

Technische Universität Berlin

Fakultät IV - Elektrotechnik und Informatik

Fachgebiet AOT

Prof. Dr. Sahin Albayrak

Masterarbeitsproposal

Hand Gesture Recognition for Human-Robot
Interaction

Sivalingam Panchadcharam Aravinth

Matrikel-Nummer 342899

Betreuer Dr. Yuan Xu

Contents

Index	II
1 Abstract	1
2 Motivation	2
3 Background	3
3.1 Robot and Artificial intelligence	3
3.1.1 NAO Vision	3
3.1.2 Extending NAO with Depth Camera	4
3.2 Hand Gesture	4
3.3 Hand Gesture Recognition	5
3.3.1 Gesture Modeling	5
3.3.2 Gestural Taxonomy	6
3.3.3 Gesture Recognition	7
4 Goal	8
5 Approach	9
5.1 Sensing	10
5.2 Feature Extraction	10
5.3 Modeling and Classification	10
5.4 Gesture Analysis and Recognition	11
5.5 Human-Robot Interaction	12
6 Schedule	13
7 Details	14
Bibliography	15

1 Abstract

Human-robot interaction (HRI) has been a topic of both science fiction and academic speculation even before any robots existed [1]. HRI research is focusing to build an intuitive, and easy communication with the robot through speech, gestures, and facial expressions. The use of hand gestures provides an attractive alternative to complex interfaced devices for HRI. In particular, visual interpretation of hand gestures can help in achieving the ease and naturalness desired for HRI. This has motivated a very active research concerned with computer vision-based analysis and interpretation of hand gestures. Important differences in the gesture interpretation approaches arise depending on whether 3D based model or appearance based model of the gesture is used.

In this thesis, we attempt to implement the hand gesture recognition for robots by modeling, training, analyzing, and recognizing gestures by using Computer Vision and Machine Learning techniques. Additionally, 3D based gesture modeling with skeletal points tracking will be used. On the one side, gestures will be used command the robot to execute certain actions and on the other side gestures will be translated and spoken out by the robot.

We further hope to provide a platform to integrate Sign Language Translation to assist people with hearing and speech disabilities. However, further implementations and training data are needed to use this platform as a full fledged Sign Language Translator.

Keywords

Human-Robot Interaction, HRI, Computer Vision, Depth Camera, Hand Gesture, Gesture Recognition, Sign Language Translation, Hidden Markov Model, NAO

2 Motivation

Huge influence of computers in society has made smart devices, an important part of our lives. Availability and affordability of such devices motivated us to use them in our day-to-day living. The list of smart devices includes personal automatic and semi-automatic robots which are also playing a major role in our household. For an instance, Smart Vacuum Cleaner is a robot that automatically cleans the floor and goes to its charging station without human interaction.

Interaction with smart devices has still been mostly through displays, keyboards, mouse and touch interfaces. These devices have grown to be familiar but inherently limit the speed and naturalness with which we can interact with the computer. Usage of robots for domestic and industrial purposes have been continuously increasing. Thus in recent years there has been a tremendous push in research toward an intuitive, and easy communication with the robot through speech, gestures, and facial expressions.

Tremendous progress had been made in speech recognition, and several commercially successful speech interfaces have been available. However, there has been an increased interest in recent years in trying to introduce other human-to-human communication modalities into HRI. This includes a class of techniques based on the movement of the human arm and hand, or hand gestures. The use of hand gestures provides an alternative mode of communication for Human-robot interaction than the conventional cumbersome devices.

