Utilization of Visual Sensing and Face Analysis for Enhancing E-Learning

SIYANG YU

Abstract

Educational theory manifests, learning is a complicated mental process that learners participate actively, knowledge is internalized during this learner-centered process. The learning will be efficient and have excellent outcome if learning is designed to attract and suit learners well. Efforts to keep students engaged and motivated have great importance in self-paced e-learning, since this type of e-learning accommodates large number of learners with various types and characteristics. It requires much more effort and labor for instructors to design and develop a course than in conventional educational form. Thus, support for self-paced e-learning is the target in this research, which findings can be generalized to a wide variety of learning.

The first portion is devoted to learning state recognition through visual sensing [*1]. Feedback is utmost important in learning since it is the foundation to understand how education is going on. However, instructors have difficulties to know how students learn as little information can be retrieved from the places where students actually perform self-paced e-learning. Lack of feedback information hinders students to get timely and appropriate guidance, i.e. personal tutoring. In this research, we focused on automatic detection of three types of mental status, which specifically involving, concentration-distraction, difficulty-ease, and interest-boredom. They are selected according to their direct and close relation to learning as important and valuable information for instructors to understand how learners react to learning. The detected information enables teachers to provide mentoring, modify materials, conduct educational analysis etc.

A novel system is designed to estimate learners' internal states during learning through introducing pattern recognition technology which takes image processing results. A Red-Green-Blue-Depth (RGB-D) camera is mounted on an existing e-learning system. A video of a learner's face and upper-body is captured. For each interval

[*1] This portion is based on "Learning State Recognition in Self-Paced E-Learning" [1], by the same first author, which appeared in the IEICE Transactions on Information and Systems, Copyright(c) 2017 IEICE. The material in this thesis was presented in part at the IEICE Transactions on Information and Systems [1], and the figures in this thesis are reused form [1] under the permission of the IEICE.

of 30 seconds, a feature vector is calculated based on the processing of the captured RGB-D videos. It covers three categories of non-verbal information, presence information, head and facial parts information, and probability of gazing at the screen. Learning states are estimated by using the Support Vector Machine (SVM) based on the above feature vectors.

To develop the recognition component, ground truth score of learning states were gathered through self-evaluation by learners. Self-evaluation scores of 3119 samples of 30 second-interval for 1559.5 minutes were gathered from seven undergraduate students in Kyoto University. To avoid learning interruptions and obtain high reliance on self-reports, both learner's face and learning content can be shown synchronously to learner right after e-learning session, and learners provided self-evaluation about learning states by watching them. For the benefit of teachers, we used a five-level scale for learning state measurement, e.g., very concentrated, concentrated, neutral, distracted, and very distracted.

The average accuracy of leave-one-out cross validation based on above experimental data was 60.9%, 61.6% and 65.6% for concentration-distraction, difficulty-ease, and interest-boredom respectively by strict matching criterion which means an exact match between the prediction of the SVM and the ground truth. The result evaluated by lenient matching criterion was also examined. Lenient matching is considered indicative of a correct prediction when it is sufficiently close to the ground truth. Specifically, the distance between them is less than or equal to one. The performance was 90.6%, 86.2%, and 92.7% for concentration-distraction, difficulty-ease, and interest-boredom, respectively.

One critical problem is how to deal with a variety of students, i.e., inter-personal differences among learners. If target samples are taken from different participants, the performance of the system is much lower than the above results. For this purpose, a scheme that can quickly adjust to a new learner with a little effort of the learner was designed. First, it is assumed that the system has sufficient number of samples with ground truth from multiple learners as already mentioned above. Next, the system requests a new learner to give a small number of samples with self-evaluation, which are called "representative samples". Then, the system automatically selects the classifier that fits the new learner's representative samples. For this selection, several methods are considered and evaluated. Five representative samples (30 seconds intervals) were chosen randomly from a new learner's learning records, and their performance was evaluated. The results demonstrate the accuracy-based method provides better performance than the unified classifiers, while the similarity-based method does not provide good results.

It is important to note that each selected classifier demonstrated inferior performance compared to the classifier that demonstrated the best performance for the learner. Therefore, reducing the gap between the selected classifier of a new learner and the most appropriate classifier of that learner is an important objective. Further research was investigated for choosing better representative samples instead of random selection that result in better performance. The proposed method of choosing representative samples is based on the following assumptions. Frequently appearing samples can be good representative samples; a set of representative samples that covers a wider area of the feature space provides better accuracy; a set of representative samples with enough variety of classes gives better accuracy. Kernel density estimation is used to estimate the distribution of samples of a new student. Clustering is applied to samples with a large probability density. For each cluster, one sample is chosen as representative sample. In the experiments, a slight improvement in the average accuracy was observed, although it was not significant. However, the proposed method did demonstrate the advantage of avoiding bad cases. Through experiments, certain characteristics of the data and classifications were confirmed. For example, neighboring samples around representative samples were not classified well, especially with respect to the difficulty–ease state. In addition, the distribution of the representative samples displayed a significant correlation with respect to accuracy, which make them a potentially valuable indicator for future investigations of new methods.

The second portion focuses on supporting pronunciation learning in Computer Assisted Language Learning (CALL) [*2]. Pronunciation is a fundamental factor in speak-

[*2] This portion is based on "Visual Emphasis of Lip Protrusion for Pronunciation Learning" [2], by the same first author, which appeared in the IEICE Transactions on Information and Systems, Copyright(c) 2019 IEICE. The material in this thesis was presented in part at the

ing and listening. Intelligible pronunciation not only enables students to understand and be understood but also helps them to monitor their speech on the basis of input from the environment. Correct pronunciation requires the accurate position of various articulators, such as the tongue, lips and jaw. Traditional explanations given by written texts and conventional multimedia approaches do not give sufficient explanations of the articulations.

For providing more advanced supports for pronunciation learning in e-learning, the following visualization functions are considered: (a) demonstrating learners how native speakers pronounce words by showing articulation with three-dimensional (3D) information, (b) showing learners how they pronounce the words, and enable them to check their correctness and/or weakness. A prototype system for (a) was implemented. Realization of function (b) on learners' side is left for future work; however, the framework can be the same way as (a). 'Rounded pronunciation' is chosen as the learning target, because Japanese learners often have difficulties in such pronunciation with lip protrusion. It cannot be easily recognized through ordinary pictures or movies, and explanation with 3D information around the mouth provides good supports for Japanese learners.

The framework is designed as follows. Videos with 3D information are obtained by an RGB-D camera, and then pseudo-coloring, is applied to the captured images. Lip protrusion is measured as the relative depth of the lip to the other parts of the face. The tip of the nose is used as the reference point because it is steadily observed regardless of ordinary body movements, and its location is not affected by mouth movements. Pseudo-coloring is applied to the mouth region according to the measured depth. Lip subarea and non-lip subarea are differently colored. For lip subarea, it is colored vividly with an attention-grabbing color if the lips are protruded. For non-lip subarea, it is colored to provide a contrast. The lip subarea is further segmented into the upper lip and the lower lip subarea, and they are colored separately because of the physiological characteristic that upper lip bulges slightly

IEICE Transactions on Information and Systems [2], and the figures in this thesis are reused form [2] under the permission of the IEICE.

more than lower lip. The color is chosen based on the Hue-Saturation-Intensity color space. Intensity is maintained to give the original two-dimensional information as it is. Hue is modified according to the relative depth. Significant discontinuity, i.e. coarse quantization, is necessary for protrusion to be noticed. Saturation is used to emphasize lip protrusion. The maximum saturation allowed with hue and intensity is calculated. Then, saturation is determined based on the maximum value and relative depth. The color around the border of mouth area is smoothed to attenuate the effect of unnatural and drawing unnecessary attention.

Evaluation was conducted to verify how learners could improve their pronunciation by watching the enhanced videos that captured teachers' pronunciations. Forty-three students in Kyoto University with various learning experiences were gathered. They were randomly separated into two groups for comparing the effects of common video and pseudo-colored video, respectively. Improvements from baseline methods such as the symbol-based method were evaluated by taking linear regression and t-test. As the results, decreased number of incorrect pronunciations for rounded vowel showed the superiority of the pseudo-colored videos.

The experiments of learning state estimation were conducted on language e-learning materials. For future work, we need to proceed to other types of e-learning materials such as natural science, humanities, etc. Age and learning place of learners are also critical factors for learning effects. We need to further investigate how our framework contributes for diverse conditions. The similar problems are in pronunciation learning supports. Various pronunciation in various language may be difficult for learners. Much room for applications is left for future works.

Concerning the system design, the proposed system is a prototype that is a combination of a camera, screen capture device, audio recording device, and software for image processing and pattern recognition. They should be well integrated into an ordinary computer or mobile device for the convenience of learners. For this purpose, there are already some mobile phones and tablet devices that have RGB-D camera, and we can expect that those devices will be good platforms.

Contents

Abbreviations

E-learning	Electronic learning
CD	Compact Disk
DVD	Digital Video Disk
CALL	Computer Assisted Language Learning
MOOC	Massive Open Online Course
STM	Short-Term Memory
WM	Working Memory
LTM	Long-Term Memory
ZPD	Zone of Proximal Development
AECT	Association for Educational Communications and Technology
FACS	Facial Action Coding System
RGB-D	Red-Green-Blue-Depth
SVM	Support Vector Machine
3D	three-dimensional
SDK	Software Development Kit
L2	Second Language
ASR	Automatic Speech Recognition
HSI	Hue-Saturation-Intensity
MCAR	Missing Completely At Random
IPA	International Phonetic Alphabet
RBF	Radial Basis Function
EEG	Electroencephalography
PCA	Principal Component Analysis

Chapter 1

Introduction

Education has been being revolutionized by information and communication technology. Researchers, practitioners, and other people concerned with education have recognized and been excited about the extraordinary potential of using technologies in education. Nations, academic and educational institutions, universities and organizations spend colossal resources, financial, human etc., to research, develop, deploy and utilize technologies in education. Electronic learning (e-learning) has been integrated into various levels of education, from digitalized content to intelligent tutoring system, from over-head projector to virtual reality etc. As a consequence, learning outcome and efficiency are improved effectively. Researches about e-learning have advanced to wider perspectives, such as ethical issue, social aspect, which are not constrained to integration of technology and pedagogy. Technology-oriented studies also develop to more profound stage that focus on actual and specific issues in various e-learning application rather than working as media simply in a superficial level. Our study explores technology-enhanced supporting for e-learning system. Two realistic issues in practical application of e-learning system are considered and investigated intensively under a non-intrusive, non-contact, and economical circumstance. The first portion is devoted to the recognition regrading learners' internal sates during learning which are important and valuable feedback for instructors. It is helpful for educators to understand how education is going on and how learners react. The second portion focuses on supporting pronunciation learning in Computer Assisted Language Learning (CALL). A novel method for producing a visual-enhanced material for pronunciation learning is designed, which helps learners to pronounce in a correct way with referring three-dimensional (3D) information of mouth shapes in pronunciation.

