


Proceedings of the Second Symposium on Information and Communication Technology

General Chairs: [Thang Huynh Quyet](#) [Hanoi University of Science and Technology, Vietnam](#)

[Dinh Khang Tran](#) [Hanoi University of Science and Technology, Vietnam](#)

Publication of:

- Conference
- [SoICT '11 Symposium on Information and Communication Technology 2011](#)
Hanoi, Viet Nam — October 13 - 14, 2011
[ACM](#) New York, NY, USA ©2011


 2011 Proceeding


 [Bibliometrics](#)

- Citation Count: 25
- Downloads (cumulative): 4,297
- Downloads (12 Months): 290
- Downloads (6 Weeks): 50


Tools and Resources

 TOC Service:
 [Email](#)  [RSS](#)

 [Save to Binder](#)

 Export Formats:
[BibTeX](#) [EndNote](#) [ACM Ref](#)

Share: |

 [Contact Us](#) | Switch to [single page view](#) (no tabs)

[Abstract](#) [Source Materials](#) [Authors](#) [References](#) [Cited By](#) [Index Terms](#) [Publication](#) [Reviews](#) [Comments](#) [Table of Contents](#)

Proceedings of the Second Symposium on Information and Communication Technology

Table of Contents

[← previous proceeding](#) | [next proceeding →](#)

[From services composition to end users programming](#)

[Boualem Benatallah](#)

Pages: 1-1

doi>[10.1145/2069216.2069217](#)

In this talk, we will review state of the art in services composition and reuse. We discuss main issues related to simplifying composition and increasing reuse. Current APIs and composition techniques including mashups, however, aim toward developers ... [expand](#)

[Models of information diffusion in social networks](#)

[Eric Gaussier](#)

Pages: 2-2

doi>[10.1145/2069216.2069218](#)

Social networks now play a central role for sharing information and discussing different types of events. The way information spreads in such networks has often been compared to the way innovations spread in marketing or viruses spread in populations. ... [expand](#)

[Multi-GNSS environment: present and future opportunities for South-East Asia](#)

[Gabriella Povero](#)

Pages: 3-3

doi>[10.1145/2069216.2069219](#)

In the last decade, the world of Satellite Navigation has experienced an important change: while in late 90s the USA and Russia played a dominant role as only providers of global navigation satellite systems, the new century saw the arrival of new actors ... [expand](#)

[From grid to cloud computing](#)

[Ladislav Hluchý](#)

Pages: 4-4

doi>[10.1145/2069216.2069220](#)

With the advance of computational technologies, the applications running on modern distributed systems became more and more complex. Large-scale applications like environmental applications require such an amount of computation powers on demand that ... [expand](#)


SESSION: **Evolutionary computation and constraint solving**

[An approach of ant algorithm for solving minimum routing cost spanning tree problem](#)

[Nguyen Minh Hieu, PhanTan Quoc, Nguyen Duc Nghia](#)

Pages: 5-10

doi>[10.1145/2069216.2069222](#)

Full text:  [PDF](#)

Minimum routing cost spanning tree problem-MRCT is one of classical problems in network designing. In this paper, we will introduce a new


Web page classification is the process of categorizing a web page into one or more classes which have been predetermined. If we remove all HTML tags from a web page, then this process can be considered as a text classification problem. However, this ... [expand](#)

[A fully automatic hand gesture recognition system for human-robot interaction](#)

[Thi Thanh Mai Nguyen](#), [Ngoc Hai Pham](#), [Van Thai Dong](#), [Viet Son Nguyen](#), [Thi Thanh Hai Tran](#)

Pages: 112-119

doi>[10.1145/2069216.2069241](#)

Full text:  [PDF](#)


Recently, human - machine interaction (HMI) becomes a hot research topic because of its wide applications, ranging from automatic device control to designing and development of assistant robot or even smart building at sparser scale. One of the most ... [expand](#)

[A meaningful model for computing users' importance scores in Q&A systems](#)

[Pham Tuan Long](#), [Nguyen Van Dong Anh](#), [Nguyen Thi Thanh Vi](#), [Le Quoc](#), [Huynh Quyet Thang](#)

Pages: 120-126

doi>[10.1145/2069216.2069242](#)

Full text:  [PDF](#)


This paper presents a meaningful model for computing users' importance scores in Q&A systems that can stimulate the development of these systems. Since the score can be used in ranking users' expertise, it can motivate users to contribute more to Q&A ... [expand](#)

[Modeling the brown plant hoppers surveillance network using agent-based model: application for the Mekong Delta region](#)

[Viet Xuan Truong](#), [Alexis Droqoul](#), [Hiep Xuan Huynh](#), [Minh Ngoc Le](#)

Pages: 127-136

doi>[10.1145/2069216.2069243](#)

Full text:  [PDF](#)


This paper aims at modeling a brown plant hopper (BPH) surveillance network, called BPH surveillance network model (BSNM). In this model, we apply the Unit Disk Graph (UDG) technique to setup a graph with multiple surveillance nodes (light traps). An ... [expand](#)

[New protocol supporting collaborative simulation](#)

[Trong Khanh Nguyen](#), [Nicolas Marilleau](#), [Tuong Vinh Ho](#), [Amal El Fallah](#)

Pages: 137-145

doi>[10.1145/2069216.2069244](#)

Full text:  [PDF](#)

Major researches in the domain of complex systems are interdisciplinary, collaborative and geographically distributed. The purpose of our work is to explore a new methodology that facilitates scientist's interactions during the simulation process. Through ... [expand](#)