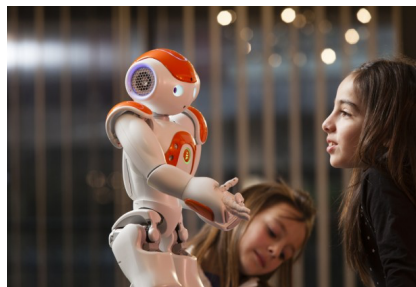


Figure 2.1: Human Robot Interaction by Aldebaran Robotics

3 Background

3.1 Robot and Artificial intelligence

Proper vision is the utmost importance for the function of any vision based autonomous robot. Areas of artificial intelligence deal with autonomous planning or deliberation for robotic systems to navigate through an environment. A detailed understanding of these environments is required to navigate through them. High-level information about the environment could be provided by a computer vision system that is acting as a vision sensor.

Computer vision is a field that includes methods for acquiring, processing, analyzing, and understanding images and, in general, high-dimensional data from the real world in order to produce numerical or symbolic information, e.g., in the forms of decisions [4].

In this thesis, we will focus on the hand gesture recognition using computer vision techniques for a humanoid robot named as NAO, as shown in the figure 3.1. Nao is an autonomous, programmable humanoid robot developed by Aldebaran Robotics. The Nao Academics Edition was developed for universities and laboratories for research and education purposes. Table 3.1 shows the specification of NAO according to Aldebaran Robotics.

3.1.1 NAO Vision

Two identical video cameras are located in the forehead of NAO. They provide a up to 1280x960 resolution at 30 frames per second. They can be used to identify objects in the visual field such as goals and balls, and bottom camera can ease NAO's dribbles. NAO contains a set of algorithms for detecting and recognizing faces and shapes. NAO can recognize who is talking to it or find a ball or, eventually, more complex objects.

Table 3.1: Nao’s hardware and software specification

Height	58 centimetres (23 in)
Weight	4.3 kilograms (9.5 lb)
Autonomy	60 minutes (active use), 90 minutes (normal use)
Degrees of freedom	21 to 25
CPU	2 x Intel Atom @ 1.6 GHz
Built-in OS	Linux
Compatible OS	Windows, Mac OS, Linux
Programming languages	C++, Python, Java, MATLAB, Urbi, C, .Net
Vision	2 x HD 1280x960 cameras
Connectivity	Ethernet, Wi-Fi
Sensors	4 x directional microphones 1 x sonar rangefinder 2 x IR emitters and receivers 1 x inertial board 9 x tactile sensors 8 x pressure sensors

3.1.2 Extending NAO with Depth Camera

3D cameras such as Microsoft Kinect and Asus Xtion are used not only for gaming but also for analyzing 3D data, including algorithms for feature selection, scene analysis, motion tracking, skeletal tracking and gesture recognition.

In this thesis, we focus on using Asus Xtion as an external hardware as shown in the figure 3.2 due to the computational limitations of the NAO to process the images from two inbuilt hd stereo cameras, that ultimately makes NAO incapable of doing an efficient gesture recognition in real time.

3.2 Hand Gesture

Human hand gestures are a means of non-verbal interaction among people. They range from simple actions of using our hand to point at, to the more complex ones that express our feelings and allow us to communicate with others. To exploit the use of gestures in HRI, it is necessary to provide the means by which they can be interpreted by robots. The HCI interpretation of gestures requires that dynamic and/or static configurations of the human hand, arm, and even other parts of the human body, be measurable by the machine [3].

