Visual Servoing for a User’s Mouth with Effective Intention Reading in a Wheelchair-based Robotic Arm

Won-Kyung Song†, Dae-Jin Kim†, Jong-Sung Kim†, and Zeungnam Bien†
† EECS, KAIST, 373-1 Kusong-Dong, Yusong-Gu, Taejon, 305-701, KOREA
† VR Center, ETRI, 161 Kajong-Dong, Yusong-Gu, Taejon, 305-350, KOREA
zbien@ee.kaist.ac.kr

Abstract
There exists the co-operative activity between a human being and rehabilitation robots because the human operates rehabilitation robots in the same environment and has the benefit of rehabilitation robots, such as manipulatory or mobile functions. Intention reading is one of the essential functions of human-friendly rehabilitation robots in order to promise the comfort and safety for all who need them. First of all, the overall structure of a new wheelchair-based robotic arm system, KARES II, and its human-robot interaction technologies are presented. Among those technologies, we concentrate on visual servoing that allows this robotic arm to operate autonomously via visual feedback. Effective intention reading, such as recognizing the positive and negative meaning of the user, is performed on the basis of changes of the facial expression around lips that is strongly related to the user’s intention while this robotic arm provides the user with a beverage. For the efficient visual information processing, log-polar mapped images are used to control the stereo camera head that is located in the end-effector of the robotic arm. The visual servoing with effective intention reading is successfully applied to serve a beverage for the user.

1 Introduction
Wheelchair-based robotic systems are mainly used to assist the elderly and the disabled who have handicaps in sensory and motor functions in limbs. Such a system consists of a powered wheelchair and a robotic arm, and has not only a mobile capability through the wheelchair but also a manipulatory function via the robotic arm, and thus makes possible the coexistence of a user and a robot in the same environment. In this case, the user needs to interact with the robotic arm in comfortable and safe ways. However, it has been reported that many difficulties exist in human-robot interactions in existing rehabilitation robots. For example, manual control of the robotic arm takes a high cognitive load on the user part while physically disabled persons may have difficulties in operating joysticks dexterously or pushing buttons for delicate movements [4]. In addition, MANUS evaluation users reported that the most difficult thing using rehabilitation robots is too many commands for a small adjustment and too many functions to keep in mind at the beginning [4]. Therefore, human-friendly human-robot interaction is one of essential techniques in a wheelchair-based robotic arm.

In this paper, we consider the wheelchair-based robotic system, KARES (KAIST Rehabilitation Engineering Service system) II, which we are developing as a service robotic system for the disabled and the elderly, and discuss its human-robot interaction techniques (Fig. 1). Among human-robot interaction techniques, visual servoing is dealt with as a major topic.

Figure 1: The wheelchair-based robotic arm and its human-robot interaction technologies.
Visual servoing actually contributes autonomous capability to a robotic arm. However, the user’s intention can be often disregarded in visual servoing, and those situations may menace the user. So, intention reading capability is desirable to execute visual servoing in order that the robotic arm may move safely and comfortably. The facial images of a user are adopted as the cue to read the simple intention, such as the positive and negative meaning of the user. Log-polar mapped images are efficiently used to control the vergence angle of the binocular camera head and to obtain the depth of a target in an eye-in-hand camera configuration. Moreover, conventional Cartesian images and log-polar mapped images are simultaneously applied to process visual information in accordance with visual tasks.

In Section 2, we present the new wheelchair-based robotic arm system. In Section 3, visual servoing for a user’s mouth with intention reading is proposed. In Section 4, experimental results and discussions are described. Finally, concluding remarks are given in Section 5.

