph-cuts optimization. *JAISE 4*, 6 (2012), 535–546.

[40] http://docs.pointclouds.org/trunk/group_features.html.

[41] J Shotton, A Fitzgibbon, M. C. T. S. M. F. R. M. A. K., and Blake, A. Real-time human pose recognition in parts from single depth images. *Computer Vision and Pattern Recognition, IEEE Conference on, Colorado, CO, 2011* (2011).

[42] Jain, H., and Subramanian, A. Real-time upper-body human pose estimation using a depth camera. *HP Technical Reports 1*, 190 (2010).

[43] Jaremko, J., Poncet, P., Ronsky, J., Harder, J., Dansereau, J., Labelle, H., and Zernicke, R. Estimation of spinal deformity in scoliosis from torso surface cross sections. In *Spine* (Jul-2001), vol. 26.

[44] Jaremko, J., Poncet, P., Ronsky, J., Harder, J., Dansereau, J., Labelle, H., and Zernicke, R. Indices of torso asymmetry related to spinal deformity in scoliosis. In *Clinical Biomechanics* (Oct-2002), vol. 17.

[45] Jonathan Starck, Atsuto Maki, S. N. A. H., and Matsuyama, T. The multiple-camera 3-d production studio. *IEEE Transactions on circuits and systems for video technology 19* (2009), 6.

[46] Kazhdan, M., Funkhouser, T., and Rusinkiewicz, S. Rotation invariant spherical harmonic representation of 3d shape descriptors. *SIGGRAPH symposium on Geometry processing* (2003), 156–164.

[47] Knopp, J., Prasad, M., Willems, G., Timofte, R., and Gool, L. V. Hough transform and 3d SURF for robust three dimensional classification. *European Conference on Computer Vision (ECCV)* (2010).

[48] L. Igual, J.C. Soliva, A. H.-V. S. E. X. J. O. V., and Radeva, P. A fully-automatic caudate nucleus segmentation of brain mri: Application in volumetric analysis of pediatric attention-deficit/hyperactivity disorder. 10(105). *doi: 10.1186/1475-925X-10-105* (2011).

[49] Lai, K., Bo, L., Ren, X., and Fox, D. A large-scale hierarchical multi-view rgb-d object dataset. *International Conference on Robotics and Automation* (2011).

[50] Lai, K., Bo, L., Ren, X., and Fox, D. Sparse distance learning for object recognition combining rgb and depth information. *Int. Conference on Robotics and Automation* (2011).

[51] Leroux, M. A., and Zabjek, K. A noninvasive anthropometric technique for measuring kyphosis and lordosis: application for scoliosis. vol. 25.

[52] M Reyes, A Claps, J. R.-J. R., and Escalera, S. Automatic digital biometry analysis based on depth maps. *Computers In Industry* (2013).

[53] Mikic, I., Trivedi, M., Hunter, E., and P. Cosman, A. B. Posture estimation from multi-camera voxel data. In *CVPR* (2001).

[54] Mitra, N., Nguyen, A., and Guibas, L. Estimating surface normals in noisy point cloud data. *International Journal of Computational Geometry and Applications 14* (2004), 261–276.

[55] Molina, J., and Martnez, J. A synthetic training framework for providing gesture scalability to 2.5d pose-based hand gesture recognition systems. *Machine Vision and Applications* (2014).

[56] Mori, G., Belongie, S., and Malik, J. Efficient shape matching using shape contexts. In *TPAMI* (2005), vol. 27.

[57] Mortensen, E. N., Deng, H., and Shapiro, L. A sift descriptor with global context. *CVPR 1* (2005), 184–190 vol. 1.

[58] N. Dalal, B. Triggs, C. S. Human detection using oriented histograms of flow and appearance, 2006. ECCV.

[59] Ni, B., Moulin, P., and Yan, S. Pose adaptive motion feature pooling for human action analysis. *International Journal Of Computer Vision* (2015).

[60] Parizeau, M., and Plamondon, R. A comparative analysis of regional correlation, dynamic time warping, and skeletal tree matching for signature verification. *IEEE TPAMI 12*, 7 (1990).

[61] Parsa, K., Angeles, J., and Misra, A. Pose-and-twist estimation of a rigid body using accelerometers. In *International Conference on Robotics and Automation - ICRA* (2001), vol. 3.

[62] Pedregosa. Scikit-learn: Machine learning in python. In *Journal of Machine Learning Research* (2011), pp. 2825–2830.

[63] P.O. Ajemba, N.G. Durdle, D. H., and Raso, V. J. Validating an imaging and analysis system for assessing torso deformities. In *Computers in Biology and Medicine* (March2008), vol. 38.

[64] Potmesil, M., and Chakravarty, I. A lens and aperture camera model for synthetic image generation. vol. 15.