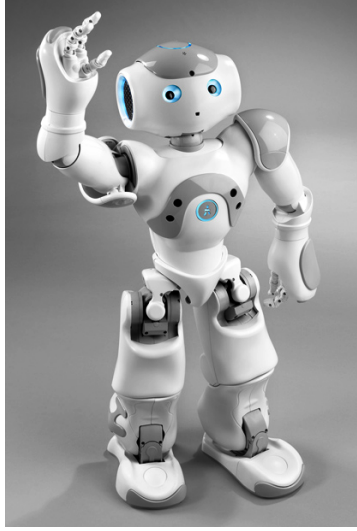


Figure 3.1: NAO

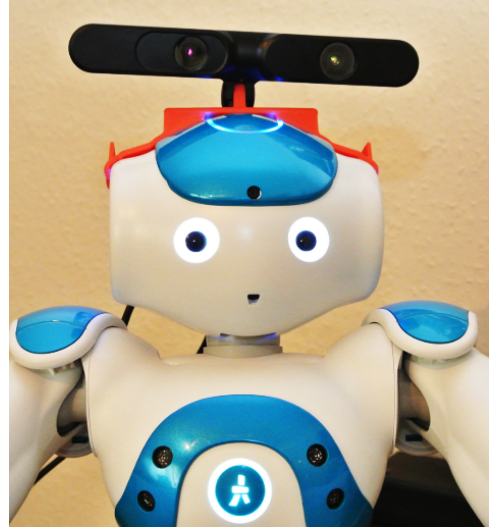


Figure 3.2: Asus Xtion mounted on NAO

3.3 Hand Gesture Recognition

Initial attempts to recognize hand gestures resulted in electro-mechanical devices that directly measure hand and/or arm joint angles and spatial position using sensors [5]. Glove-based gestural interfaces require the user to wear such a complex device that hinders the ease and naturalness with which the user can interact with the computer controlled environment.

Even though such hand gloves are used in highly specialized domain such as simulation of medical surgery or even in the real surgery, the everyday user will be certainly deterred by such a sophisticated interfacing devices. As an active result of the motivated research in HRI, computer vision based techniques were innovated to augment the naturalness of interaction.

3.3.1 Gesture Modeling

Figure 3.3 shows various types of modeling techniques used for Gesture modeling [5]. Selection of an appropriate gesture modeling depends primarily on the intended application. For an application that needs just hand gesture to go up and down or left and right, a very simple model may be sufficient. However, if the purpose is a natural-like interaction, a model has to be sophisticated enough to interpret all the possible gesture. The following section discusses various gesture modeling techniques which are being used by the existing hand gesture recognition applications.

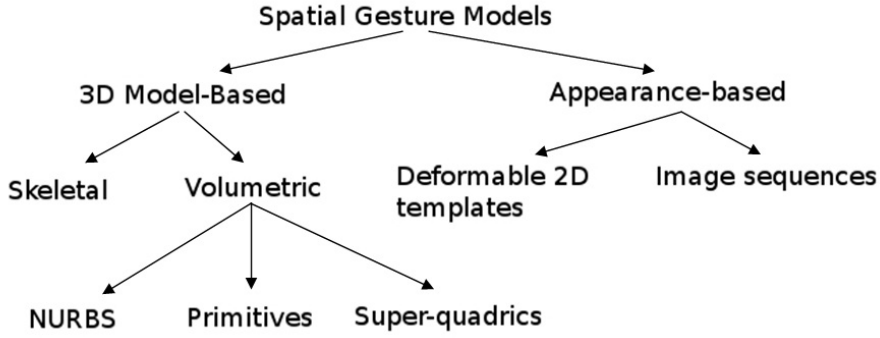


Figure 3.3: Different gesture modeling

Appearance based models don't use a spatial representation of the body, because they derive the parameters directly from the images or videos using a template database.

3D hand based model approach uses volumetric or skeletal models, or even a combination of both. Volumetric approaches have been heavily used in computer animation industry and for computer vision purposes. The models are generally created of complicated 3D surfaces. The drawback of this method is that is very computational intensive.

Instead of using intensive processing of the 3D hand models and dealing with a lot of parameters, one can just use a simplified version that needs the joint angle parameters along with segment length. This is known as a skeletal representation of the body, where a virtual skeleton of the person is computed and parts of the body are mapped to certain segments [6]. The analysis here is done using the position and orientation of these segments and the relation between each one of them.

In this thesis, we focus on Skeletal based modeling algorithms which are faster because the detection program has to focus only on the significant parts of the body. Moreover, it allows to do pattern matching against a template database.

3.3.2 Gestural Taxonomy

Several alternative taxonomies have been suggested that deal with psychological aspects of gestures [5]. All hand/arm movements are first classified into two major classes as shown in the figure 3.4.