2 New Wheelchair-based Robotic Arm System with Interaction Technologies

Major users of a wheelchair-based robotic arm system are the elderly and the disabled who have handicaps in sensory and motor functions in limbs. Sometimes, the elderly who have hemiplegia can use a wheelchair-based robotic arm as needed. Specially, people with spinal cord injury are candidates for users among the disabled [9]. For instance, people with C-2 to C-4 lesion only able to use eye, facial, mouth, and head movements. They do not have use of their hands, arms, or legs. It is very difficult to give independent capability to those people via classical rehabilitation treatments. Therefore, those people need special engineering systems, such as rehabilitation robots, to guarantee the independence with improvement of life quality.

Tasks of a wheelchair-based robotic arm have been specified in accordance with surveys of factories in which the disabled work, and some interviews with disabled people with spinal cord injury, such as C-4 and C-5 lesion. We consider the necessary activities of daily living via a wheelchair-based robotic arm. The eight principally defined tasks of a robotic arm are as follows [1]: (1) hitting an abdominal region, (2) serving a meal, (3) serving a beverage, (4) shaving, (5) wiping a face with a wet towel, (6) picking the object on a floor, (7) turning on/off a switch, and (8) opening/closing a door.

Two factors are essential in intelligent human-robot interaction technology; one is intention reading of the user, and the other is autonomous capability of the robot. First, intention reading allows the user to command to the robotic arm possibly by using the bio-signal [5], the haptic suit [7], or the “eye-mouse” device [13] which utilizes eye movements. Second, the autonomous capability in controlling the robotic arm is needed to realize the user’s commands as well as to accommodate sensory feedback. Autonomous operations for specified tasks make the user to handle rehabilitation robots easily. A typical example is visual servoing-based or compliance-based control of a robotic arm.

A user’s intention can be inferred from various sources such as: sound, voice, hand movements, body movements, bio-signals of muscles, eye movements, facial expression, and so on. We are developing several human-robot interaction techniques including a haptic suit, bio-signal recognition, and an eye-mouse [1]. Appropriate techniques are selected, depending on the user’s degree of disability, and are integrated in synergistic or complementary ways.

Fig. 2 shows the input/output relation of subsystems for human-robot interaction. For example, the task to serve a meal for a user is conducted as follows: First, the eye-mouse reads the user’s intention, and transfers processed information, such as the position of a desired food, to the visual servoing module. Next,
the visual servoing module controls the robotic arm to
spoon up the food and moves it to the user's mouth.
Consequently, the intention reading of the user and
autonomous capability of the robotic arm should well
be coordinated to perform the task successfully.

3 Visual Servoing for a User’s Mouth with Intention Reading

Among various human-robot interaction techniques,
we concentrate on the visual servoing that is intro-
duced to control the pose of the end-effector of a
robotic arm relative to a target using visual feedback.
Visual servoing in a wheelchair-based robotic arm usu-
ally renders many difficulties: To mention a few, it
is difficult to process visual information in real-time,
to be robust for varying illumination, and to process
data under vibration of a robotic arm [1]. A biologi-
cally inspired vision system is selected for a break-
through of the above difficulties in conventional visual
servoing. A major idea of the biologically inspired vi-
sion system is to increase attention on the important
parts of images. This strategy can be implemented
by non-uniform sampling in images. We call it space
variant vision [10]. The log-polar mapping is one of
the well-known space variant vision techniques, and
has following advantages: maintaining high resolution
on the part of interest, a wide field-of-view, and fast
processing speed. The log-polar mapping of an image
from conventional Cartesian coordinates \((x,y)\) to log-
polar coordinates \((\xi, \gamma)\) is defined as Eq. (1) and Eq.
(2) [2].

\[
\xi = \log_b(\rho/\rho_0), \quad \rho = \sqrt{x^2 + y^2} \tag{1}
\]

\[
\gamma = q\eta = q \arctan(y/x) \tag{2}
\]

The eye-in-hand configuration takes images from the
standpoint of the robot, and shows various advan-
tages for a wheelchair-based robotic arm [1]: easy
camera calibration, less vibration effect, various im-
age sequences for orientation and resolution, and so
forth. Comparing with monocular vision, binocular
vision has the following advantages [1]: easy to get
depth information through vergence movements [2], an
increasing user’s safety, and easy to mimicking visual
process of animals. We have implemented a binocular
color camera head [12], which is small and lightweight,
for an eye-in-hand configuration and have done the
depth estimation using vergence control.

