Interactions of Perceptual and Conceptual Processing: Expertise in Medical Image Diagnosis

Junya Morita [a,*] Kazuhisa Miwa [b] Takayuki Kitasaka [b]
Kensaku Mori [b] Yasuhito Suenaga [b] Shingo Iwano [c]
Mitsuru Ikeda [d] Takeo Ishigaki [c]

[a] *Graduate School of Human Informatics. Nagoya University. Furo-cho. Chigusa-ku. Nagoya. Japan.*

[b] *Graduate School of Information Science. Nagoya University. Furo-cho. Chigusa-ku. Nagoya. Japan.*

[c] *Graduate School of Medichine. Nagoya University. 65, Tsurumai-cho, Showa-ku, Nagoya. Japan.*

[d] *School of Health Science, Nagoya University. 1-20, Daikou-Minami 1. Higashi-ku. Nagoya. Japan.*

Abstract

In this study, we selected medical image diagnosis as a task to investigate how expertise influences the relations between perceptual and conceptual processing. In an experiment, participants, namely five novices and five experts, made diagnoses on thirteen CT images. We obtained two types of data concerning verbal protocols and manipulating computational systems. The segments related to perceptual and conceptual processing were extracted from these data, and the interrelations of the two components were analyzed. Consequently, we confirmed three salient features in the experts: (1) the experts verbalized more types of findings and more types of hypotheses than novices; (2) the experts generated several hypotheses in the early phases of the task; and (3) they newly verbalized many perceptual features during conceptual activities, and verbalized conceptual words during perceptual activities. These results suggest that expertise in medical image diagnosis involves not only the development of both perceptual and conceptual processing, but also the development of an ability to connect the two components.

Key words: Cognitive Experiment, Expertise, Protocol Analysis, Medical Image Diagnosis

1 Introduction

Human cognitive systems consist chiefly of two components: one for perceptual processing, which extracts information from the external world, and the other for conceptual processing, which retrieves and uses knowledge in the memory. Integration of the two components can be considered as a central foundation of human cognition (e.g., Nisser, 1976; Simon and Lea, 1978).

This topic, which has been mainly discussed in the fields of cognitive science and psychology, is beginning to become inseparable from the research field of human-computer studies because recent developments in information technology make it difficult to perform human cognitive tasks without computational devices. In a situation where one works with computational devices, the connection between these two components provides the basis for interactions between information presented by computational devices and knowledge retrieved from a human memory system.

This paper addresses how perceptual and conceptual processing are related to each other in a real-world cognitive task with computational devices. In particular, we focused on expertise as a factor in this relationship because the past studies repeatedly pointed out that expertise in a specialized field is one of the most influential factors in the differences affecting human cognition. We believe that investigations of the influence of expertise on the relations between perceptual and conceptual processing provide basic data for future developments of user interfaces tailored to individuals.

We chose medical image diagnosis (i.e., radiological diagnosis) as a task to address the above questions. In this task, a physician makes a diagnosis while viewing medical images such as radiographs, computed tomography (CT) images, or magnetic resonance imaging (MRI) images. We believe that the task is suitable for investigating the question because it is a typical cognitive activity involving human-computer interactions in which expertise influences the two components (perceiving features from medical images and retrieving physiological knowledge).

To clarify the goals of the present study, we briefly review (1) theoretical studies on the relations between perceptual and conceptual processing, (2) experimental studies on learning that alternates the relations of the two components, and (3) experimental studies on medical image diagnosis.

* Corresponding author. Present address: School of Knowledge Science. Japan Advanced Institute of Sceience and Technology. 1-1 Asahidai, Nomi, Ishikawa, Japan.
Email address: j-morita@jaist.ac.jp (Junya Morita).

1.1 Interactions of perceptual and conceptual processing

The relations between perceptual and conceptual processing have been widely discussed in psychology and cognitive science. They have commonly acknowledged the existence of mutual influence between the two components: the *bottom-up process* in which perceptual processing drives conceptual processing, and the *top-down process*, where conceptual processing drives perceptual processing.

For example, Neisser (1976) proposed the *perceptual cycle theory*, the aim of which was to combine the bottom-up and top-down aspects of human activities, and explained human activities in a complex and dynamic context. According to this theory, the human perceptual system is composed of an iterative cyclic process comprising three activities: extracting features from environments, remembering concepts (or schemata) from extracted features, and searching features in environments.

The models of scientific discovery (e.g., Simon and Lea 1978) or abductive reasoning (e.g., Johnson and Krems, 2001) also assume interactive cycles of components as a basis of human cognition. For example, Simon and Lea (1978) proposed *the dual-space search model*, which explored the interactions of data search and hypothesis generation. The dual-space search model has so far guided a large amount of studies employing computational and psychological methods (e.g., Dunbar, 1993; Klar and Dunbar, 1988; Klar, 2000; Kulkarni and Simon, 1988; Miwa 2004; Okada and Simon, 1997).

Related discussions can be found in the literature of cognitive architectures (ACT-R: Anderson and Lebiere, 1998; CPM-GOMS: Gray et al., 1993; or EPIC: Kieras and Meyer, 1997), which have recently been applied to complex human behaviors with computational devices, such as the manipulation of a modular phone, web-page searching, and flight operations (e.g., Anderson,et al., 2004; Brumby and Howes, 2004; Byrne, 2001; Fu et al., 2004; Salvucci, 2005; Taatgen, 2005). These architectures implement not only traditional modules of production systems but also perceptual modules to take external information into the systems. The cognitive process represented by such architectures is based on interactions between the perceptual and cognitive modules.

1.2 Learning and development of expertise in other fields

Based on the above theories, many experimental studies have been conducted to investigate the effects of learning on the relations of the two components. These studies have repeatedly confirmed the shift from the top-down to bottom-

up process as an effect of extensive training.

For instance, in the experiment conducted by Goldstone et al. (2000), subjects were required to learn correspondences from subtle visual features with conceptual categories. Through an extended period of training, the subjects recieved a conceptual feedback, and began to react immediately to a perceptual stimulus and to retrieve the category name directly. Such a learning process, called *perceptual learning*, could be considered as the shift from the top-down to the bottom up process.

Similar shifts have also been demonstrated in simulation studies which employ ACT-R architecture dealing with complex manipulations of computational devices. For example, Taatgen (2005) conducted a simulation study employing ACT-R architecture on the Air Traffic Control task, in which a model is required to both detect planes in the radar screen and identify the types of planes. In his study, the model learned the task through compiling declarative knowledge into procedural production rules. After the compilation, the model could directly evoke the production rules from the environment. Consequently, the time required to accomplish the task decreased significantly. The learning mechanism of ACT-R is characterized by shifts from the top-down process, in which declarative knowledge controls the firing of production rules, to the bottom-up process, in which production rules are directly evoked by the environmental information.

In addition to the above laboratory and computer-simulation studies, there are findings that confirmed the shift in the development of real-world expertise (Chase and Simon, 1973; Dreyfus and Dreyfus, 1986; Larkin et al., 1980; Patel and Groen, 1986).

For example, Larkin et al., (1980) investigated problem solving in physics, and confirmed that experts used forward reasoning based on highly compiled knowledge. In contrast, novices tended to use means-end analysis while frequently verbalizing abstract goals. It was also confirmed that in the case where the novices failed to solve the problem, they verbalized abstract physic laws in the early phase of problem solving.

Furthermore, Patel and Groen (1986) investigated a process of clinical diagnosis. In their experiment, expert physicians read texts in which clinical conditions of patients were described, and made a diagnosis on the patients. As a result of their experiment, it was shown that the process of physicians making an accurate diagnosis involved forward reasoning from findings to a diagnosis. On the other hand, in cases where the physicians made an inaccurate diagnosis, the process contained a backward reasoning strategy, beginning with a high-level hypothesis. This study indicates negative relations between a top-down process (i.e., means-end analysis strategy and backward reasoning)

and final performance in a task.

1.3 Expertise in medical image diagnosis

As the previous subsection showed, it has been repeatedly confirmed that there are the shifts from the top-down to the bottom-up process in the development of expertise. However, in the area of medical image diagnosis, it has also been confirmed that more complicated factors are involved in the development of expertise. So far, many researchers have conducted studies to identify cognitive factors in medical image diagnosis (see Woods, 1999a, 1999b, as reviews).

First, many researches have confirmed that expertise makes detection of abnormal regions fast and accurate. For example, Myles-Worsley et al. (1988) demonstrated that expert radiologists could discriminate abnormal X-ray films from normal ones within 500 msec. Also, Sowden et al. (2000) showed that expert radiologists could detect subtle changes of density in X-ray films, and explained that the learning process underlying medical image diagnosis is closely related to the mechanisms of perceptual leaning that have been extensively investigated in laboratory studies.

On the other hand, many researchers have agreed on the interactive aspects of a diagnostic process (e.g., Krupinski, 2003; Manning, Gale and Krupinski, 2005). For instance, Kundel and Nodine (1983) and Nodine and Kundel (1987), who conducted studies with eye movements data in X-ray film diagnosis, hypothesized an interactive model of the top-down and bottom-up processes. The other researchers conducted experimental studies that manipulate advance information such as clinical charts or advice from computational supporting systems (Alberdi et al., 2004; Crowley et al., 2003; Lesgold et al., 1988; Norman et al., 1992). For example, Norman et al. (1992) demonstrated that perceived features are dramatically influenced by clinical charts in which any previous disease of the patients is indicated.

Additionally, some studies directly pointed out positive relations between a top-down process and the final performance of a task. For example, Peterson (1999) observed that medical students who formed a hypothesis in the initial phases exceeded the other students in accuracy of final diagnosis. Similarly, Norman et al. (1999) conducted experiments to test the positive effects of a top-down process manipulating instructions that prompted medical students to make a hypothesis prior to observing electrocardiograms. From the results of their experiments, they confirmed the positive effects of the instructions, showing that the medical students who made a hypothesis were superior in their final diagnostic performance to the other students.

More importantly, Lesgold et al. (1988) demonstrated the changes of a di-

agnostic process due to the development of expertise by conducting protocol analysis studies for X-ray film diagnosis. Particularly, they confirmed that the process of expert diagnosis is characterized by an iterative cycle between the bottom-up and top-down processes, where an initial hypothesis is immediately triggered after a first glance at medical images, followed by a search for abnormalities in them.

1.4 The present study

According to the above previous studies, the development of expertise in medical image diagnosis involves not only a shift to the bottom-up process, but also a shift to the top-down or cyclic process. Apparently, these features of expertise are complicated compared with the findings on the learning process or the development of expertise in areas other than medical image diagnosis. In the present study, in order to understand this complexity, we conducted further protocol analysis studies on the development of expertise in real-world medical image diagnosis. Although many protocol analysis studies have been undertaken on medical image diagnosis (e.g., Azevedo and Lajoie, 1998; Lesgold et al., 1988; Raufaste et al., 1998; Rogers, 1996), our study is distinguished from those studies by three important differences.

