Samejima *et al. Robomech J (2018) 5:6*
https://doi.org/10.1186/s40648-018-0103-0

ROBOMECH Journal

RESEARCH ARTICLE

Open Access

Visual support system for remote control by adaptive ROI selection of monitoring robot

Shota Samejima[1*], Khusniddin Fozilov[2] and Kosuke Sekiyama[3]

Abstract

A visual assistance system has become attractive as a technique to improve the efficiency and stability of remote control. While an operator controls a working robot, another autonomous monitoring robot evaluates a suitable viewpoint to observe the work site, and dynamically moves to the optimal viewpoint for monitoring. Choosing the observation region (ROI: region of interest) is equivalent to deciding the action of the following autonomous monitoring system. Therefore, we focus on ROI detection in our visual support system. We propose an ROI selection method to identify the most suitable observation point and interobject relations. The monitoring robot detects a gestalt of the scene in order to identify the relations between objects. Such an adaptive ROI in real time improves the efficiency of the remote control. The experimental results indicate the effectiveness of the proposed system in terms of execution time and number of errors.

Keywords: Region of interest (ROI), Monitoring robot, Visual support, Gestalt psychology

Introduction

Autonomous intelligent robots have been employed in various challenging situations. Remote-controlled robots are mostly used when control accuracy and precision are necessary. In the 2015 DARPA robotics challenge, most of the robots were semiautonomous. Autonomous operation has already been implemented in posture control and route planning; however, overall human control remains unreplaceable [1, 2]. The most common applications for teleoperated robots include missions under unstructured or unknown environments such as search-and-rescue operations at disaster sites [3], deep sea or space exploration [4], and dealing with radioactive waste.

Various researchers have worked on enhancing operability by providing the operator with the necessary information to reduce operation time and errors. Errors usually occur because of distance uncertainties or blind spots (occlusions resulting from a sensor's location or surrounding objects). The following work focused on providing additional information to the remote-controlled robot to avoid blind spots and improve measurement accuracy. Barnes et al. [5] developed a control system that switches between remote control and autonomous control for autonomous obstacle detection and avoidance. Nielsen et al. [6] developed an operation interface using a 3D map based on 3D SLAM. Both works attempted to reduce error outbreaks by providing additional information; however, a large amount of information will also increase the load on an operator.

Multirobot systems (MRS) have been attracting attention for practical missions because a number of robots can be deployed to cover large areas and cooperatively complete tasks that a single robot cannot accomplish. MRS is also expected to enhance remote-control tasks. One of the features of MRS is that each mobile robot can share the information through mutual communication. Several robots can simultaneously observe the work environment from appropriate viewpoints. Hence, it is possible to reduce the number of blind spots and dynamically find the region of interest (ROI) to assist with remote control. In the visual support based on the MRS, in order to reduce the operator's cognitive load and avoid conflict recognition, robots have to match the collected data to share the ROI. While observation from various viewpoints by monitoring robots may provide excessive

*Correspondence: samejima@robo.mein.nagoya-u.ac.jp
[1] Department of Mechanical Science and Engineering, Nagoya University, Nagoya, Japan
Full list of author information is available at the end of the article

information, narrowing the ROI to the most relevant to the current task will reduce substantial cognitive load on the operator. With regard to the ROI sharing issue, Rokunuzzaman et al. [7] proposed the concept of semantic stability to determine the appropriate ROI size, which contains semantically related objects of interest evaluated by WordNet but excludes unrelated miscellaneous visual objects from the scene. ROI sharing is accomplished through multirobot cooperation. For example, Piasco et al. [8] proposed a connecting range of view for multirobots to increase the overall observation area.

The visual support based on the MRS requires criteria to allow a multirobot to selectively collect the necessary information for remote control. In a conventional work on remote control support systems, Kamezaki et al. [9] developed a remote-control support system by using multiple cameras that switched the image depending on the distances between the robots. Maeyama et al. [10] proposed a visual support system in which a monitoring robot followed a remote-controlled robot to constantly monitor the rear area of the remote-controlled robot. Neither of these works uses selection criteria for the observation target object.

In this paper, we propose a multirobot visual support system to reduce the number of blind spots and the uncertainty in distance perception by focusing on the necessary objects for remote control. In the proposed system, multiple autonomous monitoring robots are deployed around a remote-controlled working robot, and observe the objects that are highly related to the remote-controlled robot. The following technical issues must be taken into account in order to develop a multirobot visual support system.

1. Detection of neighboring objects and interobject relations to identify an important object.
2. Understanding environmental situations for the remote-controlled working robot.
3. Intention estimation to avoid undesired actions by the operator.
4. Information sharing of observation from the monitoring robots.
5. Path planning to move to a desired observation point.

This paper attempts to solve the technical issues for "(1) object and relation recognition" and "(2) situation interpretation".

We propose the probabilistic recognition of interobject relations by applying Deep Learning based on a convolutional neural network and gestalt psychology. An autonomous monitoring robot conducts image analysis to provide visual feedback to the operator. For image processing, the Scale-Invariant Feature Transform (SIFT) [11] is a well-known image feature matching algorithm. Another method based on Deep Learning [12] can be used for object recognition, and has many other applications such as human action analysis [13]. Moreover, Deep Learning can be applied to some unlearned objects. Hamamoto et al. [14] proposed a relation recognition method by using hidden Markov models (HMMs). However, the HMMs derive only previously learned relations. Preliminary learning is difficult because the natural environment has various relations. Thus, this paper adopts human perception principles to recognize relations between objects. The engineering applications of gestalt psychology [15, 16] allow for the detection of unique objects based on the strength of the relation. Hence, this paper adopts an engineering application of the gestalt psychology to detect latent interobject relations.

In addition, the proposed multirobot visual support system estimates a situation by applying Latent Dirichlet Allocation (LDA) to the results of object recognition. LDA is a language model that probabilistically obtains topics of words in documents. It is necessary to acquire information that allows the operator to accurately understand the working situation. TF-IDF [17] and LDA [18] are used to sort information in various subject fields including text summarizing and classification. Both techniques filter information based on the appearance frequency of the information. TF-IDF is calculated on a premise without context of information, but LDA is calculated on a premise with context of information. Because of this difference, LDA possesses a higher ability for situational understanding than TF-IDF. Hence, this paper adopts LDA to understand the environmental situation.

Proposed system overview

Architecture of autonomous monitoring system

Figure 1 shows the architecture of the proposed autonomous monitoring system. This system is composed of the operator, working robot, wide-view monitoring robot, and local-view monitoring robot. The operator remotely controls the working robot. The wide-view monitoring robot observes the entire work environment. The local-view monitoring robot observes the work environment locally.

The wide-view monitoring robot selects an ROI suitable for the working environment by estimating the relations between objects. In addition, the wide-view monitoring robot decides whether the objects found inside a selected ROI are related to a dangerous situation.

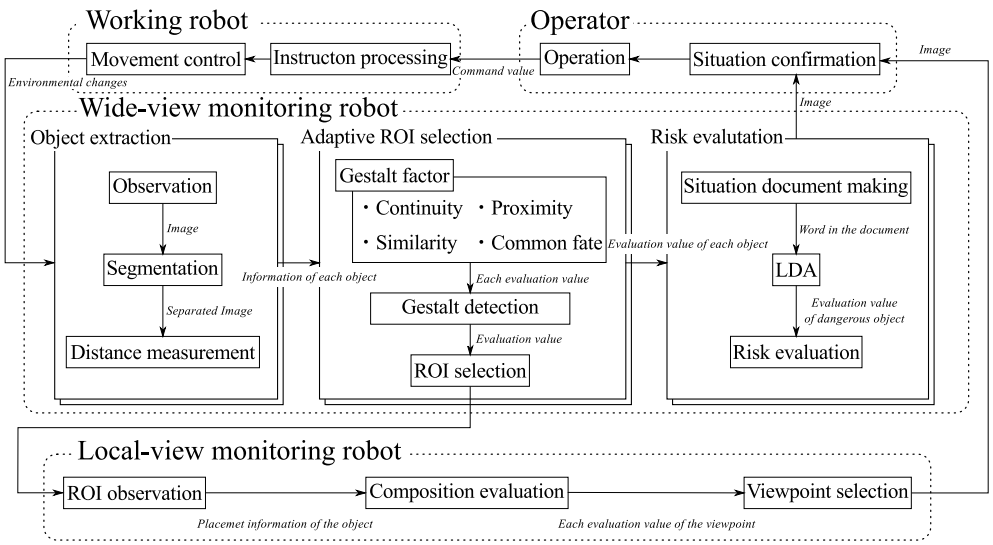

Fig. 1 Architecture of autonomous visual support system

The wide-view monitoring robot calculates the degree of relevance to the dangerous situation of the object by the LDA, and warns the operator based on the degree of relevance.

The local-view monitoring robot evaluates the relation of the object, and selects and adjusts the ROI with wide-view monitoring robot. Next, the local-view monitoring robot moves to a viewpoint depending on the interobject composition to provide visual support.

Gestalt factor as the relation evaluation item

The relations between a working robot and surrounding objects are evaluated by geometric relations and semantic relations.

Geometric relations describe how the objects are located in space. Semantic relations describe the contextual connection between the objects, that is, how their meaning or utilization purpose can be related. Geometric relations are evaluated using gestalt factors of *common fate*, *proximity*, and *continuity*. Semantic relation is evaluated by a gestalt factor of *similarity*. These gestalt factors in this paper are summarized as follows:

1. *Common fate* The common fate is the property for which the things of synchronizing motion are united.
2. *Proximity* The proximity is the property for which the objects in the neighborhood are united.
3. *Continuity* The continuity is the property for which the objects on a smooth line or curve are united.
4. *Similarity* The similarity is the property for which objects having the same elements (color, shape, semantics, etc.) are united.

Image processing methods for object recognition

General object recognition

In this paper, we adopt a technique based on depth image segmentation from an RGB-D camera [19]. The object is extracted by the object contour from the depth image. Next, the monitoring robot obtains the object's 3D coordinates by comparing the RGB image with the depth image. Once the object is extracted, it is assigned a name by the neural network. Figure 2 shows the process of general object recognition.

In this paper, a convolutional deep neural network is executed by using a library called Caffe [20]. Table 1 lists the details for the deep learning implemented in this paper. In addition, Table 1 shows the recognition precision of a main object used in the experiment. Figure 3 shows the structure of the deep learning.

