Team eR@sers 2012 in the @Home League
Team Description Paper

Hiroyuki Okada, Takashi Omori, Norifumi Watanabe and Takayuki Shimotomai
Naoto Iwahashi and Komei Sugiura
Takayuki Nagai and Tomoaki Nakamura

Address
6-1-1 Tamagawa-gakuen,Machida-shi, TOKYO 194-8610, JAPAN
E-mail
h.okada@eng.tamagawa.ac.jp

Abstract. Team eR@sers has taken part in RoboCup@Home since 2008. 2008 was the first year of the eR@sers. The eR@sers achieved a first place at the Japan Open Competition 2008, 2009, 2010, 2011 and first place RoboCup 2008 ,2010 and second place RoboCup 2009. We have imploved the ability of robots with various techniques, which are going to be applied to other robot systems or social IT systems. We introduce them and our latest research briefly in this description.

1 Team History

The Japanese Robot Team eR@sers(erasers) is the result of a joint effort of three-Japanese research groups:

Tamagawa University:the group of the College of Enginerring at Tamagawa University of Tokyo in Japan that is involved in the world championship RoboCup competitions in Four-legged since 2005. At the RoboCup 2006 Bremen , the team, FC Twaves, got the best results(best 16) in the team that participated from Japan. The members at Tamagawa University are interested in a compliant human-machine interaction architecture that is based on the machine intention recognition of the human. This work is motivated by the desire to minimize the need for classical direct human-machine interface and communication.

National Institute of Information and Communications Technology (NICT) :the group of NICT in Japan that is involved in the research of the computational mechanism which enable robots to learn the communication by language and actions through natural interaction with human.

The University of Electro-Communications:the group of the Department of Electronic Engineering, The University of Electro-Communications in Japan that takes a role in the visual processing system in this team. The main avenue of the research group is to pursuit the real human-like intelligence, including multimodal information processing.

2008 was the first year of the eR@sers. The eR@sers achieved a first place at the Japan Open Competition. And in wonderful, In RoboCup 2008 in Suzhou, China we got a first place. In 2009, the eR@sers achieved a first place at the Japan Open Competition. And in RoboCup 2009 in Graz, Austria we got a second place. In 2010, the eR@sers achieved a first place at the Japan Open Competition. And in RoboCup 2010 in Singapore we got a world championship agein.
This paper presents the main development efforts of the team in 2012.

2 Software architecture

2.1 Framework

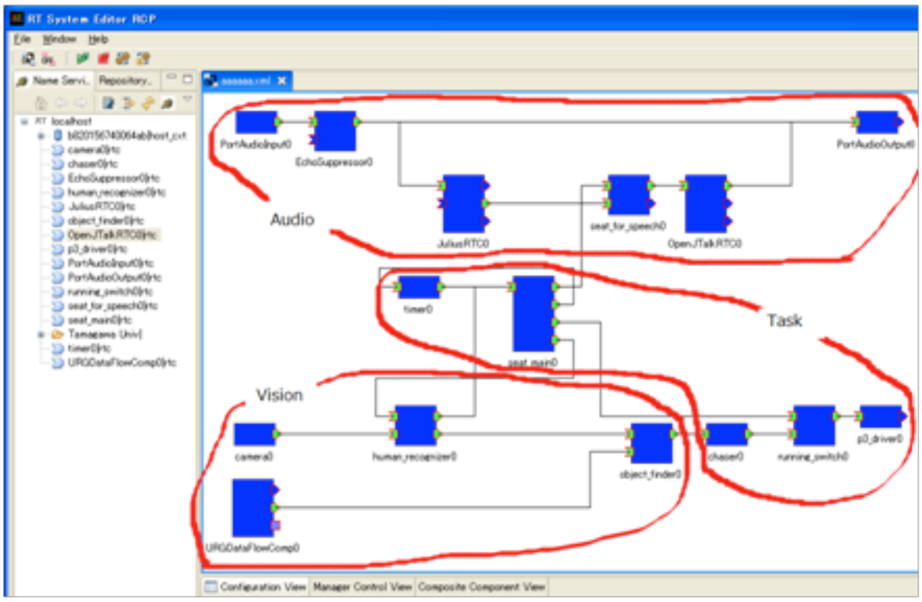


Fig. 1 eR@ser's software architecture

Fig.1. shows eR@sers software framework. In order to work on different parts of the robot system by different groups in different locations, we have used open source robotic technology middleware(OpenRTM) which is developed and distributed by Japan's National Institute of Advanced Industrial Science and Technology. Both entire robots and various usable robotic functional elements are known as RT (Robot Technology/Robotic Technology). RT-Middleware is a platform to modularise and integrate a variety of robotic functional elements as software.