1.1 About E-Learning

Although the term of e-learning has been spoken for decades, it is difficult to give an “official” definition which is commonly accepted. Many definitions were given based on a specific set of technologies and/or principles of instructional/educational design that involved which indicated the focuses. Sangrà A, Vlachopoulos D, and Cabrera N [3] summarized four categories of definitions based on their literature review. Technology-driven definitions, for example, Guri-Rosenblit S defined e-learning as “E-learning relates to the use of electronic media for a variety of learning purposes that range from add-on functions in conventional classrooms to full substitution for the face-to-face meetings by online encounters” [4]. Delivery-system-oriented definitions, typified by “E-learning is the delivery of education (all activities relevant to instructing, teaching, and learning) through various electronic media” [5]. Communication-oriented definitions accentuate communication and interactivity. In [6], communication and interaction of e-learning system were superior characteristics. Such consideration was summarized as “E-learning is education that uses computerised communication systems as an environment for communication, the exchange of information and interaction between students and instructors” in [3]. Educational-paradigm-oriented definitions were epitomized by “E-learning is defined as the use of new multimedia technologies and the Internet to improve the quality of learning by facilitating access to resources and services as well as remote exchanges and collaboration” [7]. Eventually, they came to the definition based on Delphi survey about experts’ perceptions and knowledge: “E-learning is an approach to teaching and learning, representing all or part of the educational model applied, that is based on the use of electronic media and devices as tools for improving access to training, communication and interaction and that facilitates the adoption of new ways of understanding and developing learning.” Amaral L, Leal D reached to a definition by using mathematical language, which “The process, by which the student learns through the content placed in the Internet and/or CD-Rom. The teacher, if exist, is at distance, using the internet to communicate (synchronously or asynchronously) with the students, possibly intermediated with some face-to-face moments" [8]. The latter two definitions denote e-learning becomes an umbrella term. The diversity implies e-learning should be designed according to the learning needs and requirements tightly.

Another term refers to both education and technology is educational technology. Unlike e-learning, educational technology has definitely official definition that delineated by the Association for Educational Communications and Technology (AECT), which is "Educational Technology is the study and ethical practice of facilitating learning and improving performance by creating, using and managing appropriate technological processes and resources" [9]. Although the origins of e-learning and educational technology are different, the discrimination between them is becoming more and more vague in recent years. (For example, Wikipedia has redirected "E-learning" to "Educational Technology" since 2015.) The fine discriminations are beyond the scope of this study and make no influence on the research, hence, they are deemed to be synonym.

1.2 Types of E-learning System

Taxonomy of e-learning varies if different characteristics are focused. From the separation of teachers and students, e-learning has types of technology-enhanced face-to-face learning, for instance, lecturer teaches college students with using multimedia content; distance learning, students learn by themselves with using web-based course is a typical example; and blended learning, which is a mixture of abovementioned 2 types of learning. The first type could be the upgrade of traditional classroom education. The second one breaks down the restriction of location and/or time. The last one integrates the advantages of them, that may be a proper choice for lifelong learning or career training.

If communication approach is considered primarily, e-learning could be either synchronous or asynchronous. Synchronous e-learning is about the participants, teachers and/or student peers, are able to conduct mutual interaction simultaneously which is convenient to exchange knowledge and ideas. Representative samples include virtual classroom, webinar, and online learning that equipped with real time chatting function-in text, audio, or video style. It enables participants to be a part of class from distant place which can reduce the feeling of isolation during learning. Asynchronous e-learning enables students to communicate via e-mail, blog, forum etc. Learning time is flexible in this form which allows personalized learning in some extent. Moreover, learners have time to reflect on what they have learned which enable the communications more effective and efficient.

The mainly used technology can also be the criterion for categorization. In this context, computer-based learning means people learn with the materials stored in computers, compact disk (CD) or digital video disk (DVD), without internet connection requirement. The learning materials of CALL used in Kyoto University is one perfect example. It can be used in either classroom teaching or self-paced learning conveniently. In contrast, online learning requires the connection to the server where the e-learning system is stored. The richness and the convenience of maintenance and updating of learning materials attract enormous number of people. Various learning data recorded in server are goldmine to analyze learning styles and learner's characteristics, monitor learning progress, make recommendation or tutoring, adjust and improve educational design and so forth. Massive open online course (MOOC) is the most successful case. By the end of 2017, the total number of MOOC learners reaches 81 million, more than 800 universities offer 9400 courses [10]. Besides learners are being excited to be potential students of famous university and study various knowledge, they are provided the opportunity to have credentials which may contribute to their future career and life. This is different from previous online e-learning. Not limited to that, MOOC-based degree programs have been initiated. Online graduate degree would be highly beneficial for both learners and educational institutions. Mobile learning emphasizes the mobility of learning includes mobility of learner and mobile technology. Sustained by portable devices, learners are able to learn ubiquitously in flexible time. Typical cases include spontaneous learning, people desire to use spare time effectively, urgent need of learning and so on.

Michael G. Moore discussed three types of interaction in distance education [11] which are also applicable in e-learning. Student-content interaction is the fundamental type of interaction. The education will be impossible if this interaction vanishes. Watching video, interacting with courseware etc. could be the representatives of the basic e-learning. The second type of interaction is student-instructor interaction. On the basis of interacting with content, instructor is involved to facilitate learning. The role of instructor could be active, for example, control the process of learning; or could be passive, like repliers who response to learners' requests. The third form of interaction is student-student interaction. The typical case is collaborative learning which aims to achieve better learning outcome than individual learning through sharing ideas, peer evaluation and monitoring, etc. Additional two types of interaction proposed by Anderson T. [12] do not involve students, however, they complete e-learning system which function as supporting system or module. Teacher-content interaction refers to teachers create learning materials for e-learning. High quality learning content requires good design and suitable media, will produce excellent effect. The process of creation should not be cumbersome, which means much time and complicated skills are undesirable. Authoring system is the representative. The other type of interaction is content-content interaction which concerns intelligent and adaptive supporting function. The system collects learners' data during learning, analyze, evaluate, make decision, and feedback autonomously. Learning content recommendation system is one example, which recommends learners appropriate content based on previous learning records.

Depth of interaction is another direction about interactivity for describing e-learning system. The first tier of interaction refers to learners watch and/or listen to the content what e-learning system demonstrates includes text, graphics, video, audio etc. In the second tier, interactive components are incorporated in system, such as making selection or arrangement, speaking follows content which depend on the learning design. In the third tier, e-learning system creates or simulates a specific environment or situation lead learners to immerse themselves in the context, then they are motivated to learn from the interactions that designed in the course. Educational game or game-based learning, virtual laboratory etc. fall in this type of e-learning.

Learning content-oriented perspective is another significant factor for designing e-learning system. The basic form of e-learning content is the digitalization of traditional content associated with text, figures, animation, audio, video etc. More advanced strategy of learning content design involves project-based learning, problem-based learning and so forth. Functionally speaking, the available tags for e-learning system include learning system (knowledge or content delivery system), supporting system, management system, evaluation system etc. From learning style aspect, we could have self-paced/autonomous e-learning, personalized e-learning (one-on-one tutoring), one-to-many e-learning etc. The ideas of teacher-centered and student-centered instructional design are applicable in e-learning naturally. Another interesting and promising perspective refers to intelligence of e-learning system, in this context, branches into adaptive e-learning and e-learning that manipulated by human (either instructors or learners).

Aforementioned enumeration virtually covers every characteristic about e-learning. They are not exclusive in deploying e-learning. The characteristics should be integrated according to the purpose of e-learning system and its application environment during the design.

1.3 Theoretical Foundations

Both technology and learning theory advance e-learning. This section is designated to introduce briefly the main theoretical frameworks contribute in educational design, as well as in e-learning naturally. Behaviorism was started in the early 20th century based on the behavior analysis of animals in learning experiments, e.g., Ivan Pavlov's dog experiment, and generalized to humans. In behavioral psychology, learning is considered as the change of learners' behaviors which is achieved through using reinforcements and repetitions according to the principle of "stimulus-response." Briefly speaking, a discriminative stimulus is presented to trigger a response which is reinforced by desired response is rewarded and undesired response is punished. Positive effect of behaviorism applied in teaching and learning were demonstrated generally [13]. The instructional design based on behaviorism is teacher-centered, teachers control the learning, students follow their instructions, what is right and what is wrong, then perform and reinforce accordingly. However, it is challenged and criticized by ignoring the internal processes of learning and the internal mental states involved in learning. Human learning is complex and has multiple levels, which cannot be explained only by external stimuli, especially for the high-order learning. People's thoughts, beliefs, motivations, reflections, emotions, cognitive load etc. should and must be taken into account in explaining and understanding learning. Although this pedagogical perspective is relatively superficial in conceptualization of learning and may have some flaws in learning outcome, it is still one important component in education and has excellent effects, especially for some particular educational occasions it works exceptionally effectively. Moreover, behavior modification of learners is the important measure of learning outcome.

In 1950s, a shift of learning theory started, from behaviorism to cognitivism. The cognitive process inside human head, viewed as "black box" in behavioral model, entered the spot light. Influenced by information technology, computer science and neuroscience, cognitive theory conceptualized students' learning as an information processing activity includes how the information from social environment is received, how information is processed and stored into schema (transferred into knowledge expressed as symbolic representations), and how information is retrieved upon recall. Memory works indispensably in the learning process. A widespread memory model is the two-store (dual) memory model [13], i.e., short-term (working) memory (STM or WM) and long-term memory (LTM). Information is firstly received by sensory registers, then perceived through comparison with information in LTM and enters STM (WM). After that, it could be either encoded and stored in LTM or lost. Attention plays crucial role in the process, since STM (WM) has limitation in capacity and duration. Initiative is inherent to cognitive theory indicates learners must involve themselves in their own learning actively. Therefore, learning designed based on cognitivism is learner-centered education. Instructional components are contended an important portion in learning by cognitive perspective, as well as environmental factors. Nevertheless, changes in behaviors are not overlooked, they are measured as the outcome of processing occurred in learners' mind.

Strictly speaking, constructivism is not a new learning theory but can be considered as a branch of cognitivism because they both conceive learning is cognitive process in mind. The distinction between them is how the knowledge is construed. Cognitivism contends knowledge exists externally and is independent to mind and transferred from external world into internal cognitive structure. Constructivism concur with the existence of real world, however, knowledge stems from people's individual interpretation concerns the world which is mentally constructed based on personal experience and interaction that might differ from person to person. Several key terminology introduced in Jean Piaget's and Lev Vygotsky's theories address about construction of knowledge is the core of constructivism and has profound impact on educational design. Assimilation and accommodation are two component processes while we construct knowledge based on the interactions between external reality and internal cognitive structure. Assimilation refers to incorporate new experiences into the existing framework with no modification on framework. When there is conflict, people alter their perceptions about reality to fit cognitive structure. Accommodation refers to reframe internal cognitive representations of the external world to provide consistency with the new experiences. The two processes work complementarily in our mind leads us to learning. Another key concept is zone of proximal development (ZPD). It is originally introduced by Vygotsky as "It is the distance between the actual development level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" [14], and has been expanded and developed. In simple terms, it is the territory between what a learner can do unaided and what he/she cannot do. It indicates where learning should take place and the amount of learning. Constructivism is more like an epistemology about the nature of learning [13], rather than a particular pedagogy. The framework for school reform and redesign prepared by American Psychological Association embodies constructivism learning ideas. These learner-centered psychological principles are categorized into four groups, cognitive and metacognitive factors, motivational and affective factors, developmental and social factors, and individual differences factors [15].