[65] PrimeSense Inc. *Prime Sensor NITE 1.3 Algorithms notes*, 2010. Last viewed 14-07-2011 13:19.

[66] Redondo-Cabrera, C., Lopez-Sastre, R. J., Acevedo-Rodriguez, F., and Maldonado-Bascon, S. SURFing the point clouds: Selective 3d spatial pyramids for category-level object recognition. *IEEE Computer Vision and Pattern Recognition conference* (2012).

[67] Rodgers, J., Anguelov, D., Hoi-Cheung, P., and D., K. Object pose detection in range scan data. *CVPR* (2006), 2445–2452.

[68] Ruggeri, M., Patane, G., Spagnuolo, M., and Saupe, D. Spectral-driven isometry-invariant matching of 3d shapes. *International Journal of Computer Vision (IJCV) 89* (2010), 248–265.

[69] Rusu, R., Blodow, N., and Beetz, M. Fast point feature histograms (FPFH) for 3d registration. *International Conference on Robotics and Automation* (2009), 1848–1853.

[70] Rusu, R., Bradski, G., Thibaux, R., and Hsu, J. Fast 3d recognition and pose using the viewpoint feature histogram. *International Conference on Intelligent Robots and Systems (IROS)* (2010), 2155–2162.

[71] Rusu, R., Marton, Z., Blodow, N., and Beetz, M. Learning informative point classes for the acquisition of object model maps. *Control, Automation, Robotics and Vision* (2008), 643–650.

[72] Sabata, B., Arman, F., and Aggarwal, J. Segmentation of 3d range images using pyramidal data structures,. *CVGIP: Image Understanding 57*, 3 (1993), 373–387.

[73] Saupe, D., and Vrani, D. V. 3d model retrieval with spherical harmonics and moments. *DAGM* (2001), 392–397.

[74] Shoemake, K. Animating rotation with quaternion curves. *Computer Graphics and Interactive Techniques* (1985), 245–254.

[75] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. Real-time human pose recognition in parts from single depth images.

[76] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. Real-time human pose recognition in parts from single depth images. *CVPR* (2011).

[77] Sminchisescu, C., Kanaujia, A., and Metaxas, D. Conditional models for contextual human motion recognition. *CVIU 104*, 2-3 (2006), 210–220.

[78] Steder, B., Rusu, R., Konolige, K., and Burgard, W. NARF: 3d range image features for object recognition. *International Conference on Intelligent Robots and Systems (IROS)* (2010).

[79] Tombari, F., Salti, S., and Stefano, L. D. Unique signatures of histograms for local surface description. *Proceedings of the 11th European conference on computer vision conference (ECCV)* (2010), 356–369.

[80] Toshev, A., and Szegedy, C. Deeppose: Human pose estimation via deep neural networks. *Computer Vision and Pattern Recognition, IEEE Conference on, Columbus, OH, 2014* (2014).

[81] V. Ganapathi, V., Plagemann, C., Koller, D., and Thrun, S. Real time motion capture using a single time-of-flight camera. *CVPR* (2010), 755–762.

[82] W Pichao, L Wanqing, G. Z.-Z. J. C. T. O. P. Deep convolutional neural networks for action recognition using depth map sequences. *IEEE Transactions on Human-machine Systems, 2015* (2014).

[83] W. Schwartz, A. Kembhavi, D. H.-L. D. Human detection using partial least squares analysis. ICCV.

[84] Wang, W., and Qin, X. Image inpainting algorithm based on CSRBF interpolation. vol. 12.

[85] Wohlkinger, W., and Vincze, M. Ensemble of shape functions for 3d object classification. *IEEE Robotics and Biomimetics* (2011).

[86] Wu, H., Ronsky, J., Cheriet, F., Küpper, J., Harder, J., Xue, D., and Zernicke, R. Prediction of scoliosis progression with serial three-dimensional spinal curves and the artificial progression surface technique. In *Medical and Biological Engineering and Computing* (2011), vol. 48.

[87] Y Bengio, A. C., and Vincent, P. Unsupervised feature learning and deep learning. a review and new perspectives. *CoRR abs/1206.5538, 2012* (2012).

[88] Yao, J., Ruggeri, M. R., , and Taddei, P. Automatic scan registration using 3d linear and planar features. Springer-Verlag, Ed., vol. 1.

[89] Zhang, P., Wang, Z., Zheng, S., and Gu, X. A design and research of eye gaze tracking system based on stereovision. *Emerging Intelligent Computing Technology and Applications, Lecture Notes in Computer Science 5754*, 4 (2009), 278–286.

[90] Zhang, Z. A flexible new technique for camera calibration. In *TPAMI* (2000), vol. 22.

[91] Zhou, F., De la Torre, F., and Hodgins, J. K. Aligned cluster analysis for temporal segmentation of human motion. In *IEEE Conference on Automatic Face and Gestures Recognition (FG)* (September 2008).