Manipulative gestures are the ones used to act on objects. For example moving a chair from one location to another. Manipulative gestures in the context of HRI are

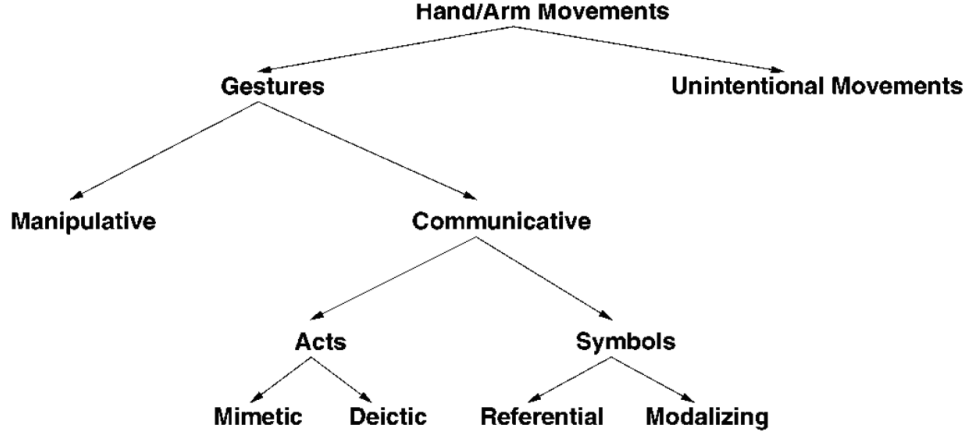


Figure 3.4: Different gesture modeling

mainly developed for medical surgery. Communicative gestures, on the other hand, have purely communicational purpose. In a natural environment they are usually accompanied by speech or spoken as a sign language. In HRI context these gesture are one of the commonly used gestures, since they can often be represented by static as well as dynamic hand postures.

In this thesis, we focus on communicative gestures in the form of symbols. They symbolize some referential action. For instance, circular motion of hand may be a referred as an alphabet "O" or as an object such as wheel or as a command to turn in a circular motion .

3.3.3 Gesture Recognition

The task of gesture recognition shares certain similarities with other recognition tasks, such as speech recognition and biometrics. Though alternatives such as DP-matching algorithms have been attempted, the most successful solutions involves feature-based statistical learning algorithm, usually Hidden Markov Models [7].

In this thesis, we focus on the recognition technique based on hidden Markov model (HMM) that is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. In simpler Markov models, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output dependent on the state is visible.

4 Goal

As described earlier, HRI research is focusing to build an intuitive, and easy communication with the robot through speech, gestures, and facial expressions. The use of hand gestures provides the ease and naturalness with which the user can interact with robots.

In this thesis, we attempt to implement the feature for NAO to recognize hand gestures and execute predefined actions based on the gesture. NAO will be extended with an external depth camera, that will enable NAO to recognize 3D hand based modeled gestures. This 3D camera will be mounted on the head of NAO, so that it can scan for gestures in the horizon. Additionally, skeletal points tracking algorithm with machine learning technique using Hidden Markov Models (HMM) will be used to recognize the gestures. With the hand gesture recognizing feature, NAO will be available to the users in two modes.

- **Command mode:** In this mode, a gesture will be recognized by NAO and related task will be executed. Even though the gesture based Interaction is real time, NAO can not be interrupted or stopped by using any gesture while it is executing a task. However, part of NAO intelligence such as voice commands can be used in such situation to stop or interrupt the ongoing task execution.
- **Translation mode:** In this mode, NAO will be directly translating the meaning of the gesticulated gestures. To achieve this, text-to-speech library of NAO will be used and recognized gesture can be spoken out loud using the integrated speaker. In future, it will allow NAO to translate a sign language to assist people with hearing and speech disabilities.

In this thesis, we planned to train NAO with few very simple gestures due to the reason that NAO has limited resources. Gestures will involve both the hands or single hand to interact with the robot.