Different understanding about learning have nurtured different learning designs and models, and promoted education to more effective and efficient manner. These theories do not work exclusively, in fact, they are blended in modern education.

From behaviorism to cognitivism and constructivism, the revolution happened in the comprehension about how learning occurs. Cognitivism and constructivism encourage the educators to design students-centered learning. Students' needs, characteristics, skills, prior knowledge, emotion and cognitive status in learning etc. need to be incorporated in design.

1.4 Strength, Weakness, and Trend of E-learning

The strength and weakness about e-learning systems are designated to discuss in this section. A general perspective is used which leads to a union of qualities concerns advantage or disadvantage considering the diversity of e-learning system, which means, some of them cross many types of e-learning, and some of them appear in a few types or specific one of e-learning. After such discussion, the trend of e-learning development is introduced, as well as personal considerations for future e-learning.

The first impression about e-learning might be the flexibility of space (location) and time for many people. Learners could access e-learning in anywhere and anytime once the content has been distributed. Geographical gap is bridged as the advances of communication technologies, especially for internet-based e-learning form. Instead of adhering to the schedule of educational institutions, learners are free to study in their convenient time.

Another freedom for learners is they can learn on their own pace according to their own learning needs, since learners are unlikely to have the same knowledge level and learning rate at the start of learning. A plenty of e-learning systems have been designed in self-paced or self-driven form that allows users to learn flexibly. Individual learning is not the only option, social interaction is also an important factor in e-learning. Collaborative learning, group learning, and one-on-one tutoring etc. these learning forms are available in e-learning too. Ways of communication equipped in e-learning offer learners the opportunity to interact with others, e.g. discuss, exchange ideas, work collaboratively, synchronously or asynchronously. In addition to peers, supports from instructors and experts in the learning community are also accessible.

The richness of learning content is one key advantage of e-learning. People are different in learning styles and preferences. Some of them are sensitive to visual information, some may prefer auditory information etc. Besides media types, representations of learning material are important too. Abstract conceptualization may be suitable for some learners, some others perform better when concrete experience is provided. These aspects are considered and implemented in e-learning. The tactics of designing learning content are not restricted to abovementioned points, learning objectives are the principal considerations in learning. Demonstration could be sufficient for grasping knowledge, however, interactivity is a necessity for mastering skills. Various levels of interaction that founded on miscellaneous technologies and ideas of design are viable in e-learning. The simple interaction includes input, selection, drag-drop, click etc., which can be implemented effortless. More sophisticated interaction such as simulation, motion sensing, and virtual reality, requires specialized knowledge and competence and/or specific devices for realization. The ideas of design, such as gamification and branching scenario, deepen the interactions. Gamification can promote engagement of participants; branching scenario is able to enrich the learning interactivity that students could gain more from their failure experiences than simply being told what are correct. Another benefit is learners can develop proficiency without risks, for example, training for experiments and machine operations. Maintenance of content in digital form is easy and quick. It is important to keep students up-to-date on knowledge.

Evaluation convenience is another noticeable advantage of e-learning. Diverse types of assessment can be employed for evaluating various learning objectives. The answers of learners are recorded easily, and most types of questions in assessment can be examined and reported automatically. Immediate feedback helps students to adjust learning timely, and also releases teachers from repetitive labor. Moreover, conducting formative evaluation becomes convenient for teachers, eventually to make appropriate learning interventions. In addition to test-like assessment, other types of data concerns learning recorded in e-learning system, such as course completion rates, learning footprints, and posts in forum, are also valuable in evaluation. They are good sources, as well as performance of students, for learning analytics or educational data mining to extract implicit information, such as learning behavior patterns. Such information is valuable and convenient for every stakeholder in e-learning, possible instances are given as follows. For learners, it is a part of e-portfolio for them to evaluate and reflect on their learning; for instructors, it benefits both student-oriented evaluation, such as supporting interventions for students who need extra support, and curriculum-oriented evaluation, for example improvement of course or development of new course; for researchers and functional groups of e-learning system, evaluation and improvement of student models and domain knowledge structure rely upon such information; for administrators, it is useful for measuring the effectiveness and efficiency of system.

The last point about advantage of e-learning included in the dissertation is cost-effective. This merit might be not apparent as other merits, as it can cost more than traditional classroom learning in deploying stage. However, long-term expense indicates e-learning is more economical. E-learning content can be easily reused, reproduced, distributed, and updated with less expense and time. Moreover, the advantage becomes more prominent along with increase of learners. Extra cost in face-to-face education, such as expense of transportation and accommodation for lecturers or learners, would be negligible in e-learning. Given learners are already familiar with the technologies which are utilized in e-learning, there is no training cost for them.

Even considering all the advantages of e-learning, the concerns and weaknesses cannot be overlooked.

E-learning is supposed to accommodate large number of learners with various types and characteristics, which implies much more effort and labor is required from instructors to design and develop a course in e-learning than in conventional educational form. To keep students engaged and motivated in e-learning is a challenge. A good instructional design is indispensable. In one hand, instructors need to ponder the best ways to combine technology and instruction which means technology can serve instruction perfectly, since the essence of e-learning is the integration of education and technology to optimize learning experience. On the other hand, instructors should have impressions about technologies at least what technology can do for learning, furthermore, training of using tools for designing and developing e-learning efficiently is necessary for instructors.

Feedback is utmost important in learning. Nonetheless, e-learning cannot compete with face-to-face education, although it can offer test-like evaluation faster to learners. Instructors can hardly know how students learn as there is little information can be retrieved from the places where students actually perform e-learning. Some learning data recorded by e-learning system could be good sources, however, they are primitive and redundant data and yet not enough. Even though teachers could get clues for who may need assistance, they still may have troubles in replying in time considering the large scale of students. Lack of feedback information hinders students to get timely and appropriate guidance, i.e. personal instruction. It is a hazard for learners' motivation and learning efficiency, also for learning content design and improvement. E-learning can be equipped with different types of communication approaches, nevertheless, it is a solo act in most of time, and the contact among people is not completely same as it is in real situation. Social interaction and emotion loss could frustrate learners. They may feel the sense of isolation. Performance of e-learning relies on self-discipline heavily, learners need to regulate time for themselves, to keep themselves focus that away from distraction. It would be optimistic if we assume all of student can do it. Among the students who can, some may perform better with people around who can motivate them well. To keep learners motivated and engaged is one important goal to which e-learning need to devote attention. Otherwise, we have to risk low learning efficiency and high probability of drop outing of learners. Another flaw of e-learning is the potential cheating behaviors of students. Without surveillance of teachers, students may cheat on the test for higher scores.

The concerns about the data recorded and collected in e-learning, such as data privacy, data ownership, and data protection, can be seen as another weakness of e-learning.

The trends of future e-learning research and development would have many orientations to work on while different characteristics and technologies are considered, for example, adaptive learning, social learning, micro-learning, and mobile learning. From the perspective of learners, e-learning should be personalized and adaptive in content, learning strategy, evaluation etc. The length of knowledge should not be long, so the fragmented learning can be realized to make best use of time. In educators' perspective, the e-learning system should be versatile that their ideas about learning could be realized without much effort and pain, from the design for one specific knowledge point to an entire course. In contrast to such intellectual work that e-learning system is not completely competent, the e-learning system should take charge of manual work as much as possible to release teachers from it. For example, automatic evaluation of quiz, feedback collection and demonstration, detection of at-risk student etc. Based on the perspective of administrators, in addition to be user-friendly in managing e-courses and e-learners, e-learning system should become more cost-effective in deployment, maintenance, and upgrade.

Among the future trends, "Intelligence" is one of the keywords relates to most of them, typical cases are adaptive learning and personalized learning. Potential users of e-learning system would have slim chance to have similar knowledge level as starting, similar learning styles and similar learner characteristics. Their learning objectives may be different too. As claimed by cognitivism and constructivism, learning is an active process refers to internalize knowledge or construct own interpretation based on outer experiences, i.e. learning content. If learning path, learning pace, content type, and learning strategy etc. are tailored according to individual's needs, interests, learning styles etc. and adapted based on real learning situation automatically, the learning outcome would be optimized. Personalized learning and adaptive learning aim for this purpose. The ideally final goal is tough and seems to be far away from current status, but lots of researches and practices are working towards it purposefully or unintentionally.

My research in pursuing the doctor degree was motivated by this consideration. I devoted myself to recognize the learning states of learners in the first portion of research, because such feedback is important and valuable for teachers to understand how learning is going on which underlies the personalized learning and adaptive learning. During self-paced e-learning, various education related data can be generated by e-learners, and captured and recorded by e-learning system, such as video, audio, learning footsteps, mouse click, keyboard stroke. These data contain details about learning, however, they are too much redundant. The research aims to extract the learning states information from the big amount of redundant data. That information would help teachers in designing and conducting personalized and adaptive learning.

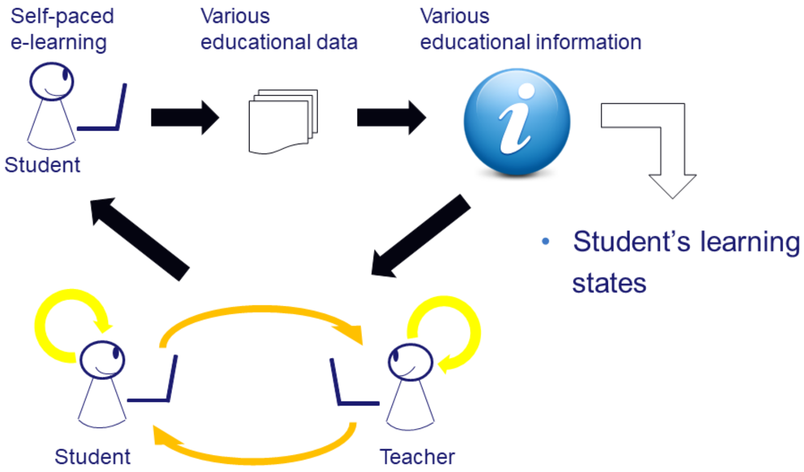


Fig. 1.1 Feedback loop of recognized learning states

The designed framework for this purpose was verified on CALL of English learning materials in Kyoto University, however, the findings can be generalized in macro perspective of e-learning. Recognition of learning state information not only can serve well for self-pace e-learning but also would be beneficial for other types of e-learning. It focuses on the commonality across e-learning. Such functionality could be expected to be integrated in all types of e-learning in the future.

The second portion of research I dedicated to was inspired by the idea of learning content adaptation. In contrast to abovementioned general perspective, the research focused on a specific e-learning. The typical problem is found in the territory of language education. The new functions to e-learning system are designed based on the particular attributes of language education and the particular characteristics of learners, and the current e-learning system is enhanced by affordable technology.

No matter which future directions for research and development of e-learning, suitable technology should be selected cautiously and integrated with instruction perfectly according to specific functionalities or issues. Furthermore, the instruments required by technology should be as low cost as possible considering the size of e-learning.

1.5 Target and Purpose of the Study

As briefly introduced in previous section, utilization of information communication technology in facilitating e-learning was investigated in two research directions. Self-paced e-learning type was selected as the target environment of investigation, specifically speaking, CALL system developed and used for self-paced learning in Kyoto University.