Working robot detection

Detecting the working robot on a work site is of great importance. To calculate various relations between the robot and environment, we have to robustly detect and recognize the robot regardless of environmental conditions. The working robot has manipulators with multiple degrees of freedom (DoF); hence, it can have a complex shape and contour. Therefore, a recognition method based on depth image segmentation from an RGB-D camera is not sufficient. We adopted a high-speed detection method object detection method based on a YOLO convolutional neural network published by Redmon et al. [21]. We trained the network to detect the working robot's cart because it has relatively stable depth information to extract the 3D point. The network architecture

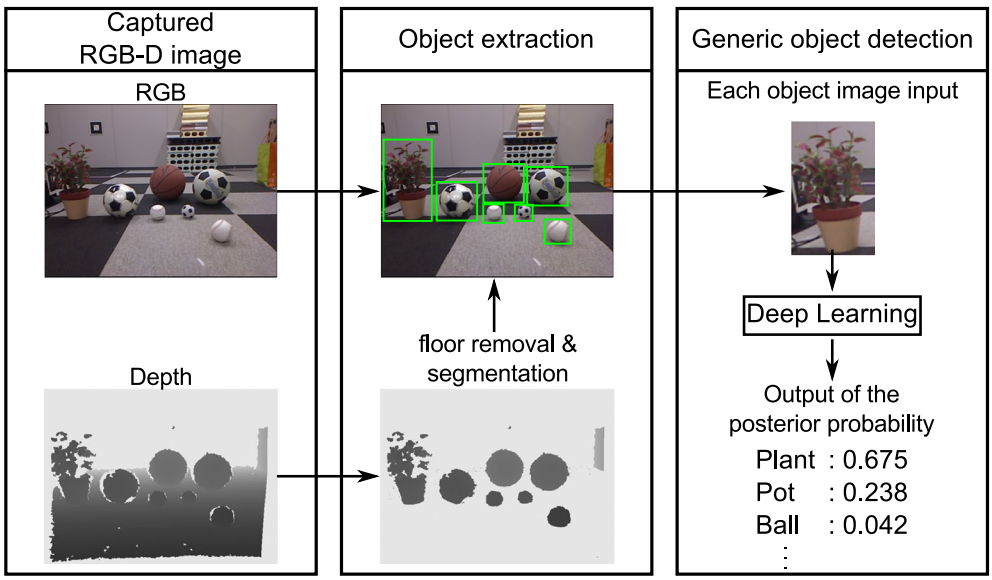

Fig. 2 Process of general object recognition

Table 1 Recognition performance of deep learning

Index items	Value
Number of learning image	Training: 9000, test: 1000
Number of class	6
Learning iteration	200,000
Pipe*	71.5, 86.6, 94.1%
Valve*	71.9, 85.6, 95.7%

* The probability for which the true name of the object is included in the Top(k) ($k = 1, 2.3$) of the output result of the deep learning

used for robot detection and the training process are displayed in Fig. 4 and Table 2.

Evaluation of the geometric relation

In this section, we will describe the calculation methods used to evaluate the geometric relations between objects. Figure 5 shows a schematic view of the objects and working robot in space. The speed of the working robot and the distance between the robot and objects have a close relation with the working task. By analyzing the speed and the distance, we can describe how the objects interact.

Evaluation of the common fate

The factor of *common fate* groups the objects if they move in the same direction with the same speed. Hence, we use *common fate* to detect situations where the working robot is carrying an object.

First, we define the variables used in the evaluation of *common fate*. $\boldsymbol{v}_r$ is the velocity vector of the working robot. $\boldsymbol{v}_o$ is the velocity vector of the object. $\boldsymbol{v}_{ro} = \boldsymbol{v}_o - \boldsymbol{v}_r$ is the relative velocity vector. When the movement of the working robot and an object are synchronized, the

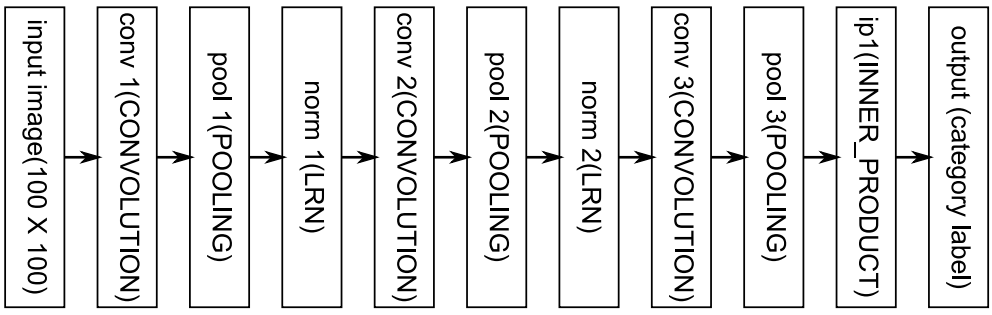

Fig. 3 Structure of convolutional deep neural network

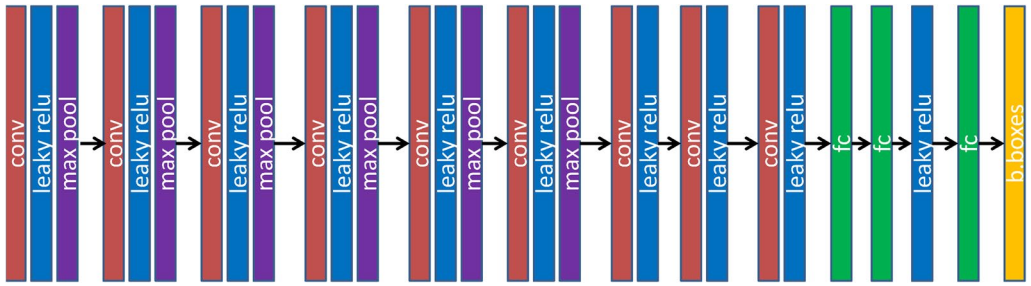

Fig. 4 YOLO-v1 network's architecture [26]

Table 2 Working object detection performance

Index items	Value
Number of learning image	Training: 1500, test: 500
Number of class	2
Learning iteration	15,000
Intersection over union	89.6%
Mean average precision	75.5%

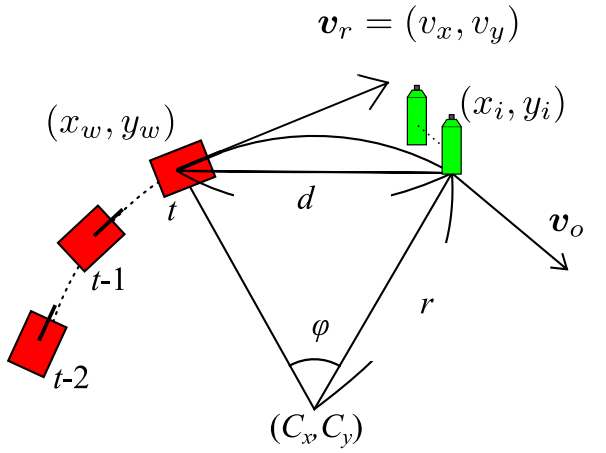

Fig. 5 Schematic view showing a geometric relation. Note that origin of positions is defined as position of wide-view monitoring robot

relative vector becomes zero. As given in Eq. (1), the *common fate*: $CF(O_i)$ of the object (O_i: i is the object number). The object number is allocated from the one closer to the monitoring robot) is calculated by binarization processing using the threshold value δ. When $CF_i = 1$, the object has a *common fate* with the working robot.

$$CF(O_i) = \begin{cases} 1, & (\|\boldsymbol{v}_{ro}\|/\|\boldsymbol{v}_r\| \leq \delta)\ , \quad (1a) \\ 0, & \text{otherwise.} \quad (1b) \end{cases}$$

Evaluation of the proximity

The factor of *proximity* groups the objects that are close to each other. When the working robot approaches an object, it is considered to have a relation with the working robot. Thus, the *proximity* $D(O_i)$ is defined by the distance d_i between the working robot and the object i, which is calculated by the respective relative position from the RGB-D camera. The smaller the distance d_i, the stronger the *proximity*. The maximum distance d_{max} varies according to the type of camera. This paper used Asus Xtion as the RGB-D camera. Therefore, the normalized coefficient R_d is defined by

$$R_d = d_{max}/2. \quad (2)$$

From Eq. (2), the evaluation of the *proximity* $D(O_i)$ is given by

$$D(O_i) = (d_i/R_d)^2. \quad (3)$$

The *proximity* is stronger when $D(O_i)$ is smaller.

Evaluation of the continuity

The factor of *continuity* groups the objects on the same line or on a curve. If an object is placed along the moving direction of the working robot, it is regarded as a relevant object. From Fig. 5, (v_x, v_y) is the velocity vector of the working robot. Let (x_w, y_w), (x_i, y_i) be the positions of the working robot and the object i, respectively. Note that the origin of the positions is defined as the position of the wide-view monitoring robot. We draw an arc connecting the positions (v_x, v_y), (x_w, y_w) and (x_i, y_i). (C_x, C_y) are the central coordinates of the arc given by Eqs. (4a) and (4b), and r is the radius of the arc given by Eq. (4c).

$$C_x = \frac{(x_i^2 - x_w^2 + y_i^2 - y_w^2)v_y + 2(y_w - y_i)(v_x \cdot x_w + v_y \cdot y_w)}{2[v_x(y_w - y_i) - v_y(x_w - x_i)]} \quad (4a)$$

$$C_y = \frac{(x_i^2 - x_w^2 + y_i^2 - y_w^2)v_x + 2(x_w - x_i)(v_x \cdot x_w + v_y \cdot y_w)}{2[v_y(x_w - x_i) - v_x(y_w - y_i)]} \quad (4b)$$

$$r = \sqrt{(x_w - C_x)^2 + (y_w - C_y)^2} \tag{4c}$$

The angle of the arc φ_i is given by

$$\varphi_i = \arccos\left(\frac{2r^2 - d_i^2}{2r^2}\right). \tag{5}$$

To normalize φ_i, the coefficient is determined as $R_c = \pi/2$ because the maximum angle is π. The evaluation of *continuity* $\Phi(O_i)$ is defined by

$$\Phi(O_i) = (\varphi_i/R_c)^2. \tag{6}$$

If the working robot's speed is zero, we cannot evaluate whether there are any objects along the line of robot's movement; hence when the (v_x,v_y) is equal to zero, $\Phi(O_i)$ is also equal to zero. The relation of *continuity* is strong when $\Phi(O_i)$ is smaller.

Geometric relation evaluation function

The geometric evaluation value E_g is defined by (7) using $D(O_i)$ and $\Phi(O_i)$ defined by Eqs. (3) and (6).

$$E_g(O_i) = \begin{cases} 1, & CF(O_i) = 1 \\ \exp\{-(\lambda_g D(O_i) + (1-\lambda_g)\Phi(O_i))\}, & \text{otherwise} \end{cases} \tag{7}$$

where the range of $E_g(O_i)$ is $0 \le E_g(O_i) \le 1$, and weight coefficient $\lambda_g = 0.7$. Larger values of $E_g(O_i)$ imply the presence of close interaction with the working robot and object *i*.

Evaluation of the semantic relation

Semantic relations can be described through understanding the meaning of the objects in space. Depending on how the objects can be used, we estimate whether the robot can interact with them.

Evaluation of the similarity

The gestalt factor of *similarity* groups the objects if they have similar features, patterns, or shapes. However, in this paper, we evaluate the relation between the objects based on their semantic connections with the ongoing task. This enables adaptive support to find and include in the ROI an object that can be used by the robot's end effector.

In this paper, the *similarity* is calculated by applying WordNet [22]. WordNet is a hierarchical dictionary of the English language. We compare the semantic connection between the words by calculating hierarchical distances. Figure 6 shows an example of WordNet's treelike hierarchical structure. By comparing the distance from two given words to their common root, we estimate how similar they are, that is, the *similarity* relation. $\text{len}(c_1, c_2)$ represents the hierarchical distance between the word c_1 and c_2 in WordNet.