First, we investigated the development of expertise in CT image diagnosis. CT images are cross-sectional images of a human body, and stacking them up reconstructs three-dimensional human anatomical structures. We chose this task because CT images have two important characteristics compared with radiographs (i.e., X-ray films), which the previous studies mainly used. The first characteristic of CT images is that they provide fine-grained pictures of the physical states of patients. Although specialized knowledge is required to interpret CT images, it would appear that it is a straightforward matter to extract the relevant features from them. Therefore, it can be considered that CT image diagnosis is a task that is unlikely to change the process of diagnosis resulting from the development of expertise. It is important for understanding the nature of interactions between perceptual and conceptual processing to investigate whether the shifts of the process could be found in the task of CT image diagnosis. The second characteristic is related to the environment in which CT image diagnosis is performed. In CT image diagnosis, physicians usually receive vast amounts of information through computer monitors and must make a large number of physical manipulations using computational devices. That is, the task is a typical cognitive activity that involves complex manipulations of computational devices. We think that it is important for us to understand the cognitive process behind such manipulations in order to develop future computer-aided diagnosis systems.

Second, we quantified the components of medical image diagnosis and attempted to demonstrate the effects of expertise on the relations of the two components. Past studies seemed to use qualitative methods of investigation to understand the process of medical image diagnosis. For example, some studies used protocol excerpts to discuss qualitative differences between experts and novices (e.g., Lesgold et al., 1988; Rogers, 1996), while others directly rated or coded types of strategy in verbal protocols data (e.g., Crowley et al., 2003; Peterson, 1999). So far, there have been only a few quantitative investigations into the interactive processes of perceptual and conceptual processing. However, in order to reach a scientific understanding of the process of medical image diagnosis, it is necessary to take an approach that quantifies verbal protocol data objectively. We believe that the objective understanding of the relationship between perceptual and conceptual processing will lead to a future framework for general human-computer studies. We also think that more fruitful relationships between users and computational devices could be constructed based on automatic methods of analyzing user behaviors. Therefore, we developed our own protocol analysis method, which analyzes data in an automatic way.

Third, in addition to the verbal protocols, we analyzed external activities involved in the process of diagnosis. As noted the above, CT image diagnosis involves complex manipulations of computational devices. Therefore, we assumed that a cognitive process involved in the task is partially externalized in the computational devices. In the present study, we captured manipulations of the systems, and distinguished types of activity concerning perceptual processing from activities associated with conceptual processing. We investigated how these two types of activity are interconnected in the process of medical image diagnosis. So far, there have been only a few studies conducting detailed analyses of these activities in medical image diagnosis. We think that analyzing these activities could contribute a deeper understanding of the relationship between perceptual and conceptual processing, leading to a theoretical framework for human-computer interactions in medical image diagnosis.

2 Method

In order to investigate medical image diagnosis in a realistic context, the experiment was performed in a room located in the radiology department at Nagoya University, where participants in our experiment usually work. The ambient room light was set to about 200 lux.

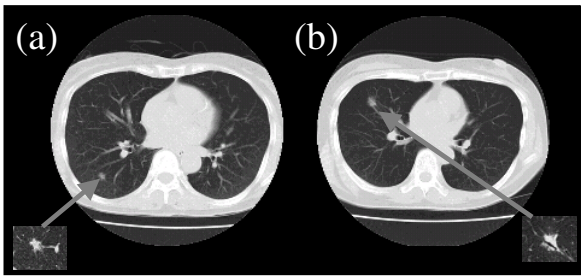


Fig. 1. Examples of case materials (a: an example of benign cases, b: an example of malignant cases). Arrows indicate locations of target lesion. The starting points of the arrows indicate enlarged images of the target lesions.

2.1 Participants

Ten participants were recruited from the radiology department at Nagoya University. They were divided into the following two groups.

- *Experts* ($n = 5$). These participants were radiologists, who each held an academic position in the radiology department. They had five to twenty years experience (*Mean*: 11 years) in medical image diagnosis.
- *Novices* ($n = 5$). These participants were residents and graduate students of the radiology department. They were physicians who had completed the degree of undergraduate medicine, and had less than two years experience in medical image diagnosis.

2.2 Task

The experimental task was to make "differential diagnoses of lung nodules (malignant/benign)." In this task, the participants were required to investigate nodular lesions and the overall state of the lung area, and to determine the pathological states of the lesions. However, determination at the differentiation level of the nodules was not required of the participants because this seemed to make the task extremely difficult.

2.3 Materials

2.3.1 Cases

We randomly chose case materials from a database, which consisted of cases whose diagnoses had already been determined by operations, biopsies, or follow-up examinations. All of the chosen cases contained at least one nodular lesion, and some of them included multiple lesions. In a later section the most

significant nodular lesion in each case was refereed to as the target lesion, which the participants were required to make a diagnosis on. They were not asked to make diagnoses on the other lesions, but they could use the other lesions to make a decision about the target lesions.

The selected cases consisted of eight benign and six malignant ones. The mean size of benign lesions was 11.64 mm (SD: 4.52), and the mean size of malignant lesions was 16 mm (SD: 2.77). The benign cases had been diagnosed as tuberculosis, organizing pneumonia, amyloidosis, or benign tumors, while the malignant cases included a variety of lung cancers, such as well differentiated carcinoma and squamous cell carcinoma.

2.3.2 CT datasets

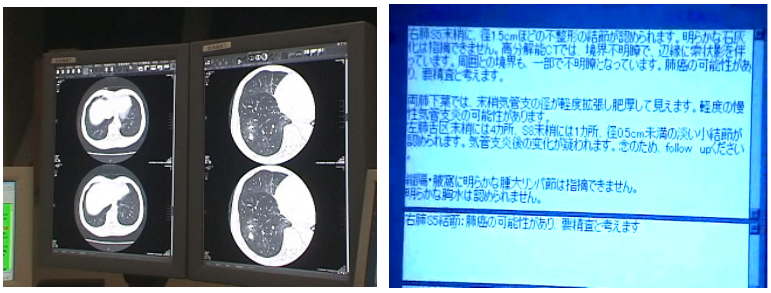
Each of the cases consisted of three types of CT datasets; we refer hereafter to these three types of CT data as *lung-window CT*, *mediastinal-window CT*, and *high-resolution CT*. Each type has the following specifications.

- *The lung-window CT*. This dataset includes CT slices with a window level of -600 H.U, a window width of 1800 H.U, and a slice thickness of 5-10 mm, and shows the overall lung area composed of 30 to 50 slices. By using this type of dataset, a physician can judge the location of a target lesion and observe base diseases of the lung area such as emphysema or interstitial pneumonia.
- *The mediastinal-window CT*. This dataset is the same as the lung-window CT dataset, except that display conditions are adjusted to show the mediastinal area clearly (window level, 50 H.U; window width, 300 H.U). By using this type of image, physicians can check for abnormalities in the mediastinal regions and the axillary regions.
- *The high-resolution CT*. This dataset focuses on a target lesion (resolution, about 300 μm; slice thickness, 0.5 to 2 mm).Usually, physicians use this type of image to investigate important features of a target lesion (*density, shape*) and its relations to lung tissues (*blood vessels, bronchi, and pleural membranes*).

2.3.3 Devices

In the experiment, the participants used the following two devices, which were the ones that they usually employed.

- *The device for viewing CT images* (see Fig. 2a).Two of the three types of CT datasets were presented on two LCDs: one is on the left and the other on the right. These were monochrome monitors with 256 gray levels and a resolution of 1200 $\times$ 1600 (Eizo Nanao Corporation, RadiForceG20). They



(a) The device for viewing CT images. CT images were presented on the two LCDs on the left and right.

(b) The device for writing medical reports. The system consisted of two text forms (findings, and impressions).

Fig. 2. Devices for the experiment.

were calibrated to the DICOM standard. In the experiment, the participants were able to freely change the types of CT datasets on the LCDs and choose the viewing distance without any constraints. In addition, each of the two LCDs was able to display a series of two slices, and the participant could select a left-right or right-left arrangement by using the workstation's mouse. The slices in the datasets are aligned from top to bottom, making it easy to observe three-dimensional structures of a patient's body by scrolling with the mouse wheel. The positions of displayed slices are indicated by the scroll bars on the LCDs. For example, the right LCD in Fig. 2a presents the slices in the high-resolution CT, displaying the upper side of a nodule. On the other hand, the left LCD in Fig. 2a shows the slices in the lung-window CT, displaying the under side of the lung area.

- *The device for writing medical reports* (see Fig. 2b). The medical reports were written using a computational device that provides two text forms. The form on the top was mainly used to write *findings* (i.e., observed abnormal features), while that on the bottom was used to write *impressions* (i.e., suspected diseases). In the experiment, all of the reports were written in Japanese.

2.3.4 Clinical histories

In this experiment, no clinical information other than CT images was presented because the previous studies indicated significant influence of clinical histories on the accuracy of diagnosis.

2.4 Procedure

The participants participated in the experiment individually. The experiment required a total of two to four hours, divided into the following four stages.

(1) *Instructions.* Each participant was given the following instructions: "imagine the situations where abnormal findings were detected as a result of screening tests. Your task is to make differential diagnoses of the detected abnormal findings." Following this, each participant was also instructed to verbalize all of their thoughts without filtering them.
(2) *Practice task.* Each participant made a diagnosis on one of the benign cases while being prompted to talk aloud. If the participant did not talk aloud for more than about ten seconds, the experimenter prompted the participants by an encouragement such as "please continue to talk aloud." The data obtained in the practice task were excluded from analysis.
(3) *Main task.* In the main task, each participant made diagnoses on thirteen cases that included seven benign and six malignant cases. In order to avoid order effects, the presentation order was randomized among the participants. For each case, the participants investigated the CT images and wrote a medical report about abnormal findings and suspected diseases. During the main task, all of the think-aloud protocols were recorded with a single MD recorder. Additionally, two digital video cameras captured the displays of the two devices.
(4) *Rating malignancy.* Following the diagnosis of each case, the participant was asked to rate how strongly s/he felt that the target lesion was malignant (0: absolutely benign to 10: absolutely malignant).

3 Data analysis

3.1 Recorded data

We obtained four types of data in the experiment: (1) the rating scores of malignancy; (2) the think-aloud data; (3) the texts written in the device for writing the reports; and (4) video records of the manipulations of the devices. For each type of data, we made the following assumptions: the first type of data was assumed as the final performance of diagnosis; the second, third, and fourth types of data were related to the process of diagnosis. In particular, verbalized contents such as the second and third types were assumed to represent the participant's thinking process, which can be analyzed by the protocol analysis method; and the fourth type of data was assumed to represent external activities in the task. Prior to presenting the results of the

experiment, we outline the analysis methods for each type of data.

3.2 Accuracy of diagnosis

In order to confirm the superiority of the experts in medical image diagnosis, we calculated the two types of score from a participant's rating score for malignancy. The first type of score is a d', a basic score given by signal detection theory. The score is computed for each participant by a set comprising the hit and false-alarm ratio ($d' = Z(Hit) - Z(FA)$). [1] This score has been used in many studies on medical image diagnosis. However, it is not the best index for the diagnostic performance because this index does not reflect how confident the participants felt with their judgments. Because of this limitation, we calculated A_z as a second type of score. It was calculated as the area under the receiver-operation curve (ROC) for each participant.