The first direction of research refers to learning state recognition through visual sensing. The designed and evaluated supporting functionality can be integrated into CALL, and is expected to work on other kinds of e-learning too. Personalized learning and adaptive learning rely on various feedback from learners to carry out personalization and adaption in e-learning. Feedback can either be proactive that learners write directly to instructors about their feelings or demands; or passive which requires to infer from observable information of learners, for instance, non-verbal information. Explicit feedback could be scores of tests or quizzes; implicit feedback includes learning paths or actions on e-learning system etc. E-learning systems could have advantages in gathering partial of feedback effortlessly, however, acquisition of other kind of feedback without too much effort is a challenging work especially for self-paced e-learning. Learners' cognitive-affective states when they perform self-paced e-learning provide significant feedback that can serve in many phases, including design, development, utilization, management, and evaluation of processes and resources for learning. However, it is a challenging task to obtain such valuable information as mentioned earlier. To handle this issue, an innovative system is designed to estimate learners' internal states during learning through introducing pattern recognition technology which takes image processing results as input. Furthermore, a crucial issue for practical application about estimating the learning states of a new learner whose characteristics are not well known in advance is explored. Because of e-learning is designed to face a variety of new students who may have various backgrounds and join at different time and places. Considering the diversity of learner characteristics, a mechanism to adapt the recognition system to new learners is desperately needed.

The second portion of research concerns pronunciation support for language learning through 3D sensing of a face. Different languages do not have the same phonetic systems. This diversity requires learners to learn new movement of articulators for new pronunciations and also necessitates a lot of practice to achieve proficiency. Difficulties and barriers brought by the diversity would be attenuated if learning materials could be more targeted on the problematic points which enable learners to be aware of the problems easily and offer more comprehensible learning information. Thus, the learning content and supporting functionality provided by language e-learning system need to be adaptive to the characteristics of target languages and target learners with the final goal of maximizing the learning performance.

Although evidences have showed the importance of visual information in correct pronunciation, most researchers and practitioners using CALL systems have paid considerable attention mainly to audio, the importance and effect of vision in articulation have not been well explored. Traditional explanations given by written texts and even conventional multimedia approaches of showing pictures or videos do not give sufficient explanations of the articulations. For this problem, a novel method for producing an enhanced visual material for pronunciation learning is designed. A typical difficulty when Japanese learners study foreign languages is the articulation of rounded and unrounded words which is selected for evaluating the proposed idea. Japanese language does not have clear distinction between rounded and unrounded vowels. However, some other languages do have clear distinction between rounded and unrounded vowels, for example, Chinese, German, and French. Lip protrusion is the key factor in pronouncing rounded vowels correctly but it is not a must in Japanese language. Towards this goal, the enhanced learning material is produced with respect to visual emphasis of lip protrusion. Three-dimensional sensing of a face, image processing technology, and pseudo coloring idea are involved.

1.6 Structure of Dissertation

The dissertation is organized as follows. Chapter 1, at first, presents a very brief introduction about the concrete purposes of the research. Then, details about research background and environment refer to e-learning are shown, include the descriptions with respect to definitions, types, theoretical foundations, strength, weakness, and trends, as well as the targets and purposes of the study. In Chapter 2, firstly, background concerning learning state recognition is introduced, then followed by the statement about selected targets of learning state in the research. The details concerning framework and implementation towards recognition of learning states are presented. The evaluation of recognition proposal is shown in the end of this chapter. Chapter 3 elaborates the investigation about the essential problem in practical application of learning state recognition, classifier selection for new learners as inter-personal difference cannot be ignored. The strategies for handling inter-personal differences and classifier selection are introduced and evaluated based on experiments. The further investigation refers to selection of better representative samples for improving classifier selection performance is explained afterwards. Chapter 4 discusses the need, importance and design of vision-emphasized learning material for language education, specifically, pronunciation learning at first. The details refer to implementation of the idea of design, experiment of verification and result are elaborated in the second half of the chapter. In the final chapter, conclusion and summary of research work is drawn. Consideration of future work and limitations in the present study are also discussed in Chapter 5.

1.7 Summary

Education has been being revolutionized by information and communication technology. E-learning has boomed over many years and become a widespread manner of learning. It is difficult to give an "official" definition for e-learning, however, the distinction between e-learning and educational technology which is another term refers to both education and technology and has the official definition given by AECT, is becoming more and more vague. Hence, they are deemed to be synonym. E-learning system has many categories if different characteristics, for example the technology, learning strategy, and idea of design that applied to e-learning, are focused. Such diversity implies e-learning should be designed according to learning needs, purposes, requirements, and application environment, with knowing the advantages and disadvantages of various types of e-learning. The development of learning theory reveals the importance of motivating and engaging students in learning. This challenge is more momentous in e-learning situation since it more relies on autonomy of learners.

Among the trends for future research and development of e-learning, the keyword of “Intelligence” is more promising. It could help e-learning system to advance the motivation and engagement of learners to a higher level. The typical cases include adaptive learning and personalized learning.

Our research was undertaken to gather important and valuable feedback for teachers to make e-learning more adaptive and personalized for e-learners, and to produce more attractive and informative e-learning materials for e-learners, which are motivated and inspired by this keyword. For the former one, a general supporting functionality concerning learning state recognition of learners throughout the learning process was investigated. For the latter research, new functionality for enhancing visual information perception in pronunciation learning was designed for CALL, due to the importance and effect of vision in correct articulation have not been well explored. An issue of rounded pronunciation for Japanese learners was selected for verification.

Chapter 2

Learning State Recognition through Visual Sensing

2.1 Background

Learning is a complex internally cognitive process as elaborated by cognitivism and constructivism. Significance of autonomy of students cannot be denied and ignored. It is crucial for any types of education to motivate learners to participate in learning proactively as possible as they can. It is even more momentous in self-paced e-learning. Intrinsic learning needs of learners and the desire to achieve their learning objectives would impel them, however, inappropriate learning design, e.g. inappropriate learning materials and learning strategy, delayed intervention, would attenuate even obliterate their motivation. In the contrast, if learning is adaptive and personalized, motivation will keep prospering and learning will be effective and efficient. Mental status of learners is important and valuable information to understand how learners react to learning. Imagining in traditional classroom education, while instruction is being conducted teacher is observing students' responses simultaneously, then adjusts steps of instruction and/or undertakes suitable interventions accordingly. This mechanism, however, is quite challenging in self-paced learning as well as in face-to-face education when class size is big.

Mental status refers to many terms, cognitive or affective, such as affect, emotion, motivation, reward, attention, which represent diverse understanding that from different communities such as psychology, cognitive science, neuroscience, engineering, computer science, sociology, philosophy, and medicine [16]. The discrimination of those terms is not the interest of this study, hence, umbrella word “cognitive-affective state” is used to represent them. The role of cognitive-affective state in education is undeniable and theorists started researching it long time ago. The influence has 2 sides, like the coin, positive and negative influence. Some cognitive-affective states could energize students that help them learn and perform more successfully and efficiently. Some other cognitive-affective states could interfere their learning, possible consequences include students do not learn strictly or methodically, distraction from learning etc. Positivity or negativity of impact is not determined by valence of cognitive-affective state simply, for instance, negative feeling such as difficulty could work positively in learning, and positive feeling like excited may interfere learning. In addition to valence of cognitive-affective state, arousal is also a key factor, for example, calm is no-arousal which could be a good state for learning, excited is arousal, it also works fine in learning with proper degree, but overly excited may lead to unsatisfactory learning outcome. Sometimes dominance is considered as the third dimension on the basis of valence-arousal cognitive-affective state model, which leads to more complicated influence analysis about cognitive-affective state upon learning. In contrast to the history of research concerning cognitive-affective state in learning, research refers to automatic computing or recognition of cognitive-affective state is relatively new. The terminology of affective computing is originated by R. W. Picard in 1995 [17]. A new direction for cognitive-affective study refers to learning in information era has been proposed and gained more and more attention.

Plenty of researches have been undertaken to study affective computing in learning environments. These researches had different affective-cognitive states as targets and utilized various modalities for recognition. Butko et al. [18] proposed an automatic facial feature extraction system that was designed based on the Facial Action Coding System (FACS). Seven GentleBoost classifiers were used to recognize the expression of being interested, thinking, tired or bored, confused, confident or proud, frustrated, and distracted on the part of a learner during interactions with a teacher. Ammar et al. [19] focused on detecting the contour of eyes, eyebrows, and mouth. Distance changes among these, opening of the eye, outdistance between the interior corner of the eye and the eyebrow, opening of the mouth in width, opening of the mouth in height, outdistance between the eye and eyebrow, outdistance between the corner of the mouth and the external corner of the eye, were used for classifying six universal emotions, i.e., joy, sadness, anger, fear, disgust, and surprised. Whitehill et al. [20] investigated the correlation between facial expressions and self-reported difficulty. The facial expression recognition was implemented based on FACS. Zakharov et al. [21] also used facial features to identify whether the affective state was positive or negative. This enabled a pedagogical agent persona to respond to a learner's action on the basis of the learner's cognitive and affective states. D'mello et al. [22] studied dialogue features extracted from conversations between learners and an intelligent tutoring system and on the classification of boredom, confusion, flow, frustration, and neutrality. Litman et al. [23] used acoustic and prosodic features of students' speech to detect negative, neutral, and positive emotions. In [24], physiological signals were used to recognize emotions developed during the learning process. Three kinds of sensor were used for skin conductance, blood volume pressure, and electroencephalography (EEG) to recognize four kinds of emotion, engagement, confusion, boredom, and hopelessness. Yang [25] used combinations of mouse operations and facial information to detect attending and responding states of students, including attentive vs. inattentive and active vs. passive, respectively. Woolf et al. [26] combined four sensors to recognize confident, frustrated, excited, and interested, using facial expressions, postures measured by the pressure from the seat cushion and back pad, a learner's hand pressure measured by a special mouse, and skin conductance.

Much research progress has been made with respect to cognitive-affective state recognition in learning environments, as well as the exploration for educational intervention based on the recognized cognitive-affective states. However, different research perspectives for affective computing in learning remain to be explored. This research is undertaken from practical application orientated perspective. Problems which have been less investigated in other perspectives are the focus and studied.

2.2 Targets of Learning State

Various cognitive-affective states have been targeted, but some of them are either not closely relevant to learning performance, or could not be used readily by teachers. In this research context, three categories of cognitive-affective states are selected on the basis of teachers' standpoint, specifically, concentration-distraction, difficulty-ease, and interest-boredom. They are selected according to their direct and close relation to learning. In this sense, happiness, sad, anxiety such emotions indirect to learning are not the targets for recognition although they could influence learning too. Abovementioned learning states of learners are important and valuable information for instructors to understand how learners react to learning. One typical example can be imagined, when teacher realized the students are lost in the educational content from the observations of their non-verbal information, e.g., they do not pay enough attention on the teacher or they show facial expression of feeling difficulty, the teacher would make appropriate adjustments in instruction to regulate students' behaviors, such as reminds students, increases the interestingness of learning material, gives more concrete or students-related examples, introduce prior knowledge to current knowledge.

