# Human-Computer Collaborative Object Recognition for Intelligent Support

Kazuaki Kondo, Hideyuki Nishitani, and Yuichi Nakamura

Academic Center for Computing and Media Studies, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, Japan.
{kondo, nishitani, yuichi}@ccm.media.kyoto-u.ac.jp

**Abstract.** This paper introduces a novel framework for collaborative object recognition, which expands the applicability and improves the accuracy of object recognition. In this framework, a system not only recognizes targets but also detects and evaluates conditions that may make recognition difficult, and tries to resolve the situation by presenting the user with information on how to alter the conditions. The user can see how to make improvements, leading to correct recognition with little effort. The system can provide a useful, easy-to-use tool. In this research, a prototype system for kitchen scenes is designed, which can achieve situation evaluation and human-computer collaboration to improve recognition. We verified the framework by observing improvements in recognition accuracy and behavior of users in our experiments.

**Key words:** Human-computer collaboration, object recognition, and intelligent support system.

## 1 Introduction

Recently, applications of automatic recognition techniques have expanded rapidly due to advanced image-based recognition algorithms and high-speed computational processing. This suggests that these techniques can handle not only highly controlled environments but also more general settings in daily life. Recognition algorithms for such general situations should adjust to dynamic change of the many factors affecting the recognition target, e.g., appearance, lighting conditions, movement, occlusion, and distinctive features. These adjustments are particularly necessary when a system or observation target includes a person. For practical applications, human actions should not be heavily constrained, even if they often cause complicated situations and unexpected accidents, which prevent recognition. However, the presence of a person in the system has positive as well as negative aspects. For example, a human's flexible collaborative and cognitive abilities can simplify or facilitate the recognition task in some applications.

In this paper, we propose a novel framework for collaborative recognition as an approach to these issues. The framework enhances the performance of image-based recognition and expands its available applications with simple assistance provided by a user. This concept is particularly suitable for the construction of a smart system that intelligently supports human activities depending on the situation. A collaborative recognition framework can construct a mutually beneficial relationship between the user and the support system as described below. The system must accurately recognize the situation to provide the appropriate support expected or required by the user. The ability to provide this support is enhanced by his collaboration. Thus, users can benefit considerably by assisting in the recognition task. For similar reasons, a human-computer interface that regards the user himself as an input device is expected to be a remarkable application. The following two scenarios are examples of possible applications.

- **Smart support of cooking activity:**
  If the system quickly and accurately recognizes the situation and the stage of a cooking task in a kitchen setting, it can provide informative support to a user who is cooking, based on the results of object recognition. Recognizing ingredients and cookware on a cutting board enables the system to teach appropriate cutting methods. By understanding the stage the user has reached, the system can provide the procedure for the next step in advance. The collaborative recognition enhances such user support scenarios.
- **Appliance control interface using gesture:**
  Gesture recognition techniques have begun to be applied to control interfaces for electrical appliances such as TVs, air conditioners, and audio equipments. However, gesture recognition for various users in general settings is still impractical. The collaborative recognition framework requires, for example, the to user move slowly, adjust his posture, and face in a particular direction, so that his gestures can be accurately recognized as control commands.

Note that some applications, such as the detection of intruders or suspicious persons with surveillance cameras, are not suited to this approach, because no collaboration is expected in these cases.

The principal approach to effective collaboration is to provide appropriate information about the recognition status to the user. This is achieved by the following two functions.

- The system not only recognizes a target, but also evaluates the current situation to detect unfavorable situations and identify them as recognition malfunctions. These situations include recognition failure, inaccurate results, and problematic situations that interfere with recognition.
- When such situations are detected, the system proposes measures to improve them and elicits collaboration. These measures are presented to the user in an intuitive and easy-to-understand manner to reduce his cognitive and physical burden.