3.3 Protocol analysis

In order to investigate the cognitive process behind the final diagnostic performance, we conducted a verbal protocol analysis. However, the amount of verbal protocols data obtained in our experiment was so large that the traditional hand-coding protocol analysis was difficult in practice. To ensure reliable coding, and to conduct a detailed quantitative analysis, we developed a semi-automatic protocol analysis method, in which the Japanese morphological analysis system ChaSen (Matumoto et al., 2000) was used. ChaSen is a standard tool for text analysis and text mining in Japan. The system automatically converts plain texts to word sequences using dictionaries of words and grammar. In this analysis, we directly described semantic tags in ChaSen's word dictionary. The coding procedure comprised the following seven stages. [2]

(1) *Transcribing the data.* Think-aloud data and texts written in medical reports were transcribed. We synchronized the written texts in the reports to the think-aloud texts. When think-aloud texts were concurrently written in reports, the duplicated sentence was deleted. Following this, the

[1] The hit and false-alarm ratios were calculated by dichotomizing the ratings at the score of 6, except for the four participants who rated more than one case as neutral (the score of 5). For these four subjects, we divide the scores into three categories (the scores less than 4, the score of 5, the scores more than 6), and calculated d_e, which is an approximation of d' in the ROC analysis (Wickens, 2002).

[2] In this analysis we elucidated a single case of a single expert due to technical failures of the experimental devices.

texts were segmented into statements, which were time-stamped according to analysis of the recorded digital video data.

(2) *Morphological analysis (1).* The texts obtained through the above procedure were input into ChaSen. ChaSen then analyzed the texts with the default word dictionary, and output 104 473 words.

(3) *Selecting the words.* Most of the words output by the above procedure were syncategorematic terms (e.g., prepositions), or words that did not directly relate to the diagnostic activities (e.g., conjunctions, fillers). Therefore, these kinds of words were eliminated from the subsequent analysis.

(4) *Creating a new dictionary.* We created a new word dictionary comprising the words selected by the above procedure. Additionally, technical terms that were not appropriately discriminated by the default dictionary were registered into the dictionary.

(5) *Marking semantic tags.* A semantic tag was labeled in each of the words. The tags were divided into the following four main categories.
- *Percept.* This tag indicates a vocabulary of perceptual features, which can be directly observed from the CT images (323 words).
- *Concept.* This tag indicates a word concerning physiological or pathological features on the CT images, such as a disease name or a method of surgery (148 words).
- *Region.* This tag indicates a word concerning lung area or an organization of the lung, which is a technical term of anatomy (165 words).
- *Goal.* This tag indicates a word concerning the type of CT image or a word relating to the task that the physicians performed (19 words).

All of the labeling was performed by the first author. Following this, the third author labeled all of the words with the above tags independently. The first and third authors agreed on 91% of the tags. Furthermore, the sixth author, who is an expert radiologist, checked the labeling from the viewpoints of radiological validity.

Of the four main categories, we focused on *Percept* and *Concept*. These two categories can be regarded as verbal outputs of perceptual and conceptual processing, respectively. Thus, the first author divided the words tagged as *Percept* and *Concept* into several *subcategories*, which represent dimensions of perceptual features or detailed semantic meanings (Percept: Density / Shape / Number / Inside / Size / Category / Relation / Dist. / Others; Concept: Malignant / Benign / Others / Artifact / Surgery / Clinical / Forward / Judge). We also divided the words tagged as the two main categories into several *objects* that indicate anatomical regions mainly used by the words (Nodule / Lung / Br. / Overall / Others). The labels of *objects* are related with the labels of *Region*, but more closely connect perceptual and conceptual processing. Definitions of these tags are shown in Table 1.

(6) *Morphological analysis with the new dictionary.* After deletion of the default dictionary, morphological analysis was again conducted with the new dictionary. ChaSen with the new dictionary output 13 984 words.

Table 1
Definitions of subcategories and objects. The numbers of words registered in the dictionary are in parentheses. Br. stands for "the bronchus".

	Names of tag	Definitions (examples)
Subcategories of Percept	Density (36)	Types of words meaning density or thickness (e.g., Ground-Glasse-Opacity, unevenness, brightness).
	Shape (89)	Types of words meaning shape or silhouette of nodules (e.g., borderline, spicula).
	Number (24)	Types of words meaning the number of objects (e.g., single, multiple, many).
	Inside (9)	Types of words related to qualities of inside nodules (e.g., solid pattern, pneumatic, cavity).
	Size (35)	Types of words related to size (e.g., large, centimeters, thick).
	Category (21)	Types of words related to medical categories of densities (e.g., mass, nodule, cyst).
	Relation (54)	Types of words meaning relations among multiple objects (e.g., catch up, cramp up, pass over).
	Dist. (27)	Types of words meaning distributions of multiple objects (e.g., granular, sectional).
	Others (28)	Types of words that could not be allotted to the above categories (e.g., choke up, salient).
Subcategories of Concept	Malignant (22)	Types of words that are possible for use in diagnosis of the target region, and indicating lung cancer (e.g., lung cancer, adenocarcinoma, carcinoma).
	Benign (31)	Types of words that are possible for use in diagnosis of the target region, and indicating diseases other than lung cancer (e.g., lung cancer, tuberculosis).
	Others (38)	Types of words that refer to disease other than nodules (e.g., pneumonectasia, interstitial pneumonitis, heart infarction).
	Artifact (13)	Types of words that refer to densities caused by artifacts of X-ray photography (e.g., gravity, breath).
	Surgery (15)	Types of words that refer to densities caused by artifacts of X-ray photography (e.g., post surgery, cut off).
	Clinical (16)	Speculations on clinical conditions of a patient (e.g., young, smoking, man).
	Forward (7)	Future treatment of a patient (e.g., follow-up, biospy).
	Judge (13)	Types of words related to a participant's knowlege or judgement (e.g., benign-or-malignancy).
Object	Nodule (284)	Perceptual features or disease related to nodules (e.g., spicula, lung-cancer).
	Lung (48)	Perceptual features or diseases that do not consist of nodules (e.g., satellite region, emphysema).
	Br. (18)	Perceptual features or diseases related to the bronchus (e.g., occlusion, bronchial infection).
	Others (66)	Perceptual features or diseases related to the mediastinal region, abdominal region, or armpits (e.g., hepatic cysts, fatty liver).
	Overall (55)	Ambugious words that could not be allotted to the above categories (e.g., abnormal, disease, pathologize).

Table 2 shows an example of the output words.

(7) *Marking with New.* After completion of the above procedures, we marked the words that had not appeared in the previous word sequence with a tag, *New* (see the seventh column of Table 2). This tag indicates the initial appearance of words in the process of diagnosis. In addition, we used the tag to count the number of unique words in each case for each participant.

3.4 Analysis of the external activity

As noted earlier, CT image diagnosis involves manipulations of computational systems. Therefore, the cognitive process involved in the task could have been investigated by analyzing the manipulation of the systems. However, in the present experiment, we could not directly obtain the manipulation log from the systems because the systems were set up in a room used for normal medical services. It was not allowed to install any software to obtain the manipulation log. Therefore, we used the digital video data, which was captured by the camera fixed on the device for viewing CT images, to analyze the manipulations of the systems. The analysis procedures comprised the following five stages. [3]

(1) *Automatic detections of changing CT slices.* As shown in Fig. 2a, scroll bars on the LCDs have a relatively high density compared to the background, making it possible to detect positions of the scroll bars using a simple threshold processing method. Fig. 3 presents results of the processing. The line graph in the figure corresponds to the movement of the scroll bars. The horizontal axis corresponds to the positions of the scroll bars, while the vertical axis represents the task timelines (minutes).

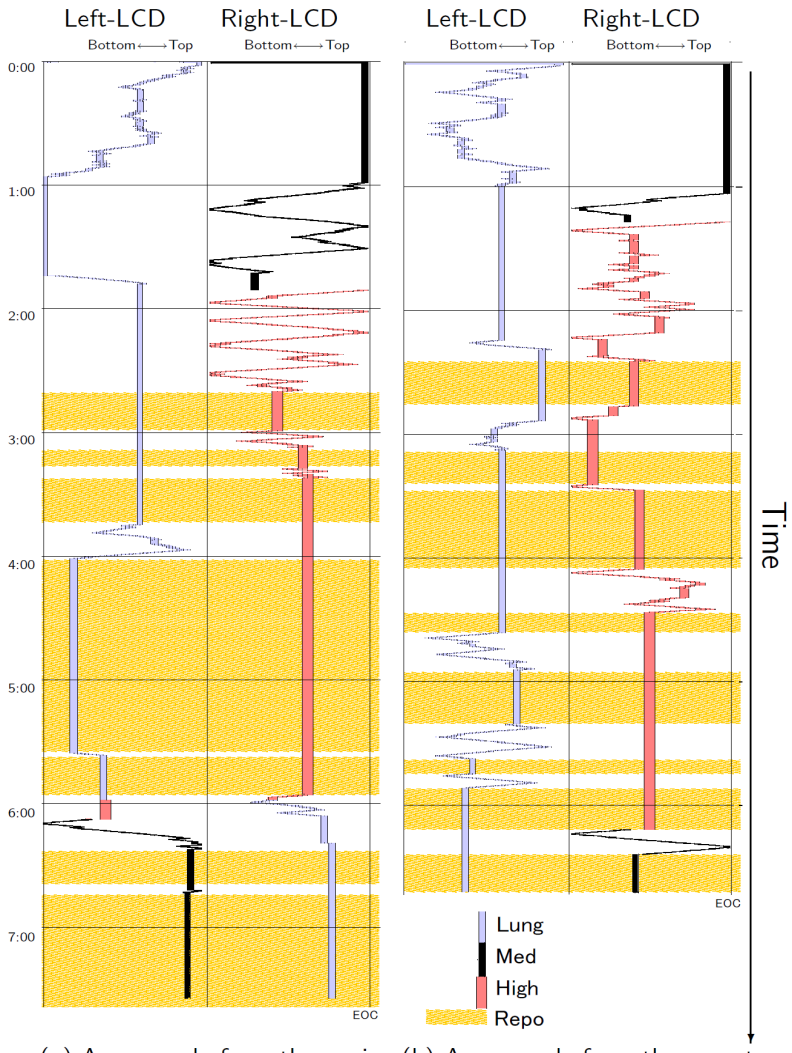
(2) *Automatic detections of CT dataset switching.* After drawing a graph for each case, we detected the time points in which a different type of CT dataset appeared on the LCD. As shown in Fig. 3, there are gaps in the line graph (e.g., about 1:40 in the right display for the novice, and about 6:20 in the right display for the expert). These gaps represent the time points in which the participant switched CT dataset types on the LCD.

(3) *Coding of the CT dataset types.* For each time-point detected by the above procedure, the types of switched CT datasets were coded by observing the original video data. In Fig. 3, the CT dataset types are distinguished by line colors: the light gray lines represent the lung-window CT; the dark gray lines represent the mediastinal-window CT; and the black lines represent the high-resolution CT.

[3] In addition to the above single case of a single expert, another case of a single novice was excluded from this analysis due to technical failures of the DV cameras.