2.2.1 Concentration-Distraction

Concentration-distraction could be considered as the most important learning state index as it is the foundation for learning happens. Some other close related terms like attention control, flow, engagement, have also been intensively investigated in education. As in previous introduction about cognitivism, STM(WM) is limited in capacity and duration, attention plays like a filter to control the information that enters in human system. The significance of concentration in education has been investigated and emphasized in many studies. A substantial literature has revealed that there is positive correlation between student concentration and academic performance and achievement, which reviewed in [27]. The higher the concentration level students have, the better the learning performance they can attain. To make good use of student's power of concentration is crucial for development and meaningful learning, which is accordant with the core of Montessori philosophy [28]. It will be great help if teachers are able to have concentration information of students.

2.2.2 Difficulty-Ease

Difficulty-easiness can be a good index of learning material appropriateness. According to the zone of proximal development, appropriate content is that which a learner can understand, but does not understand yet. The learning efficiency and outcome will deteriorate, if the learning content is not suitable for learners. Difficulty of learning content plays a crucial role in maintaining students' concentration. As [29] described by setting appropriate challenges-tasks are neither too difficult nor too simple for ones' abilities-individuals are in the state of flow which is a state of concentration so focused. If teachers understand students' feeling to content, they could adjust to keep the tasks in the zone of proximal develop of students by setting appropriate challenges.

2.2.3 Interest-Boredom

In order to motivate students, in addition to appropriate challenge, interestingness is also necessary. In the meta-analysis of more than 50 studies which was conducted in [30], a positive relationship between interest and academic achievement has been found. Shirey [31] reported that information can be learned readily by college - age students, if it was interesting to them. From this viewpoint, interest-boredom is an important factor for learning efficiency. Interestingness is another important factor in motivating students to remain in concentration. When peoples experience interest, their attention is directed and they are engaged in activity [32]. If learners have strong interest in learning, learning efficiency improves. Teachers need to design learning to enable students to participate as actively as possible.

It is widely recognized that they are mutually related. Concentration on a target can foster interest. Interest can be referred to as an important driving force which can result in concentration. However, concentration can be affected not only by interest but also by a variety of factors both inside and outside of learners, e.g., motivation, fatigue, stimulus strength, and time and place of learning. For difficulty and interest, Silvia [33] reported that both low and high in difficulty may cause low rating in interest, which implies that interest captures an aspect different from difficulty. Therefore, our scheme deals with those three as separate indices. Its advantage is clear if we think of their usage. Difficulty is an important feedback that is helpful for keeping the level of the learning materials adequate, e.g., learning materials need to be easier if learners feel too much difficulty and vice versa. Interestingness is also good information to make learning materials attractive. Teachers get good feedback how much learners are interested in each portion. Concentration is helpful for knowing the attitude of a learner, and useful for evaluating a learner. We take the same approach for the estimation of those three internal states.

2.3 Framework and Implementation

As introduced in background of this chapter, a variety of modalities to sense learners' behaviors can be considered. One noteworthy issue in practical application that must be considered is the applicability of modality and equipment for raw data measurement. The choice must provide rich information, but more importantly, must be easily integrated into practical e-learning environments without imposing additional constraints on learners or learning environments. Visual sensing of learners is a good choice, because it is non-intrusive and non-contact, and a small camera can be integrated easily with existing e-learning systems with current technologies, without heavy constraints and cost is preferable.

An innovative system is designed to estimate learners' internal states during learning through introducing pattern recognition technology which takes image processing results. Furthermore, a crucial issue for practical application on estimating learning states of a new learner whose characteristics are not well known in advance is explored. Because of e-learning is designed to face a variety of students. Considering the diversity of learner characteristics, a mechanism to adapt the recognition system to new learners is desperately needed.

The overview of proposed scheme is illustrated in Figure 2.1. There are three principal components. At each e-learning site, a Red-Green-Blue-Depth (RGB-D, color and depth) camera is mounted on an existing e-learning system, as shown on the top left in the figure. A video of a learner's face and upper-body is captured by a Kinect camera in RGB-D format. Learning states are estimated through the module as in the bottom portion of Figure 2.1. In this component, visual features are obtained through the processing of RGB-D images. Subsequently, they are input to Support Vector Machine (SVM) to recognize learners' internal states. Details will be given in the following sections. The detected information is summarized and presented to teachers in a comprehensible manner, as depicted in the top right portion of Figure 2.1. This information enables teachers to provide mentoring, modify materials, or conduct educational analysis etc. For instance, teachers can find and pay additional attention to those learners who need more assistance, and provide personal tutoring, recommend learning content and strategy accordingly. Teachers can also design and adjust materials to keep learners motivated, e.g., by setting tasks a step ahead of learners' current skill levels, increase the interestingness. Various educational evaluations, such as diagnostic, formative, and summative assessment, can benefit from such information too, as well as learning analytics. Actual design of this utilization is left for future work.

Some ideas about the user interface design in the third component can be considered, and the examples are shown as follows. The simple visualization of learning states recognition information is demonstrated by Figure 2.2, Figure 2.3, and Figure 2.4 from different view of points. Figure 2.2 shows the overall information about one learner, how many learning sessions have been conducted, the basic information about one learning session, and the links for other information browsing. Figure 2.3 demonstrates the view of a specific learning state index. The pie-chart shows the distribution of scores, with the clicking on the chart, the corresponding learning segment's links are provided for further browsing of detailed information. Figure 2.4 shows the view of 3 learning states together.

For browsing actual learning scene, there are two main requirements for the brows-

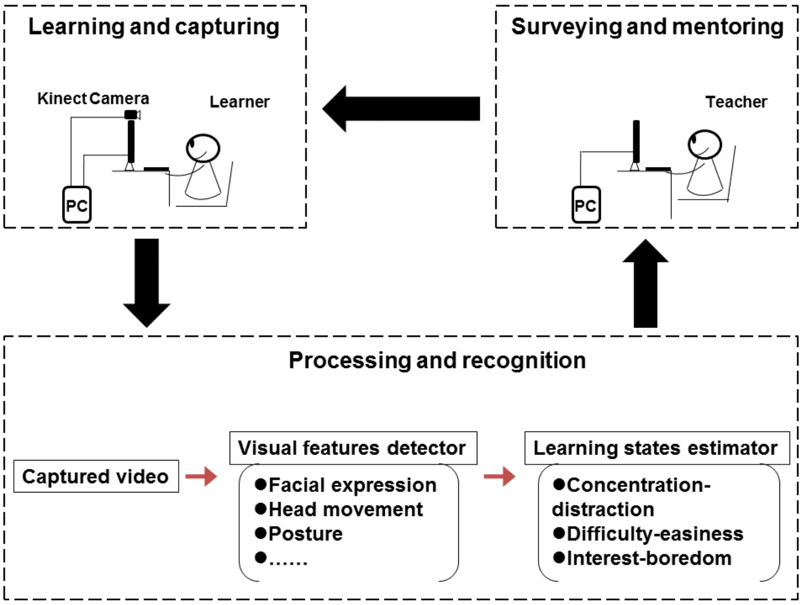
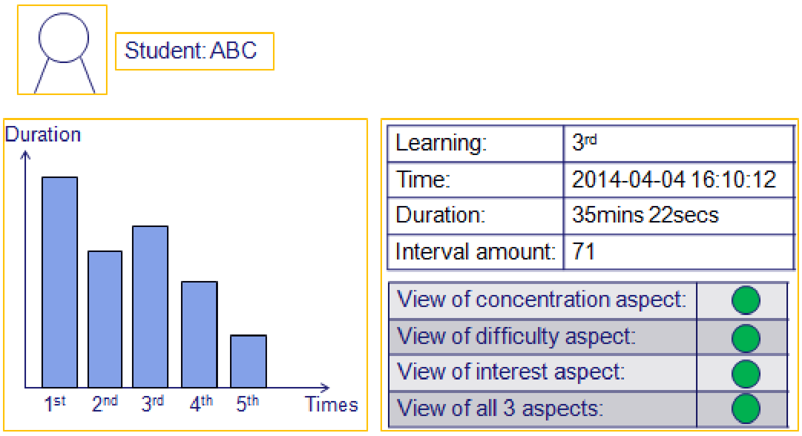


Fig. 2.1 Overview of scheme

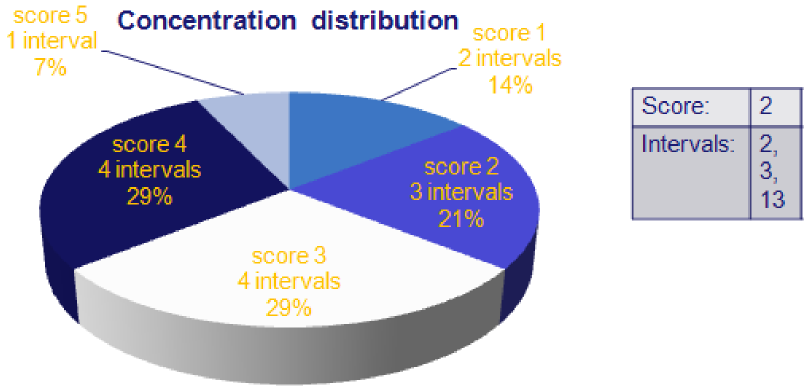


Learning:	3rd
Time:	2014-04-04 16:10:12
Duration:	35mins 22secs
Interval amount:	71

View of concentration aspect:	●
View of difficulty aspect:	●
View of interest aspect:	●
View of all 3 aspects:	●

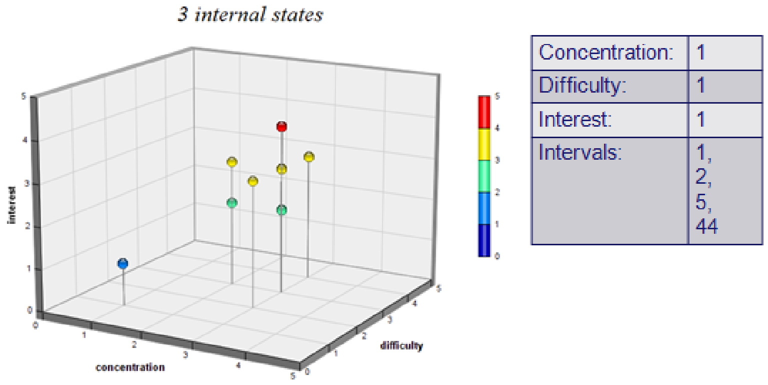
Fig. 2.2 Portal of information browsing

ing interface: when teachers look at the interface at the first sight, it should give teachers intuitive understanding of learning state about the learning; it would be very beneficial and efficient if the interface can provide various temporal resolution in browsing. Teachers can grasp global impression with high temporal resolution, as well as when teachers want to check status in detail, the advanced information of the particular period can be shown. Figure 2.5, Figure 2.6, and Figure 2.7 gives an example of design that meets the above requirements. There are 3 main portions; the uppermost area shows the general information, and the middle presents a learner's learning



Score:	2
Intervals:	2, 3, 13

Fig. 2.3 View of a specific learning state



Concentration:	1
Difficulty:	1
Interest:	1
Intervals:	1, 2, 5, 44

Fig. 2.4 View of 3 learning states together

state information. The estimated learning state scores-illustrated by the height of the bars, and representative face images are shown together to provide teachers intuitive understanding. The representative face images could be the images that present the learner's face at particular time, or more sophisticated rating system could be possible. In the latter case, the representative images could be the face image at the best, average, and worst case in the target period depending on the rating system. The lowermost portion presents the main learning content mostly shown in corresponding period.

