Paper:

Visual Perception for a Partner Robot Based on Computational Intelligence

Indra Adji Sulistijono[*,**], **and Naoyuki Kubota**[***,****]

[*]Dept. of Mechanical Engineering, Graduate School of Engineering, Tokyo Metropolitan University
1-1 Minami-Osawa, Hachioji, Tokyo 192-0397, Japan
E-mail: indra-adji@ed.tmu.ac.jp
[**]Electronics Engineering Polytechnic Institute of Surabaya - ITS (EEPIS-ITS)
Kampus ITS Sukolilo, Surabaya 60111, Indonesia
[***]Dept. of System Design, Tokyo Metropolitan University
1-1 Minami-Osawa, Hachioji, Tokyo 192-0397, Japan
E-mail: kubota@comp.metro-u.ac.jp
[****]PRESTO, Japan Science and Technology Agency (JST)
[Received March 10, 2005; accepted June 10, 2005]

We propose computational intelligence for partner robot perception in which the robot requires the capability of visual perception to interact with human beings. Basically, robots should conduct moving object extraction, clustering, and classification for visual perception used in interactions with human beings. We propose total human visual tracking by long-term memory, *k*-means, self-organizing map, and a fuzzy controller is used for movement output. Experimental results show that the partner robot can conduct the human visual tracking.

Keywords: partner robot, computational intelligence, visual perception, *k*-means, self-organizing map

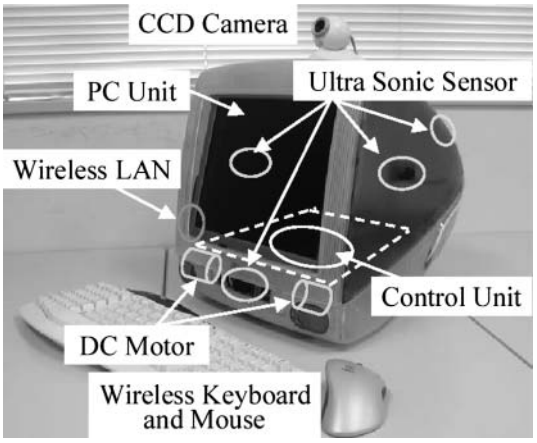

Fig. 1. Partner robot MOBiMac.

1. Introduction

Among different next-generation human-friendly robots, next-generation personal robots are designed to entertain or assist human beings and must be able to recognize human beings, interact with human beings in natural communication, and learn to interact with human beings. Visual perception is especially important because vision may include much information for interaction with human beings.

In visual tracking research, the problem of human visual tracking has received great attention. Image-based human tracking may play an important role in next-generation surveillance and human computer interfaces. Determining the positioning of the body in a video stream is difficult because of significant variations in appearance throughout the sequence [18, 19]. The robot may specify human intention using a built map, because a task depends on environmental conditions. Robots should learn environmental maps and human gestures or postures to communicate with human beings. In previous studies, different image processing methods for robotic visual perception have been proposed such as differential filters, moving-object detection, and pattern recognition [15–17]. For pattern recognition, robots require patterns or templates, but human-friendly robots cannot know patterns or templates beforehand, so robots must learn patterns or templates through interaction with human beings. For this, we propose visual tracking in which robot extracts a human being from an image taken by a built-in CCD camera. Since we assume a human being moves, a moving object becomes a candidate. Long-term memory (LTM) is used for this extraction. A differential filter detects a moving object and a color combination pattern is extracted by *k*-means. The robot classifies the detected human being by using self-organizing map (SOM) based on the color pattern, then moves toward the detected human being.

This paper is organized as follows: Section 2 explains vision-based partner robots, and proposes human clustering based on visual perception. Section 3 presents experimental results, and Section 4 summarizes conclusions.

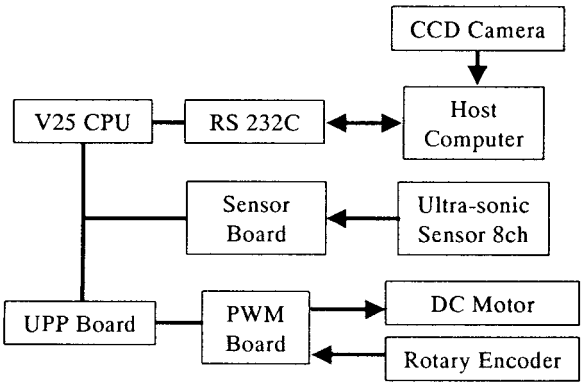

Fig. 2. MOBiMac control architecture.