Table 2
Examples of coding for verbal protocol data. The first column shows time points of verbalizations for each word. The second to sixths columns show outputs of ChaSen (Main = main category, Sub = subcategory). The seventh column shows *New*, which was coded after morphological analysis.

Time	Word	Synonymous words	Main	Sub	Object	New
25	S8	S8	Region	Lung		new
25	S10	S10	Region	Lung		new
26	right of	right	Region	Lung		new
26	S10	S10	Region	Lung		
32	light	light	Percept	Density	Nodule	new
32	nodular density	nodule	Percept	Category	Nodule	new
42	spiculation	spicula	Percept	Shape	Nodule	new
42	surrounding area	surround	Percept	Dist.	Lung	new
42	attach	attach	Percept	Relation	Overall	new
48	surrounding area	surround	Percept	Dist.	Lung	
48	GGO	grand-glass-opacity	Percept	Density	Nodule	new
48	attach	attach	Percept	Relation	Overall	
52	part of	localized	Percept	Dist.	Lung	new
52	GGO	grand-glass-opacity	Percept	Density	Nodule	
55	pleura	pleura	Region	Lung		new
55	connect	connect	Percept	Relation	Overall	new
64	atelectasis	atelectasis	Percept	Others	Br.	new
64	attach	attach	Percept	Relation	Overall	
•	•	•	•	•	•	•
•	•	•	•	•	•	•
•	•	•	•	•	•	•
325	abnormal	abnormal	Concept	Others	Overall	new
325	largement	largement	Percept	Size	Others	
336	mm	mm	Percept	Size	Nodule	
336	largement	largement	Percept	Size	Others	
336	transfer	transfer	Concept	Malignant	Nodule	new
336	possibility	possibility	Concept	Judgement	Others	new
351	pleural effusion	pleural effusion	Region	Lung		
369	bronchus	bronchus	Region	Br.		
369	calcification	calcification	Percept	Density	Nodule	
372	lymphatic node	lymphoglandula	Region	Mediastinum		
372	calcification	calcification	Percept	Density	Nodule	
379	malignant	malignant	Concept	Malignant	Nodule	
391	lung cancer	malignant	Concept	Malignant	Nodule	
399	right lung	right lung	Region	Lung		
399	S8	S8	Region	Lung		
399	nodular density	nodule	Percept	Category	Nodule	
399	lung cancer	malignant	Concept	Malignant	Nodule	
418	light	light	Percept	Density	Nodule	
418	nodular density	nodule	Percept	Category	Nodule	



(a) An example from the novices(b) An example from the experts

Fig. 3. Examples of analyses of external activity (a: An example from the novices, b: An example from the experts). The vertical axis represents the task timeline (minutes), and the horizontal axis represents positions of the scroll-bars in each display. Types of activity are distinguished by lines and background colors.

(4) *Semi-automatic detection of which LCD the participants focused on.* We semi-automatically detected on which of the LCDs the participants focused. We considered the display in which the scroll bar was moving as the LCD under focus. Additionally, we considered that when neither of the two scroll bars moved, the participants were manipulating the device for writing the reports. In Fig. 3, the backgrounds of such sequences are painted in gray.

(5) *Categorizing external activity.* By taking the above steps, we categorized each participant's activity into the following four types.

- *Observing the lung-window CT.* This type of activity was defined by the time points at which the lung-window CT was displayed on the LCD under focus. At these time points, the participant was assumed to have

been observing the lung-window CT.

- *Observing the mediastinal-window CT*. This type of activity was defined by the time points at which the mediastinal-window CT was displayed on the LCD under focus. At these time points, the participant was assumed to have been observing the mediastinal-window CT.
- *Observing the high-resolution CT*. This type of activity was defined by the time point at which the high-resolution CT was displayed on the LCD under focus. At these time points, the participant was assumed to have been observing the high-resolution CT.
- *Writing the reports*. This type of activity was defined by the time points at which none of the two scroll bars moved. At these time points, the participant was assumed to have been writing the report.

We used the above categories to analyze the relationship between perceptual and conceptual processing. First, we distinguished the first three types of activity (observing each CT dataset) from the last type (writing reports) and then assumed that the former was related to perceptual processing and the latter was related to conceptual processing. Specifically, the first type (observing the lung-window CT) was considered as an activity to observe overall impressions of the lung, the second type (observing the mediastinal-window CT) was an activity to check abnormal features in regions other than the lungs, and the third type (observing the high-resolution CT) was an activity to observe the nodule intensively. The assumption comes from normative analysis about the experimental task. In the experiment, the participants observed the physical states of patients through the two LCDs. In other words, they could not observe the features without manipulating CT datasets displayed on the LCDs. Moreover, when the participants wrote reports, they summarized perceptual features that had already been found in earlier activities, and they made decisions about what disease was affecting the patients. In the experiment, the participants were required to write their decisions about the disease in the final reports. That is, in the task, the novel perceptual features were found in observing each type of CT dataset, and the final decisions were made in writing reports.

3.5 Statistical tests

As a result of the above analysis, we obtained several dependent measures for each case of each participant (10 subjects $\times$ 13 cases). We aggregated the dependent measures across participants to generate a mean for each case. The statistical tests for all the dependent measures, except for the accuracy scores (d', A_z), used case materials as the units of analysis ($n = 13$). For all tests, we set the significance level at .05, marking effects with * ($p < .05$) or ** ($p < .01$). In the tests for the final task performance, we used one-tailed distributions

Table 3
Task performance [Mean (SD)]

	Novices	Experts	One-tailed t-test
d'	0.661 (0.729)	1.392 (0.273)	$t(8) = 1.87*$
A_z	0.630 (0.106)	0.792 (0.106)	$t(8) = 1.77$
Required Time (seconds)	548.87 (86.50)	494.95 (116.58)	$t(12) = 2.13*$

because it would be by definition unlikely that the novices performed better than the experts. In the other tests, we used two-tailed distributions.

4 Results and discussion

4.1 Task performance

Table 3 summarizes the differences in task performance between the experts and the novices, showing the two types of accuracy score and time required to finish each case. The difference of d' between the experts and the novices reached a significant level, indicating that the experts made their diagnoses more accurately than the novices did. In addition, we obtained marginally significant differences in A_z between the experts and novices ($p = 0.057$). As for the time to finish a case, we also confirmed that the experts made their diagnoses faster than the novices did. Taken together, these results indicate the superiority of the experts over the novices in the performance of medical image diagnosis.

In the following subsections, we show the results concerning the process of diagnosis underlying the final performance. Sections 4.2 and 4.3 respectively present the results concerning verbal protocol analysis and the analysis of external activity. In section 4.4 we show results of the analysis of the relations between the two types of data.

4.2 Results of verbal protocol analysis

4.2.1 Overall pattarns of tagged words

Outlining the results of the verbal protocol analysis, Fig. 4 presents the number of words tagged in the four main categories. The figure describes two values in each bar: the upper value represents the mean of the total number of words

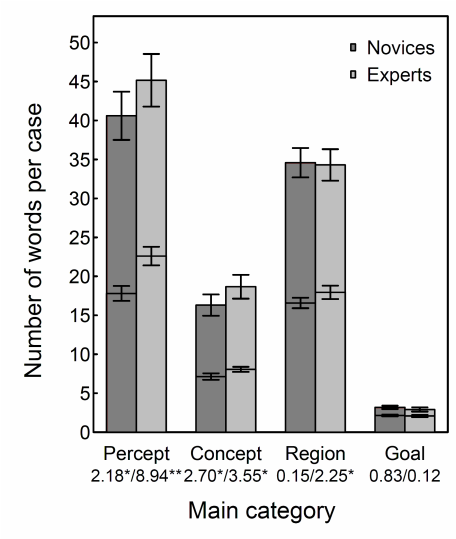


Fig. 4. Main categories of verbalized words. The upper value of each bar represents the mean number of verbalized words, while the lower value of each bar represents the mean number of newly verbalized words (marked as *New*). The error bars represent a one standard error of the mean. Values under each label indicate t-values for the difference between the experts and the novices (two-tailed, paired, $n = 13$). The left and right values indicate the results for the number of total words (the upper values of the bars) and new words (the lower values of the bars), respectively.

per case, while the lower value denotes the mean number of words that were marked with *New*.

The upper values in the figure demonstrate that the experts exceeded the novices in the total amounts of words tagged as *Percept* and *Concept*.[4] Such results imply that expertise improves their ability to verbalize their thinking process concerning perceptual and conceptual processing. Furthermore, the results confirm that the differences were not caused by verbalizations of duplicated words, but by verbalizations of varieties of words, because the same differences in expertise were observed in the lower values of the figure.

Roughly speaking, the above results are consistent with previous findings in the field of medical image diagnosis. For examples, Lesgold et al. (1988) confirmed that experts verbalized more types of finding and more hypotheses than novices. Rufaste, Eyrolle, and Mariné (1998) also revealed that semantic networks constructed from experts' verbalizations were richer than those from novices' verbalizations.

[4] Since there were differences among the four categories in the number of words registered in the dictionary, we did not conduct a statistical analysis for comparison among the categories.

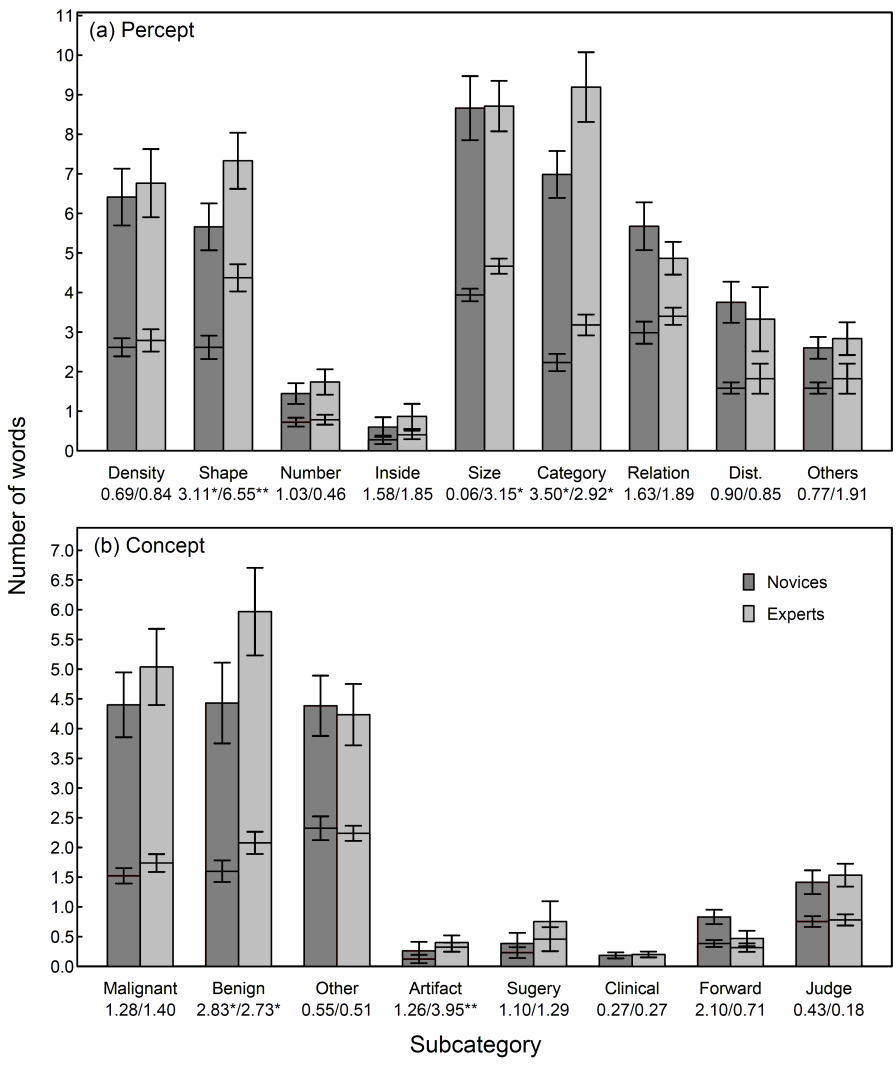


Fig. 5. Subcategories of verbalized words (a: Percept, b: Concept). The notations of values are the same as in Fig. 4.