We decompose those eight tasks defined in Section 2
into several subtasks to find the valuable role in vi-
sual servoing. According to the analysis of the tasks
[1], the most important subtasks for visual servoing
are “grasping an object” and “approaching a user’s
mouth”. Therefore, if we perform a scenario, such as
“grasping a cup on a table” and “bring the cup to the
user’s mouth”, other tasks should be covered by the
extension of the scenario. Specially, the user’s mouth
is an interesting target because the mouth is non-
stationary and the deformable characteristics in “ap-
proaching the user’s mouth”. Also, the user’s mouth
may express a satisfaction degree of the serving, and
thus this fact can be used to read the user’s intention.
For example, opening the mouth means that the user
wants to drink a beverage, on the other hand, clos-
ing the mouth means that the user does not want to
drink any more. If a robotic arm system recognizes
the mouth shape during the serving a beverage, then
positive or negative meaning of the user can be ob-
tained. Although the user’s distinguishable meanings
are only two kinds, this information is very effective
in actual tasks.

In order to perform “approaching the user’s mouth”,
following steps of visual information processing are de-
sired: (1) detecting the mouth on the face, (2) center-
ing the mouth on the face, (3) tracking the mouth on
the face, and (4) recognizing changes of the mouth
shape. According to the comparison analysis for log-
polar mapped images and conventional Cartesian im-
ages, it is found that log-polar mapped images are
very efficient in most of the detecting, centering, and
tracking of the target through area based matching
techniques because the small-sized images and higher
weights in the central region of images. However, those
images are not suitable to recognize changes of the
mouth shape because changes of the shape are very
sensitive to the centering. Therefore, changes of the
mouth shape are recognized in conventional Cartesian
images after finding the mouth region via log-polar
mapped images.

It is assumed that the hair and jaw are located in the
upper and lower part in images, respectively. Open-
ing/closing the mouth is easy to check the visual in-
formation along the vertical line from the jaw to the
hair in images under this assumption.

3.1 Detecting the mouth on a face

In order to detect the mouth on a user’s face, the lo-
cation of the user’s face should be found in images. In
this work, we have adopted HLAC (Higher-order Lo-
cal Autocorrelation) [6] in log-polar mapped images with the I11213 color space. FCM (Fuzzy C-Means Clustering) is added to the face discrimination process, and thus we can deal with many uncertainties in that decision process. According to analyses of the HLC based detection, smaller log-polar mapped images have higher separation capability measured by a separation index \( SI \) in Eq. (3) because of reduced high frequency components. That is, the higher recognition rate can be achieved through only tiny-sized images. Thus, it is very effective to make a real-time working system.

\[
SI = \frac{\sum_{i=0}^{N_f-1} ||y^{(i)} - \bar{y}_n||}{\sum_{i=0}^{N_n-1} ||y^{(i)} - \bar{y}_n||}
\]  

where, \( y^{(i)} \) is the i-th face vector, \( N_f \) is the number of the face vectors, \( y^{(j)} \) is the j-th non-face vector, \( \bar{y}_n \) is the mean value of non-face vectors, and \( N_n \) is the number of the non-face vectors.

Next, the facial region should be detected to extract the mouth. Here, we also adapts the I11213 color space for the effective facial region detection. These color bands are expressed as \( I_1 = (R + G + B)/3 \), \( I_2 = R - B \), and \( I_3 = (2G - (R + B))/2 \). Assumptions for this process are as follows:

- A near frontal face,
- Natural background (i.e., not necessary to make a simple background, but a different background with facial color), and
- Under the fluorescent lamp (i.e., common laboratory environment).

First, given facial images are converted to the I11213 color space. Second, conventional image processing techniques such as binarization, logical OR operation, and morphological operation are used to choose the face candidate region. At last, a projection-based approach is used to extract the facial region with the mouth.