5 Approach

The figure 5.1 shows the flow diagram of hand gesture recognition system that is going to implemented in this project work. Each block contains different software components that are executed sequentially. However, training phase must be carried out before this system is available for recognition. Finally, these components will be integrated into NAO.

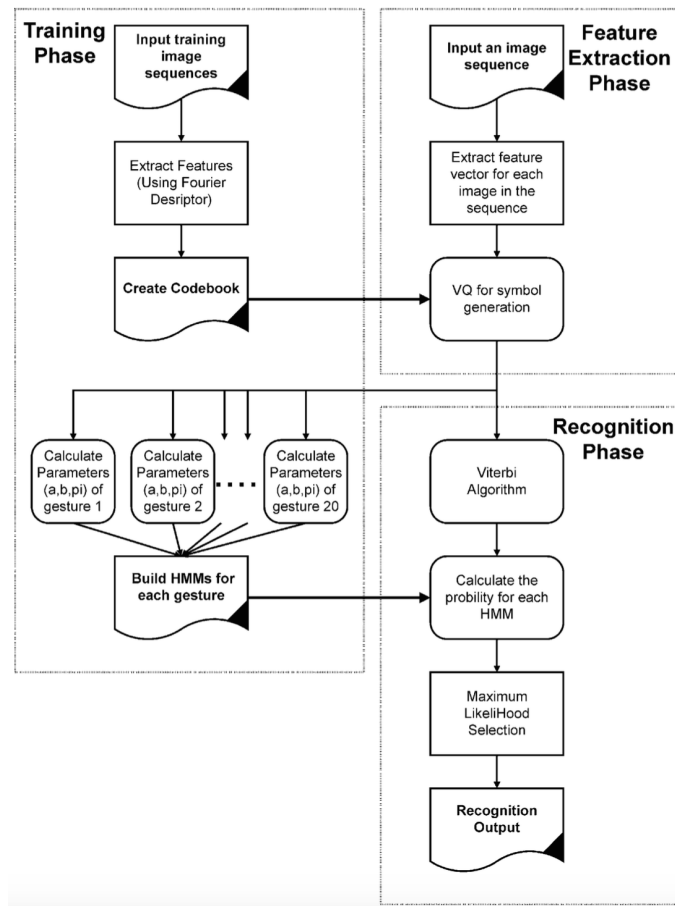


Figure 5.1: Flow diagram of hand gesture recognition system



Figure 5.2: Depth image of 3D Camera

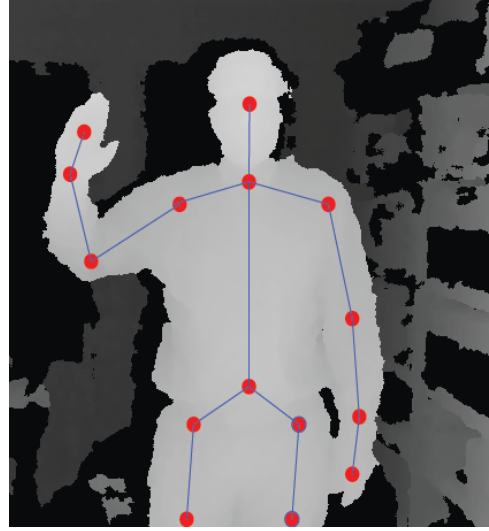


Figure 5.3: Skeleton tracking

5.1 Sensing

3D camera records 30 frames of color image as well as depth image per second and outputs as a data package. Figure 5.2 shows the single frame of depth image taken from Microsoft Kinect where darker gray values represent the farther distance and lighter gray values represent the closer distance from the camera.

5.2 Feature Extraction

Output package from sensor data will be inputted to feature detection and extraction unit. OpenNI is a software component that will track the anatomical landmarks of the human body from the package and extract significant joint angle parameters along with segment length and present them three dimensionally as shown in the figure 5.3. Finally, only joints of both the arms will be picked out from the array of features, since it will be the significant joints need for hand gesture recognition.