In RT-Middleware, an RT functional element is modularized as a software component called an RT-Component, and a robot is implemented by combining RT-Components on RT-Middleware. RT-Components have data ports and service ports to communicate with other components, and you can integrate various components easily by standardizing these interface specifications. All RT-Components have a common state machine inside, and you can manage a large number of components in higher-level application programs through integration. Moreover, since RT-Components have standard interfaces to alter internal parameters, you can reuse them without recompiling.

All components are connected through the "naming server" with GigE and have subscription information that describes required information for the processing in each component. The "Vision Module" is responsible for object/face detection and recognition tasks. The "Audio Module" provides speech recognition and text to speech functionalities. The "Task Module" works as a controller for each scenario of the @Home competitions. We have seven task modules in total, which are switched in accordance with the scenario.

2.2 Vision system.

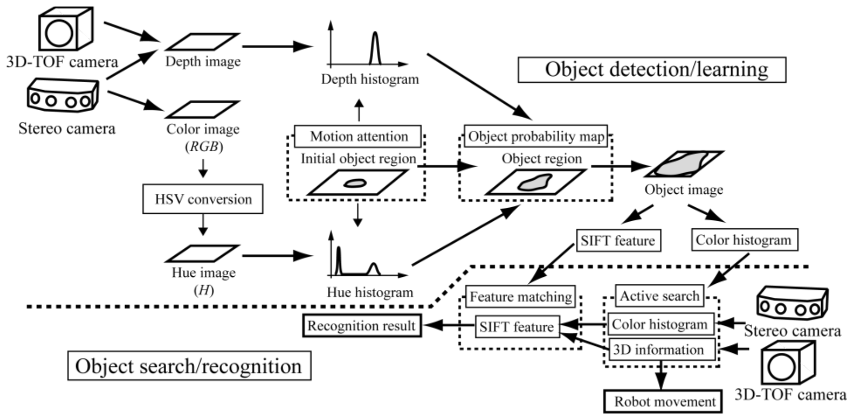


Fig. 2 Processing in the vision module

Two color images are grabbed simultaneously with 320x240 pixels resolution at 10 frames per second. These images are used for the stereo processing to obtain depth information. Both depth and color information are utilized for object/face detection, learning and recognition tasks.

We also use 3D-TOF (time of flight) camera using NIR to capture very accurate 3D information. All of these cameras are calibrated each other so that the corresponding pixels can be found easily.

Object detection and learning. The vision system needs to extract objects from visual scene to learn objects in real time. A problem in extracting multiple objects from unlabeled images is that it is impossible to tell which part of the image corresponds to each individual object, and which part is irrelevant clutter. This system solves the problem using a priori knowledge that the segments with synchronous motion are parts of an identical object. Hence the motion detector is first employed in the object detection subsystem.

As shown in Fig.2. the motion detector extracts the initial object region at first. Then, the object information such as color (hue) and depth is taken from the region. In particular, hue and depth histograms are taken from the region and normalized. Since these two histograms can be considered as probability density functions of the target object, the object probability map of each component at each pixel location can be easily obtained. The weighted sum of these two object probability maps results in the object probability map. The map is binarized followed by the connected component analysis, and then final object mask is obtained.

In the learning phase, object images are simply collected, and then histogram and SIFT features are extracted. This information is used for the object search and recognition. It should be noted that the SIFT features are normalized using accurate 3D information in order to cope with affine distortion.

Object search/recognition. In order to recognize objects, we take two-step strategy. At first, the entire input image is scanned with calculating color histogram of a rectangular area in the image. Each histogram is collated with the histogram of the target object, and the area that gives the highest similarity is output as a position where the target object is detected. When the detected object is far away from the robot, the robot moves toward the object. When the object is close enough, a feature point matching is carried out in the detected region. We use the object matching algorithm that consists of two main stages: the selection of local scale invariant feature(SIFT), and the matching of constellations of such key points.