SESSION: Speech processing

[Emotional speech classification using hidden conditional random fields](#)

[La The Vinh](#), [Sunqyoung Lee](#), [Young-Koo Lee](#)

Pages: 146-150

doi>[10.1145/2069216.2069246](#)

Full text:  [PDF](#)


Although there have been a great number of papers in the area of emotional speech recognition, most of them contribute to the feature extraction phase. Regarding classification algorithm, hidden Markov model (HMM) is still the most commonly used method. ... [expand](#)

[Speech enhancement using combination of dereverberation and noise reduction for robust speech recognition](#)

[Tien Dung Tran](#), [Dang Khoa Nguyen](#), [Thi Anh Xuan Tran](#), [Quoc Cuong Nguyen](#), [Huu Binh Nguyen](#)

Pages: 151-158

doi>[10.1145/2069216.2069247](#)

Full text:  [PDF](#)


In this paper, we describe a speech enhancement approach for robust speech recognition. This approach consists of two stages to solve both current problems of speech recognition: reverberation and noise. Firstly, speech signal is dereverberated by suppression ... [expand](#)

[Blind speech separation combining DOA estimation and assignment problem approaches](#)

[Vuong Hoang Nam](#), [Nguyen Quoc Trung](#), [Tran Hoai Linh](#)

Pages: 159-164

doi>[10.1145/2069216.2069248](#)

Full text:  [PDF](#)

In this paper, we propose an effective method for blind speech separation of convolutive mixtures in the frequency domain. The main difficulty in a frequency approach is the permutation problem. In the proposed method, we use two previous approaches ... [expand](#)

[Non-uniform unit selection in Vietnamese speech synthesis](#)

A fully automatic hand gesture recognition system for human - robot interaction

Thi Thanh Mai Nguyen, Ngoc Hai Pham, Van Thai Dong, Viet Son Nguyen, Thi Thanh Hai Tran
MICA Center, HUST - CNRS/UMI 2954 - Grenoble INP
Hanoi University of Science and Technology
1 Dai Co Viet St., Hanoi, Vietnam
{Thanh-Mai.Nguyen, Ngoc-Hai.Pham, Van-Thai.Dong, Viet-Son.Nguyen, Thanh-Hai.Tran}@mica.edu.vn

ABSTRACT

Recently, human - machine interaction (HMI) becomes a hot research topic because of its wide applications, ranging from automatic device control to designing and development of assistant robot or even smart building at sparser scale. One of the most important questions in this research field is finding out an efficient and natural method of HMI. Among several channels of communication, hand gestures have been shown to be an intuitive and efficient mean to express an idea or to control something. For a successful hand gesture based interaction between human and robot, a vocabulary of hand gestures needs to be defined. To resolve this problem, in this paper, we propose a framework to study the behavior of Vietnamese in using of hand gesture in communication with robot. This study allows designing a hand gesture vocabulary for human - robot interaction (HRI) applications. In the literature, there are no works similar to ours. This makes our twofold contributions: (1) a proposed framework for hand gesture recognition: hand gesture vocabulary design, feature extraction for hand posture representation, hand posture classification, and hand gesture database construction; (2) some experimentations are realized to evaluate the defined hand gesture set that can be used in general situation of HRI. The experiment results show that the defined hand gesture set satisfies the both criteria: intuitiveness and recognisability.

Categories and Subject Descriptors

Image/Video Processing, Object Detection and Recognition, Face and Gesture Analysis, Vision for Graphics and Robotics.

General Terms

Image Processing and Computer Vision

Keywords

Visual recognition, Haar like feature, Cascaded Adaboost Classifier, Hand Gesture, Wizard of Oz technique.

1. INTRODUCTION

Gesture is an intuitive and efficient mean of communication between human and human in order to express information or to interact with environment. In Human Computer Interaction (HCI),

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SoICT 2011, October 13–14, 2011, Hanoi, Vietnam.

Copyright 2011 ACM 978-1-4503-0880-9/11/10...\$10.00.

hand gesture can be an ideal way that a human controls or interacts with a machine. In that case, the machine must be able to recognize human hand gesture. Recently, hand gesture recognition becomes a hot research topic in the HCI and Computer Vision field due to its wide applications such sign hand language, lie detection, game, e-learning, human-robot interaction, etc.

In a human robot interaction application, to be able to interact with human through hand gesture, the robot needs to understand hand gestures. The recognition will be performed by learning examples of gestures of interest and then recognizing a given new gesture. For a successful hand gesture based interaction between human and robot, a vocabulary of hand gestures needs to be defined and a gesture based protocol of communication should be understood by both human and robot. Then a system for hand gesture recognition must be built so that it could be integrated in the robot for an automated interaction.

This paper proposes a framework for designing hand gesture set and a system for hand gesture recognition. For hand gesture vocabulary, we use the Wizard of Oz technique, which has been considered as an essential tool for the design of multimodal interface. For hand gesture recognition, we inspire the idea proposed by Viola and Jones [1] which has been shown to be very successful for object detection and classification. Our main contributions are:

- A wizard of oz framework for designing hand gesture vocabulary, which can be re-used to design other interaction modality interface (e.g. speech);
- A common vocabulary of hand gestures which can be used for any human robot interaction application without needing to be redefined;
- A database of hand gestures commonly used for human robot interaction community;
- A fully automated hand gesture recognition system which is available to be integrated into robot for human robot interaction

This paper is organized in five sections as follows: In Section 2, we analyze some works relating to the hand gestures acquisition, and hand posture classification. In following part (Section 3), a frame work for hand gesture recognition is proposed that includes the hand gesture vocabulary design, feature extraction for hand posture representation, hand posture classification, and hand gesture database construction. Section 4 represents some experimentation that is performed to evaluate the automatic hand gesture recognition system. Conclusions and future works are discussed in Section 5.