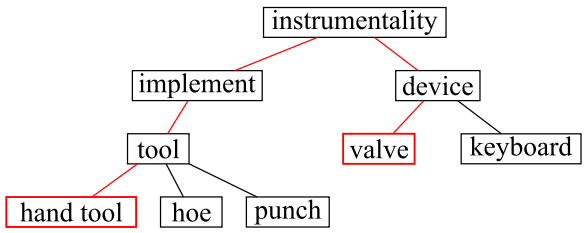

Fig. 6 Example of WordNet

When $\text{len}(c_1, c_2)$ equals 0, the evaluation value of the *similarity* is 1. A similarity between two words c_1 and c_2 is defined as

$$\text{sim}(c_1, c_2) = 1 - \exp\left[-\left\{\frac{\ln(\text{len}(c_1, c_2))}{2DR_s}\right\}^2\right]. \tag{8}$$

where D is the maximum hierarchical distance of WordNet. Equation (8) is derived by Leacock et al. [23]. In addition, R_s is the normalized coefficient that is defined by

$$R_s = \frac{1}{2}\underset{c_1,c_2 \in \boldsymbol{C}}{\text{argmin}}(\ln(\text{len}(c_1, c_2)/2D)). \tag{9}$$

where the range of $\text{sim}(c_1, c_2)$ is $0 \le \text{sim}(c_1, c_2) \le 1$. $\text{sim}(c_1, c_2)$ indicates that a semantic relation is strong when the value approaches 1. For example, from Fig. 6, when c_1 is *hand_tool* and c_2 is *valve*, the hierarchical distance is $\text{len}(hand\_tool, valve) = 5$, and *similarity* is $\text{sim}(hadn\_tool, valve) = 0.735$. The normalized coefficient is $R_s = 1.84$.

Expected value of the semantic relation

In this section, the evaluation value of the *similarity* $E_s(O_i)$ between an object O_i and the tools is defined, where the tools are the objects that are handled by the robot's end effector. In our work, the names of the tools n_w are predefined. Algorithm 1 shows the calculation process of the expectation with regard to the semantic relation. By using deep learning, the robot acquires the name of the object from the camera image. Next, several object name candidates are outputted by deep learning. Therefore, it is necessary to define the *similarity* evaluation when there is more than one candidate for the object name. The expected value of the semantic relation $E_s(O_i)$ is defined by Eq. (10) where $p(n_k|O_i)$ are the Top-k probability values in which each object has a name n_k (k is the name number).

$$E_s(O_i) = \sum_{k=1}^{K} \frac{p(n_k|O_i)}{\sum_{l=1}^{K} p(n_l|O_i)} \text{sim}(n_k, n_w). \tag{10}$$

The range of $E_s(O_i)$ is $0 \leq E_s(O_i) \leq 1$. Larger values of $E_s(O_i)$ imply the presence of close interaction with the working robot and an object.

Algorithm 1 Expectation calculation of similarity evaluation value

```
The floor removal of Depth image
Object segmentation
for i = 0 to N do
    Calculate probability of object name by Deep Learning
    A − 0, B = 0 # Initialization of A and B
    for k = 0 to K do
        A = A + p(n_k|O_i)sim(n_k|n_w)
        B = B + p(n_k|O_i)
    end for
    E_s(O_i) = A/B
end for
```

ROI selection based on the relation evaluation

ROI selection

By using geometric and semantic evaluation values defined in Eqs. (7) and (10), the unified relation evaluation is calculated by

$$E(O_i) = \lambda E_g(O_i) + (1 - \lambda)E_s(O_i), \tag{11}$$

where the range of $E(O_i)$ is also $0 \leq E(O_i) \leq 1$ with the weight factor coefficient $\lambda = 0.75$. The weight coefficient was experimentally determined. When the value of $E(O_i)$ approaches 1, the system interprets it as the presence of a close interaction with the neighboring object.

By using a threshold in the relation evaluation value, the system decides which objects are included in the ROI. The threshold T is calculated in real time by applying a discriminant analysis method. The threshold varies from 0 to 1 virtually, and objects are classified into two sets, that is, the set α inside the ROI or the set β outside the ROI. This is defined by

$$\alpha = \{O_i | E(O_i) \geq T\}, \tag{12a}$$

$$\beta = \{O_i | E(O_i) < T\}. \tag{12b}$$

The set α includes n_α objects and has the average of the evaluation value $\bar{E}_\alpha$ with variance σ_α^2. Likewise, the set β includes n_β has the average of the evaluation value $\bar{E}_\beta$ with variance σ_β^2. A within-class variance σ_W^2 and a between-class variance σ_B^2 are defined by

$$\sigma_W^2 = \frac{n_\alpha \sigma_\alpha^2 + n_\beta \sigma_\beta^2}{N}, \tag{13a}$$

$$\sigma_B^2 = \frac{n_\alpha (\bar{E} - \bar{E}_\alpha)^2) + n_\beta (\bar{E} - \bar{E}_\beta)^2}{N}. \tag{13b}$$

where E is the mean of the evaluation value of all objects in the working environment, and N is the total number of objects. Based on the degree of separation, threshold T is defined as

$$T = \underset{T \in 0,1}{\text{argmax}} \left(\frac{\sigma_B^2}{\sigma_W^2} \right). \tag{14}$$

Validation of usability of ROI selection system

In this section, we determine whether the proposed ROI selection system is coherent with human choices. We compare the ROI selected by a human operator with the proposed system's evaluations. To estimate the system's ROI selection, we asked three laboratory members to watch a video for the experiment. Three subjects watched a recording where a working robot carried an object (a valve or a pipe) to a destination while avoiding obstacles. Next, for each frame, the subjects selected which objects they would like the working robot to focus on or observe from a closer distance. Similarly, the proposed ROI selection system analyzes the video and selects the objects. The reliability of the proposed ROI selection system is evaluated by comparing the participants' answers with proposed system's answers.

Figure 7 shows the set of objects classified based on each answer. $n(A_j \cap C_j)$ is the number of objects that were selected as inside-ROI by the subjects and the proposed system in Frame$_j$. $n(B_j \cap D_j)$ is the number of objects that were selected as outside-ROI by the subjects and the proposed system in Frame$_j$. The agreement degree P is calculated by

$$P = \frac{\sum_{j=1}^{N} \{n(A_j \cap C_j) + n(B_j \cap D_j)\}}{\sum_{k=1}^{N} n(C_k \cup D_k)}, \tag{15}$$

where N is the number of video frames.

The agreement degree P with the first subject S_1 was 0.875. The agreement degree P with the second subject S_2 was 0.829. The agreement degree P with the third subject S_3 was 0.840. In all cases, the agreement degree was beyond 0.8. Overall, the average of the agreement degree between the subjects was 0.83.

Table 3 shows the agreement degree P based on gestalt factors used to evaluate the interobject relations. When the proposed system uses only one gestalt factor, as we can see from Table 3, the ROI selection using only a proximity factor scores the highest with $P = 0.798$. On the other hand, a system evaluation based on a combination of similarity and common fate factors scored the lowest with $P = 0.470$. The reason is that the similarity and common-fate factors estimate the relation when the robot is interacting with a specific object using the manipulators or an end effector. Hence, these factors are inferior to the

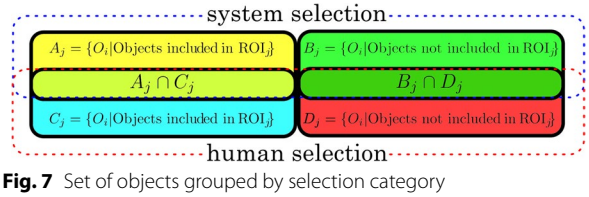

Fig. 7 Set of objects grouped by selection category

Table 3 Agreement degree values based on gestalt factors used for ROI selection

Gestalt factor	P	Gestalt factor	P
Proximity	0.798	Continuity	0.613
Similarity	0.475	Common fate	0.575
Proximity, continuity	0.856	Proximity, similarity	0.802
Proximity, common fate	0.792	Continuity, similarity	0.567
Continuity, common fate	0.632	Proximity, continuity, similarity	0.868
Similarity, common fate	0.470		
Proximity, continuity, common fate	0.859	Proximity, similarity, common fate	0.797
Continuity, similarity, common fate	0.602	Proximity, continuity, similarity, common fate	0.875*

*The agreement degree is highest. Therefore, these gestalt factors are used in proposed system

factors of proximity and continuity that evaluate the relation between the robot and multiple objects in space. In addition, according to results given in Table 3, the proposed system using four gestalt factors has the highest agreement score with $P = 0.875$ and can be used to represent the human choices.

Situation interpretation based on the LDA

Understanding the situation at a work site is of great importance in providing suitable visual support. During teleoperation, the operator is already preoccupied with controlling the working robot and has to react to various changes in environment. The working situation and dynamic relations between objects are closely linked in the remote-control tasks. The situation interpretation of a still image that does not include a dynamic relation change can narrow the choices of the situation but cannot arrive at the only correct answer. Therefore, we propose a situation interpretation system based on data of the working robot's actions.

In addition, an assessment of the working situation requires various intertask relations to correctly estimate the working situation. For example, a semantic relation is more important than a geometric relation when determining whether the robot is approaching an object or is about to collide with that object. The movement of the robot at the time of a collision resembles that at the time of the approach operation, but the semantic connection between the target object and the robot at the time of collision is different than that at the time of the approach operation. From the above, a situation interpretation system requires data that represents various relations for dynamic image analysis.

Preparing of training data

In this paper, we use latent dirichlet allocation (LDA) [18] to predict potential interactions between objects. This paper employs the GibbsLDA++: AC/C++ Implementation of Latent Dirichlet Allocation [24] developed by Xuan-Hieu et al. as the learning code of the LDA topic model. Time-series data of gestalt evaluation values proposed in the previous section are used as the training data. Time-series data for gestalt factor evaluation are acquired from the video of a robot's movement under various conditions, e.g., carrying an object, maneuvering around an obstacle, and collisions.

Learning process by the LDA

Table 4 lists the parameters used for the LDA topic model. The target document $\boldsymbol{\theta}$ is time-series data over gestalt evaluation values. Classification target topic $\boldsymbol{\phi}$ is the choice made under each situation. Optimizing the distribution of time-series data over gestalt evaluation values, and the distribution of the situation choices, are called LDA learning in this paper.

The parameter of learning target date distribution is α, and the parameter of classification target topic distribution is β. Each parameter of Dirichlet distribution is updated by

$$\alpha^{\text{new}} = \alpha \frac{\sum_{d=1}^{D} \sum_{k=1}^{K} \Psi(N_{dk} + \alpha) - DK\Psi(\alpha)}{K \sum_{d=1}^{D} \Psi(N_d + \alpha K) - DK\Psi(\alpha K)}, \tag{16a}$$

Table 4 Parameters used in LDA topic model

Parameter	Details
θ	Probability that time-series data of gestalt evaluation value belongs to situation choice k
ϕ	Probability that gestalt evaluation value w appears in situation choice k
Z	Situation choices
W	Time-series data of gestalt evaluation value
w	Gestalt evaluation value
D	Number of time series documents
K	Number of topics
V	Number of the kinds of data in time-series data
α	Parameter for distribution over situation choices
β	Parameter for distribution over gestalt evaluation value

Table 5 Learning condition

Parameter	Details
Number of classification topics	3
Name of topics	Transportation, avoidance, collision
Number of learning data	20 cases per topic
Size of 1 learning data	3–5 (kB)
Hyperparameter initial value α	0.5
β	0.1
Learning iteration	20,000

$$\beta^{\text{new}} = \beta \frac{\sum_{k=1}^{K}\sum_{v=1}^{V}\Psi(N_{kv}+\beta) - KV\Psi(\beta)}{V\sum_{k=1}^{K}\Psi(N_k+\beta V) - KV\Psi(\beta V)}. \quad (16b)$$

where N_{dk} is the number of documents assigned to the topic k in time-series data d, N_{kv} is the number of documents v assigned to the topic k in all time-series data, and N_d is the number of documents included in time-series data d. $\Psi(x)$ is a Digamma function.