4.2.2 Development of expertise in each of perceptual and conceptual processing

As noted earlier, we divided the words tagged as *Percept* and *Concept* into further categories such as *subcategories* and *objects*. Using these coding results, we show detailed features of expertise in each of perceptual and conceptual processing.

Fig. 5 shows the numbers of words tagged as each *subcategory*. The results for the *subcategories* of *Percept* (Fig. 5a) confirmed that the experts verbalized more words tagged as *Shape*, which represents overall features of nodules, and *Category*, which is defined by multiple perceptual features. These results suggest that the development of expertise changes the dimensions of perceptual features that they focused on. The results are consistent with the findings

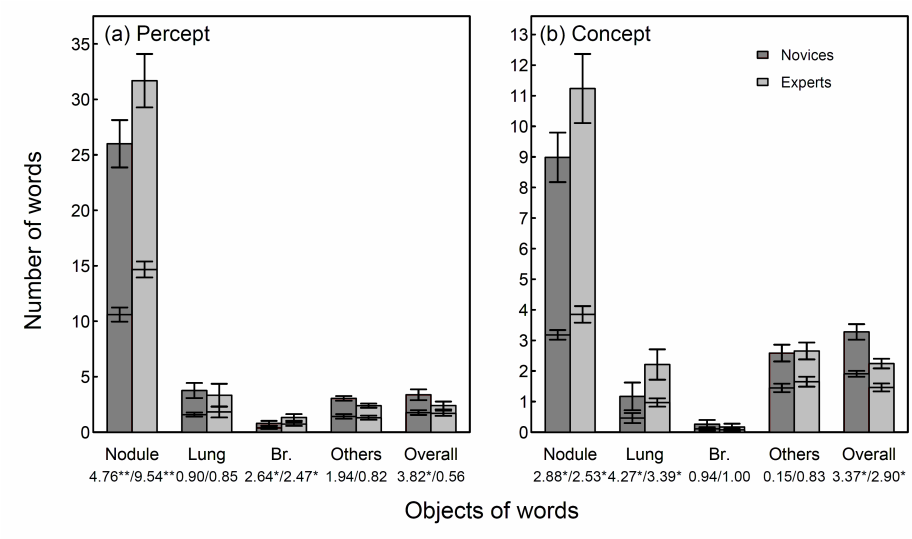


Fig. 6. Objects of verbalized words (a: Percept, b: Concept). The notations of values are the same as in Fig. 4.

obtained in the studies on perceptual learning (e.g., Goldstone et al., 2000), which indicate that the dimensions that are focused on change dramatically through extensive training in a category learning task. For the *subcategories* of *Concept* (Fig. 5b), we confirmed that the experts exceeded the novices in *Benign*. Since there is no significant difference in *Malignant*, the result indicates that the novices paid relatively less attention to the possibility of a target lesion being benign.

Fig. 6 shows the numbers of words from the viewpoint of *objects*. Roughly speaking, common patterns can be observed in *Percept* (Fig. 6a) and *Concept* (Fig. 6b): the experts verbalized more words tagged as *Nodule* than the novices, whereas the novice verbalized more words tagged as *Overall* than the experts. This expert-novice difference can be explained by the requirements of the task. Since the task was differential diagnosis of lung nodules, the results imply that the experts tuned their process to the specifics of the experimental task. This interpretation is also consistent with the findings obtained in Lesgold et al. (1987), in which experts limited their efforts to the task-relevant regions. Furthermore, the results indicate that the differences observed in Fig. 4 were not caused by verbalizations about task-irrelevant lesions.

4.2.3 Time transitions of verbalized words

Our main interest in this investigation is to understand the relations between perceptual and conceptual processing in medical image diagnosis. In order to understand how the two components are interrelated in the process of diagnosis, we investigated transitions in the amount of verbalization that occurs with the development of a diagnosis. In this analysis, we defined four phases of diag-

nosis, dividing the time to finish diagnosing each case into equal time intervals. The problem with this analysis is that the time lengths of the phases were not the same for each of the different cases and different participants. However, the aim of this analysis was to extract patterns of the relations between perceptual and conceptual processing. Although this analysis has limitations, we think that dividing the process by equal intervals is very useful for detecting patterns from the data.

Fig. 7 shows the number of words tagged as *Percept* (Fig. 7a), and *Concept* (Fig. 7b) in each phase. As the previous graphs illustrated, Fig. 7 depicts two types of values: the upper value represents the total number of words, while the lower one denotes the number of newly verbalized words in each phase. For each of the measures, the results of statistical tests could be summarized as follows.

(1) *The total number of words tagged as Percept* (The upper values in Fig. 7a). For both the novices and the experts, the values decreased from the initial phase to the final phase. The novices' value decreased from Phase 1 to 4 gradually whereas the experts' value decreased sharply from Phase 3 to 4.
(2) *The number of new words tagged as Percept* (The lower values in Fig. 7a). Similar to the above results, a decreasing pattern with time was confirmed for both the novices and the experts. Additionally, it was confirmed that the experts' value decreased more gradually, exceeding the value of the novices in Phase 2 to 4.
(3) *The total number of words tagged as Concept* (The upper values in Fig. 7b). For both the novices and the experts, the values increased from Phase 3 to 4. In addition to this, the experts' value decreased from Phase 1 to 2.
(4) *The number of new words tagged as Concept* (The lower values in Fig. 7b). The experts' value decreased from Phase 1 to 2. On the other hand, there was no difference across phases for the novices.

The above four results revealed characteristics of the novices and the experts in the relations between perceptual and conceptual processing. The novices decreased their verbalizations of *Percept* from the initial to later phases (see 1, 2) and increased their verbalizations of *Concept* in the later phases (see 3). This pattern indicates a one-way, bottom-up process, in which the outputs of perceptual processing are sent to the conceptual processing. On the other hand, the salient features of the experts appear in the verbalizations of *Concept* in the initial phase (Phase 1), which are distinguished from the verbalizations of *Concept* in the last phase (Phase 4). In Phase 1, the experts verbalized many words classified into *Concept* not only from among the total number of words (the upper value) but also from among the number of new words (the lower value). In Phase 4, however, they verbalized many words

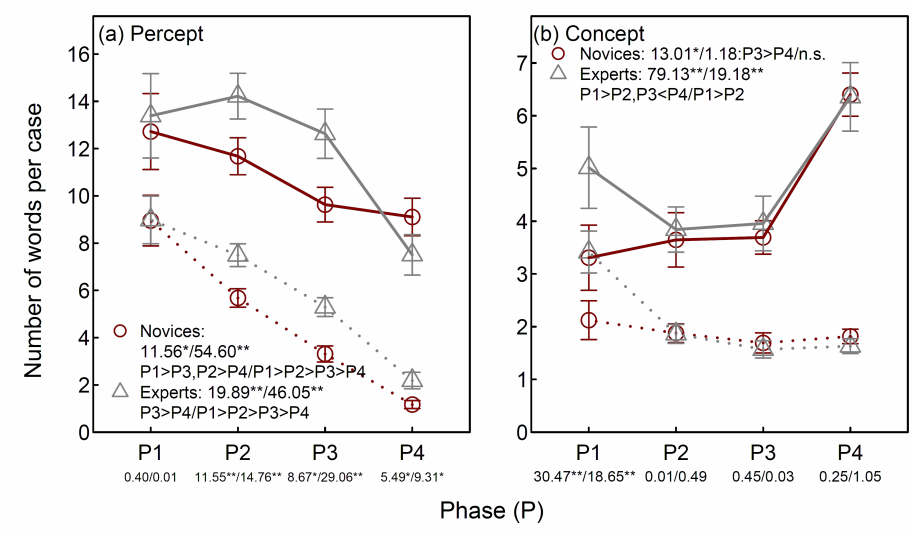


Fig. 7. The task timelines of verbalizations (a: *Percept*, b: *Concept*). The upper (solid) and lower (broken) lines represent mean numbers of total words and new words, respectively. For each value of each tag type, 2 (the expertise) $\times$ 4 (the phases) analysis of variance (ANOVA) revealed a significant interaction of the expertise and the phases: the number of *Percept*, $F(3, 36) = 6.50$*; the number of new *Percept*, $F(3, 36) = 4.24$*; the number of *Concept*, $F(3, 36) = 3.01*$; the number of new *Concept*, $F(3, 36) = 6.96$*. Following the above effects, simple main effects of expertise at each phase were investigated. The results (F values, $df = 1, 12$) are noted under the labels. The left and right values indicate the results for the number of total words and new words, respectively. Similarly, the simple main effects of the phases at each group (F values, $df = 3, 36$) are indicated in the legends for each of the number of total words and new words. In addition, the legends indicate the pairs that have the least difference among the pairs in which significant differences were confirmed by Tukey HSD.

classed in *Concept* from among the total number of words, but not from the number of new words. Therefore, the predominant appearance of *Concept* in Phase 1 can be considered as multiple initial hypotheses, whereas the higher frequencies of words classified into *Concept* in Phase 4 can be considered as a few alternative hypotheses that had already been refined in the previous phases. These characteristics of the experts' process can be interpreted as the cyclic process of perceptual and conceptual processing: this finding is consistent with the previous studies on medical image diagnosis (e.g., Lesgold et al., 1988).