2. Visual Perception

2.1. A Partner Robot

We developed the partner robot MOBiMac (**Fig.1**) to be used as a personal computer and partner. Two CPUs are used for PC and robotic behaviors. The robot has two servomotors, four ultrasonic sensors, and a CCD camera (**Fig.2**). The ultrasonic sensor measures 2000mm, so the robot can capture different behavior, such as collision avoidance, human approach, and visual tracking.

The robot takes an image from the CCD camera to extract a human being and, if it detects them, it extracts the related color patterns, conducting visual tracking (**Fig.3**), detailed below.

2.2. Differential Filter

To detect a human being, we use a psychology field [4–9]. Centering on selective attention research leading to figure-background organization. Visual perception is organized into a central object, called a figure and its blurred surroundings, called a background. Our visual system operates flexibly and adaptively to perceive the environment using bottom up and top down. Bottom-up processing depends directly on external stimuli, while top-down processing is influenced by expectations, stored knowledge, context, and etc.

We use expectation-based LTM to detect a human being considered to be moving. The robot has LTM used as background image $GI(t)$ consisting of $\mathbf{g}_i (i = 1, 2, \ldots, l)$. Temporal image $TI(t)$ is generated using differences between the image at t and the $GI(t-1)$. Each pixel has belief value b_i satisfying $0.01 < b_i < 0.99 (i = 1, 2, \ldots, l)$, and this is updated as follows,

$$\begin{cases} b_i \leftarrow \alpha \cdot b_i & \text{if } d_i > \gamma \\ b_i \leftarrow \alpha^{-1} \cdot b_i & \text{otherwise} \end{cases} \quad \ldots\ldots\ldots \quad (1)$$

where d_i is the maximal difference in i-th pixel between current and background image, α is a discount, l is the number of pixels and γ is the different value threshold. If the pixel color change is small, belief becomes large.

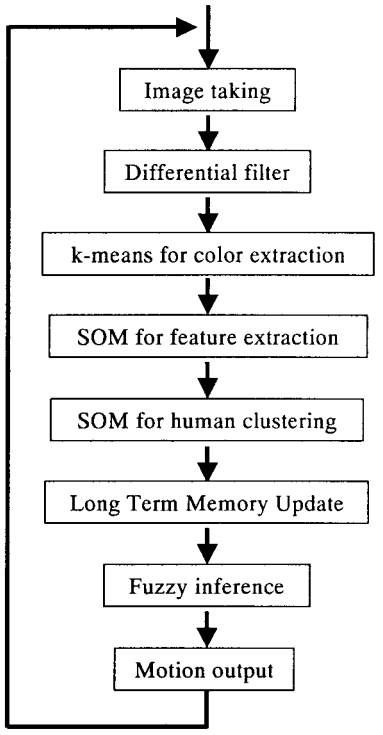

Fig. 3. Visual tracking architecture.

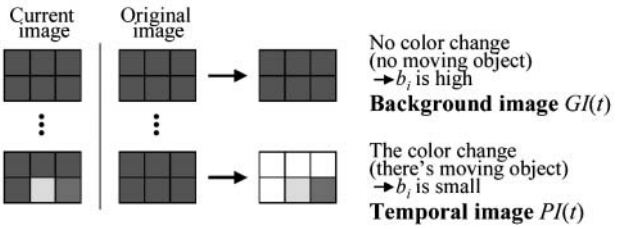

Fig. 4. LTM.

Each $GI(t)$ pixel is updated as follows,

$$\mathbf{g}_i \leftarrow b_i \cdot \mathbf{g}_i + (1 - b_i) \cdot \mathbf{p}_i \quad \ldots\ldots\ldots \quad (2)$$

where $\mathbf{p}_i$ is the color RGB of the i-th pixel at image t. If there are no moving objects in front of the robot, a background image with high belief is obtained, so the robot understands that a moving object or a human being is in front of the robot using $TI(t)$. **Fig.4** shows LTM.

2.3. Color Extraction

Our proposed learning consists of three clustering stages, i.e., color clustering by k-means, feature extraction by SOM, and human clustering by SOM.

Different unsupervised learning methods have been proposed [11, 12]. In batch learning, a set of all data is required beforehand, but incremental learning updates design parameters when new data is given to the learning system. A color pattern of a human being is extracted