Table 5 shows the learning conditions. The situation is estimated based on the situational choices with respect to the highest value among the topic distributions with respect to the time-series data for each neighboring object. The probability that time-series data d is assigned to topic k is calculated as

$$\theta_{dk} = \frac{N_{dk}+\alpha}{N_d+\alpha K}. \quad (17)$$

Danger warning based on topic classification result

By applying LDA topic modeling to the data based on a gestalt time series, the system creates a sentence using information regarding to the working robot and each object around it, for example, "the working robot is approaching an obstacle" or "the working robot is near Object-A". Next, the created sentences for a given frame in a video sequence are grouped as one document, and topic classification is applied.

The system will classify the document by topics. From the topic that was predefined as dangerous, the system will extract the object name and classify it as a dangerous object. In our system, collision is set as a dangerous topic. A warning alert is set based on the probability value of the collision, which is calculated by estimating the object's relation evaluation value.

A warning is indicated by using different color signals. The color signal is displayed on the screen presented by the monitoring robot. The color level CL is determined by (18) based on the probability of dangerous situation choice k_{danger}.

$$CL = \begin{cases} Red, & \theta_{dk_{danger}} \geq 0.7, \quad (18a) \\ Yellow, & 0.3 \leq \theta_{dk_{danger}} < 0.7, \quad (18b) \\ Green, & \text{otherwise.} \quad (18c) \end{cases}$$

Confirmation experiment of the situation interpretation

This section examines whether the proposed system can correctly classify moving images containing only single situations. The outlines of the moving image sequences show that each situation is summarized as follows: transportation, avoidance, and collision.

Classification accuracy is evaluated based on the correct answer ratio and average selection probability. The correct answer rate CR is the rate at which the correct answer option is selected from all video frames. The correct answer rate CR is calculated by

$$CR = \frac{N_{cc}}{N_f}. \quad (19)$$

where N_{cc} is the number of frames that contain the correct answer choices, and N_f is the total number of video frames. The average selection probability is the average value of the selection probability $\theta_{dk_{correct}}$ of the correct answer choice $k_{correct}$ in each frame. The average selection probability AP is calculated by

$$AP = \frac{\sum_{n=1}^{N_f}\theta_{dk_{correct}i}}{N_f} \quad (20)$$

where $\theta_{dk_{correct}}i$ is the selection probability of the correct answer choice in frame i.

Table 6 lists the correct answer rate CR and average selection probability AP for each choice. The value of correct answer rate CR in all options exceeds 0.85. Thus, it can be said that the proposed system has a high situational interpretation capability. In addition, since the value of the average selection probability is about 0.7–0.8, the threshold value in Eq. (18) is appropriate.

Control of local monitoring robot

The local monitoring robot is an omni-wheel robot which is composed of three motors and can move to any direction while turning at any rotary speed by controlling

Table 6 Topic classification result

Topic	*AP*	*CR*
Transportation	0.821	1.0
Avoidance	0.724	0.875
Collision	0.857	0.894

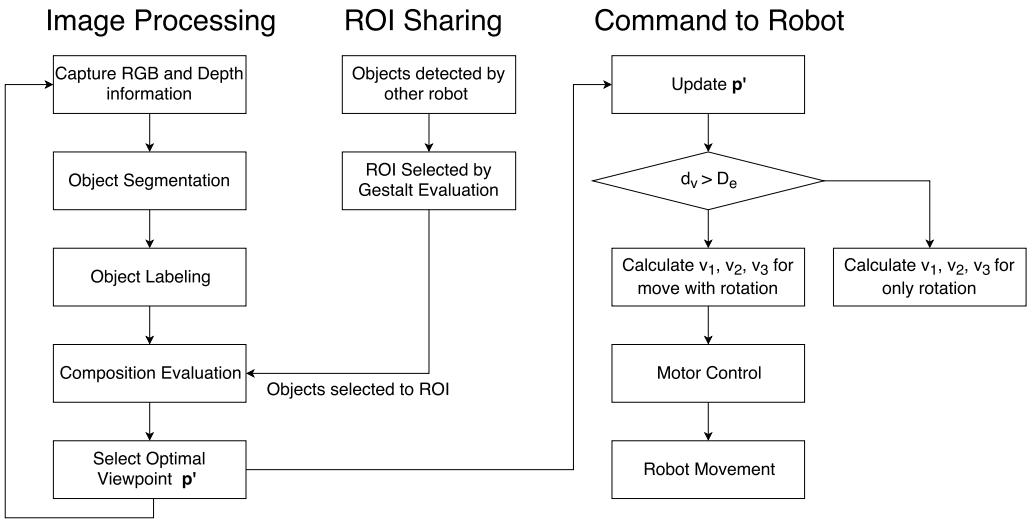

Fig. 8 Flowchart of robot control

the speed of three motors. A flowchart of robot control is shown in Fig. 8. The monitoring robot receives the information regarding to objects that were selected to ROI and moves to the suitable position for observing. In this paper, we used the viewpoint selection method developed by Ito et al. [25]. In this method, depending on the position of objects in space, candidate viewpoints are generated. The robot estimates the composition and gives an evaluation score for each candidate viewpoint. The viewpoint with the highest score is selected as the optimal viewpoint.

Experiments

ROI selection in a working environment

This section evaluates the proposed ROI selection system in a working environment. The robot performs a transportation task. The transportation object is a pipe. The destination is set near the valve. Experiment snapshots are shown in Fig. 10. (The object numbers are allocated from the one closer to the monitoring robot.) Figure 9 shows the changes in relation evaluation value E over time. Table 9 shows how the proposed system interprets situation around the working robot.

Figure 10a shows the snapshots from the wide-view monitoring robot at frame 128. The wide-view monitoring robot outputs three images: the wide view on the right that contains the number of objects detected, and also the color signal to warn the operator according to Eq. (18), a graph displaying the objects' name with evaluation value, and the selected ROI for the respective frame. From Table 7 and Fig. 9, we can see that for frame 128, the evaluation value for the pipe is the highest and is greater than threshold T from Eq. (14). Therefore, it is selected as the ROI for frame 128. Moreover, we

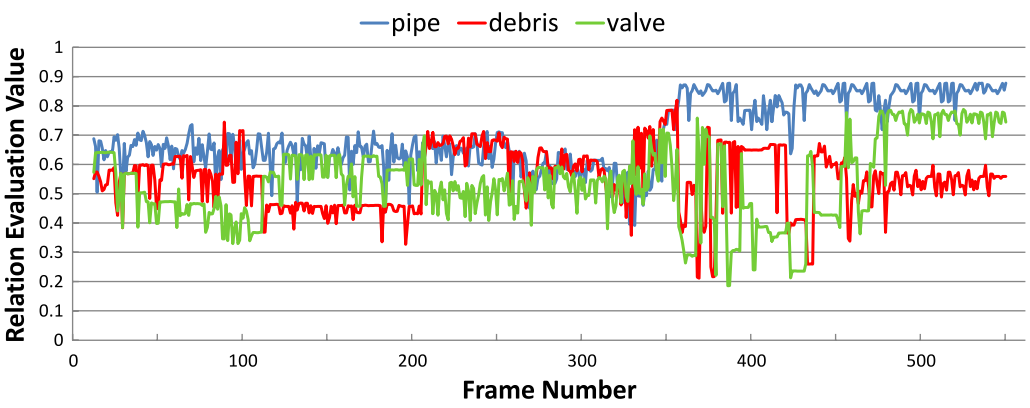

Fig. 9 History of relation value throughout task

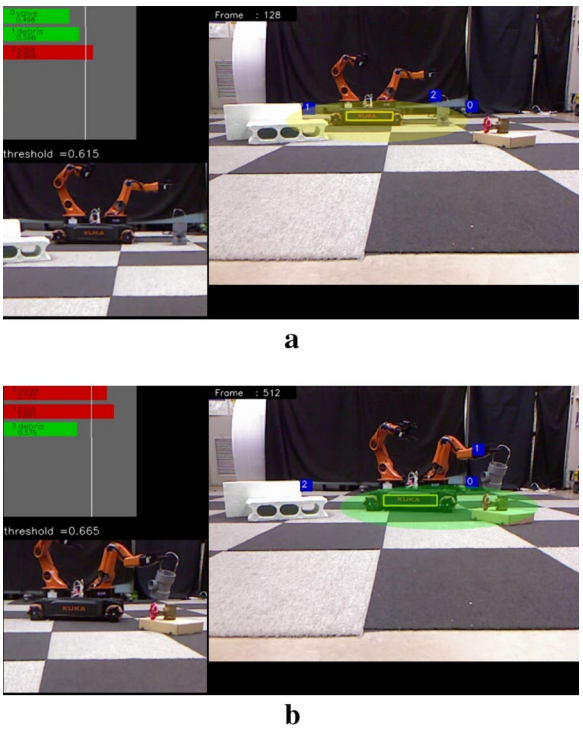

Fig. 10 ROI selected during transportation task. **a** Frame = 128, **b** Frame = 512

can see from Table 9a the robot is moving towards pipe and valve. Thus from Eq. (18) the color signal shown in Fig. 10a is yellow.

Similarly, in Fig. 10b and Table 8, the pipe and valve are included in the ROI. From Fig. 9, we can see that after frame 350, E for the pipe is around 0.9. Thus, we can say that is the moment when the pipe is picked. As for the valve from frame 460 E increases and reaches value over 0.7. The proximity and continuity values for the valve are very small. Thus results in a high geometric relation. Therefore, the valve is selected as the ROI for frame 512. From Table 9b, the robot is still; hence according to Eq. (18) the color signal shown in Fig. 10b is green.

The above result indicates that the proposed ROI selection system can dynamically select an ROI. Overall, the proposed system was able to extract the important objects by estimating their relations using gestalt factors.

Validation of adaptability to unlearned objects

Compared to the risk recognition system that we proposed in the past [26], the advantage of the risk recognition system based on the time-series data of a gestalt evaluation is that it can cope with unlearned objects. The performance of the proposed system was evaluated based on the system's ability to warn the operator in an environment that contains both unlearned and learned objects.

This experiment was conducted under the following conditions:

- Case A is a condition where five learned objects are detected as obstacles.
- Case C is a condition where eight unlearned objects are detected as obstacles.

Table 10 lists the conditions and number of objects used for each case. The remote control for all cases was to transport a valve while avoiding the obstacles.

Figure 11a, b show corresponding images of the working robot approaching the obstacles in the case of A and B, respectively. Table 11a, b show the sentences created by LDA learning to evaluate the potential risks. The proposed system warns the operator using color signals in case of danger according to Eq. (18). Green indicates Safe, yellow indicates Caution, and red indicates a Warning.

In case A, the working robot is around objects that are known to the system. From Fig. 11a and Table 11a we can see a situation where the robot is still and not moving. Therefore, since there is no change in the robot's position, the system outputs the Green color, stating that the situation is Safe.