4.2.4 Summary

We analyzed the verbal protocols data using the quantitative methods in which semi-automatic segmentations and tagging were conducted. The analysis successfully reconfirmed the expert-novice differences observed in the previous

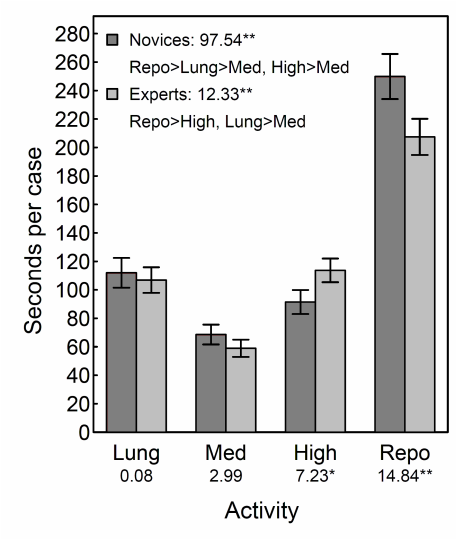


Fig. 8. The amount of each type of external activity; *Lung* = observing the lung-window CT, *Med* = observing the mediastinal-window CT, *High* = observing the high-resolution CT, and *Repo* = writing the reports. A 2 (the expertise) × 4 (the activity types) ANOVA revealed a significant interactions of the expertise and the activity types, $F(3, 36) = 12.33^{**}$. Similar to Fig. 7, the simple main effect of expertise at each activity type (F values, $df = 1, 12$) is shown under each label. In addition, the legend indicates simple main effects of activity types at each group (F values, $df = 3, 36$) and results of multiple-comparisons (Turkey-HSD).

studies, revealing the effects of expertise not only in the total amount but also in the time transitions of verbalizations. However, there are some limitations, requiring further analysis of the above results. It is reasonable to consider that the process of medical image diagnosis involves an implicit process that is difficult to verbalize. In particular, think-aloud methods cannot be considered sufficient for investigations of perceptual processing. Ericsson and Simon (1980) considered that one of the difficult domains to apply the think-aloud method is a task requiring perceptual processing, and they discussed the possibility that the method changes cognitive processes. It was also implied from Fig. 4 that the differences may depend on the differences in ability to verbalize the thinking process between the experts and the novices. This verbalization ability could affect the patterns of verbalization in Fig. 7. Taking the above limitations into account, we considered that it is necessary to investigate the relations between the perceptual and conceptual processing using additional methods that compensate for verbal protocol analysis.

4.3 Results on the external activity

4.3.1 Case studies

Prior to presenting quantitative results, we qualitatively discuss how the participants performed in the task. Fig. 3, explained in an earlier section, shows qualitative features of activity, indicating common and differential patterns of the examples.

The common patterns of the examples are order relations of activity: first, both the expert and the novice observed the lung-window CT slices from the top to bottom; second, they checked the mediastinal-window CT, and observed the high-resolution CT. After around three minutes, they started writing the reports, and repeated writing reports and observing images until they had reached the end of the case.

One of the differences between the examples is a type of activity on which they spent proportionally more time. Although the total time to finish diagnosing the case was longer for the novice, he started writing the report at almost same time as the expert did. This indicates that the novice proportionally spent more time writing the report. More interestingly, compared with the novice, the expert frequently repeated writing the report and observing the images, as can be observed from the figure in which the diagram representing the expert has many breaks in the gray background. Importantly, this impression is consistent with the results shown in Fig. 7, suggesting that the experts used a cyclic process of perceptual and conceptual processing.

4.3.2 Overall patterns of external activity

In order to present the overall tendencies of external activity, we present Fig. 8, which depicts the time (seconds) to engage in each type of activity. As the figure illustrates, the novices spent more time writing the reports, whereas the experts spent proportionally more time observing high-resolution CT images. These results are consistent with the impressions observed in Fig. 3. In addition, we can find a correspondence between the above results and those in Fig. 6. Since high-resolution CT is the dataset focusing on target nodular lesions, it can be argued that the experts tuned their methods of diagnosis to the specifics of the experimental task (differentiation of lung nodules).

4.3.3 Time transitions of external activity

As noted earlier, we assumed that the types of activity, defined in our study, correspond respectively to perceptual and conceptual processing. Therefore,

investigations on transitions of activity would shed light on the relations between perceptual and conceptual processing. Fig. 9 shows the transitions occuring in each activity. From each of the four figures, the following observations can be made.

(1) *Observing the lung-window CT* (Fig. 9a). The values for the experts and the novices peaked at Phase 1, and decreased from Phase 1 to 2. It was also confirmed that the value for the novices was higher than that for the experts in Phase 1.
(2) *Observing the mediastinal-window CT* (Fig. 9b). There were remarkable differences between the experts and the novices. While the value of the novices peaked at Phase 2, that of the experts was distributed to Phases 1 and 4.
(3) *Observing the high-resolution CT* (Fig. 9c). The novices' value decreased from Phase 2 to 3. In addition to this effect, the experts' value increased from Phase 1 to 2. The effects of expertise were confirmed in Phases 3 and 4, indicating that the experts exceeded the novices in the later phases of diagnosis.
(4) *Writing the reports* (Fig. 9d). Both the novices and the experts increased their values from Phase 1 to 4. Additionally, the effects of expertise were confirmed in Phases 3 and 4. Contrary to the results obtained from the high-resolution CT, the novices exceeded the experts in the later phases of diagnosis.

The above results confirm the impressions from Fig 3 that the order relation of activity for both the experts and the novices (from Lung through Med or High to Repo). Interestingly, this relation seem to be more rigid in the novices because the novices more intensively observed the lung-window CT in the earlier phase (see 1), and spent more time writing the reports in the later phases (see 4). The rigidity of the novices is also supported by the results of observing the mediastinal-window CT (see 2), in which the value of the novices peaked at a single phase. More interestingly, there was a contrast in the effects of expertise between observing the high-resolution CT (see 3) and writing the reports (see 4). This contrast indicates that the experts spent relativity more time observing the high-resolution CT in the later phases in which writing the reports was a dominant type of activity. That is, the contrast suggests that the experts switched a perceptual activity (observing the high-resolution CT) to a conceptual activity (writing the reports) more frequently. The following analysis is aimed at confirming this interpretation directly.

4.3.4 Number of switchings between types of external activity

In order to confirm the above interpretation, we counted the number of switchings from one activity to another. If expertise made a diagnostic process cyclic,

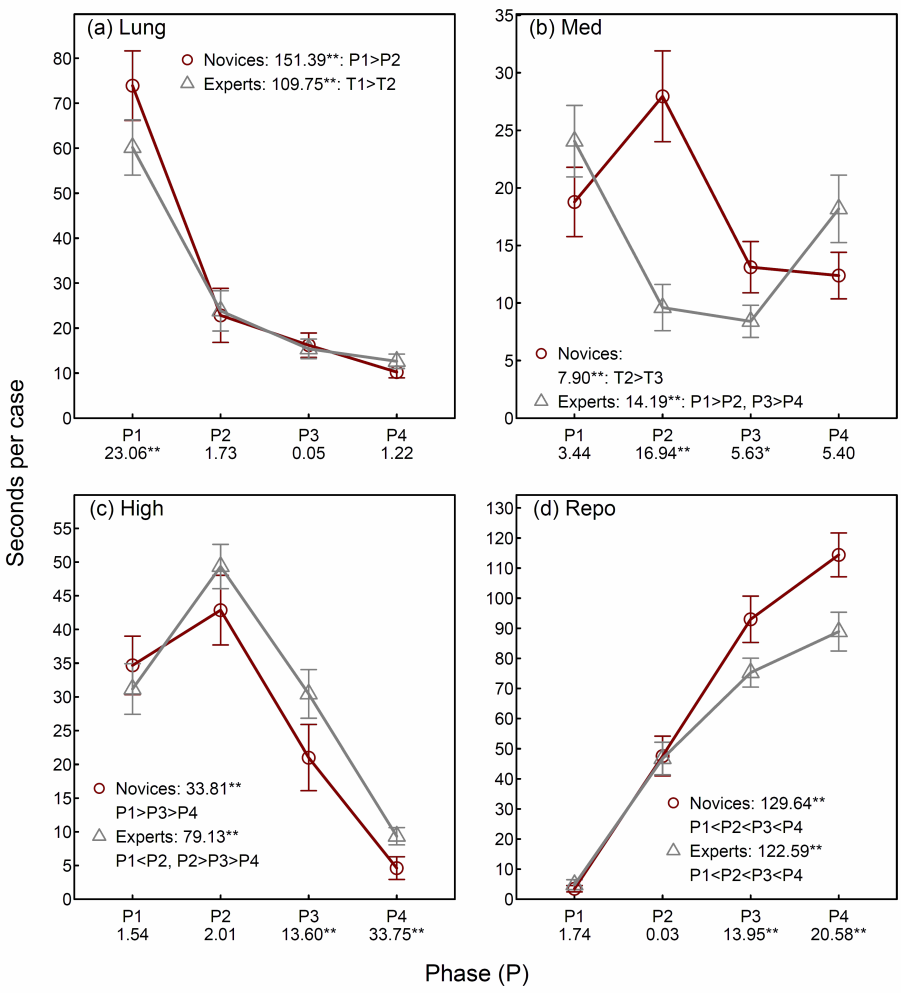


Fig. 9. The task timelines of external activity. For each type of external activity, 2 (the expertise) $\times 4$ (the phases) ANOVA revealed a significant interaction of the expertise and the phases: Lung, $F(3, 36) = 10.80$*; Med, $F(3, 36) = 12.98$*; High, $F(3, 36) = 5.57$*; Repo, $F(3, 36) = 10.56$*. The notations of values in the legends and the labels are the same as in Fig. 7.

there would be differences between the experts and the novices in the number of switchings. Fig. 10 shows how many times the participants switched from each type of activity. In all four types of activity, it was confirmed that the numbers of switchings were larger for the experts than for the novices. Among these effects, the difference in writing the reports is especially important because the effect represents that the experts provided more shifts from conceptual processing to perceptual processing throughout the diagnostic process.

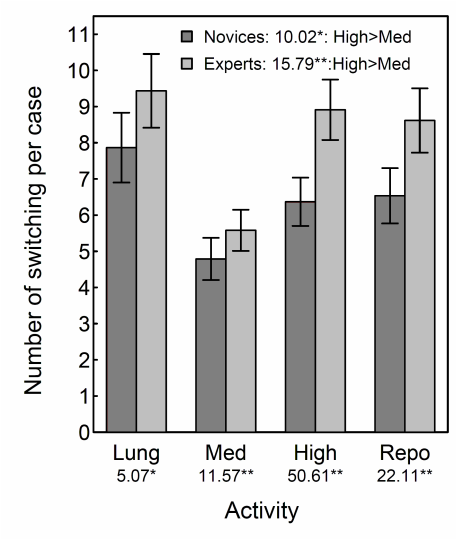


Fig. 10. Number of switchings of external activity. A 2 (the expertise) $\times$ 4 (the activity types) revealed a significant interactions between the expertise and the activity types, $F = 3.06$*. The notations of the values in the legends and the labels are the same as in Fig. 7.

4.3.5 Summary

Summarizing the above results, we can characterize the relations between perceptual and conceptual processing for both the experts and the novices. The results characterize the process of the experts as a cyclic one of perceptual and conceptual processing. In contrast, the process of the novices can be characterized by a one-way, bottom-up process from perceptual to conceptual processing. Importantly, these characteristics are consistent with the results of verbal protocol analysis. Therefore, the analysis of external activity supports the results of verbal protocol analysis. Some results on the external activity, however, do not correspond to the results of verbal protocol analysis, or vice versa. For example, we did not show results for external activity corresponding to the initial hypothesis that clearly confirmed in the verbal protocol analysis. On the other hand, we did not show the results concerning the switching of processing in the verbal protocol analysis, which was confirmed in the analysis of the external activity. The next section provides a further analysis of the above limitations.