## 2 Related Studies

Several approaches related to collaborative recognition have been followed. Semi-automated approaches[4, 7] and semi-supervised learning[1] are conceptually similar to collaborative recognition, because they also require intelligent help provided by human to obtain accurate results. However, a collaborative recognition scheme has some different characteristics from those conventional approaches. First, a collaborative recognition system evaluates whether user's help is necessary based on degree of good condition and reliability of estimated results. Then the system adaptively asks users to assist when it is need, unlike the conventional approaches request it in ready determined timings. Contents of assists are also different. Conventional semi-automatic approaches mainly request cognitive and fixed tasks such as correctly labeling data. The collaborative recognition scheme requests adaptive tasks to the situation including physical support. Furthermore, since target applications are interactive scenarios as listed in the previous section, we should focus on how to design the real-time interactions.

A collaborative recognition system tries to improve its performance by providing feedback regarding the observed information. This concept is also used in active sensing approaches[5, 6], in which the next sensing method is determined by the current sensing results. However, differences between the two approaches arise with the presence of a user in a setting. For example, in collaborative recognition, many configurations allow the user to change the scene, he can provide intelligent collaboration and evaluation, he knows the correct answers in some cases, and he can skip unnecessary recognition tasks.

Interactive object recognition with an artificial agent[3], proposed by Ozeki et. al., is the approach most similar to our proposal. We theoretically and systematically discuss a collaborative recognition concept based on their early ideas to construct a general framework. In this paper, relations between recognition failure and degree of good condition, a framework of recognition improvement, and suitable interactions for well corporation are proposed and designed to apply the collaborative recognition scheme into actual recognition tasks.

The relationship between conventional one-way recognition and collaborative recognition is also interesting. These two approaches complement rather than conflicting with each other. As collaboration resolves unexpected situations and competent recognition reduces the degree of burden to collaborate, the combination can create an advanced recognition framework.

## 3 Collaborative Object Recognition

### 3.1 Human-Computer Collaboration Model

In this paper, we focus on collaborative object recognition in a kitchen setting for application to a cooking support scenario. The system's recognition framework is illustrated as a loop-back model including a user, as shown in Fig. 1(a). The recognition process recursively proceeds as follows: (A) the system recognizes a target object and a situation, (B) the system provides informational feedback to the user, and (C) the user improves the conditions that prevent recognition. The following assumed conditions are required for successful operation.

- **Assumption1:** The system can (uniquely) discriminate a target object under good conditions.
- **Assumption2:** The user can improve the conditions.

– **Assumption3:** The user can evaluate the results of object recognition.

One-way object recognition without information feedback satisfies only the first assumption; the performance deteriorates as environmental conditions deteriorate. Most conventional automatic recognition approaches correspond to this configuration. The second assumption is that the user can collaborate in the recognition task. It can be satisfied in many human support applications. The third assumption is easily satisfied due to the cognitive capabilities of humans. Of course, assumptions 2 and 3 both require the user to be present in the setting.

We designed smart information feedback to enhance the collaboration loop, in terms of burden on the user. We propose to reduce the two main burdens caused by imposed collaborations and inappropriate support based on incorrect recognition results. Unfortunately, these burdens have a trade-off relationship, as shown in Fig. 1(b), because the user's collaboration reduces the number of recognition failures. However, it should be possible to construct a better trade-off relationship in which easy collaboration addresses many recognition failures. The collaborative object recognition system installs the following two information feedback mechanisms to reduce the burden of collaboration.

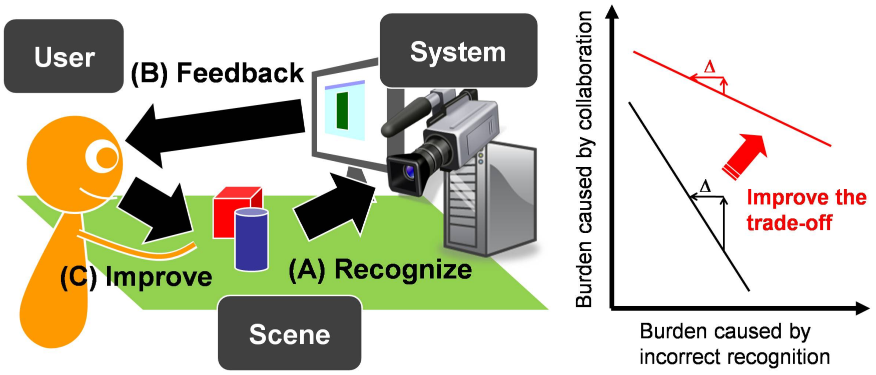
– **Recognition State Reporting** This function helps the user evaluate the need for collaboration. The state consists of recognition failures, inaccurate recognition results, problematic situations, and also instances of correct recognition. The first three indicate that the system has failed at recognition and requests the user's collaboration, which triggers the collaboration loop. Information on correct recognition reassures the user that collaboration is not required.
– **Improvement Measure Suggestion** When the above unfavorable situations occur, the system proposes measures for improvement to help the user evaluate and select collaborative activities. Because it is hard for a typical user who does not understand the recognition algorithm to determine how to act for recognition improvement, the measures are intuitively and concretely proposed to the user.