In case B, the working robot is surrounded by more objects, and all obstacles are unknown to the system. However, the object detection method based on depth image segmentation allows us to extract all objects in space. Next, the system evaluates the relations between all objects and the robot. Because we know the 3D positions of the unknown objects, we are able to evaluate the geometric relations. Table 11b shows the sentences created by proposed system for case B.

We already mentioned that the factors of proximity and continuity are sufficient to select the ROI. Next, the data based on the gestalt-evaluation time series is classified

Table 7 Each evaluation value at *Frame* = 128

Object name	Common fate	Proximity	Continuity	Geometric relation	Semantic relation	Evaluation value
Debris	0	0.54	0.46	0.597	0.481	0.568
Valve	0	0.73	0.99	0.447	0.649	0.498
Pipe	0	0.38	0.17	0.728	0.492	0.669

Table 8 Each evaluation value at *Frame* = 512

Object name	Common fate	Proximity	Continuity	Geometric relation	Semantic relation	Evaluation value
Debris	0	0.72	0	0.604	0.488	0.575
Valve	0	0.28	0	0.817	0.636	0.772
Pipe	1	–	–	1	0.524	0.881

Table 9 Sentences created by LDA learning at Frame = 128 and Frame = 512

Object number	Sentences
(a) Frame = 128	
0	Robot approaches to valve; robot is moving
1	Robot moving away from debris; robot is moving
2	Robot approaches to pipe; robot is moving
(b) Frame = 512	
0	Robot is near valve; robot is standing
1	Robot has pipe; robot is standing
2	Robot is near debris; robot is standing

Table 10 Environmental conditions for each case

	Case A	Case B
Number of target object	1	1
Number of goal point object	1	1
Number of learned obstacle object	5	0
Number of unlearned obstacle object	0	8

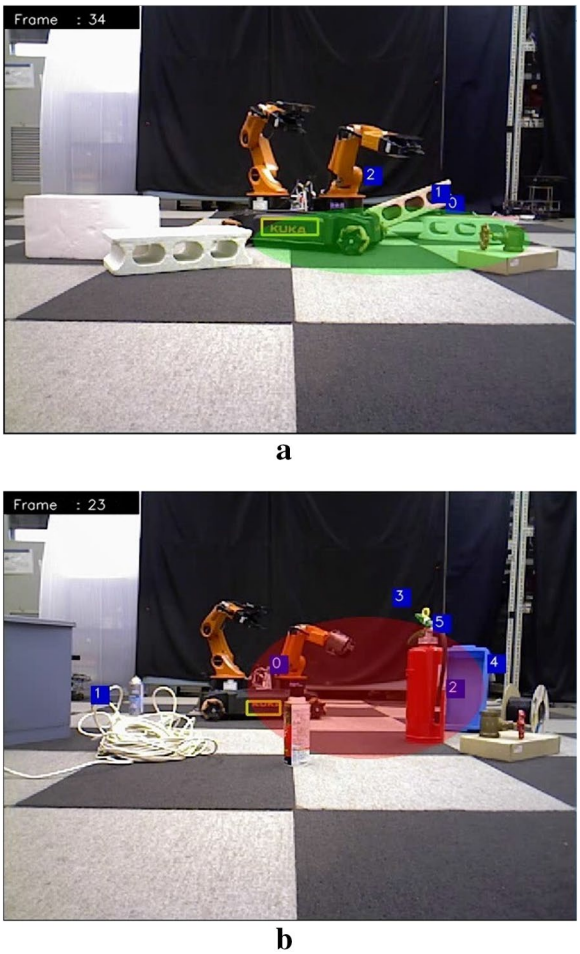

Fig. 11 State of danger warning under each condition. **a** Warning of danger under case A, **b** warning of danger under case B

by LDA learning for risk evaluation. As a result, the system is able to detect the unknown obstacles and warn the operator. From Fig. 11b, we can see that the color signal turned Red when the working robot started to move in the direction of the obstacles.

Therefore, even in an environment where unlearned objects are placed, the proposed system successfully estimates the risks and sends a notification using colors.

However, the following problems were revealed: The first problem is a decline in the number of FPSs when the number of detected objects increases as a result of increased calculation costs. The second problem is that the monitoring robot may not be able to recognize small objects. In Fig. 11b, the monitoring robot could not stably recognize the spray can at the center of the screen.

Verification of improvement of operability in remote control

In this section, the ability for operability improvement of the remote control with the proposed visual support system is evaluated. To evaluate the effectiveness of the proposed system, we performed an experiment. Our goal was to validate the improvements in robot teleoperation by using the proposed system. We have to note that, the target objects, destinations and task scenarios described below are unknown to proposed system.

Environment description

The experiment took place in our laboratory. The KUKA youBot with two 5DoF manipulators was used as a teleoperation robot, hereinafter called a working robot

Table 11 Sentences created by LDA learning for Case A and Case B

Object number	Sentences
(a) Case A	
0	Robot is near valve, robot is standing
1	Robot is near debris, robot is standing
2	Robot is near debris, robot is standing
(b) Case B	
0	Robot approaches to unknown; robot is moving
1	Robot approaches to unknown; robot is moving
2	Robot approaches to valve; robot is moving
3	Robot approaches to unknown; robot is moving
4	Robot approaches to unknown; robot is moving
5	Robot approaches to unknown; robot is moving

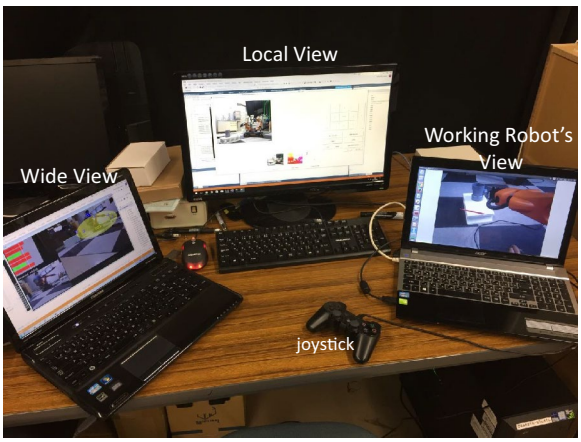

Fig. 12 User interface

(Robot-W). One of the working robot's manipulators was equipped with a camera and was programmed to track another arm's end effector. Thus, the operator could see the manipulator's end effector and perform various tasks while relying on the images from the camera arm. The proposed support system consisted of two robots: a fixed wide-view monitoring robot (Robot-M_W) and a moving local-view monitoring robot (Robot-M_L). Both monitoring robots were equipped with RGB-D cameras. The task site contained both known and unknown system objects.

To create conditions close to a real-life situation, all participants remote controlled the working robot behind a screen, relying only on real-time images from the cameras outputted to monitors, as can be seen in Fig. 12. The target object and destination are unknown to proposed system.

Participants

Six subjects who were members of our laboratory participated in the experiment. All participants were males between 24 and 30 years old. In addition, all participants had some experience related to robotics. However, only half of the subjects were familiar with the working robot's controls. Therefore, those who were not familiar with the robot were introduced to manipulators and practiced for 40 min one day before the experiment. All participants were given about 5 min to practice immediately before the experiment.

Task

The teleoperation task was a simple Peg-in-Hole manipulation. As target objects, a container with pencils and a case with holes were chosen. The experiment tasks are described below:

1. Approach the container with pencils, and pick up a pencil.
2. Carry the pencil to the destination.
3. Peg the pencil into one of the holes.
4. Repeat the process three times.

Evaluation methods

Each participant had to complete the tasks twice:

- Relying on camera arm + two fixed cameras.
- Relying on camera arm + developed support system.

To examine the improvements in operability, we measured the success rate, time spent on the task, and the number of operational mistakes between two conditions given above. The success rate was evaluated as 100% if the operator managed to peg three pencils during three attempts.

If only 2/3, 1/3, or 0/3 were pegged, then the success rate was evaluated as 66, 33, or 0%, respectively. The number of operational mistakes was counted as the number of collisions between the working robot and surrounding objects. If owing to collision the target objects were dropped, the task was considered complete with a success rate of 0%. The time spent on the task was measured from the moment the operator moved the working robot until the third pencil was pegged.

Figure 13 shows the experiment environment. There were several obstacles around the working robot. We can also see that the pencils in the container were not perpendicular; hence, the operator had to adjust the manipulator to pick up the objects. Neither the container nor the case with designated holes was fixed, so in case of mistakes, they could fall or flip over.

The views from each robot and support images provided by monitoring robots in this experiment are summarized below:

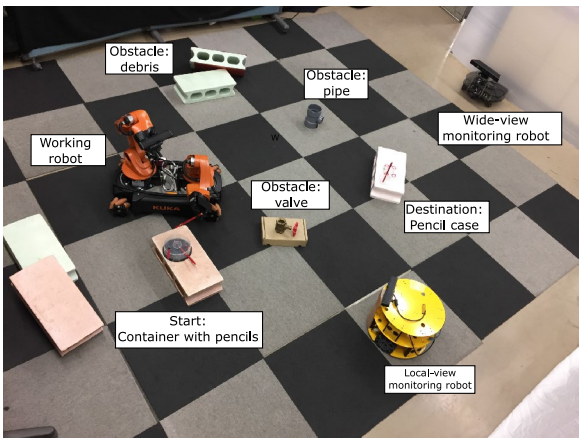

Fig. 13 Experiment environment

Figure 14a shows the view sent from Robot-W's camera arm. The camera arm was programmed to track the second arm's end effector. Therefore, the operator was able to remote control Robot-W and perform various manipulations using only the camera arm. However, the view was limited, and the operator had to constantly change the end effector's direction to see the surroundings. To visually confirm that the object was grasped, the operator might need to also move Robot-W's cart and determine how to approach the objects to get a suitable view. Since the view from the camera arm was not intuitive to operate the robot, and the camera was able to capture only a limited area, the teleoperation process became very difficult and time consuming.

Robot-M_W observed the entire working environment and selected the ROI by evaluating the relation between Robot-W and the surrounding objects. Hence, the movement of Robot-M_W was limited compared to that of Robot-M_L. In addition, Robot-M_W evaluated the possible risks and warned the operator with color signals. Figure 14b displays the image sent from Robot-M_W. As we can see, Robot-M_W output the image that contained the color signal and output the global position of Robot-W in the environment.

The colored ellipse in the larger window is an example of a color signal that the system output based on the risk evaluation method [26].

Robot-M_W selected the ROI after evaluating the inter-object relations in the given scene, and shared it with Robot-M_L. Since Robot-M_W focused on grasping the entire scene, the target objects could sometimes be blocked by Robot-W or its manipulators. Yet, by sharing the ROI, the proposed system was able to provide a suitable support from an optimal viewpoint.

Figure 14c shows the images sent by Robot-M_L. Robot-M_L received information on the objects that were selected for the ROI by Robot-M_W, determined a new, combined ROI, and autonomously moved to the optimal viewpoint. As a result, the local-view-monitoring robot observed the working robot and provided visual support from a closer range.

Figures 15 and 16 show the snapshots from the experiment. We can observe the changes in Robot-M_L's position. From Fig. 16b, we can see that Robot-M_L had a control interface. Since the observation region of Robot-M_L was limited, the robot had to move a lot while tracking Robot-W. Hence, in case of error, the operator could also remote control Robot-M_L.

Results

Figure 17 shows a comparison of the remote-control efficiency and accuracy with/without the proposed visual assistance.