4.4 Correspondence between verbalized words and external activity

The previous two sections presented the results of analysis for the two types of data: verbalized data and data concerning external activity. Each type of data is assumed to have different correspondences to perceptual and conceptual processing. In verbalized data, the words tagged as *Percept* and *Concept* were

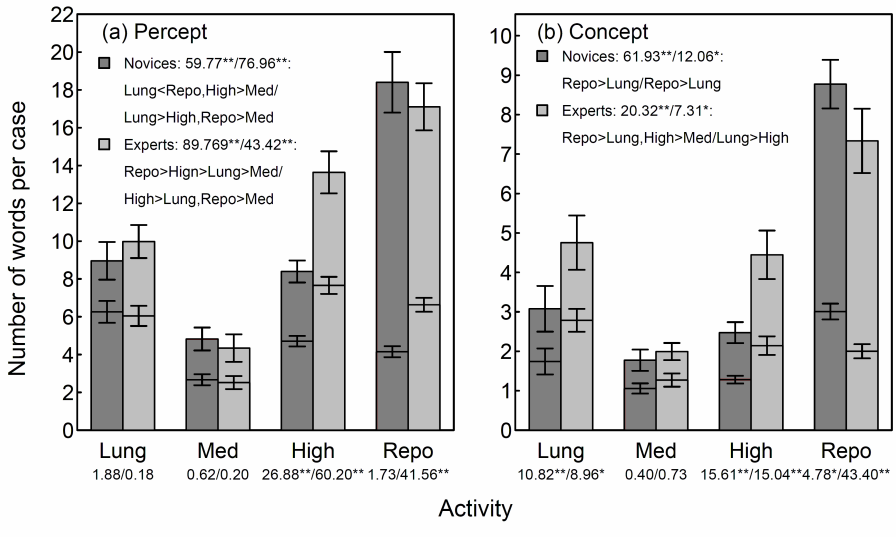


Fig. 11. Correspondence between the verbalizations and the external activity. The upper and lower values of each bar represent mean numbers of total words and new words, respectively. For each value of each tag type, 2 (the expertise) $\times$ 4 (the activity types) ANOVA revealed a significant interactions between the expertise and the activity types, the number of *Percept*, $F(3, 36) = 18.41$**; the number of new *Percept*, $F(3, 12) = 18.41$**; the number of *Concept*, $F(3, 36) = 8.48$*; the number of new *Concept*, $F(3, 36) = 11.07$*. The notations of values in the legends and the labels are the same as in Fig. 7.

regarded as the verbalized outputs of perceptual and conceptual processing. On the other hands, regarding external activity, observations of CT images were considered a type of activity concerning perceptual processing, and the writing reports was seen as a type of activity concerning conceptual processing. We considered that connecting the two analyses would bring further insight to the relations between perceptual and conceptual processing.

Fig. 11 shows number of words verbalized when engaging in each type of external activity. For each of *Percept* (Fig. 11a) and *Concept* (Fig. 11b), the total number of verbalized words was shown as the upper value, and the number of new words was shown as the lower one. The results can be summarized as follows for each measure.

(1) *Number of words tagged as Percept* (The upper values of Fig. 11a). The experts' value exceeded that of the novices in observing the high-resolution CT.
(2) *Number of new words tagged as Percept* (The lower values of Fig. 11a). In addition to the above effect, the value of the experts exceeded the value of the novices in writing the reports.
(3) *Number of words tagged as Concept* (The upper values of Fig. 11b). The value of the experts exceeded the value of the novices in observing the lung-window and high-resolution CT, whereas the value of the novices

exceeded the value of the experts in writing the reports.

(4) *Number of new words tagged as Concept* (The lower values of Fig. 11b). The obtained effects were the same as in 3 above.

The above results revealed detailed characteristics in the diagnostic process of the experts and the novices. Since the novices' value exceeded that of the experts only in writing the reports (see 3 and 4), the novices' process can be considered a one-way process, in which they summarized observed features and considered the final diagnosis while writing the reports.[5] In contrast, the experts' process could be characterized as follows: the experts verbalized more words tagged as *Concept* while observing the CT images (see 3 and 4), and they verbalized more new words tagged as *Percept* while writing the reports (see 2). These features indicate close connections between perceptual and conceptual processing in the process of the experts' diagnosis. The former feature indicates the close connection from the activity concerning perceptual processing to the outputs of conceptual processing, while the latter indicates the close connection from the activity concerning conceptual processing to the outputs of perceptual processing. Especially, the latter feature provides direct evidence of interactivity of the experts' process. Since there are no perceptual features on the device for writing the reports (see Fig. 3b), the newly verbalized features in this activity cannot be considered as the features perceived from the medical images. Instead, they must be features recalled from when trying to decide on a final diagnosis.

5 General discussion

In this section, we discuss the implications and limitations of the present study. First, we address the implications for understanding the development of expertise in medical image diagnosis, and then deal with the relations of our results to general theories of human cognition. Finally, we note some limitations of the present study.

5.1 Contributions to the field of medical image diagnosis

Many researchers have so far investigated perceptual and conceptual processing in medical image diagnosis. We believe that our study contributes to fur-

[5] It appears that, regardless of the category, the number of verbalizations in writing the reports is higher in the novices than the experts. Therefore, we compared the novices with the experts in the number of words written in the final reports. As a result, we did not find any difference between the experts and the novices in the number of words in the final reports.

ther understanding of the cognitive process in medical image diagnosis.

A first contribution of our study is in the experimental task in which CT images were used as the materials. As noted earlier, compared to radiographs, which the prior studies mainly adopted, CT images provide features that can be clearly observed by physicians. Therefore, it can be concluded that we have confirmed the cyclic process in a task that is unlikely to change due to the development of expertise. We speculate that one of the possible reasons causing this cyclic process is *uncertainty*. Regardless of visual clarity, features on medical images only *statistically* relate to actual diseases (a related discussion is in Sharples et al., 2000). In situations where a single feature is not enough to decide a correct concept, it could be useful to use a cyclic process, where multiple hypotheses are proposed and they drive the search for additional features in the images (a similar discussion is in Norman et al., 1999).

A second contribution of our study is in the analysis method for the two types of data. We showed transitions of the two components, perceptual processing and conceptual processing, in each of the two types of data. Our methods are contrasted with the previous studies, which discussed the nature of expertise based on protocol excerpts or direct coding for strategy types. We think that the approaches adopted by the previous studies are superior in demonstrating qualitative features of the diagnostic process, but inferior in objectivity. Our approach successfully provides support for the previous findings with respect to the quantitative data.

Furthermore, we consider that the advantages of our approach are not limited to reconfirming previous findings; we believe that the semi-automatic quantitative analysis sheds light on the underlying mechanisms of human cognition. A third contribution of the study comes from the advantages of our analysis methods. We considered that our study revealed that the relations between perceptual and conceptual processing in medical image diagnosis are not *iterative* but *interactive*. The term *iterative relations* means here that the two components cyclically repeat but do not influence each other. On the other hand the term *interactive relations* means here that the two components cyclically repeat and do influence each other. Our results, especially the results shown in Fig. 7 and Fig. 11, clearly indicate that the interactive relations are involved in the process of medical image diagnosis. For example, Fig. 7b indicates that the experts generated multiple initial hypotheses after a first glance at medical images, and they finally narrowed the search to a few alternative hypotheses. Furthermore, Fig. 11 presents a strong connection between perceptual and conceptual processing in the process of the experts' diagnosis. Fig. 11a illustrates that the experts generate new perceptual features while writing the reports, and Fig. 11b indicates that they verbalized conceptual words while observing the images. These results provide evidence of the interactive relations, and shed light on the cognitive mechanisms in medical image

diagnosis.

Finally, we would like to assert a contribution of our study from the viewpoint of cognitive engineering. This study was performed as part of a larger project that is being conducted in collaboration with radiologists, researchers of cognitive science, and researchers of image-processing engineering, with the aim of developing intelligent systems that support the diagnostic process (details in Morita et al., 2004, 2005). So far, image-processing engineering has developed elaborate tools that mainly support physicians' perceptual processing (e.g., Mori et al., 2000). We believe that the combination of image-processing engineering and cognitive scientific analysis will make it possible to create innovative tools for supporting the interactive process in medical image diagnosis.

5.2 Relationships with the other areas of learning and expertise

Through the evolution of cognitive science and psychology, the interactions between perceptual and conceptual processing have been frequently investigated. These studies pointed out that the development of expertise accelerates the shift from the top-down to bottom-up process. Although such a shift seems to contradict our results, we consider the contradictions not to be acute. We agree that some aspects of medical image diagnosis could not be explained by previously discussed learning mechanisms, such as perceptual learning or compilations. However, it is clear that the basic learning mechanisms underlie the interactive process observed in the present study.

In particular, we think that the processing represented by the ACT-R architecture has similar features to the process observed in the present study. As mentioned in section 1.2, the model constructed by Taatgen (2005) implemented learning mechanisms of compilation, which creates new production rules by collapsing two production rules into a single rule. In the initial state of his model, declarative knowledge controlled the firings of production rules. However, after the compilation, production rules could be directly evoked by the environmental information, and declarative knowledge no longer involved firings of production rules.

This can be viewed as a study that contradicts our results. Contrary to his compilation model, the experts in our study frequently verbalized conceptual knowledge compared with the novices (see Fig. 4). This conflict depends on differences in characteristics existing among the tasks. Unlike the air traffic control task, the task used in our study cannot be accomplished without declarative knowledge because medical reports should contain verifications of the diagnosis connecting conceptual knowledge with observed features.

Even though the above contradiction exists, we can find a commonality between our results and his study. Importantly, he showed that compilation not only eases the retrieval of declarative knowledge, but also promotes time-sharing of parallel behaviors of multiple modules. The air traffic control task consists of multiple sub-tasks including perceiving airplanes on a monitor, clicking on one of the planes on the monitor, and pressing keys. Each of these sub-tasks can be performed by the manual or visual modules of the ACT-R architecture. In the initial state of his model, the tasks were conducted serially from the sub-tasks concerning the visual module to ones related to the manual module. After the compilation of declarative knowledge, these sub-tasks were performed simultaneously, and the two modules interactively exchanged information with each other. This parallel process is consistent with our results presented in section 4.4; the experts retrieved conceptual knowledge while observing images and wrote reports while considering perceptual features. Using the terms of the ACT-R architecture, we can imply that such an activity is a parallelization of the perceptual and declarative modules. Although further investigations is required to judge whether such a process can be replicated in the ACT-R architecture, our results have sufficient meaning for this theory.

There are also some contradictions, though not serious ones, between our study and the previous studies on expertise in areas other than medical image diagnosis. For instance, Dreyfus and Dreyfus (1986) distinguished intuitive holistic understanding of experts from analytical thinking of novices by conducting case studies on expertise in nursing activities. Although their distinction seems to be inconsistent with our claim, they also acknowledged the cyclic processes of experts in medical image diagnosis. Their explanation is that close examination tends to follow holistic decision-making in a situation where important decisions are made.