2. RELATED WORKS

In the literature, there are two main approaches for acquisition of hand gestures: glove-based approach [2] and vision-based approach [3], [4]. Glove-based approach requires the human to wear a specific glove. This approach is very fast and accurate. But the human feels uncomfortable when wearing a sensor-based glove. In addition, the glove is very expensive. To overcome this drawback, vision-based approach uses camera to capture visually hand gestures. The image/video of hand gesture captured by camera will be next processed by hand gesture recognition system using image processing/computer vision techniques. With the rapid increasing of technology, camera becomes more and more cheap and is ubiquitously used, so the vision-based approach is more economical and convenient for human - robot interaction in ubiquitous perceptive environment than glove-based approach.

Vision-based hand gesture recognition approaches could be divided into two categories: 3D hand model-based approaches and appearance-based approaches [5]. In this paper, we will analyze only some typical approaches of each category, mostly the ones proposed in the context of human – robot interaction, and then evaluate their performance in order to compare with our proposed framework.

3D hand modeling approach represents explicitly hand posture through parameters as angles at junction point and hand pose [6], [7], [8], [9]. When hand gesture is represented by a 3D model, the 3D parameters of the model will be learnt. The recognition will be carried out by projecting the 3D model on the image that we need to detect hand. The recognition consists of looking for a transformation that minimizes the difference between points on contour lines of image in question and the projected model.

We found that the 3D hand model-based approach offers a rich description that allows a wide class of hand gesture. This approach is ideal for realistic interactions in virtual environment. However, it has some disadvantages. First, at each frame, the initial parameters of the model have to be close to the solution; otherwise it is liable to find a suboptimal solution (i.e. local minima). Secondly, the fitting process is very sensitive to noise. Thirdly, as 3D hand models are articulated deformable objects with many DoFs (Degrees of Freedom), a very large number of images are required to cover all hand configurations under different views. Finally the 3D model approach cannot handle the inevitable self-occlusion of the hand when mapping it to 2D plan. For these reasons, most works based on 3D model are hand tracking because the hand configuration doesn't change a lot between two consecutive frames.

Appearance-based approaches use image features to model the visual appearance of the hand. In the following, we will study some methods using visual features for hand modeling and recognition. Appearance models do not give an explicit representation about hand structures, but lead a lot of methods for hand recognition. Features range from global (e.g. entire image, PCA (Principal Component Analysis) technique) to local (e.g. Haar like feature), static to motion features, numeric (e.g. oriented histogram) to structured features (e.g. elastic graphic, ridge and blob). Global features give a simple learning and recognition but they are not robust to occlusions. Local features as Haar like give the good results of recognition. In addition, the computation is very fast. However, Haar like features require more examples for training and depend strongly of hand subjects to be learnt. Ridge and blob are good features for representing structured objects but blob and ridge extraction is quite sensitive to noise (when connecting ridge at intrinsic scale) and time - consuming.

In [10], the author used the entire image of the hand as input of a neural network to learn and recognize a set of command gestures in a robot controlling application. The network has four layers (the input layer, two hidden layers, the output layer). The number of inputs and outputs equal to the sample resolution of hand posture (20 x 20, 18 x 20 or 18 x 30). To recognize hand posture in the image, an active window containing face will be determined. The skin detection will locate hand in the image based on the relative position between hand and face and hand posture will be classified using the trained neural network. The classification is carried out by computing the distance of the example to the posture set. To evaluate the algorithm, the authors have built a database of six posture classes corresponding to A, B, C, 5, Point, V. Each posture is used to execute a command in a system for locating and tracking individual speaker (LISTEN). A thousand images have been taken in different conditions of lighting, background, view, scale. When the images are of uniform background, the recognition rate obtains 94% while it is reduced significantly to only 75% in case of complex background. The algorithm has been also evaluated with Jochen database of grayscale images with resolution 128 x 128, of ten postures taken from 24 different subjects on white-black or complex background [11]. The recognition rate attains 93.7% with simple background images and 84.4% with complex background images.

The hand posture classification using neural network has following advantages: the neural network can model more complex distribution of data than traditional methods. However, it has some drawbacks: it is not able to describe the model of the data. This is why one cannot explain why in some cases it works well but not other cases. In addition, it is very hard to extract rules from neural network to help analyzer to explain the results. As all others methods, if the training data are not significant, the network cannot give good response.

In [12], the authors used Cascaded Adaboost for hand posture classification using Haar like features. Each Cascaded Adaboost classifier is trained to recognize a posture class. To classify postures, the parallel structure of Cascaded Adaboost will be built. In this work, four postures are considered. The dataset are taken with webcam at resolution 320 x 240. Each posture class has 450 samples taken at different scale, angle view on simple background. 500 non-hand images are taken as negative images to train the Cascaded Adaboost classifiers. The recognition rate is about 98% for each posture class when testing with 100 images. The algorithm works well even there is a rotation of 15 degree of the hand.

In [13], Haar like features extracted from candidate image window are inputs of Cascaded Adaboost. Six classes of postures have been learnt: closed, side point, victory, open, Lpalm, Lback. These postures are taken in varying conditions. To evaluate the algorithm, the authors have taken 2300 images of right hand of ten men and ten women with two different cameras in indirect hand lighting condition. Images are then normalized to rotation and size. The authors showed that the Cascaded Adaboost built from 100 weak Adaboost classifiers gives recognition rate 92.23% with false recognition rate 1.01×10^{-8} .