### 3.2 Centering and tracking the mouth on a face

First of all, the segmentation of a facial region in color images is performed in order to center the target in the central region of images or to track the target. It is difficult to obtain the region of a user’s mouth through the color information because the lips color is similar to the facial color and the lips shape is deformable. However, the color information of the lips can be used to segment when the mouth is so near the camera. Thus, the overall facial region is extracted by R-B (or HSL) color space, and then mouth location can be roughly estimated.

Next, the image centering, which is to locate the detected mouth on the central region of images, is required to track the target. Then the target should be tracked through the robotic arm control because the user’s mouth on the face is frequently moved in accordance with the facial or body movements. Log-polar images are suitable for centering and tracking due to small-sized images, higher weighting in the center of the field-of-view, and rotation/scale invariance via polar topology. We normally use the Weiman’s algorithm [14] to center the target. In general, the face corresponds to an intermediate range target or a close range target. Intermediate range targets are those that subtend a substantial number of pixels, but may not overlap the center of the field of view in log-polar mapped images, and the camera position can be moved to the direction that enlarges pixel counts of the target, that is, diminishes the position of the center of gravity of the target area along \( \gamma \)-axis. When the camera closes to the target, the target looks large in the field of view of log-polar mapped images, and is nearly centered. In this case, the maximum bulge (the maximum value \( \gamma \) of the target) is the feature to control the camera. According to the above strategy, we can obtain the control commands for the camera in the log-polar coordinates, and those commands can be changed to the conventional Cartesian coordinates. Using the Jacobian matrix for Eq. (1) and Eq. (2), the relation between the Cartesian coordinates \((x, y)\) and the log-polar coordinates \((\xi, \gamma)\) is represented as follows:

\[
\begin{bmatrix}
\delta x \\
\delta y
\end{bmatrix} = \rho \begin{bmatrix}
\log(k) \cos(\eta) & -\sin(\eta)/q \\
\log(k) \sin(\eta) & \cos(\eta)/q
\end{bmatrix} \begin{bmatrix}
\delta \xi \\
\delta \gamma
\end{bmatrix}
\]

We have estimated approximated depth information of the target through the vergence control or the size of the segmented target. This depth information is used to control the position of the camera on the end-effector of a robotic arm.

### 3.3 Recognizing changes of the mouth shape

In general, conventional Cartesian images are more effective than log-polar mapped images to recognize the mouth shape exactly. Faces or facial elements are efficiently found through log-polar mapped images [3].
However, small-sized log-polar mapped images may have insufficient information to recognize the shape such as mouth openness. Thus, a window is selected in log-polar mapped images, and then the window-based approach is used to recognize the mouth shape in Cartesian images.

In order to recognize changes of the mouth shape in conventional Cartesian images, we adopt Gabor filter banks [8] and a fuzzy logic-based decision maker that can handle uncertainties in real situations and can extend the whole intention reading scheme as humans do. First, we select the linguistic variables for the mouth openness. One linguistic variable is for the open status of the mouth, and the other variable is for the closed status of the mouth. Second, fuzzy rule bases are constructed, such as Rule #1: If MOUTH OPENNESS is LOW, then INTENTION is NEGATIVE. Rule #2: If MOUTH OPENNESS is HIGH, then INTENTION is POSITIVE. Rule #1 includes turning the user’s head. Finally, effective features are extracted from given facial images. Three geometrical features are the height ratio of the mouth to the face, the area ratio of the mouth to the face, and the response of Gabor filter banks. Humans can recognize the degree of the mouth openness from not only global features such as the height and the area ratio of the mouth to the opponent’s face but also local features such as the shape of the mouth and other deformed facial parts. The Gabor features are formulated as follows. If the size of an extracted mouth image is $W \times H$, then the response of the Gabor filtered image is expressed as $I_G (i, j)$. By the vertical projection of Gabor-filtered images, we can reduce the amount of information. In this case, the vertical projected values $y_{proj}(j)$ can be expressed as Eq. (5).