In addition, even in areas other than medical image diagnosis, it has been pointed out that the development of expertise facilitates the shift to the interactive process of multiple components. In particular, we believe that the process presented in this paper is similar to those of design and scientific discovery. For example, in studies on design, many researchers have confirmed interactive relations between discovering novel visual features in sketches and conceptualizing design goals (Goldschmidt, 1994; Kavakli and Gero, 2001; Suwa et al., 2000). In addition, in the area of scientific discovery, Dunbar (1993) demonstrated that an interactive process of data search and hypothesis generation relates to high task performance.

Although we successfully demonstrated expert-novice differences in the process of medical image diagnosis, our findings have several limitations. One of these comes from accidental errors in the analysis method. Compared with conventional hand-coding methods, our semi-automatic methods might introduce not a few errors in the results for the verbal protocols and the external activity. However, it is difficult to imagine that the obtained results are artifacts of our methods because it is reasonable to suppose that randomized errors were introduced in both novice and expert groups. Therefore, we believe that the obtained differences between the experts and the novices are sufficiently reliable.

A second limitation is in the generalization of our results. The experiment was conducted in a single institute, and the number of participants was limited. Therefore, we should be careful about applying our findings to other institutes. We think that the generalizations need to be based on the correspondences with previous findings. As noted earlier, the previous studies pointed out the effects of expertise in the number of verbalizations and in the relations between perceptual and conceptual processing. Therefore, we believe that these findings are applicable indeed to other institutes.

A related limitation pertains to the stages of expertise. Many researchers have argued that the development of expertise in medical image diagnosis is not a monotonic function (Lesgold et al., 1988; Raufaste et al., 1998): those studies showed that physicians who had intermediate experience in medical image diagnosis or expert physicians who did not belong academic institutes tended to use a bottom-up process. Therefore, we could not apply our results to all medical experts.

Finally, we did not show relations between the diagnostic process and the diagnostic performance. Therefore, it is unclear whether the observed process is preferable or not for gaining high performance. In addition, we cannot reject the possibility that the interactive process was caused by the difficulty of the task. Although the case materials were randomly chosen, the final diagnostic performance was not sufficiently high. In easier tasks such as the nodule detection task, the experts might possibly use a bottom-up process.

Despite the above limitations, we can assert that our study contributes a deeper understanding of the interactive process of perceptual and conceptual processing.

6 Conclusion

This paper presented experimental studies on the relationship between perceptual and conceptual processing and demonstrated the effects of expertise on the relations in medical image diagnosis. We believe that understanding this topic is important for constructing a theoretical basis for human-computer studies because the integration of these cognitive components can be regarded as a central foundation of human-computer interactions. In future studies, the analysis method and the data presented in this paper are expected to be used to construct a detailed theory such as cognitive models and applied to develop innovative computer-support systems.

Acknowledgments

This work was supported in part by the Grant-In-Aid for 21st COE program provided by the Ministry of Education, Culture, Sports, Science and Technology of the Japanese government. The authors thank Dr. Fukushima, Dr. Kamioka, Dr. Okada, Dr. Kawase, Dr. Tachi, Dr. Kawai, Dr. Sawaki, Dr. Satake, Dr. Ito of Nagoya University Hospital for participating in the diagnosis experiment.

References

Alberdi, E., Povyakaio, A.A., Strigini, L., Ayton, P., 2004. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. Academic Radiology 11 (8), 909–918.

Anderson, J.R., Lebiere, C., 1998. The Atomic Components of Thought. Mahwah, NJ: Lawrence Erlbaum.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y., 2004. An integrated theory of the mind. Psychological Review, 111(4). 1036-1060.

Azevedo, R., Lajoie, S.P., 1998. The cognitive basis for the design of a mammography interpretation tutor. International journal of artificial intelligence in education 9, 32–44.

Brumby, D.P., Howes, A., 2004. Good enough but I'll just check: Web-page search as attentional refocusing. In: Proceedings of the sixth International Conference on Cognitive Modeling. 46–51.

Byrne, M.D., 2001. ACT-R/PM and menu selection: applying a cognitive architecture to HCI. International Journal of Human-Computer Studies 55, 41–84.

Chase, W.G. and Simon, H.A., 1973. Perception in Chess. Cognitive Psychology, 4, 55-81.

Crowley, R.S., Naus, G.J., Stewart, J., Friedman, C.P., 2003. Development of visual diagnostic expertise in pathology: An information-processing study. Journal of

the American Medical Informatics Association 10 (1), 39–51.
Dreyfus, H.L., Dreyfus, S. 1986. Mind over machine. Blackwell.
Dunbar, K., 1993. Concept discovery in a scientific domain. Cognitive Science 17(2), 397–434.
Ericsson, K.A., Simon, H.A. 1980. Verbal reports as data. Psychological Review 87, 215–251.
Fu, W., Bothell, D., Douglass, S.D., Haimson, C., Sohn, M., Anderson, J., 2004. Learning from real-time over-the-shoulder instructions in a dynamic task. In: proceedings of the sixth International Conference on Cognitive Modeling. 100–105.
Goldschmidt, G., 1994. On visual design thinking: the vis kids of architecture. Design Studies 15 (2), 158–174.
Goldstone, R.L., Steyvers, M., Spencer-Smith, J., Kersten, A., 2000. Interactions between perceptual and conceptual learning, In: E. Diettrich and A.B. Markman (Eds.) Cognitive Dynamics: Conceptual Change in Humans and Machines. pp. 191–228.
Gray, W.D., John, B.E., Atwood, M.E., 1993. Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world performance. Human-Computer Interaction 8, 237–309.
Johnson, T.R., Krems, J.F., 2001. Use of current explanations in multicausal abductive reasoning. Cognitive Science 25, 903–939.
Lesgold, A., Rubinson, H., Feltovitch, P., Glaser, R., Klopfer, D., Wang, Y., 1988. Expertise in a complex skill: Diagnosing X-ray pictures. In: M. Chi, R. Glaser, and M. Farr., (Eds.), The nature of expertise. Erlbaum, Hillsdale, NJ. pp. 311–342.
Kavakli, M., Gero, J.S., 2001. Sketching as mental imagery processing. Design Studies 22 (4), 347–364.
Kieras, D., Meyer, D.E., 1997. An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. Human-Computer Interaction 12, 391-438.
Klar, D., Dunbar, K., 1988. Dual space search during scientific reasoning. Cognitive Science 12, 1–48.
Klar, D., 2000. Exploring science: The cognition and development of discovery processes. Cambridge, Mass.: MIT press.
Kundel, H.L., Nodine, C.F., 1983. A visual concept shapes image perception. Radiology 146, 363–368.
Krupinski, E., 2003. The future of image perception in radiology: Synergy between humans and computers. Academic radiology 10, 1–3.
Larkin, J.H., McDermott, J., Simon, D.P., Simon, H., 1980. Model of competence in solving physics problems. Cognitive Science 4, 317–345.
Manning, D.J., Gale, A., Krupinski, E.A., 2005. Perception research in medical imaging. British Journal of Radioly 78, 683–685.
Matumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., Asahara, M., 2000. Japanese morphological analysis system ChaSen version 2.2.1. Nara institute of science and techonology.
Miwa, K., 2004. Collaborative discovery in a simple reasoning task. Cognitive Systems Research 5 (1), 41–62.
Mori, K., Hasegawa, J., Suenaga, Y., Toriwaki, J., 2000. Automated anatomical

labeling of the bronchial branch and its application to the virtual bronchoscopy system. IEEE Transactions on Medical Imaging 19, 103–114.

Morita, J., Miwa, K. Kitasaka, T., Mori, K., Suenaga, Y., Iwano, S., Ikeda, M., Ishigaki, T., 2004. Chance discovery in Image diagnosis: Analysis of perceptual cycles. In: Proceedings of the first European Workshop on Chance Discovery The sixteenth European Conference on Artificial Intelligence. 162–171.

Morita, J., Miwa, K. Kitasaka, T., Mori, K., Suenaga, Y., Iwano, S., Ikeda, M., Ishigaki, T., 2005. Expertise in interactions of perceptual and conceptual processing. In: Proceedings of the twenty-seventh Annual Conference of Cognitive Science Society 1541–1546.

Myles-Worsley, M., Johnston, W.A., Simons, M.A., 1988. The influence of expertise on X-ray image processing. Journal of Experimental Psychology: Learning, Memory, and Cognition 14 (3), 553–557.

Neisser, U., 1976. Cognition and reality. W.H. Freeman and Company.

Nodine, C.F., Kundel, H.L., 1987. The cognitive side of visual search in radiology. In: J. K. O'Regan and A. Levy-Schoen., (Eds.), Eye Movements: From Physiology to Cognition. New York: Elsevier Science, pp. 573-582.

Norman, G.R., Brooks, L.R., Coblentz, C.L., Babcook, C.J., 1992. The correlation of feature identification and category judgments in diagnostic radiology. Memory & Cognition 20 (4), 344–355.

Norman, G.R., Brooks, L.R., Colle, C.L., Hatala, R.M., 1999. The benefit of diagnostic hypotheses in clinical reasoning: Experimental study of an instructional intervention for forward and backward reasoning. Cognition and Instruction 17, 433–448.

Okada, T., Simon, H., 1997. Collaborative discovery in a scientific domain. Cognitive Science 21, 109–146.

Patel, V.L., Groen, G.J., 1986. Knowledge based solution strategies in medical reasoning. Cognitive Science 10, 91–116.

Peterson, C., 1999. Factor assciated with success or failure in radiological interpretation: diagnositic thinking approaches. Medical educations 33, 251–259.

Raufaste, E., Eyrolle, H., Mariné, C., 1998. Pertinence generation in radiological diagnosis: Spreading activation and the nature of expertise. Cognitive Science 22, 517–546.

Rogers, E., 1996. A study of visual reasoning in medical diagnosis. In: Proceedings of the eiteenth annual conference of the cognitive science society. 213–218.

Salvucci, D.D., 2005. A multitasking general executive for compound continuous tasks. Cognitive Science 29, 457–492.

Sharples, M., Jeffery, N.P., du Boulay, B., Teather, B.A., Teather, G., du Boulay, G.H., 2000. Structured computer-based training in the interpretation of neuroradiological images. International Journal of Medical Informatics 60, 263–280.

Simon, H.A., Lea, G., 1974. Problem solving and rule induction: unified view. In L. Gregg, (Eds.), Knowledge and Cognition. Erlbaum, Hillsdale, NJ. pp. 105–127.

Sowden, P.T., Davies, I.R.L., Roling, P., 2000. Perceptual learning of the detection of features in X-ray images: A functional role for improvements in adults' visual sensitivity? Journal of Experimental Psychology: Learning, Human perception and performance 26 (10), 379–390.

Suwa, M., Gero, J., Purcell, T., 2000. Unexpected discoveries and S-invention of

design requirements: Important vehicles for a design process. Design Studies 21, 539–567.

Taatgen, N., 2005. Modeling parallelization and flexibility improvement in skill acquisition: From dual task to complex dynamic skills. Cognitive Science 29, 421–455.

Wickens, T.D., 2002. Elementary signal detection theory. Oxford university press.

Woods, B.P., 1999a. Visual expertise. Radiology 211, 1–3.

Woods, B.P., 1999b. Decision making in radiology. Radiology 211, 601–603.