5.3 Modeling and Classification

In order to use this hand recognition system, all chosen gestures must be observed and the system must be trained. Therefore, a set of simple gestures will be chosen and observed for training. Each gesture is isolated in time and gesticulated for

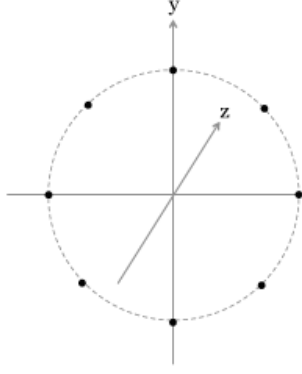


Figure 5.4: HMM States of circular gesture

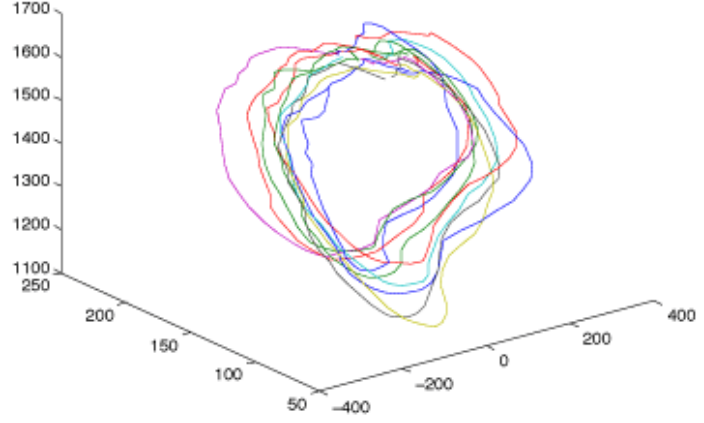


Figure 5.5: Observations of circular gestures

certain duration. However, sensors provides 30 frames of discrete states of gesture per second.

For example, a gesture is gesticulated by simply drawing a circle in the air and its ideal states are shown in the figure 5.4. It will be modeled as a position of hand passing through 8 states of the circle. Each state is a point in space with x, y and z axis data. This approach makes it possible for us to reduce our observation data to sequential 3D points and focus on the recognition task without processing all those pixels. Each trained model can then be used to determine with what probability a given gesture appears in test data. Therefore, the trained Hidden Markov models will be used to recognize gestures.

5.4 Gesture Analysis and Recognition

This step contains the analysis of gesticulated gesture and finding out the likelihood of that gesture with trained data, known as gesture recognition. Figure 5.5 contains colored rings of noisy data of gestures that represent instances of a real circle gesture. Sensing and feature detection module will produce 60 observations of the circle gesture for 2 seconds, since the depth camera records at 30 frame per second. To decide whether a given set of 60 observations contains a circle gesture, we need to first determine the likelihood that the hand passed through the eight states of the gesture in the expected sequence.

Discrete HMM is a finite set of M possible output symbols and a sequence of N hidden states which reveal some probability. To reduce our real gesture data to a

workable number of discrete output symbols and states, we can use any clustering algorithm to cluster the 3D points of all our training data for circle gesture sequences into N clusters and label them. That is to say, every point in the training data represents an output symbol that is closely tied to one of the 8 true states of the model.

Looking at the labeled data, we can estimate how likely it is that a hand passed through the 8 clusters in the same sequence as a circle gesture, and if the likelihood is high enough, then the gesture is considered to be recognized.