Figure 17c shows the success rate in each case. On average, when the support system was not used, the success rate was 58%. When the support was enabled, the success rate increased to 80.3%.

Furthermore, there was also a decrease in the number of operational mistakes and execution time by using the proposed system.

The average time spent on a task without support was 1456.1 s (24 min), whereas the execution time with support was 835 s (13 min). The average of operational mistakes without support was 8, and the number of mistakes when support was enabled was only 3, which is more than two times less.

Since the view from each camera was outputted on a different display, participants had a difficulty in choosing which view is the most appropriate at a given time during remote control operation. An experienced operator can efficiently use static cameras to position the robot in environment and adjust the camera arm to the desirable view. However, it requires practicing and high concentration.

By using Gestalt psychology, we are able to determine ROI that will correspond to human operator choices. Our system selects the ROI for an operator that reduces the time required to analyse the image from cameras. Next, monitoring robots cooperatively determine the optimal observation angle and move to observe the selected ROI from an optimal angle of view. By dynamically providing the only relevant information from a suitable viewpoint, our adaptive visual support method helps operators to effectively teleoperate working robot in a shorter period of time.

From these results, we can say that the proposed system significantly enhanced the teleoperation process.

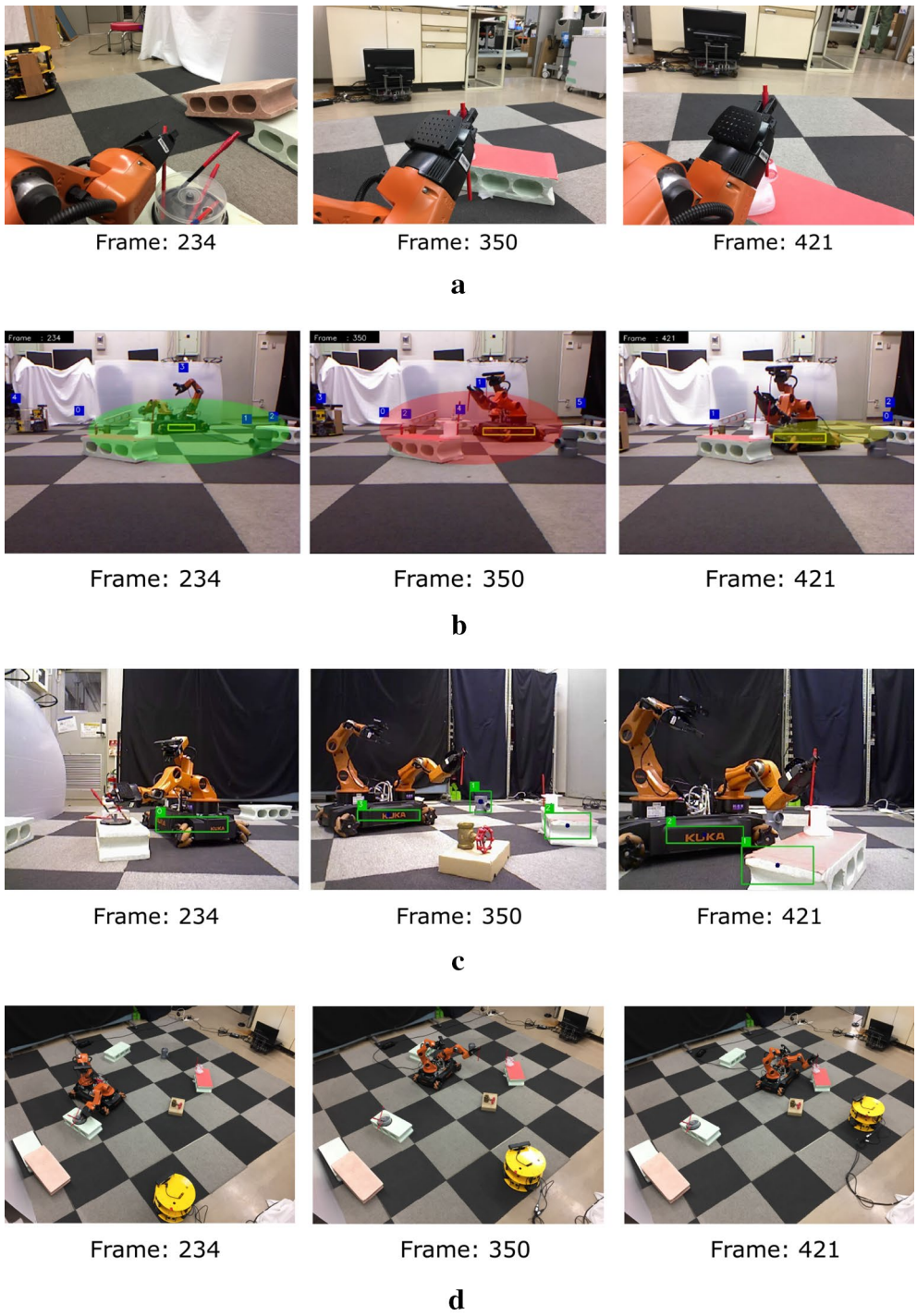

Fig. 14 Images sent by each robot. **a** View from working robot Robot-W, **b** view from Robot-M_W, **c** view from Robot-M_L, **d** view from top

However, owing to task complexity, it is difficult to completely eliminate the mistakes. The Peg-in-Hole process requires precise robot control. By providing the operator with both global and local views, we were able to decrease the time required for the operator to decide how to approach the target to pick or peg an object. To achieve accurate control, the operator must comprehend the information provided from the available viewpoints. This generally requires more time. Therefore, there is

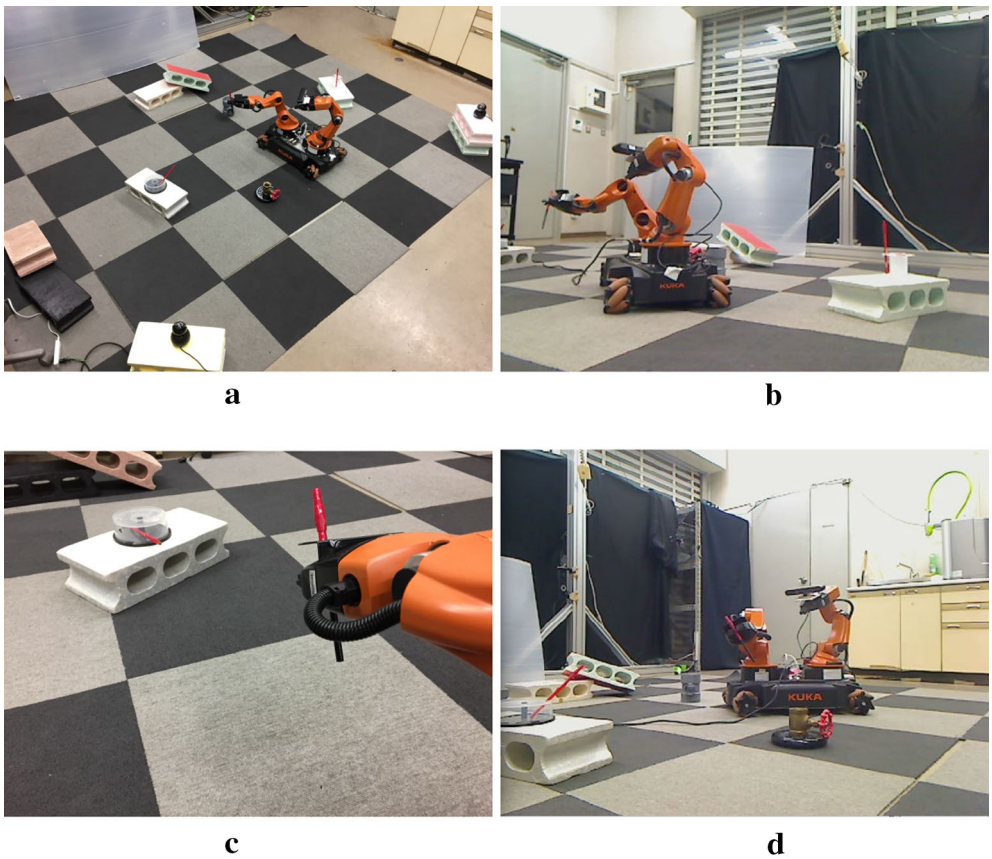

Fig. 15 Snapshot of experiment without support. **a** Working robot under operation, **b** view from static camera one, **c** view from working robot, **d** view from static camera two

a trade-off between the time spent and the number of mistakes.

Questionnaire survey

To evaluate the usability of proposed system we conducted a questionnaire survey. Our goal was to evaluate an overall user experience and also to analyze operator's feedback on proposed visual support system.

In total 6 laboratory members participated in the survey. The task scenario was the same as in the verification of improvement in operability experiment. Each participant had to control the working robot and perform Peg-in-Hole manipulation twice:

- Relying on camera arm + two fixed cameras.
- Relying on camera arm + developed support system.

After both of the tasks were completed, operators answered the questionnaire. Survey included questions on control experience and also various aspects of local view and wide view monitoring robots. Evaluation was made based on 5 point Likert scale:

- Strongly agree = 5.
- Agree = 4.
- Neutral = 3.
- Disagree = 2.
- Strongly disagree = 1.

Each participant had 4 trials and total number of questionnaires collected was 24.

First of all, participants were asked whether they felt that the remote control was easier and the distance perception improved when the proposed support system was used. Figure 18 shows the survey results. Regarding to distance perception more than 60% of responses were positive, with the average score of 3.83. Next, similarly the majority of the responses was in agreement with the statement that remote control became easier with the average score of 4.41, where 50% of the answers were in strong agreement with the statement.

Next, the participants evaluated various aspects of proposed system and how they used them during execution of chosen task scenario. Figure 19 shows the evaluation

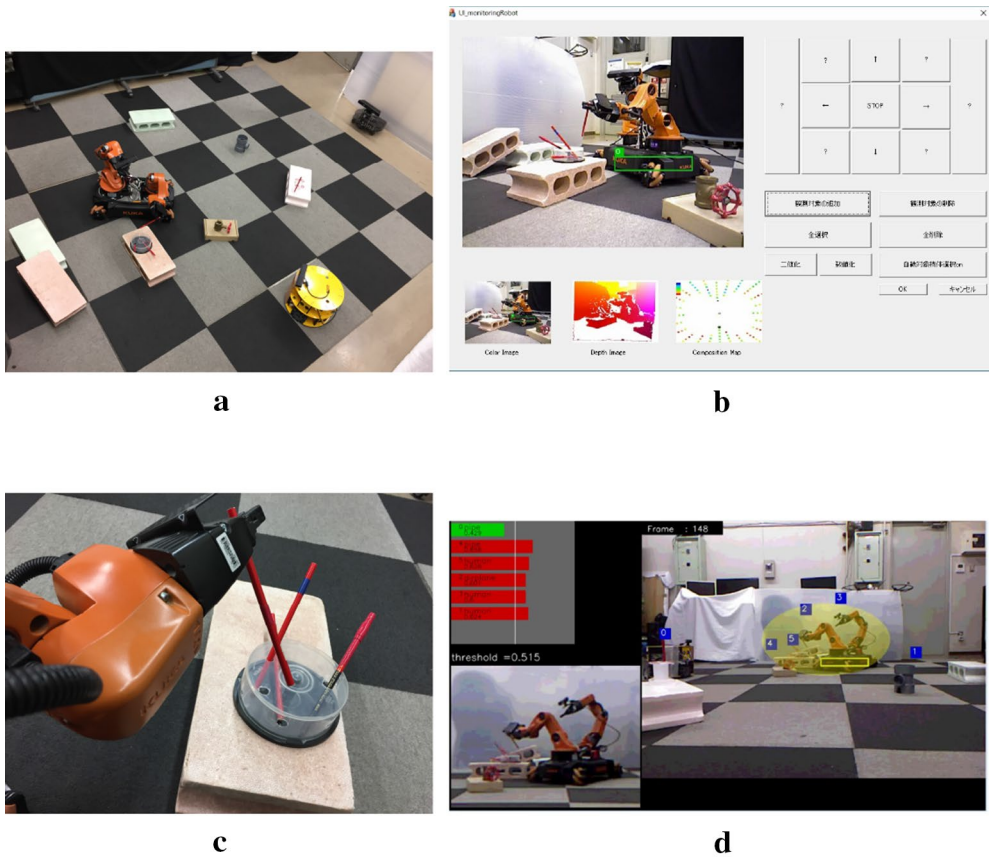

Fig. 16 Snapshot of experiment with support. **a** Working robot under operation, **b** control panel of local-view monitoring robot, **c** view from working robot, **d** view from wide-view monitoring robot

results for visual support provided by local view and wide view monitoring robots.