As Haar like features are not invariant to rotation, strong change in illumination and scale, [14] computed SIFT (Scale Invariant Feature Transform) features and used SIFT descriptor as input vector for Adaboost. Three hand postures are considered: Palm, Fist and Six, which are acquisitioned in different conditions. 642 training images are taken from Massey dataset [15]. 450 images of the Fist posture and 531 images of the Six posture are taken at different lighting conditions. Background image are taken from

830 images from internet and 149 images taken by the authors. To evaluate the algorithm, 275 images have been taken from webcam 320 x 240. The recognition rate obtained with Adaboost is 95.4% in average. The authors showed that the combination of SIFT - Adaboost is better than Haar - Adaboost, mostly when there is noise and the background is complex. The SIFT - Adaboost works still well when the hand rotates an angle of 40 degree. We found that Adaboost is a good classification method for hand posture classification. SIFT - Adaboost is still better than Haar - Adaboost due to some interesting properties of SIFT features. However, SIFT is very time consuming therefore not suitable for real-time application.

SVM (Support Vector Machine) has been used for object classification and recognition. In [16], the authors used active contour algorithm to determine the human body and hand contour lines. Some heuristics are used to determine hand on this active contour. The feature vector is then built from the coordinates of the four points: top of head and fingertip, feet position and the shoulder. Two subjects participated into experimentation. Five videos are captured. The SVM is used to recognize "Pointing" gesture and obtained recognition rates 71% when hand is observed from side camera and 94.4% in case of frontal camera.

Techniques for hand gesture recognition depend of features and the method used for hand representing. Cascaded Adaboost is a good accompany of Haar like features, first to reduce the number of features to be considered, secondly, to reject rapidly non-candidate hand gestures. The recognition rate is about 92%. The combination of SIFT and Cascaded Adaboost gives better recognition rate than Haar like features and Cascaded Adaboost. However, SIFT is time-consuming and not widely used for real-time applications. SVM is a typical classifier for all types of numeric features. However, currently the using of SVM for hand gesture recognition does not obtain good result. Neural network is used in some cases and give quite good results.

For dynamic hand gestures, HMM (Hidden Markov Model) is commonly used in signal processing and also in hand gesture recognition [17], [18]. Generally, each state of HMM represents a configuration of hand (hand posture) and the number of nodes in HMM represents key configurations representing dynamic hand gesture. The using of ANN (Artificial Neural Network) in the choosing of the best state that fits the observation data is a good solution in the framework of HMM [19]. It gives a better recognition rate (90%) than using only HMM. However ANN is time consuming.

In conclusion, there is no exact answer for the question: which method is the best for hand gesture recognition. The recognition rate reported in the cited papers is not evaluated on the same database. In general, each work has been evaluated on a database built by the authors themselves according to the application context. Some of databases are published for research use. But it is necessary to rebuild database for a specific application. In addition, this gesture set, as proposed by researchers, is imposed for human without considering if they do this in a comfortable manner or not. Beside, the methodology for designing and building a hand gesture database has not been mentioned yet in all related scientific papers.

For these reasons, in this paper, we would like to deal with two main problems: (1) designing a common hand gestures for research use in human robot interaction (HRI); (2) building an automated hand gesture recognition system for HRI applications.

3. PROPOSED FRAMEWORK FOR HAND

GESTURE RECOGNITION

We would like build a fully automated hand gesture recognition system for human - robot interaction applications. The term "fully automated" means that no manual intervention needs to be required during the communication between human and a robot. In many applications such as assistant robot in a library or museum; static hand gesture is enough to explain a control or an attitude of the human to the robot. Therefore, in this paper, our framework will deal only with static hand gestures.

3.1 Proposed framework for hand gesture recognition

We propose a framework consisting of two main processes: training and recognition (Figure 1). Training process aims to learn hand gesture classifiers while the other process performs the classification of a given hand gesture into a predefined hand gesture category. The whole system is composed of the following main modules:

- **Feature Extraction** module extracts significant and discriminant features for hand representation.
- **Classifier Learning** module learns classifiers for each hand gesture category.
- **Hand detection and gesture recognition:** Given a new image, this module scans all extracted regions at different scales and it will determine if the candidate region contains a hand and which hand gesture category it belongs to.

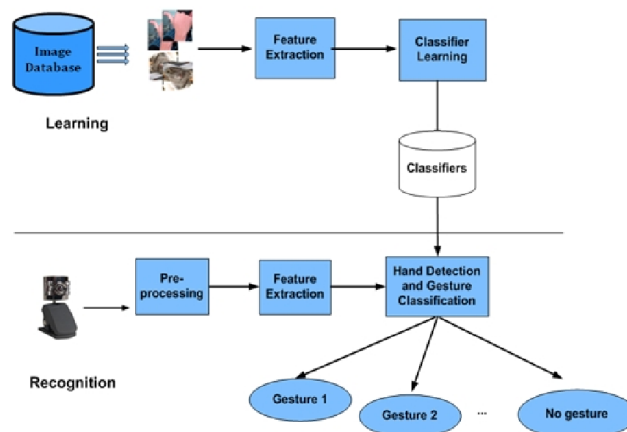


Figure 1: Proposed framework for hand gesture recognition

In order to realize a fully automated hand gesture recognition system, we identify the following tasks:

- **Task 1:** Define a hand gesture set that is considered in human - robot interaction.
- **Task 2:** Choose the best features for hand representation and a convenient method for feature extraction.
- **Task 3:** Identify a method of hand gesture classification which is suitable to the above hand representation.
- **Task 4:** Build a hand database for training and evaluation of the recognition system.