### 3.2 Recognition Improvement Framework

Here we explain a framework of recognition improvement using information feedback. Let $R$ be the degree of good conditions. It can be expressed as

$$R = f(\mathbf{x}), \quad \mathbf{x} = [x_1, x_2, ...] \tag{1}$$

assuming that the recognition algorithm does not change; $x_i$ represents factors that affect recognition performance, e.g., the state of a target object, user behavior, and environmental conditions. When $R$ is small, recognition probably fails. In this case, it is effective to estimate the change $\Delta\mathbf{x}$ that maximizes $R$ by a measure for improvement $S$, which can then be applied to the scene in order to achieve correct recognition.



(a) Loop-back model of collaborative object recognition

(b) trade-off between the two burdens

**Fig. 1.** Concept of collaborative object recognition

$$S = \Delta\mathbf{x}_{max} = argmax\{f(\mathbf{x}+\Delta\mathbf{x})\} \tag{2}$$

However, the collaborative recognition framework does not aim the best situation $R_{max}$ by $\Delta\mathbf{x}_{max}$ to certainly acquire correct recognition result, but an improvement on the current situation to boost the probability of correct recognition. This is because $\mathbf{x}$ is eclectic and highly complicated, preventing correct formulation of $f(\mathbf{x})$ and estimation of the optimal solution $\Delta\mathbf{x}_{max}$. Measures for improvement $\{S\}$ are expressed as a set of $\Delta\mathbf{x}$ that increase $R$.

$$\{S\} = \{\Delta\mathbf{x}; \Delta R = f(\mathbf{x}+\Delta\mathbf{x}) - f(\mathbf{x}) >> 0\}. \tag{3}$$

Although $\{S\}$ does not always achieve correct recognition with a single collaboration, only a few collaboration loops are enough. This assertion is based on the following two assumptions.

- The system can correctly recognize objects under good conditions even if they are not optimal.
- There are few obstacles to recognition.

These correspond to supposing that the first assumption listed in section 3.1 is satisfied at a high level, but they are not severe conditions when using recent advanced recognition techniques.

### 3.3 Information Feedback Algorithm

The above conceptual framework forms the following approximate algorithm for estimating a set of $S_i = \Delta\mathbf{x}$. First, it detects unfavorable situations in terms of 1. recognition failures and 2. problematic situations. Then, measures for improving each detected situation are evaluated. Then, measures for improving each detected situation are evaluated.

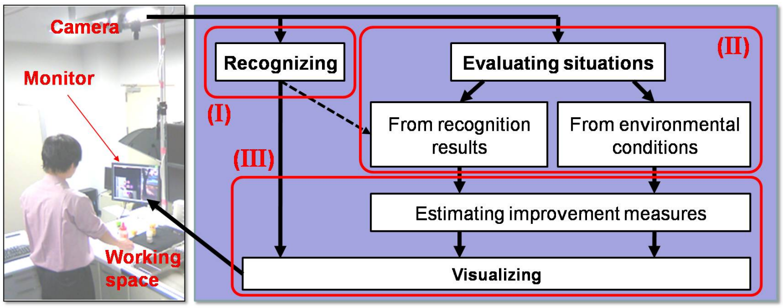


**Fig. 2.** Snapshot and processing flow of the prototype system. (I)-(III) are the implemented modules explained in section 4.2-section 4.4