From Figure 19a, c we can see that operators mostly relied on wide view monitoring robot during Navigation with average score of 3.71 whereas only 25% of participants with score of 2.22 used wide view monitoring robot during Peg-in-Hole manipulation. Furthermore, the majority of survey results on ROI relevance and danger warning using color signals were also positive, with the average score of 3.59 and 3.29 respectively.

Figure 19b, d show the questionnaire results on local view monitoring robot evaluation. The number of positive responses on usage of local view monitoring robot was high not only for Peg-in-Hole manipulation but also for Navigation with the score of 4.42 and 4.62 respectively. The optimality of selected viewpoint was also evaluated, and the survey results show that slightly more than 50% of the answers were in agreement with the optimality statement with the average score equal to 3.33.

Furthermore, we also investigated the movement of local view monitoring robot and the problems regarding to optimal viewpoint selection. Figure 20 shows the survey results on the local view monitoring robot's movement and chosen viewpoints for two cases: during Navigation and Peg-in-Hole manipulation. As we can see from Fig. 20a, b, the monitoring robot mostly moved in desired direction of operator, with the percentage of positive responses equal to 70 and 60% respectively. However, the survey revealed that the robot's movement was not smooth and stable, since 50% of answers were in agreement with the statement that the robot's movement was jerky during both Navigation and Peg-in-Hole manipulation tasks, with the average evaluation scores equal to 3.3 and 3.42 respectively.

We also found out that during navigation the local view monitoring robot's observation was within an acceptable distance, because the participants disagreed with the statement of need to observe from closer range. On the other hand, in case of Peg-in-Hole manipulation, the number of positive and negative responses varied for the same statement regarding to the observation range: 40% negative, 30% neutral and 30% positive.

The opinion of participants also largely diverted on the statement whether the objects of their interest were in

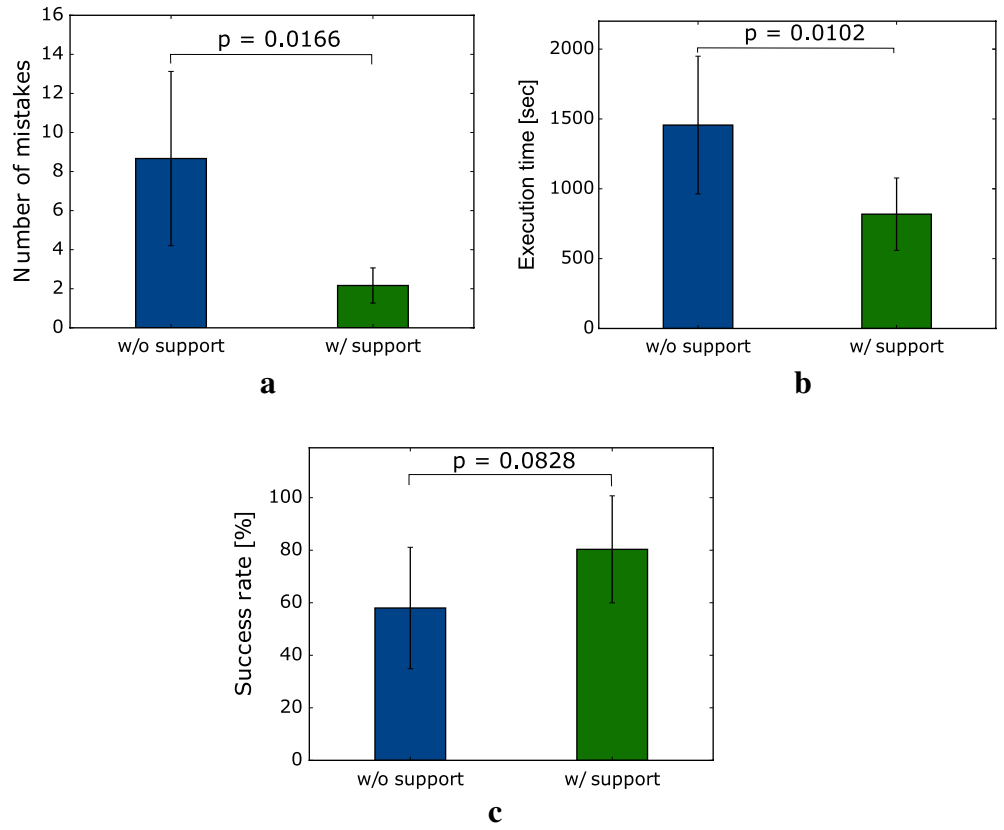

Fig. 17 Evaluation of effectiveness of proposed autonomous visual support system by execution time and number of errors. **a** Number of errors occurrence, **b** execution time, **c** success rate

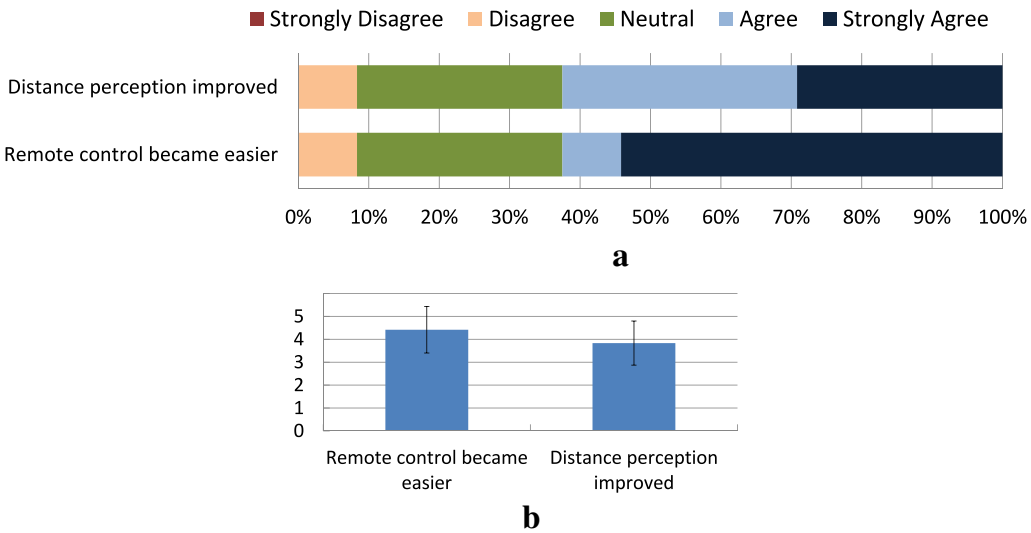

Fig. 18 Survey results on overall system evaluation. **a** Response frequency, **b** average score

the middle of the screen both during Navigation and Peg-in-Hole tasks. Some of the participants commented that particularly during the Peg-in-Hole task, they wanted the local view monitoring robot to focus more on working robot's end-effector rather than on target objects. Therefore, we need a further investigation on what kind

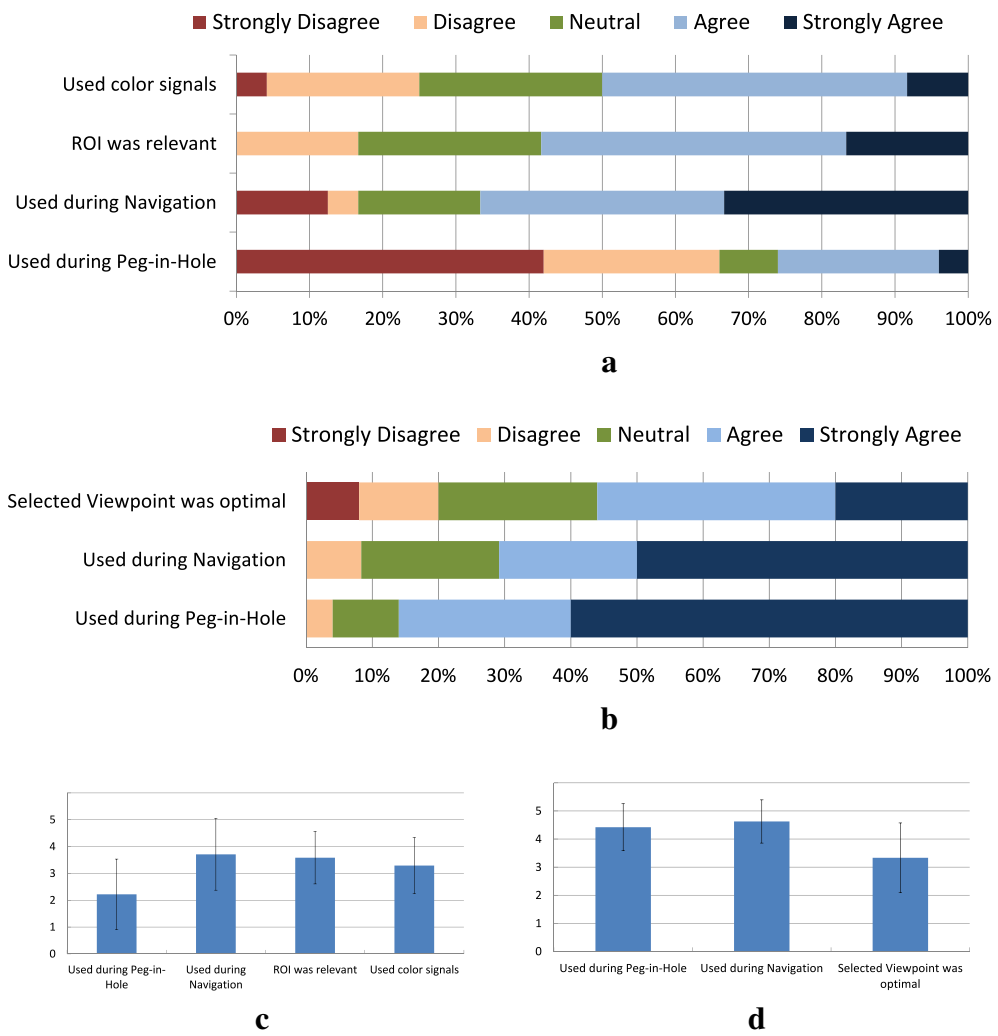

Fig. 19 Survey results on the evaluation of support by wide view and local view monitoring robots. **a** Response frequency for wide view robot, **b** response frequency for local view robot, **c** average score for wide viewrobot, **d** average score for local viewrobot

of visual support the operator wants during various task scenarios.