In the following subsections, we will describe in more detail each task and identify our contributions in each task.

3.2 Hand gesture vocabulary design

3.2.1 Framework of hand gesture vocabulary design

The framework of designing hand gesture vocabulary is presented in Figure 2. It consists of four main blocks: (1) definition of interaction scenarios; (2) human – robot interaction (HRI) observation in each scenario by camera; (3) hand gestures extraction and analysis; (4) definition of hand gestures set. In the second block, a set of people will be invited to participate into interaction with the robot without knowing that their interaction is registered (we refer to the Wizard of Oz technique - an efficient way to examine user interaction with robot). This allows obtaining the most natural HRI.

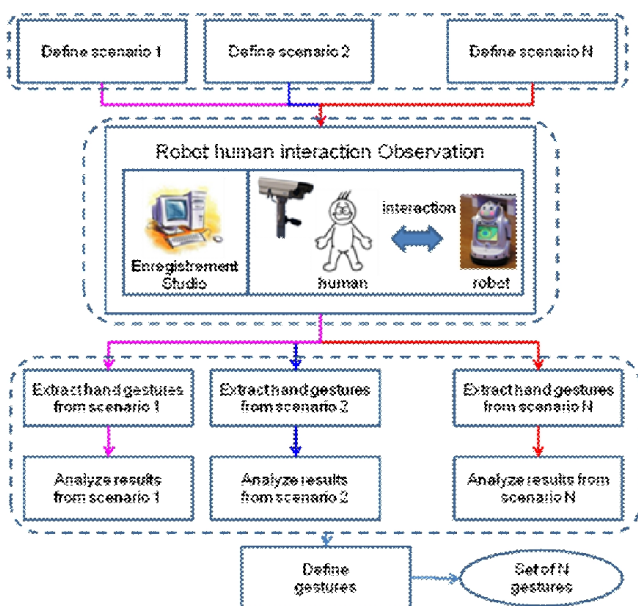


Figure 2: Framework of designing hand gesture vocabulary

3.2.2 Definition of HRI scenarios

In order to study the behaviors of Vietnamese in communication with robot and to build a set of hand gestures, we define a series of HRI scenarios in a simulated library context. It needs to be noted that this simulated context is not a special context, so the HRI studied in this context can be used and extended too many other contexts. The scenarios must be basic and simple which allow subjects play them easily and exactly.

The simulated library is a room of size 3m x 3m in which we equip some tables, chairs, bookshelves. All are similar to a reading room in the library so that the human can feel as in a real library. In this context, the scenarios are played by two actors (a human and an assistant robot in the library). We focus on some general and basic demands that a human needs and often uses in the library, e.g. human wants to look for a room, a book/paper/magazine, or he/she wants to know more about the history of library, the library map, the book/paper/magazine brief ... or even wants to know about the robot's services in the library, etc.

To define interaction scenarios, we invent situations and assign roles to a human and a robot. The scenario can start with a human entering into the library, learnt that there is a service robot, he looks around the room to find the robot, then calls the robot coming near to him to ask some services like looking for a book; asking to know more about the book; looking for a room; etc. During the playing, the human can do anything (by gesture or voice) to explain his demand or his attitude to the robot. Once all demands are responded/refused, the human feels (un)happy to pass

the time in the library, he ends the interaction with the robot and goes outside. Figure 3 extracts a frame of a scenario in which a human is interacting with the assistant robot in the library.



Figure 3: An example of scenario in which a human asks the robot to know more about the abstract of a book to which his hand is pointing. The robot will answer the human by synthetic voice using Vietnamese speech synthesis system.

Although these scenarios are played in the context of a library with library specific operations, we will only study behaviors of human interacting with the robot in the most five common situations: *call the robot*; *point to something for a service*; *agree or disagree with the robot's answer*; *finish the interaction*.

3.2.3 HRI observation

Once scenarios are defined, we start filming the scene with three cameras to assure that all in the scene are visible. In order to study the hand gesture set of Vietnamese in HRI, a multimodal corpus (video/audio) was built with twenty-two native Vietnamese people (eleven males and eleven females) with a mean age of 23. There are fourteen right-handers, and eight left-handers. These people have the same awareness and knowledge level.

To be able to obtain the natural HRI, we say to the human that we would like to *test the robot's abilities*, i.e. *the performance of speech and gesture recognition system embedded on robot while interacting with human*. People do not know that robot is controlled by an anonym technician in another room.

All people are asked to express five different demands above by using their voice and hand gestures. All twenty two peoples play two times all the defined scenarios, yielding 66 video files (22 subjects x 3 cameras). All videos files are recorded in the same format avi, sampled at 25 fps with resolution 352 x 280. After selecting and editing, we have obtained 850 clips (corresponding to 459 scenarios) that only present one hand gesture per one scenario.

3.2.4 Hand gesture analysis and proposed hand gesture vocabulary













All 850 clips selected from the database are analyzed. The analysis should answer to the following questions:

- Which gestures are used in each scenario?
- How are gestures characterized?

A hand gesture is defined as a sequence of movements of hand postures. In general, a gesture is composed of three phases:

preparation; execution; finish, but we are interested only in the execution phase.

Table 1: The hand gestures are using to express the five different commands: Call robot, point to an object, Agree or Disagree with robot’s answer, and Finish the interaction with robot.