$$y_{proj}(j) = \sum_{i=0}^{W-1} I_G(i, j), \ 0 \leq j < H \quad (5)$$

To reduce the offset of $y_{proj}(j)$, the absolute values of the derivative of the projected values $dy_{proj}(j)$ are given by Eq. (6).

$$dy_{proj}(j) = |y_{proj}(j) - y_{proj}(j-1)|, \ 0 \leq j < H \quad (6)$$

In addition, to normalize $dy_{proj}(j)$ with respect to the size of a mouth region and the different intensity of images, Gaussian weighted sum is applied. Thus, the Gabor feature $f_G$ is given by Eq. (7).

$$f_G = \sum_{j=1}^{H-1} \frac{w_G(j)dy_{proj}(j)}{\sum_{j=1}^{H-1} w_G(j)} \quad (7)$$

where, $w_G(i)$ means the Gaussian weights.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Success ratio</th>
<th>Positive false ratio</th>
<th>Negative false ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>94%</td>
<td>12%</td>
<td>0%</td>
</tr>
<tr>
<td>0.6</td>
<td>90%</td>
<td>8%</td>
<td>12%</td>
</tr>
<tr>
<td>0.7</td>
<td>76%</td>
<td>0%</td>
<td>48%</td>
</tr>
</tbody>
</table>

Table 1: The face detection results.

4 Experimental Results

The experiment procedure of the detection is as follows: first, the face discriminant space is constructed through 50 sampled images including 25 facial images and 25 non-facial images with complex backgrounds. In this step, we can get a transformation matrix to make a 2D vector representation for each image. Second, other test images are acquired and transformed into 2D vector representation. Next, we can allocate membership values to each image like Eq. (8).

$$\mu^{(i)} = e^{-\frac{||y^{(i)} - \overline{y}_f||^2}{\sigma^2}} \quad (8)$$

where, $\mu^{(i)}$ is a membership value for $i$-th image, $\overline{y}_f$ is the mean value of the face vectors from pre-defined discriminant space, $y^{(i)}$ is the $2D$ vector representation of each image, and $\sigma$ is the standard deviation of the face vector distribution. Finally, the existence of the face in the image is determined by comparing each membership value with pre-defined threshold. Table 1 shows the experimental result for face detection with 50 test log-polar mapped images ($64 \times 128$). Here, a success ratio means a ratio that is correctly classified results. A positive false ratio and a negative false ratio are ratios that are incorrectly classified results as the face and non-face, respectively. According to the trends of Table 1, the suitable threshold is in $[0.5, 0.6]$. We have experimented visual servoing with intention reading (Fig. 3). The major image processing is performed in two Genesis boards, Matrox, that have TMS320C80 processors, and the robotic arm is PowerCube, Amtec. The centering errors are less than ±20 pixels in Cartesian images as shown in Fig. 4, and these values are sufficiently small to perform the defined tasks. Average processing speed of visual information is 12 frames/s. Fig. 5 shows the extracted intention for a sequence of 10 images. When the user opens one’s mouth, the positive intention is.
higher. Thus, the user’s intention is successfully extracted through the proposed procedures. Sometimes, the user can yawn or talk with others during intention reading. We will add more visual cues for facial components like eyes in the future research.

5 Concluding Remarks

A novel wheelchair-based robotic arm and its human-robot interaction techniques have been considered. Visual servoing with effective intention reading has been successfully applied to the wheelchair-based robotic arm. Although log-polar mapped images are suitable for the detection, centering, and tracking, conventional Cartesian images have some merits in recognizing changes of the mouth shape. In order to be a more efficient system for visual servoing, more generalized mappings are desired in accordance with tasks.

Acknowledgments

This work is supported by the Ministry of Science and Technology of Korea as a part of Critical Technology 21 Program.

References