Conclusion and future work

In this paper, we proposed a support system with the "perception of latent interactions" and "situation estimation" as constituent elements to support remote operation. We realized an adaptive ROI selection system based on the perception of latent interaction between objects using the principles of gestalt psychology. We also developed a system that notifies the operator of danger by a situation estimation system using LDA.

By adaptively selecting ROI and sharing it monitoring robots can adapt to changes in environment and understand the situation around the working robot. Once ROI is determined, the monitoring robot moves to observe the scene from suitable angle. As a result, the proposed system provides an adaptive visual support to the operator by outputting the most relevant information from an optimal viewpoint. Experimental results show that the proposed system succeeded in reducing the number of errors and the operation time and increasing the operation's success rate.

A problem revealed during the experiments is a compromise between the execution time and the number of errors. To decrease the number of errors, careful remote control is required. However, to reduce the execution time, the operator must act faster, which may lead to operational errors. In real-life conditions, control accuracy can be vitally important compared to the time spent. Therefore, a certain time cost might be unavoidable.

From the participants feedback collected through questionnaire survey, it was revealed that more work on monitoring robots navigation is required. To enhance the remote control process, monitoring robots should

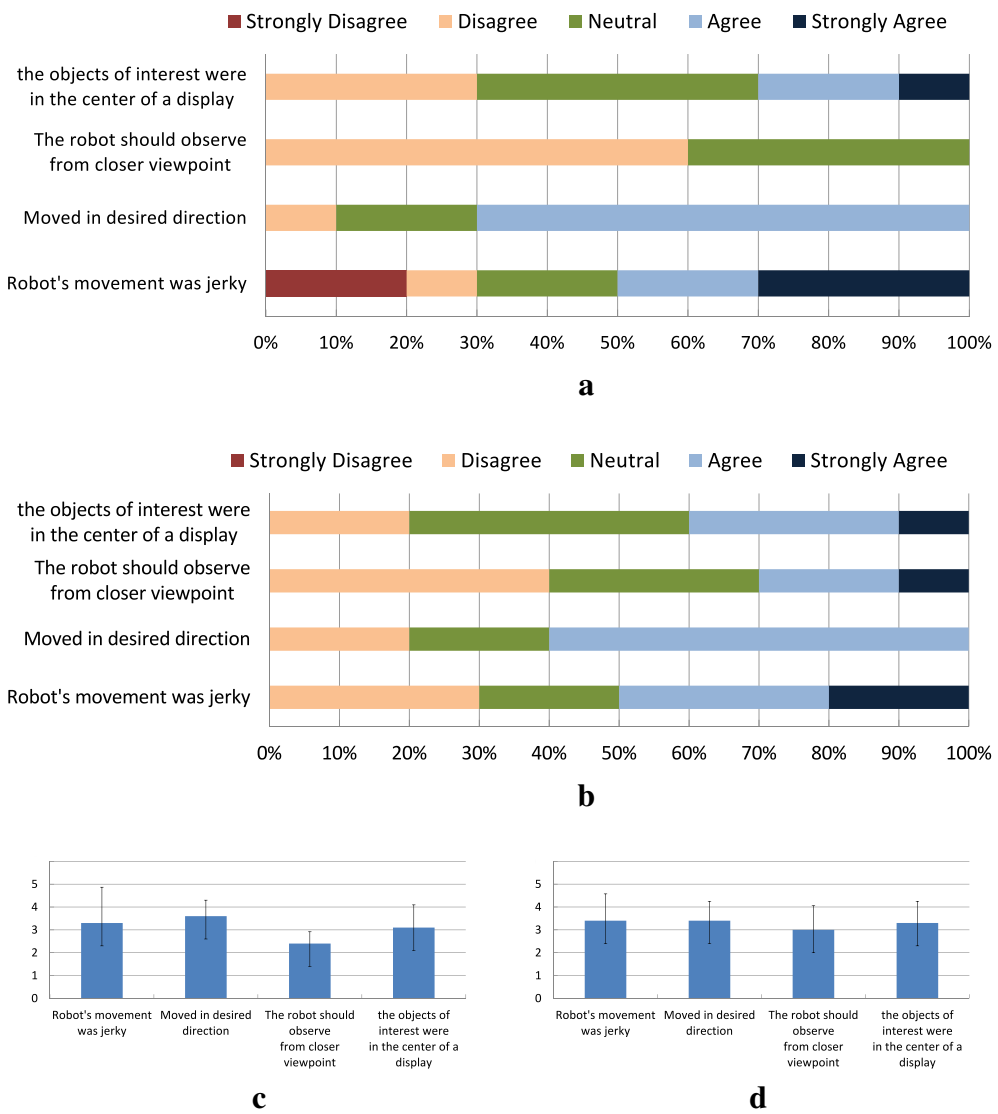

Fig. 20 Survey results on the evaluation of Local View Monitoring Robot's movement. **a** Response frequency for navigation, **b** response frequency for Peg-in-Hole task, **c** average score for navigation, **d** average score for Peg-in-Hole task

provide visual support not only from suitable angles but also move and adapt to environmental changes in a smooth motion with appropriate velocity. Therefore, to deploy the team of monitoring robot's in real-life environments, the velocity control and adaptability to changes in terrain is of great importance.

For our future work, to simplify the control task, we will develop a navigation system for remote control in a dynamic environment. Complex tasks require high concentration by the operator; hence, the longer the execution time, the heavier the mental load on the operator. A navigation system that adapts to the environment can decrease the execution time without affecting the accuracy of the remote operation.

Furthermore, to provide an intuitive control experience, we will integrate an intention evaluation method into our system. Depending on operator's skill level and also on the task scenario the requirements for visual support will vary. By evaluating the operator's intention, the monitoring robots can adjust the position to avoid unnecessary actions. Understanding operator's intention can help to develop a better user interface and to enhance the teleoperation through providing the most relevant information.

Authors' contributions
SS carried out the main part of this study and drafted this manuscript. KhF implemented all experiments and revised the manuscripts. KS contributed concepts and revised the manuscript. All authors read and approved the final manuscript.

Author details
[1] Department of Mechanical Science and Engineering, Nagoya University, Nagoya, Japan. [2] Department of Mechanical Engineering, Nagoya University, Nagoya, Japan. [3] Department of Micro-Nano Systems Engineering, Nagoya University, Nagoya, Japan.

Acknowledgements
This work was supported by JSPS KAKENHI Grant Number JP16H02880.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
Not applicable.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Funding
Not applicable.

Publisher's Note

Received: 24 April 2017 Accepted: 1 March 2018
Published online: 15 March 2018

References

1. Lim J, Lee I, Shim I, Jung H, Joe HM, Bae H, Sim O, Oh J, Jung T, Shin S et al (2017) Robot system of DRC-HUBO+ and control strategy of team KAIST in DARPA robotics challenge finals. J Field Robot 34(4):802–829
2. Johnson M, Shrewsbury B, Bertrand S, Wu T, Duran D, Floyd M, Abeles P, Stephen D, Mertins N, Lesman A et al (2015) Team IHMC's lessons learned from the DARPA robotics challenge trials. J Field Robot 32(2):192–208
3. Rohmer E, Yoshida T, Ohno K, Nagatani K, Tadokoro S, Konayagi E (2010) Quince: a collaborative mobile robotic platform for rescue robots research and development. In: Proceedings of the 5th international conference on the advances mechatronics (ICAM2010), RSJ. pp 225–230
4. Saltaren R, Aracil R, Alvarez C, Yime E, Sabater JM (2007) Field and service applications-exploring deep sea by teleoperated robot—an underwater parallel robot with high navigation capabilities. IEEE Robot Autom Mag 14(3):65–75
5. Barnes DP, Counsell MS (1999) Haptic communication for remote mobile manipulator robot operations. In: American Nuclear Society, proc. 8th topical meeting on robotics & remote systems. Citeseer
6. Nielsen CW, Goodrich M, Ricks RW et al (2007) Ecological interfaces for improving mobile robot teleoperation. IEEE Trans Robot 23(5):927–941
7. Rokunuzzaman M, Umeda T, Sekiyama K, Fukuda T (2014) A region of interest (roi) sharing protocol for multirobot cooperation with distributed sensing based on semantic stability. IEEE Trans Syst Man Cybern Syst 44(4):457–467
8. Piasco N, Marzat J, Sanfourche M (2016) Collaborative localization and formation flying using distributed stereo-vision. In: 2016 IEEE international conference on robotics and automation (ICRA), Stockholm, Sweden, pp 1202–1207
9. Kamezaki M, Yang J, Iwata H, Sugano S (2015) Visibility enhancement using autonomous multicamera controls with situational role assignment for teleoperated work machines. J Field Robot 33(6):802–824
10. Maeyama S, Okuno T, Watababe K (2016) View point decision algorithm for an autonomous robot to provide support images in the operability of a teleoperated robot. SICE J Control Meas Syst Integr 9(1):33–41
11. Lowe DG (1999) Object recognition from local scale-invariant features. In: The proceedings of the seventh IEEE international conference on computer vision, 1999, vol. 2. IEEE, New York, pp 1150–1157
12. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
13. Park E, Han X, Berg TL, Berg AC (2016) Combining multiple sources of knowledge in deep CNNS for action recognition. In: 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, New York, pp 1–8
14. Hamamoto K, Morooka K, Nagahashi H (2004) Motion recognition by combining hmm and reinforcement learning. In: 2004 IEEE international conference on systems, man and cybernetics, vol. 6. IEEE, New York, pp 5259–5264
15. Sarkar S, Boyer KL (1996) Quantitative measures of change based on feature organization: Eigenvalues and eigenvectors. Comput Vis Pattern Recognit 1996:478–483
16. Iqbal Q, Aggarwal JK (1999) Applying perceptual grouping to content-based image retrieval: building images. In: Computer vision and pattern recognition 1999, vol. 1. pp 42–48
17. Salton G, Fox EA, Wu H (1983) Extended Boolean information retrieval. Commun ACM 26(11):1022–1036
18. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
19. Filliat D et al (2012) Rgbd object recognition and visual texture classification for indoor semantic mapping. In: 2012 IEEE international conference on technologies for practical robot applications (TePRA). IEEE, New York, pp 127–132
20. Jia Y et al (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia. ACM, New York, pp 675–678
21. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, Real-Time Object Detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, Nevada, pp 779–788
22. Miller GA (1995) Wordnet: a lexical database for English. Commun ACM 38(11):39–41
23. Leacock C (1998) Combining local context and WordNet similarity for word sense identification. WordNet Electron Lex Database 49(2):265–283
24. Phan X-H, Nguyen C-T (2007) Gibbslda++: Ac/c++ implementation of latent dirichlet allocation (lda). Tech rep
25. Ito M, Sekiyama K (2015) Optimal viewpoint selection for cooperative visual assistance in multi-robot system. In: 2015 IEEE/SICE international symposium on system integration (SII), IEEE, New York, pp. 605–610
26. Samejima S, Sekiyama K (2016) Multi-robot visual support system by adaptive ROI selection based on gestalt perception. In: 2016 IEEE international conference on robotics and automation (ICRA). IEEE, New York, pp 3471–3476