The time unit in the interface can be changed to any multiple of intervals on the request of the teacher. If it is set to 6 minutes, teachers can browse 1 hour's status by 10 periods. It is much more efficient than watching through the captured videos of a learner. Figure 2.6 and Figure 2.7 give the examples of different temporal resolution.

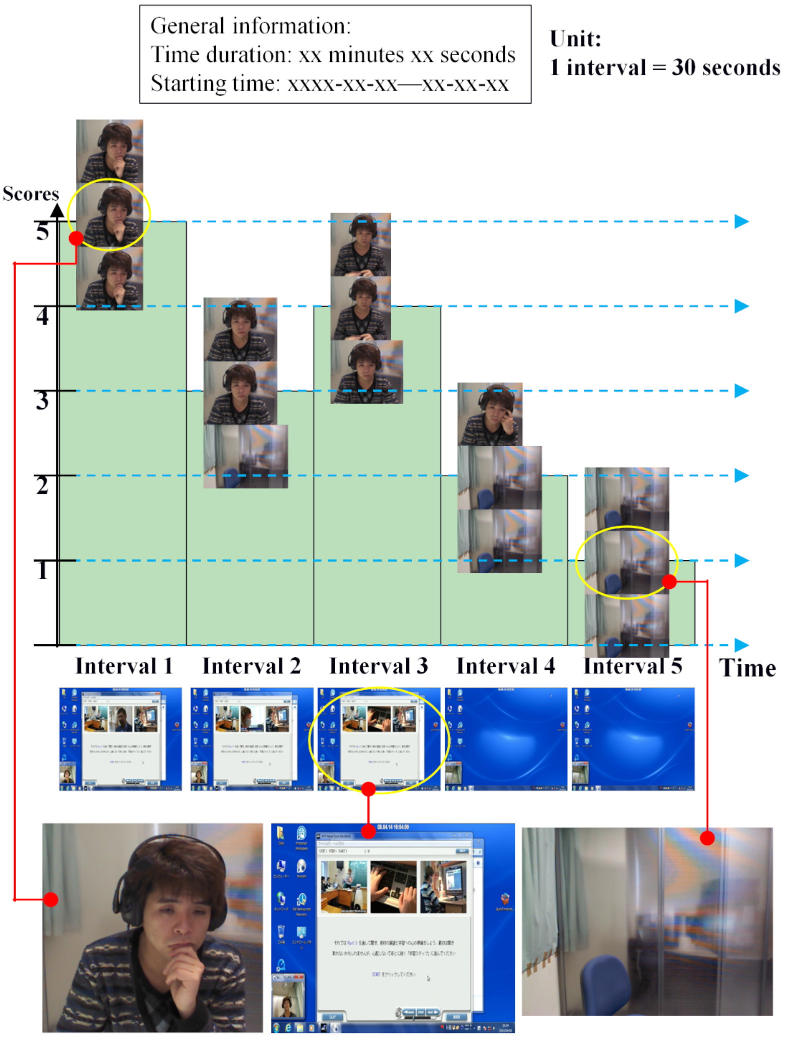


Fig. 2.5 Example of learning state browsing design

The highest, average, and lowest learning state score of the period are denoted by the corresponding face images.

The view point from specific learning content could also be possible, since the same learning content would be displayed for a certain period in most of time, especially the learning content is difficult for the learners. Teachers may have interest in how long has been spent on a particular learning content, and how the learners feel and react to it. Such information would be helpful for learning content improvement and personalized tutoring.

Figure 2.8 demonstrates the prototype of enhanced e-learning system, it is also

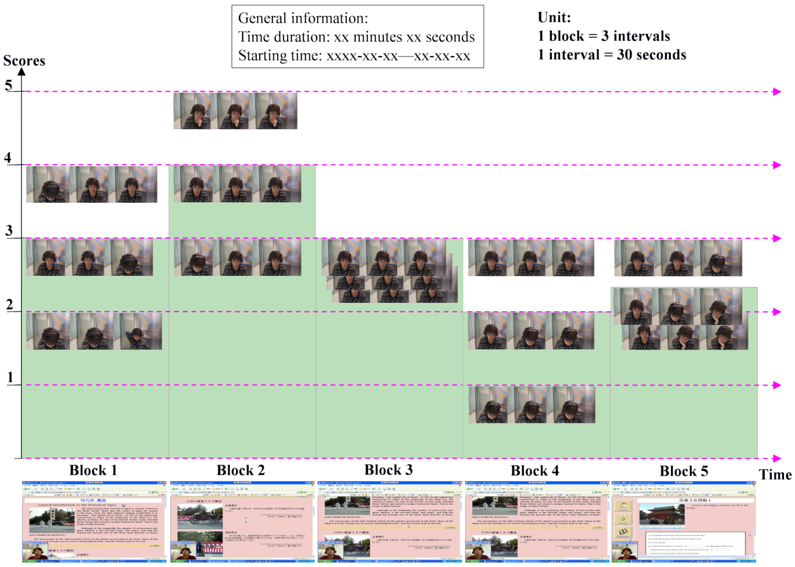


Fig. 2.6 Example of different temporal resolution A

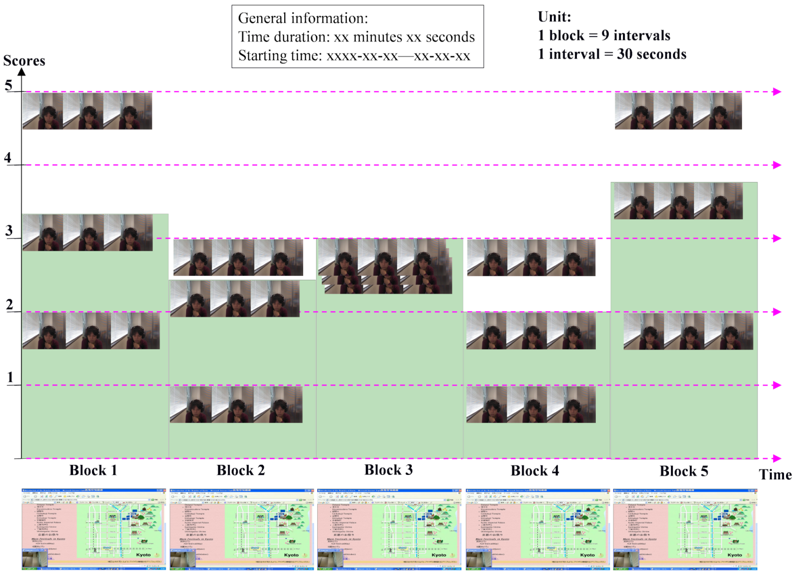
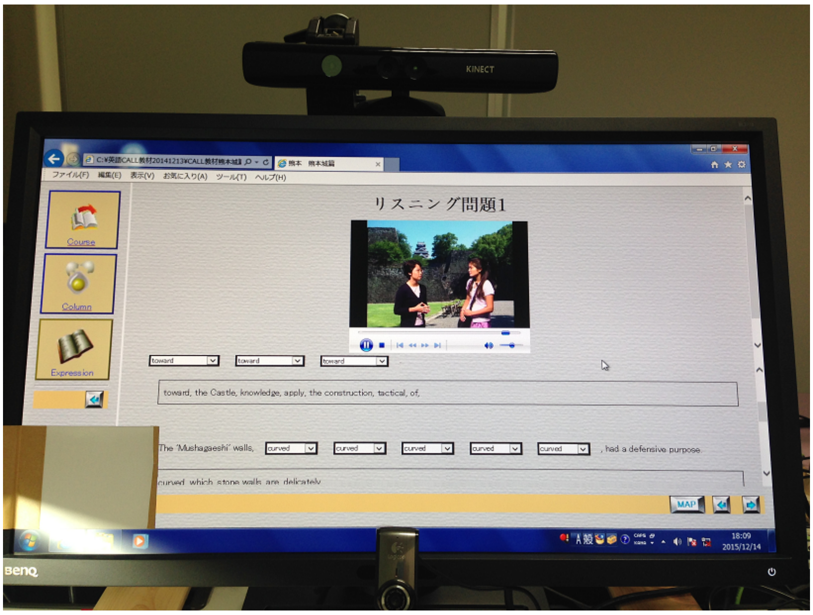


Fig. 2.7 Example of different temporal resolution B

the environment of experiment for verifying the proposal. As e-learning contents, we chose multimedia English learning materials for the CALL system at Kyoto University. These materials are used widely by students not only at Kyoto University but also in several other universities. Learners are university students expected to have self-regulatory e-learning. A Kinect camera is located above the monitor for recording learners' video and estimating internal states. The computer screen is captured continuously using a frame grabber. A small webcam is placed at the foot of the monitor to capture the learner's face independently of the Kinect camera. Both the learner's face and the screen capture can be shown synchronously to the learner right after the e-learning session, as shown in Figure 2.9, and the learners provide self-evaluation according to requirement by watching them. The details about self-evaluation will be introduced in the next section.

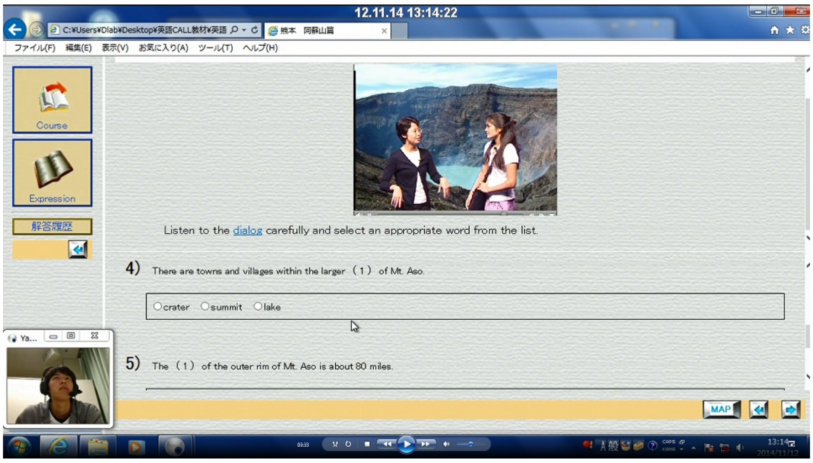


toward, the Castle, knowledge, apply, the construction, tactical, of,

The 'Mushagaeshi' walls, curved curved curved curved curved , had a defensive purpose.

curved which stone walls are delicately

Fig. 2.8 E-learning interface


12.11.14 13:14:22
C:¥Users¥Dlab¥Desktop¥英語CALL教材¥英語
熊本 阿蘇山篇
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
Course
Expression
解答履歴
Ya...
MAP
13:14
2014/11/12

Listen to the dialog carefully and select an appropriate word from the list.

4) There are towns and villages within the larger （1） of Mt. Aso.

○crater ○summit ○lake

5) The （1） of the outer rim of Mt. Aso is about 80 miles.