Type	Illustration	Description	Per.
Call1		hand open, wave, hand hollow down	92%
Call2		hand open, wave, hand hollow up	8%
Point1		open, hand point, not change hand shape	23%
Point2		close, forefinger points, not change hand shape	77%
Agree1		fingers close, but the thumb up	61%
Agree2		fingers open, but forefinger and thumb close	30%
Agree3		fingers close, but forefinger and middle finger make the victory symbol	4%
Agree4		hand clap	5%
Dis1		Fingers open, hand moves left, then right, then left, not change hand shape	82%
Dis2		close, forefinger points down, not change hand shape	18%
Stop1		Fingers open, hand moves left, then right, then left, not change hand shape	94%
Stop2		fingers close, but forefinger and middle finger make the victory symbol	6%

The analyzed results show that:

- There are two interesting differences between the human - human interaction and the HRI: (1) Vietnamese use hand gesture more often to impress robot than they do in human – human; (2) the performing time for one gesture in the case of HRI is longer than the one of human – human interaction, because they need to keep the gestures until obtaining the robot’s response. Therefore, in almost scenarios, the amplitude and speed of hand movement are categorized into average group, not narrow and fast group as we expected.
- For each scenario (the same context and the same command), several types of hand gestures are used. For example, in order to call the robot to come, Vietnamese use

two different hand gestures (call1 and call2), but to express an agreement with robot’s answer, they present four different ones (agree1, agree2, agree3, agree4) (see Table 1).

The designing of a hand gesture vocabulary needs to satisfy two following criteria: (1) the comfortableness for human when doing it; (2) the recognisability for system when observing it.

The first criterion is assured by choosing hand gestures that are mostly used by human when interacts with the robot. As analyzed results in Table 1, five following hand gestures maybe selected: Call1 (92%), Point2 (77%), Agree1 (61%), Dis1 (82%), Stop1 (94%). But it is easy to realize that Vietnamese people use a same hand gesture to express two different commands: disagree with robot (Dis1) and end the interaction (Stop1). To assure the recognisability of the system (hand gestures need to be distinct), we have decide to choose the gesture Dis2 for expressing a disagreement with robot’s answer. In conclusion, five following hand gestures (the bold one in Table 1) will be selected in the hand gesture: Call1 (92%), Point2 (77%), Agree1 (61%), Dis2 (18%), Stop1 (94%).

3.3 Feature extraction for hand posture representation

As analyzed in the related works section, Haar-like feature is a good candidate for hand posture representation due to its simplicity and efficiency for computation. Therefore, in our work we proposed to use Haar-like features. Haar-like feature is composed of "black and white" rectangular features characterized by a corner, size (width, height) and orientation (00 or 450) and a value. The value is the difference between the sum of all "white" pixel values and the one of all "black" pixel values. These values are computed in a very fast manner using integral image technique. For each image, the algorithm of computing Haar-like feature is described in the Figure 4. The output of the algorithm is a set of Haar-like features representing the input image.

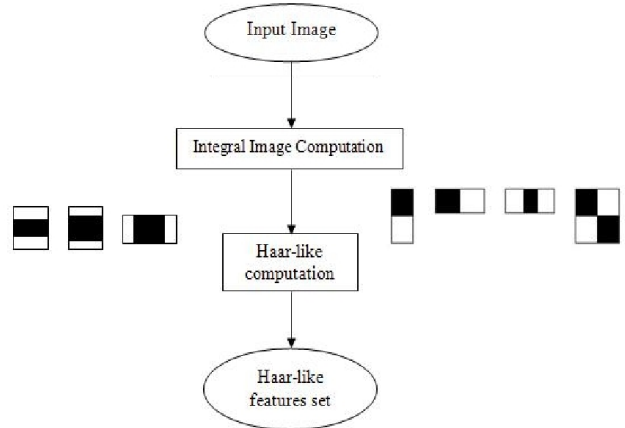


Figure 4: Algorithm of Haar-like features extraction.

3.4 Hand posture classification

Although the technique of integral image gives a quick computation of Haar-like feature, the number of features computed for one image is very big. This number is much bigger than the size of the image - over complete problem. For example with the image of size 22 x 22, the number of features is about 100000. If we represent an image by a very big number of features like this, the searching for correspondence will be not efficient. In addition,

among extracted Haar-like features, it is not true that all features are significant and discriminated for posture classification.

The Adaboost algorithm has more advantage than other learning machine techniques because it discards lots of non-significant features for hand detection. Therefore, at classification phase, in order to reduce the computation time, only a little number of features (7 - 35 in our experiment) will be used.

The main advantage of the integration of Adaboost in a cascaded architecture is that it allows improving the execution speed of classifier because the Adaboost integrated cascade can reject rapidly all candidates which are not hand posture. A candidate will be classified into one category if the candidate has passed all the cascade layers.

To reject as soon as possible all negative examples, the architecture cascade will be used. If the classifier has K stages, f_i , d_i are max false alarm rate, and min detection rate of the stage i , respectively. The false alarm rate F and detection rate D of whole cascade will be calculated as following:

$$F = \sum_{i=1}^K f_i, \quad D = \sum_{i=1}^K d_i$$

In reality, each stage can have different max false alarms and min detection rates. The next stage has the smaller false alarm rate and bigger min detection rate than the previous ones. However, the number of features to be learnt will be bigger. The bigger number of features is, the longer computation time is. In practice, for simplifying, we set f_i and d_i to the same value, hence in this case, $F = f_i^K$, and $D = d_i^K$.