1. Recognition failures $E_i(i = 1, 2, ..., m)$ are detected by evaluating the final and intermediate results of the main image-based recognition algorithm. Effective measures $S_{E_i} = \{S_{E_i1}, S_{E_i2}, S_{E_i3}, \ldots\}$ for recognition failure $E_i$ are selected from an already constructed database.
2. Problematic situations $\{C_k\}(k = 1, 2, \cdots, p)$, such as motion-based blurring, specular reflection, and occlusion, are simultaneously detected by other image processing algorithms that evaluate the environmental conditions and state of a target object. In a similar way to the recognition failure, measures $S_{C_k} = \{S_{C_k1}, S_{C_k2}, S_{C_k3}, \ldots\}$ are selected for $C_k$.

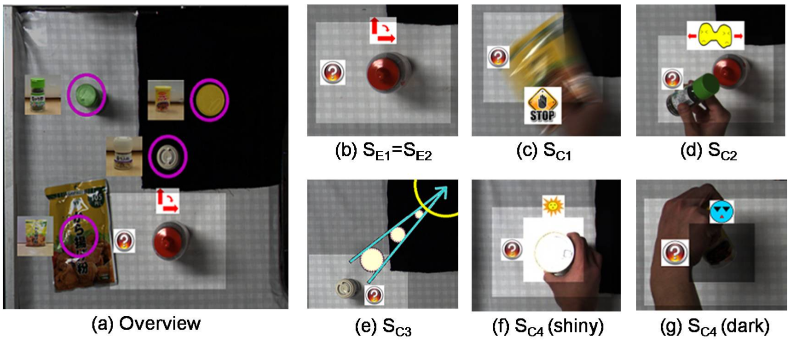
Then, choose one method $S_p$ most likely to be effective from the estimated $\{S\} = \bigcup_i S_{E_i}, \bigcup_k S_{C_k}$, and suggest it to the user.

Here we discuss the effects of $S_{E_i}$ and $S_{C_k}$ to choose the best one. Generally, $S_{E_i}$ does not always improve recognition, because a recognition failure does not reveal its obvious reasons. However, if the estimated measure is appropriate, it can directly eliminate the unfavorable situation, which improves recognition greatly. On the other hand, problematic situations can be accurately detected by the basic image processing algorithms, and their measures for improvement may not conduct correct recognition result because these are essentially for situation improvement, not for direct eliminating reasons for the failure. We select $S_{C_k}$ ahead of $S_{E_i}$ to make reliable improvements rather than unstable ones. Note that the selection priority also depends on the user's learning level, physical ability, and environmental constraints and so on.

## 4 Prototype System of Collaborative Object Recognition in a Kitchen Setting

### 4.1 System Overview

We implemented a prototype system based on the collaborative object recognition framework to validate its concept and performance. Object recognition in a



(a) Overview

(b) $S_{E1}=S_{E2}$

(c) $S_{C1}$

(d) $S_{C2}$

(e) $S_{C3}$

(f) $S_{C4}$ (shiny)

(g) $S_{C4}$ (dark)

**Fig. 3.** Example of displayed information feedback

kitchen setting was assumed. As we noted earlier, cooking activity is worth supporting and satisfies assumptions 2 and 3 in section 3.1. The prototype system performs accurate and quick object recognition under the following conditions. Objects are ingredients, seasonings, and cookware.

- Many objects enter and leave the scene.
- Unknown objects either do not appear or need not be recognized.
- Objects are moved by users' hands.
- Environmental conditions, object properties, and the relative positions of users' hands and objects may cause situations unfavorable for recognition.

We chose a compact camera and a flat-panel monitor as the sensing device and display device as shown in Fig. 2, respectively, so that the system can be easily implemented in ordinary kitchens. Images captured by the camera are processed to detect unfavorable situations and estimate measures for their improvement. The monitor presents feedback information in the form of simple illustrations and short explanations overlaid on the input image, as shown in Fig. 3. This feedback continues until a target object is uniquely discriminated. The modules shown in Fig. 2 are explained in detail in the following subsections.