Fig. 2.9 Reviewing for self-scoring

2.3.1 Scoring of Learning States

To develop the recognition component, ground truth score of learning states are indispensable. A variety of methods have been designed for collecting the data refer to cognitive-affective states in learning process. Based on the data origin, they can be categorized into two categories roughly, judgements provided by other people and self-evaluation by learners themselves. Some previous studies used evaluation by trained judges as ground truth ([26], [34], [35]). However, such judgments are often different from self-evaluation, and consequently, the ground truth indicates how learners look more than how they feel. Also, the facial expressions may be different in computer interaction compare to human interaction. The social protocol for judging facial expression, e.g. FACS, might not work confidently in self-paced e-learning. In contrast, we use learners' self-evaluation as ground truth. To avoid learning interruptions and obtaining high reliance on self-reports, we integrate several methods which summarized in [36] by comparing their advantages and limitations. However, self-reporting introduces the problem of a tendency of participants to average their ratings [36], as well as the issue of social desirability.

In most previous work, two-level evaluation was used, e.g. attentive vs. inattentive or boredom vs. neutral, as was suitable for their purposes. It is often difficult for humans to evaluate themselves quantitatively. However, multiple-level measurement is commonly used in psychometrics to measure attitudes for analysis, e.g., the Likert scale or semantic differentials. For the benefit of teachers, we use a five-level scale for learning state measurement as shown in Table 2.1. Nevertheless, this introduces ambiguity and differences among persons. This problem will be discussed in the following chapter.

Each learning period is segmented into intervals of a specified length, which is 30 seconds in our experiments. The learners will be required to evaluate themselves for every interval right after e-learning session with watching the captured learning videos. Seven participants, undergraduate students with no experience in learning with the specified e-learning materials, were gathered in the verification experiment.

Table. 2.1 Five-level scale of learning states

5	4	3	2	1
Very concentrated	Concentrated	Neutral	Distracted	Very distracted
Very difficult	Difficult	Neutral	Easy	Very easy
Very interesting	Interesting	Neutral	Boring	Very boring

Each participant participated in an average of nine sessions, each of which was approximately 30 minutes in length. We obtained samples for 1,559.5 minutes, i.e., 3,119 valid samples of intervals with self-evaluation scores as ground truth.

2.3.2 Visual Feature Sensing

For each interval, a feature vector for describing the non-verbal information of learners is obtained based on automated analysis of videos capturing learners' behaviors.

Low-level features, including three-dimensional (3D) head pose (position and angle), facial parts movements, and body area and distance, are obtained firstly from processing raw RGB-D images of learners' face and upper body by using Kinect for Windows Software Development Kit (SDK). Head pose is obtained using face detection, and movements of mouth and eyebrows are obtained using facial parts detection. These detections are based on an active appearance model [37]. Head pose is indicated in terms of translation and rotation angles in camera coordinates. Movements of mouth and eyebrows are indicated by displacements from the neutral position and shape of mouth and eyebrows. The body area is represented by the number of pixels of upper body. The body distance shows how far the learner is from the monitor screen, i.e., e-learning system.

Three categories of intermediate feature are obtained using low-level features.

(1) Presence information: Presence of the learner in front of the screen and

the distance between the learner and the screen. Presence is the attendance of the learner in front of screen. We use face detection and body area measurement results for delineating different status of learners, as only use face detection is insufficient. Specifically, when the algorithm fails to detect the face, and the body size is smaller than the threshold, a learner is considered as "absent," otherwise, the learner is "present." It would be helpful to recognized a learner sits in front of e-learning system but looks other sides and real absence.

(2) Head and facial parts information: These features are obtained directly from the head pose and facial parts movements. The lips and eyebrows movements are expressed as numerical weights varying between extrema, i.e. -1 and +1 which 0 means neutral. In lips movement, +1 means mouth opens completely; -1 represents mouth is closed which is like 0. For eyebrows movement, -1 means eyebrows raise almost all the way; +1 means eyebrows are fully lowered to the limit of the eyes.

(3) Probability of gazing at the screen: We do not use commercial eye-gaze tracking systems because of the cost. Alternatively, gazing direction, i.e., a line of sight, is estimated by face orientation based on training samples. If we assume that a learner is looking straight forward, then gazing direction is the same as face orientation, and the gazing target is the "intersection" of the line of sight and the environment. However, this does not always hold. To cope with this problem, we use the statistics of samples collected beforehand, in which a participant looked "inside" and "outside" of the screen with changing directions and conditions. The monitor screen area is quantized into 10 x 10 small regions called cells. The probability of "gazing at the monitor screen" for each cell is calculated based on Bayes' theorem, using the following formula:

$$P(A|X_i) = \frac{P(X_i|A)P(A)}{P(X_i)} \tag{2.1}$$

where A indicates that a participant looks inside the screen, and X_i indicates that the participant's line of sight intersects the i^{th} cell.

A feature vector with 33 elements is calculated based on intermediate features of each interval, as listed in Table 2.2. Together with the self-evaluation score as ground

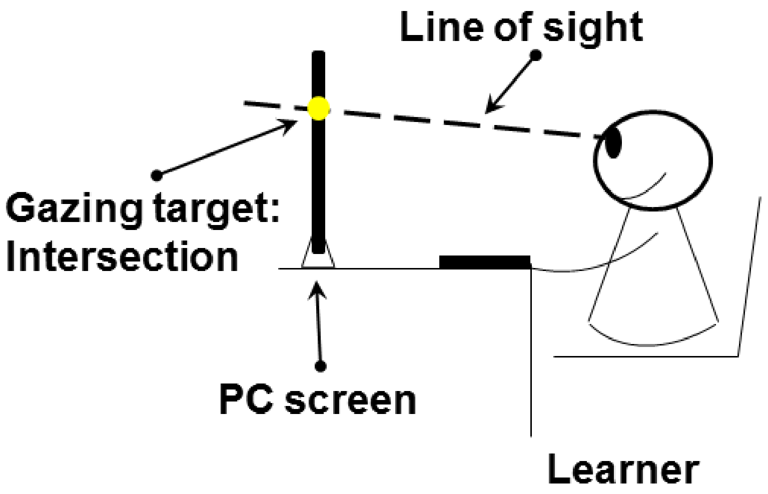


Fig. 2.10 Gazing target: intersection of line of sight and screen

truth, they comprise one training sample.

2.3.3 Classification by Support Vector Machine

SVM is deployed for classification, specifically, the one-against-one method for handling multiclass classification in LIBSVM [38] is applied. A radial basis function (RBF) is used as the kernel. Before SVM training and classification, each feature in the vector is linearly scaled into range [-1, +1] to avoid numeric problem. Grid-search and cross validation are adopted for parameters selection for RBF (Figure 2.11).

We first examined the potential performance of simple classification accuracy by the following schema:

(E1) A classifier is trained using samples from one participant. Target samples, i.e., recognition targets, are from the same participant.

(E2) A classifier is trained using samples from one participant. Target samples are taken from a different participant.

(E3) A classifier is trained using samples from all participants, excluding samples from the participant considered as a new learner. Target samples are from the excluded participant. The trained classifier in here is called the “unified classifier” for the new learner.

Table. 2.2 Thirty-three-element feature vector

Feature source:	Feature items:
Presence information	Present proportion Distance (Max, Average, Min)
Head and facial parts information	Face detection successful proportion Lips movement (Max, Average, Min) Eyebrows movement (Max, Average, Min) Head pose angle Pitch (Max, Average, Min) Head pose angle Yaw (Max, Average, Min) Head pose angle Roll (Max, Average, Min) Head position X-coordinate (Max, Average, Min) Head position Y-coordinate (Max, Average, Min) Head position Z-coordinate (Max, Average, Min)
Probability of gazing at screen	High probability proportion Moderate probability proportion Low probability proportion Consider as zero probability proportion

The samples used for training a classifier for the new learner are called “prototype samples,” and the learners who provide prototype samples are called “prototype learners.”

E1 estimates the upper bound performance of our method. In other words, the performance we can expect if the learner is well-known. E2 shows the performance degradation caused by inter-personal differences. E3 estimates the potential performance of the unified classifier for a new learner. For each scheme, the leave-one-out method is used for cross validation.

The criteria of the matching are as follows:

1. Strict matching criterion: This requires an exact match between the classification result provided by the SVM and the ground truth.

2. Lenient matching criterion: This allows a classification result to be equal or

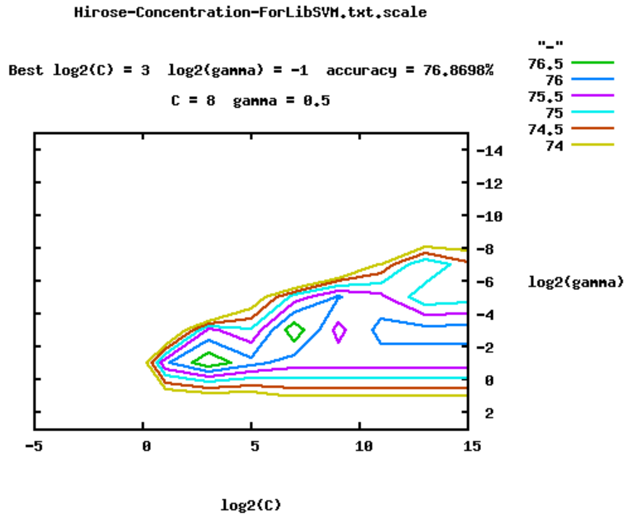


Fig. 2.11 Grid-search for parameters selection

nearly equal to the ground truth, i.e., the score difference must be at most one.

Lenient matching criterion is introduced to examine how estimated scores are close to the ground truth. If the performance of lenient matching is satisfactory, an estimated score can be useful even if it does not perfectly match to the ground truth. For example, they can be helpful for looking over the states of learners. Another aspect of lenient matching is concerning difficulty of introspection. Because introspection may contain fluctuation caused by human nature, strict matching criterion could be too severe, and teachers may feel that lenient matching criterion is reasonable. Tolerance of lenient matching is reasonable for learners with less stability.

Figures 2.12 and 2.13 show the results for the strict and lenient matching criteria, respectively. The average values are shown by thick bars, and the ranges between the best case and the worst case are shown as thin lines.

From the E1 results, we can see that the average accuracy of strict matching is approximately 60%. One of the primary reasons for the performance degradation in this case is the similarity of the behaviors among different internal states. Figure 2.14, Figure 2.15, and Figure 2.16 show three samples with different self-evaluation scores for concentration-distraction. Figure 2.14 is chosen as the reference, for which the L1

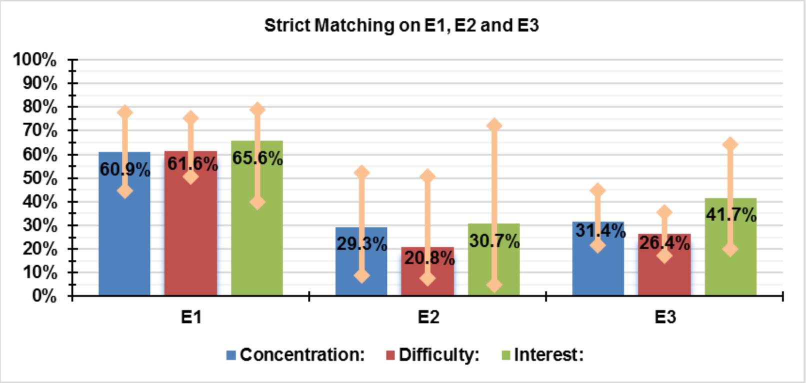


Fig. 2.12 Strict matching performance on E1, E2, and E3 conditions.