Figure 5 represents the architecture of Cascaded Adaboost classifiers for hand posture classification.

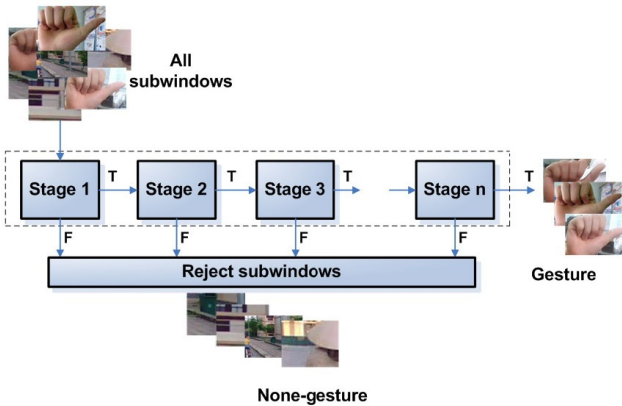


Figure 5: Architecture of Cascaded Adaboost classifiers for hand posture classification

Given a new image, one sub-window is scanned across the image at multiple scales and locations. Initially, the size of the sub-window is set by the size of positive samples, then it is increased up with a ratio s defined until the size of the sub-window is less than or equal to the size of the image. The smaller s value is the higher accuracy is. But the small s value will make an increasing of the computing time. The sub-window is also scanned across location. Subsequent locations are obtained by shifting the sub-window some number of pixels (β). The values of s and β will affect to the detector speed as well as the accuracy. In our experiment, we set s , and β value to 1.1 and 1, respectively. For

each sub-window, if the area in the sub-window has the same features with the positive images, it is a region containing the gesture.

3.5 Hand gesture database construction

In order to build a hand gesture database, we have set up a room for recording session in indoor environment (Figure 6). In order to have a several points of view, we use two cameras. As we want all video flux coming from two cameras are synchronized, we have developed a recording tool that allows to record and display videos coming from camera and the PC screen content. This allows actors to know roughly trigger sources for each gesture. Finally, we have recorded a hand gestures database for 20 Vietnamese participants including 10 females and 10 males with the age from 20 to 30 year old. As results, our hand gesture database contains 1200 videos (3 video for each gesture x 20 subjects x 2 backgrounds (uniform and complex) x 2 cameras x 5 gestures) with 5 seconds of duration.

4. EXPERIMENTATION

This section contributes to evaluate:

- If the hand gesture set defined by the Wizard of Oz technique satisfies the recognisability.
- If proposed hand gesture recognition system is automated and reliable enough to be applied in real human robot interaction application. While evaluating the performance of the proposed system, the recognisability of hand gesture set will be deducted.



Figure 6: Recording setup for building the training database

4.1 Dataset and parameter setting

In our experimentation, we define the two different experiments: Dependent subject experiment and independent subject experiment. In the both experiment, we always use our entire hand gesture database for training the recognition system that included 1200 positive images (60 images per persons x 20 subjects) and 1500 negative ones for one hand gesture. The 1200 positive images set of each hand gesture is built under the same neon light condition but with two different back ground: 600 images in the uniform back ground and 600 ones in the complex back ground.

In the dependent subject experiment, we collect a small database of two subjects (one man and one woman, 500 positive images, and 50 negative one for each hand gesture, and for one subject) in the training corpus to test the system.

As mention in [16], we use the same parameter setting of Cascaded Adaboost classifier in our recognition system: stage number is 30, the max false alarm F is 0.5, and the min detection rate D is 0.995.

For all five different hand gestures, we use the training corpus to train the five different classifiers.

4.2 Criteria for performance evaluation of the recognition system

The system was evaluated by two criteria: the recognition capability and the computation time.

To evaluate the recognition capability of the system, we used recall and precision rate. In a classification task, the precision and recall for a class are defined as follows:

- Precision = $TP/(TP+FP)$
- Recall = $TP/(TP+FN)$

where:

- TP (True Positive) is the number of items correctly labeled as belonging to the positive class.
- FP (False Positive) is the number of items incorrectly labeled as belonging to the positives class.
- FN (False Negative) is the number of items which were not labeled as belonging to the positive class but should have been.

Precision can be seen as a measure of exactness or fidelity, whereas recall is a measure of completeness. In even simpler terms, a high recall means you haven't missed anything but you may have a lot of useless results to sift through (which would imply low precision). High precision means that everything returned was a relevant result, but you might not have found all the relevant items (which would imply low recall).

To evaluate the computing time of the system we calculated the number of frames that the system can recognize per second.

4.3 Experimental Results

To evaluate the performance of the system, we have conducted two experiments:

- The first experiment aims to evaluate the performance of the system when recognizing the hand gestures of two subjects which have participated in the training system. We call "Dependent subject experiment".
- The second experiment evaluates the performance of the system with hand gestures coming from four new subjects. We call in this case: "Independent subject experiment".

In both experiment, for each subject, we take 500 positive images and 50 negative images of his/her hand gesture to evaluate.

Table 2 shows precision and recall for each class of hand gesture. The average of recall and precision rate of all gestures is 88% for both dependent subject experiments and independent subject experiments. The results show that the proposed gesture set satisfying recognisability criterion. The processing time is about 18 fps on a dual core 2.66 MHz, RAM 2GB PC system. It allows the use of the gesture set real-time video applications.