### 4.2 (I) Recognition Algorithm

A recognition module distinguishes the target object by calculating similarities to all registered objects based on the generalized Mahalanobis distances in feature space. The features are image-based global and local features of the target object in the input image. We select the size, degree of circularity, and color histogram as global features, and SIFT descriptors[2] as local features. Objects were assumed to be already registered with their corresponding feature vectors. When the most likely candidate has a prominent similarity score, this indicates a unique but temporary recognition result. It becomes a determined result after the same unique result continues for a definite period of time, which is indicated on the monitor with a circular symbol. Other cases are regarded as recognition failures and indicated by a question mark.

### 4.3 (II) Detection of Unfavorable Situation

- **Recognition Failures**
  We assumed the following recognition failures, listed in Table 1 with their detection criteria.
  - **Multiple candidates ($E_1$):**
    When multiple registered objects have high similarity scores, even the most likely candidate is regarded as an inaccurate result. The multiple candidates are displayed as a recognition failure.
  - **No candidate ($E_2$):**
    When all similarities are below a given threshold, no candidate for the target object is considered to exist. The system displays a sign indicating that no registered object corresponds to the target.
- **Problematic Situations**
  Imaging processing algorithms other than the recognition algorithm detect various problematic situations. We considered the four problematic situations listed in Table 1, based on the performance and characteristics of the recognition algorithm. Those four situations are detected by the following image processing methods using distinct criteria.
  - **Object moves rapidly ($C_1$):**
    Track the target object and the user's hands to estimate their velocity and detect rapid movements. The hand tracking algorithm uses a human skin color model and geometric constraints on the user's arms.
  - **Objects are close or occluded ($C_2$):**
    Recognize positional relationships between the target object and other objects in the scene to detect adjacency or occlusion. A combination of the tracking algorithm mentioned above, foreground extraction, and checking feature fusions of multiple objects, are used to detect this situation.
  - **Object is similar to background ($C_3$):**
    When any shadow occurs near the regions once extracted as background, the background is regarded as foreground and considered to correspond to an object similar to the background.
  - **Object is too shiny or too dark ($C_4$):**
    Detect intensity saturation on the target object to evaluate the shine on the object. Darkness is evaluated by a similar algorithm.

  Extracting such situations from actual recognition accidents is also important, but not installed in the current implementation.

### 4.4 (III) Proposal of Measures for Improvement

Generally, several measures will be available to improve each unfavorable situation. In this paper, we assume the one-to-one correspondences listed in Table 1 to distinguish their causative links. Although a reasonable measure can correspond to each problematic situation, improvement of recognition failures depends on

**Table 1.** Expected unfavorable situations and their improving measures.

| $i$ | Recognition failures $E_i$ | Improving measures $S_{E_i}$ |
|---|---|---|
| 1 | Multiple candidates | Show another view of the object |
| 2 | No candidate | same as the above |
| 3 | Two candidates without any bad situation | Provide an alternate decision interface |
| $k$ | Difficult-to-recognize situations $C_k$ | Improving measures $S_{C_k}$ |
| 1 | Object moves rapidly | Stop it |
| 2 | Objects are close or occluded | Separate them |
| 3 | Object is similar to background | Put it on another background |
| 4 | Object is too shiny or too dark | Move it under good lighting |

an ambiguously estimated measure, as discussed in section 3.3. We evaluated the measures based on our experience as described below.

- **Multiple candidates or no candidate ($S_{E_{1,2}}$):**
  Multiple candidates and no candidate denote insufficient information for unique discrimination and a difference in appearance from the registered objects, respectively. The system suggests that the user shows another view of the target object.
- **two candidates without any bad situation ($S_{E_3}$):**
  Recognition may fail without a unfavorable situation being detected. This is a complex situation that only the user can understand. In this case, the system provides an alternate decision interface to the user. However, to limit the user's cognitive burden, it works only for a case with two candidates and has the lowest presentation priority.

Because the detection processes for the two types of unfavorable situations run independently and some problematic situations occur simultaneously, multiple measures for improvement may be proposed. In this case, the system selects one measure that has the highest presentation priority among all the proposed measures. The priorities $P$ of the measures are related in magnitude by $P(S_{C_1}) > P(S_{C_2}) > P(S_{C_3}) > P(S_{C_4}) > P(S_{E_1}) = P(S_{E_2}) > P(S_{E_3})$ in this implementation. The selected measure is displayed on the monitor, as shown in Fig. 3.