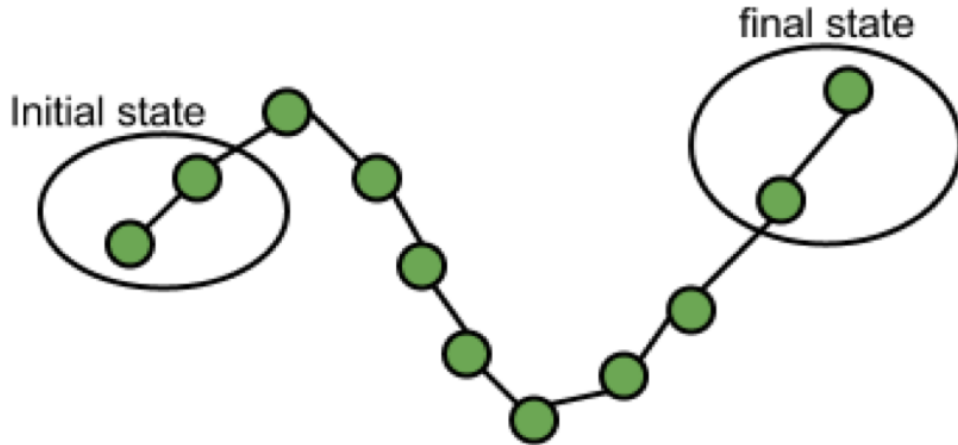


Figure 5.6: shows the states of Arabic sign language to convey "Hello". It is consisted of 11 states of Markov models and if there is higher likelihood of the hand position passed through states sequentially from initial to final, then it is recognized as "Hello".

5.5 Human-Robot Interaction

Finally, the recognized gesture will be interpreted by NAO to execute an specified task. The motion of NAO will be defined by the application Choregraphe in Command Mode. For example. circle gesture would ask NAO to turn around. However, NAO will also be available in Translation Mode by using ALTextToSpeech Library to translate the recognized gesture.

6 Schedule

- **October - Preparation:** This includes the analysis of relevant papers as well as the consideration of the already available components.
- **November - Setup:** Software and Hardware environment setup. Getting to know the tools that will be used.
- **December / January - Implementation:** Proposed functionalities will be implemented.
- **January / February - Testing:** This includes testing, bug fixing and improving the implementation.
- **February / March - Documentation:** This includes the documentation of the code and writing up the thesis.

7 Details

- Language of the Master Thesis: English
- Document processing system: LaTeX with TexStudio in Mac OSX
- Operating system: Linux Ubuntu, Mac OS X
- Software components: OpenNI, OpenCV, NAO Choregraphe
- Hardware components: NAO, Asus Xtion
- Programming Language: C++, Python
- Advisor: Dr. Yuan Xu
- Professor: Prof. Dr. Sahin Albayrak

Bibliography

- [1] *Human Robot Interaction*. Wikipedia. 1
- [2] CHANG-YI KAO, CHIN-SHYURNG FAHN: *A Human-Machine Interaction Technique: Hand Gesture Recognition Based on Hidden Markov Models with Trajectory of Hand Motion*. In *Advanced in Control Engineering and Information Science*. Elsevier Ltd., 2011.
- [3] HO-SUB YOON, JUNG SOH, YOUNGLAE J. BAE HYUN SEUNG YANG: *Hand gesture recognition using combined features of location, angle and velocity*. In *Pattern Recognition*. Elsevier Science Ltd., 1999. 4
- [4] XIA LIU, KIKUO FUJIMURA: *Hand Gesture Recognition using Depth Data*. In *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004. 3
- [5] VLADIMIR I. PAVLOVIC, RAJEEV SHARMA, THOMAS S. HUANG: *Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review*. In *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 19, No. 7. IEEE, 2007. 5, 6
- [6] LOREN ARTHUR SCHWARZ, ARTASHES MKHITARYAN, DIANA MATEUS NASIR NAVAB: *Human skeleton tracking from depth data using geodesic distances and optical flow*. In *Image and Vision Computing*. Elsevier Ltd., 2011. 6
- [7] FENG-SHENG CHEN, CHIH-MING FU, CHUNG-LIN HUANG: *Hand gesture recognition using a real-time tracking method and hidden Markov models*. In *Image and Vision Computing*. Elsevier Ltd., 2001. 7