5. CONCLUSIONS AND FUTURE WORK

This paper studies the behavior of Vietnamese in using of hand gesture in communication with robot. The study has been carried out through a wizard of oz framework. The proposed framework is general and could be used for all other studies aiming finding out other interaction methods (e.g. using speech). Results obtained

from this study are a hand gestures set commonly used in particular by Vietnamese for five different commands that all HRI applications should concern. The results obtained in the two experiments show that the hand gesture set is recognisability by the recognition system. The high values of recall and precision indicate that the proposed five hand gestures are distinct, and the proposed hand gesture automatic recognition system can distinguish easily the designed hand gesture set. It could be conclude that these five hand gestures have been designed satisfying comfortableness and recognisability criteria. It allows stating that this hand gesture set can be reused in the domain without requiring redesigning it. On the other hand, the first results on the performance of proposed system allows us to confirm that our hand gesture automatic recognition system could be use in many factual applications of human - robot interactions.

In the future, we will build and test real human robot interaction applications using this set of hand gesture combining with other modality such as speech. The human - robot interaction in fact always is multi-modalities. Human often use instinctively not only the both speech and gestures but also the face, body in his/her interaction with another. Hence, understanding not only the human's hand gesture, but also the others human modalities could help robot to interact with human in natural way.

Table 2: The recognition results of the both experiments: Independent subject experiment and Dependent subject experiment

Posture	Illustration	Dependent subject experiment		Independent subject experiment	
		Recall	Precision	Recall	Precision
Call		89%	93%	93%	96%
Agree		67%	72%	74%	76%
Disagree		98%	96%	93%	88%
Point		92%	95%	89%	87%
Stop		95%	89%	94%	95%
Mean		88%	89%	88%	88%

6. ACKNOWLEDGMENTS

The research leading to this paper was supported by the National Project ĐTĐL.2009G/42 "Study, design and develop smart robots to exploit multimedia information", under grant 42/2009G/HĐ-ĐTĐL. We would like to thank the project and people involved in this project.

7. REFERENCES

- [1] P. Viola, M. Jones. *Robust Real-time Object Detection*: Cambridge Research Laboratory Technical Report Series; 2001.

- [2] D. J. Sturman, D. Zeltzer. *A Survey of Glove-based Input*. IEEE Computer Graphics 1994;14(1).
- [3] S. Marcel, O. Bernier, J. Viallet, D. Collobert. *Hand Gesture recognition using Input/Output Hidden Markov Models*. FG '00 Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition. Washington, DC, USA: IEEE Computer Society, 2000: 456-.
- [4] V. I. Pavlovic, R. Sharma, T. S. Huang. *Visual interpretation of hand gesture for human-computer interaction: a review*. IEEE Trans on Pattern Analysis and machine Intelligence 1997;19(7):677-95.
- [5] P. Garg, N. Aggarwal, S. Sofat. *Vision Based Hand Gesture Recognition*. World Academy of Science, Engineering and Technology 2009;49:972-7.
- [6] A. C. Downton, H. Drouet. *Image Analysis for Model-Based Sign Language Coding*. Proc Sixth Int'l Conf Image Analysis and Processing, 1991: 637-44.
- [7] J. M. Rehg, T. Kanade. *DigitEyes: vision-based hand tracking for human-computer interaction*. Proceedings of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects. Austin, TX, USA 1994: 16 - 22
- [8] A. J. Heap, D.C. Hogg. *Towards 3-D hand tracking using a deformable model*. In 2nd International Face and Gesture Recognition Conference. Killington, USA, 1996: 140-5.
- [9] B. Stenger, P. R. S. Mendonca, R. Cipolla. *Model-Based 3D Tracking of an Articulated Hand*. In proc British Machine Vision Conference. Manchester, UK, 2001: 63-72, .
- [10] S. Marcel. *Hand Posture Recognition in a Body-Face centered space*. CHI'99 Conference on Human Factors in Computer Systems. Pittsburgh, PA USA, 1999: 15-20.
- [11] <http://www-prima.inrialpes.fr/FGnet/data/10-Gesture/gestures/main.html>.
- [12] Q. Chen, N. D. Georganas, E. M. Petriu. *Hand Gesture Recognition Using Haar-Like Features and a Stochastic Context-Free Grammar*. IEEE Transactions on Instrumentation and Measurement 2008;57(8):1562-71.
- [13] M. Kolsch, M. Turk. *Robust Hand Detection*. International Conference on Automatic Face and Gesture Recognition. Seoul, Korea, 2004: 614-9.
- [14] C. C. Wang, K. C. Wang. *Hand Posture recognition using Adaboost with SIFT for human robot interaction*. Proceedings of the International Conference on Advanced Robotics (ICAR'07). Jeju, Korea, 2008.
- [15] <http://www.massey.ac.nz/fdadgost/xview.php?page=farhad>.
- [16] Z. Cernekova, N. Nikolaidis, I. Pitas. *Single Camera pointing gesture recognition using spatial features and support vector machines*. European Signal Processing Conference (EUSIPCO-2007). Poznan, Poland, 2007: 130-4.
- [17] G. Rigoll, A. Kosmala, S. Eickeler. *High Performance Real-Time Gesture Recognition using Hidden Markov Models*. In International Gesture Workshop Bielefeld. Bielefeld, Germany: Springer-Verlag, 1998: 69-80.
- [18] A. Ramamoorthya, N. Vaswania, S. Chaudhurya, S. Banerjeeb. *Recognition of dynamic hand gestures*. Pattern Recognition 2003;36:2069 - 81.
- [19] A. Corradini. *Real-Time Gesture Recognition by Means of Hybrid Recognizers*. GW '01 Revised Papers from the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction. London, UK, 2001: 34-46.