# 5 Experimental Validation

## 5.1 Configurations

In the experiment, a subject lets a recognition system discriminate a target object with his collaboration. Subjects use two recognition systems implementing the following approaches for the same recognition problem.

- **Proposed approach (collaborative recognition):**
  The prototype system described in section 4 recognizes a target object and displays information about the recognition status and measures for improvement.

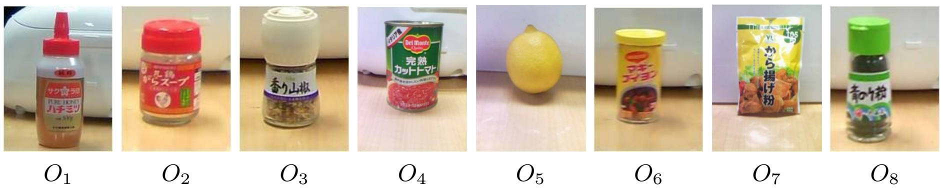


**Fig. 4.** Eight recognition targets used in the experiments.

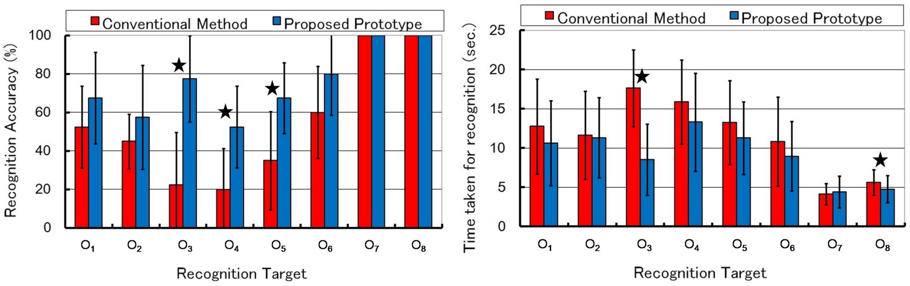
- **Conventional approach (one-way recognition):**
  This system discriminates a target object as the most likely candidate with the highest similarity and displays the result. Conventional one-way recognition actually does not report even the recognition result. It is reported to the subjects here, because they cannot perceive recognition failures if they do not receive any feedback, which is not a fair experimental configuration. However, they must evaluate the reasons for recognition failure and how to address them on their own with little information.

The evaluation targets are recognition performance and the effect of information feedback. Recognition performance is quantitatively evaluated by accuracy and time taken for correct recognition. These reflect the degree of the burden caused by incorrect recognition and that caused by collaboration, respectively. The effect of information feedback is qualitatively evaluated by observing and analyzing the subjects' behavior induced by the displayed information. The other experimental parameters are

- Twenty objects, such as ingredients and seasonings, were registered in advance.
- The eight recognition targets shown in 4 were selected from the registered objects. The same eight targets are used for all subjects.
- Subjects try to have the system recognize them one by one.
- Recognized objects were kept in the scene to create various unfavorable situations.
- Eight subjects, all novices at image-based object recognition and both recognition systems, each did five trials.

### 5.2 Results and Discussions

To provide a common criterion for evaluation, a successful recognition was defined as a correct discrimination of the target object within a time limit $L_t$. We assumed $L_t = 10$ s, considering the requirements for practical use. The recognition accuracy $r$ is calculated as $r = \frac{N_c}{N_{total}}$, where $N_c$ is the number of successes and $N_{total}$ is the number of trials. Figure 5(a) shows the $r$ values for each recognition target $\{O_i\}$. It shows that the recognition accuracies improved and that the collaborative recognition approach reduces the burden caused by



(a) Recognition accuracy

(b) Time taken for correct recognition

**Fig. 5.** Quantitative results of the experiments. Significant difference between two averages for each target object are signed with star markers, as results of t-test with 5% significance level.