Face detection/recognition. The face detection algorithm utilizes the cascade of boosted classifiers with haar like features. We also use skin tone detection and 3D information to check the detected face region, which prevents from detecting non-face region. In the detected face region, eyes and mouth are detected. These three points are used to normalize the size of the face region. As for face recognition, Embedded HMM (EHMM) is employed. EHMMs consist of a set of super states (super-states) and normal states (embedded-states), which are embedded in super-states. The super-states model the vertical direction, while the embedded states model the horizontal one.
EHMMs can absorb the changes of appearance to some extent thanks to its elastic matching property: nevertheless representing each face with a single model is not feasible. Multiple models are incorporated to cope with this problem. The system generates multiple EHMMs by observing a face from different view points and links them together. Since the vision module always checks the object's identity, the system can distinguish the face change from the change of appearance involving the rotation. This function makes it possible to connect multiple EHMMs that are associated with a single person.

Additional functionalities. We have further advanced functionalities concerning object recognition. In general, image matching based object recognition cannot deal with unseen objects. We, as human beings, solve this problem by categorization. So we have developed a robot that can categorize objects in an unsupervised manner using multimodal information. The algorithm is based on graphical model, which we call object concept model. The system can infer the function of the unseen object through its category. We strongly believe that these categorization abilities play a central role for the robot intelligence when robots work in the real home environment.

2.3 Speech Processing

Our speech technologies include noise robust speech recognition, high quality text-to-speech conversion, and many-to-one voice conversion.

Speech Recognition. For speech recognition, HMM-based speech recognition software developed at Advanced Telecommunication Research Laboratories (ATR) is used. Speech recognition system should be robust enough to recognize speech in noisy environments with various speaking styles. For acoustic modeling, the most important problem to solve is how to efficiently capture contextual and temporal variations in training speech and properly model them with fewer parameters. MDL-SSS creates speaker-independent models by data-driven clustering with contextual information based on minimul description length (MDL). This leads to high performance in large vocabularly continuous speech recognition. For front-end processing of speech recognition, adaptive noise reduction technique achieves the robustness of speech recognition against background noise. Gaussian mixure models for both speech and noise are used to form Wiener filter, and they are adapted according to input acoustic signal. To estimate the non-stationary noise sequences a particle filter-based sequential noise estimation method is used. In the proposed method, the particle filter is defined by a dynamical system based on Polyak averaging and feedback. A switching dynamical system is also applied into the particle filter to cope with the state transition characteristics of non-stationary noise. The method improves speech recognition accuracy even if noise is non-stationary.

Text-to-Speech Conversion. The concatenative text-to-speech conversion system, developed at ATR and named XIMIRA, is used. XIMERA is based on corpus-based technologies. The prominent features of XIMERA are as follows:

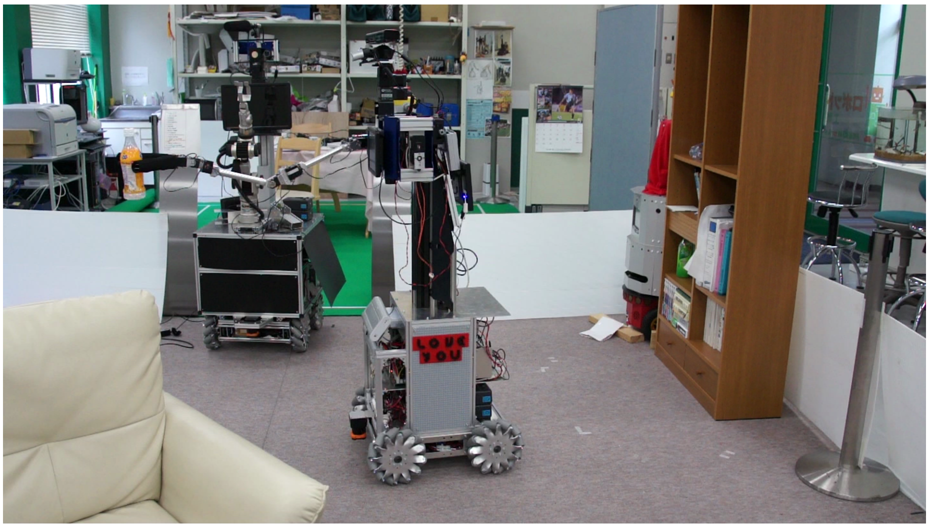
- large corpora (a 110-hours corpus of a Japanese male, a 60-hours corpus of a Japanese female, a 20-hours corpus of a Chinese female, and a 16-hours corpus of an English male)
- HMM-based generation of prosodic parameters;
- cost function for segment selection optimized based on perceptual experiments.

The result of evaluation test showed that XIMERA outperformed commercial TTS systems currently available in the market.