incorrect recognition. Here we discuss the results for each object in terms of its characteristics. Objects $O_7$ and $O_8$ were both perfectly discriminated by both approaches, because their packaging designs contain rich textures, and no similar objects were registered, preventing unfavorable situations. Objects $O_3$ and $O_4$ are examples of difficult-to-recognize objects that are similar to the background and tend to be shiny. In most cases, the conventional approach failed to discriminate them because of difficulties in feature detection. Collaborative recognition smoothly eliminated such difficulties by proposing $S_{C_3}$ and $S_{C_4}$, which greatly improved the recognition accuracy. The upper views of some objects, including $O_1$, $O_2$ ($O_5$,$O_6$), are all red (yellow) circles without any texture. Additional views ($S_{E_{1,2}}$) or an alternate decision by the user ($S_{E_3}$) worked well for those objects. The recognition accuracy improved less than in the case of $O_3$ and $O_4$, because the conventional approach might stochastically select a correct answer from multiple candidates. Unfavorable situations $S_1$ and $S_2$ often occurred regardless of the target object. The subjects followed the suggestions for improving each situation, which seemed to affect the results in Fig. 5(a) considerably.

The time taken for correct recognition, shown in Fig. 5(b), reflects the degree of the user's burden because of collaboration. These results and those for recognition accuracy indicate a reduction in the burden caused by incorrect recognition and a relatively small collaboration burden. The collaborative recognition system successfully constructed an excellent trade-off relationship between them. This was a promising result, contrary to our expectation that users would require a long time to understand the presented information and act appropriately. However, the above consideration is optimistic and not entirely reliable, because subjects were not forced to hurry during the recognition trials.

The following discussion summarizes the analysis of the subjects' behavior when using our prototype system. All subjects noticed the feedback and tried to improve unfavorable situations accordingly. They spontaneously tried other measures that were not suggested by the system, e.g., removing the object and then returning it to the scene. There was few case, various unfavorable situations prevented recognition. In most of these cases, a few interactions resulted in correct recognition results. This result confirms that the implemented recognition algorithm and the experimental configurations satisfied the assumptions addressed in section 3.2. On the other hand, collaborations sometimes failed due to an unforeseen situation. For example, subjects became confused and continued to change an object's position or stare at the monitor inactively when a correct answer was not included in the displayed recognition candidates. They often had trouble eliminating an indicated unfavorable situation, because ineffective measures for improvement were proposed. Although the subjects stopped collaborating and were slow to react in the beginning, these cases decreased as they adjusted to the system and understood the meaning of the feedback.

## 6 Conclusion

We propose a novel framework for collaborative object recognition involving human-computer interaction. The key concept is to provide information feedback consisting of recognition status and suggestions for improving unfavorable situations. It elicits effective collaboration from a user while giving only a small amount to the user's burden. The framework was validated by experimental trials with a prototype system that simulates image-based object recognition in the kitchen. We are currently attempting to enhance the proposed framework with additional functions such as case-based detection of unfavorable situations and estimation of measures for improvement. We are also considering how to provide appropriate information feedback according to how the recognition is progressing and the user's learning level.

## References

1. Nigam, K., McCallum, A. K., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Machine Learning, Vol. 39, pp. 103-134(2000).
2. Lowe, D. G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110(2004).
3. Ozeki, M., Miyata, Y., Aoyama, H., Nakamura, Y.: Collaborative Object Recognition through Interactions with an Artificial Agent. In: International Workshop on Human-Centered Multimedia, pp. 95–101(2007).
4. Suh, B., Bedersona, B. B.: Semi-automatic photo annotation strategies using event based clustering and clothing based person recognition. Interacting with Computers, Vol. 19, Issue 4, pp. 524-544(2007).
5. Takemura, N., Miura, J.: View Planning of Multiple Active Cameras for Wide Area Surveillance. In: IEEE International Conference on Robotics and Automation, pp. 3173–3179(2007).
6. Shibata, M., Yasuda, Y., Ito, M.: Moving Object Detection for Active Camera based on Optical Flow Distortion. In: 17th World Congress 2008, International Federation of Automatic Control, pp.14720–14725(2008).
7. Guttmann, M., Wolf, L., Cohen-Or, D.: Semi-Automatic Stereo Extraction From Video. In: International Conference on Computer Vision, pp. 417-424(2009).