Voice Conversion. Voice conversion is a technology for converting a certain speaker's voice into another speaker's voice. Many-to-one voice conversion realizes the conversion from arbitrary user's voice as source to a target speaker's one. In our robot, arbitrary user's input voice is converted into robot's voice. Eigenvoice conversion is applied in the voice conversion method. Using multiple parallel data sets consisting of utterance pairs of the user and multiple pre-stored speakers, an eigenvoice Gaussian mixture model (EV-GMM) is trained in advance. Unsupervised adaptation of the EV-GMM is available to construct the conversion model for arbitrary source speakers in many-to-one VC using only a small amount of their speech data.

3 Hardware

This year we take two robots "tam@home"(Fig.3) and "DiGORO"(Fig.4) to the competition.



The main hardware components of "tam@home" are:

- Omni-Directional moving base by mecanum wheel
- Hokuyo UTM-30LX laser scanners
- 6D Arm
- Kinect Camera
- Pan-Tilt Unit PTU-D46 (Directed Perception)
- ShotGun-Microphone(Sanken CS-3)

Fig. 4 DiGORO

The main hardware components of "DiGORO" are:

- Laser range finders (HOKYO URG-04X, UTM-30LX)
- Upper Body Humanoid Robot HIRO(KAWADA Industries, JAPAN)
- Four on board PCs (Core2Duo, Atom)
- Two CCD cameras
- 3D TOF camera SwissRanger SR4000 (MESA Imaging)
- Pan-Tilt Unit PTU-D46 (Directed Perception)
- ShotGun-Microphone(Sanken CS-3)

4 The contents of the web site

Our relevant publications, technical reports, as well as videos and pictures are available in :

Official website:
https://sites.google.com/site/erasers2050/home/
http://robot.lab.tamagawa.ac.jp/robocup/index.php?athome

Photos and Videos of the robot
https://sites.google.com/site/erasers2050/photos-movies/

5 Reference

1. Noriaki ANDO, Shinji KURIHARA, Geoffrey BIGGS, Takeshi SAKAMOTO, Hiroyuki NAKAMOTO, "Software Deployment Infrastructure for Component Based RT-Systems", Journal of Robotics and Mechatronics, Vol.23, No.3, pp.350-359(2011)
2. Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi, "Multimodal Categorization by Hierarchical Dirichlet Process", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp1520-1525, California, Sep.2011
3. Takaya Araki, Tomoaki Nakamura, Takayuki Nagai, Kotaro Funakoshi, Mikio Nakano, and Naoto Iwahashi, "Autonomous Acquisition of Multimodal Information for On-line Object Concept Formation by Robots", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.1540-1547, California, Sep.2011
4. Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi, "Bag of Multimodal LDA Models for Concept Formation", IEEE International Conference on Robotics and Automation, pp.6233-6238, Shanghai, May2011
5. Tomoaki Nakamura, Takaya Araki, Takayuki Nagai, Naoto Iwahashi, "Grounding of Word Meanings in LDA-Based Multimodal Concepts ", Advanced Robotics, 25, pp.2189-2206(2011)
6. Tomoaki Nakamura, Komei Sugiura, Takayuki Nagai, Naoto Iwahashi, Tomoki Toda, Hiroyuki Okada, and Takashi Omori, "Learning Novel Objects for Extended Mobile Manipulation", Journal of Intelligent and Robotic Systems, pp.1-18, Jul. (2011)
7. Muhammad Attamimi,Attamini Mizutani,Tomoaki Nakamura, Komei Sugiura,Takayuki Nagai,Naoto Iwahashi,Hiroyuki Okada and Takashi Omori: Learning novel objects using out-of-vocabulary word segmentation and object extraction for home assistant robots,Proc.of ICRA2010, pp.745--750,(2010).
8. Nakamura, T., Nagai, T., Iwahashi, N.:Multimodal Object Categorization by a Robot.Proc. Int. Conf. Intelligent Robots and Systems, (2007).
9. S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J-S, Zhang, H. Yamamoto, E. Sumita and S. Yamamoto: The ATR Multilingual Speech-to-Speech Translation System, IEEE Trans. on Audio, Speech, and Language Processing, Vol.14, No.2, pp.365--376, 2006
10. M. Fujimoto and S. Nakamura: Sequential Non-Stationary Noise Tracking Using Particle Filtering with Switching Dynamical System, Proc. Int. Conf. Audio Speech, and Signal Processing, Vol.I, pp.769-772, 2006.