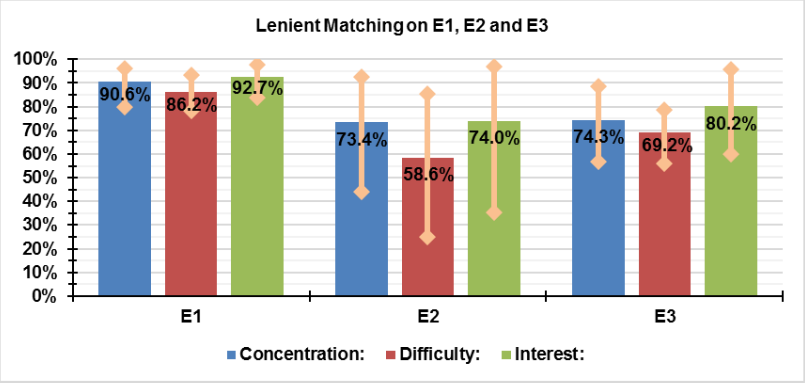


Fig. 2.13 Lenient matching performance on E1, E2, and E3 conditions.

distances to the other samples of the same student are calculated. Figure 2.15 and Figure 2.16 are the closest samples. The average variance among the three feature vectors is 0.3508. During these three intervals, the learner retained appearances as in the shown images, with only small movements. Consequently, the feature vectors for these intervals are similar, while the self-evaluation scores are different.

Another reason is the dissimilarity of behaviors for the same self-evaluation score. An example is shown by Figure 2.14, Figure 2.17, Figure 2.18, and Figure 2.19. Four representative images result in four different situations with score one for concentration-distraction producing diverse feature vectors. The reference sample is the sample shown in Figure 2.14. In Figure 2.17, the learner does not look at the screen. Figure 2.18 indicates that the learner is napping. And Figure 2.19 indicates that the learner is almost out of field. The feature vectors for those three samples

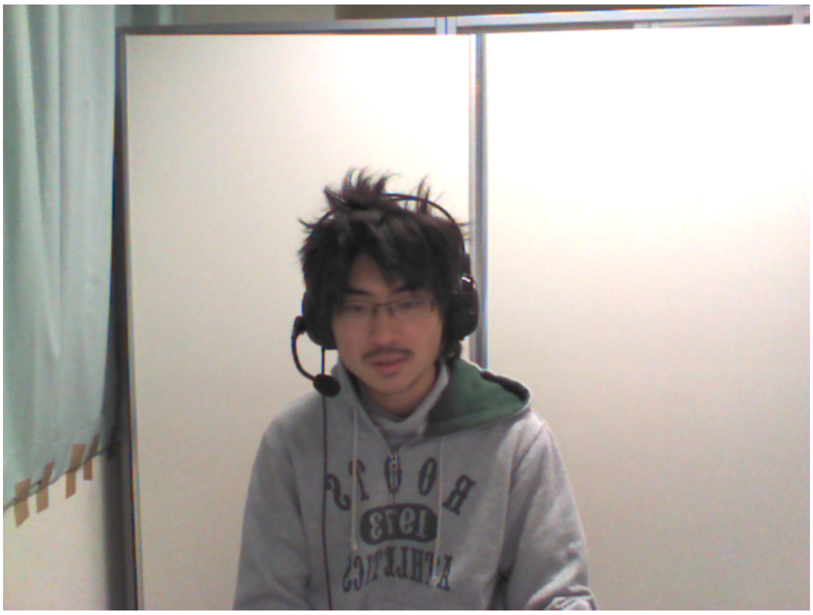


Fig. 2.14 Similar appearances but different self-evaluation scores of the same learner. Reference sample, self-evaluation score 1.

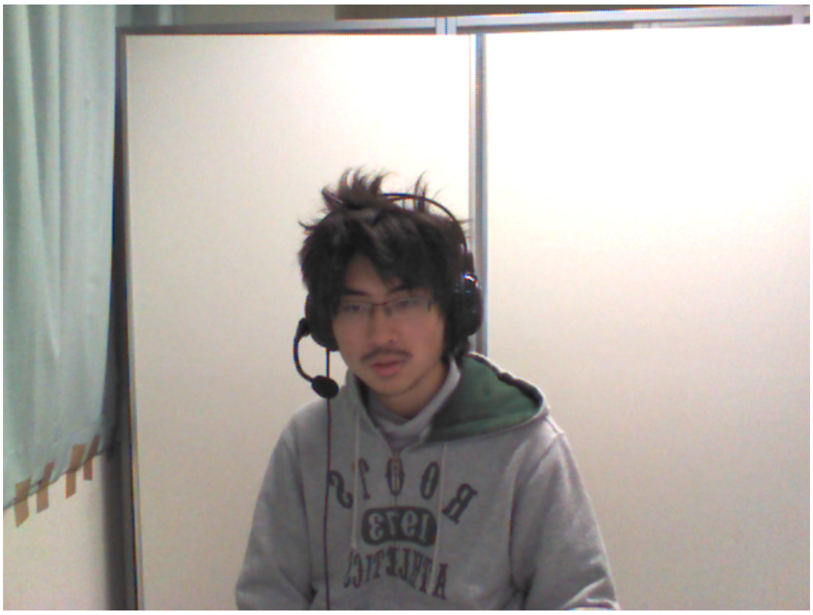


Fig. 2.15 Similar appearances but different self-evaluation scores of the same learner. Closest sample 1, self-evaluation score 3.

are much different from those for the first sample in the L1 distance metric. The average variance among them is 21.1846.

Another possible reason is the ambiguity of self-evaluation. Self-evaluation appears

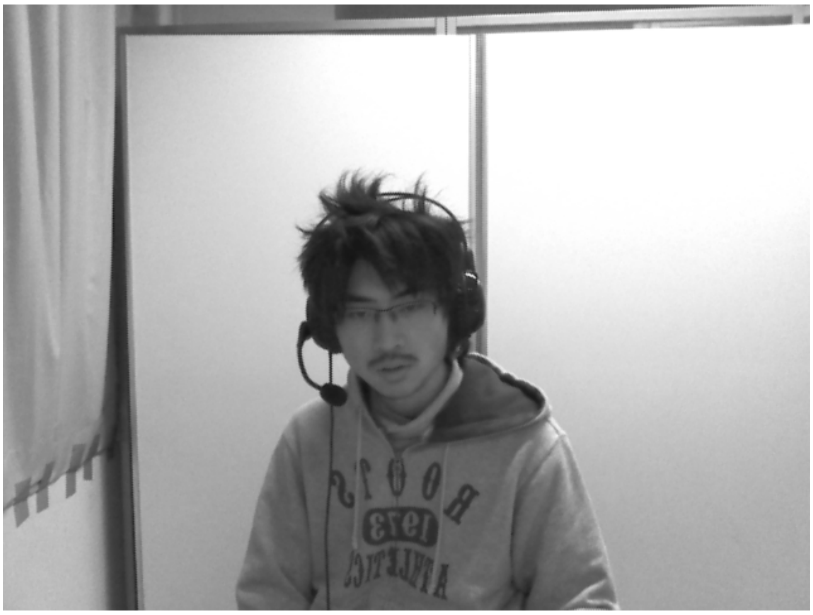


Fig. 2.16 Similar appearances but different self-evaluation scores of the same learner. Closest sample 2, self-evaluation score 4

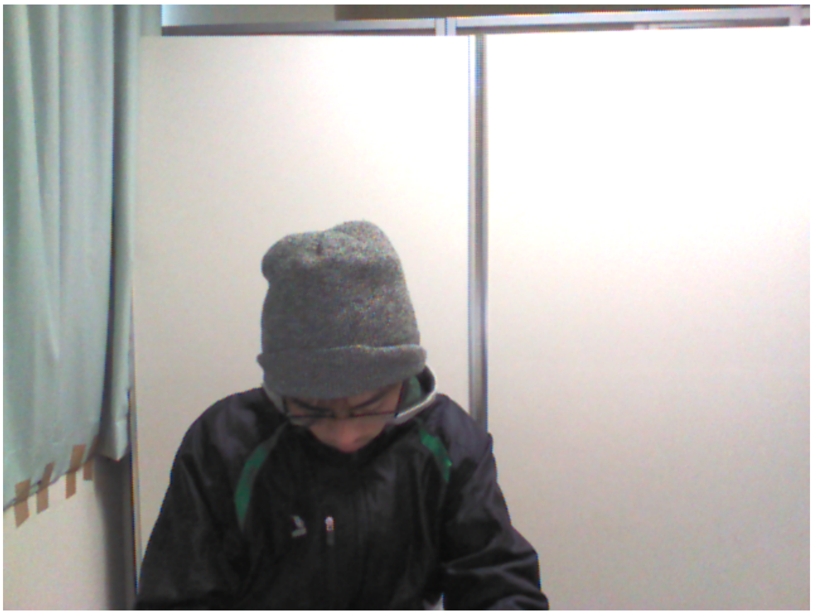


Fig. 2.17 Different situations for same self-evaluation score: score 1 in concentration-distraction, case 1

to fluctuate because of the difficulty of introspecting. From this point of view, we cannot expect perfect performance. However, all of those drawbacks are relaxed in the lenient matching cases. We obtain approximately 90% accuracy for all three pairs

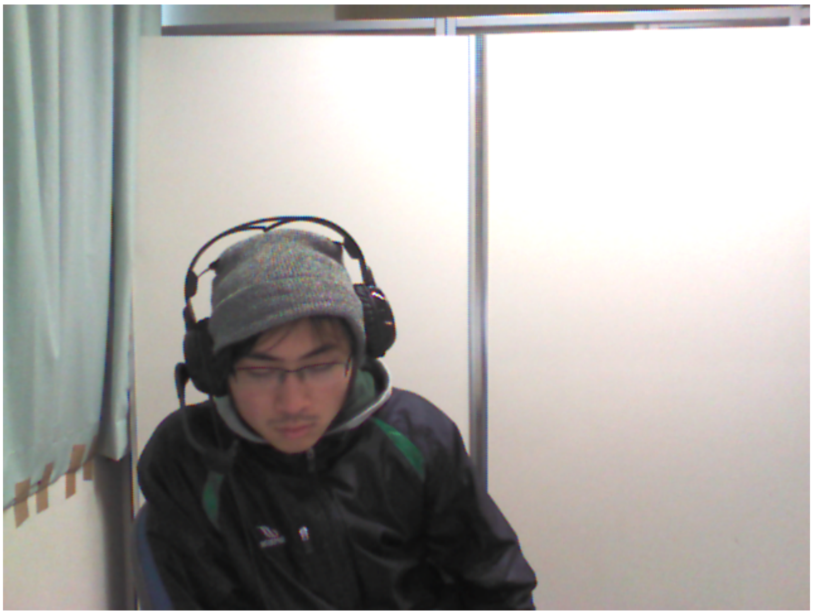


Fig. 2.18 Different situations for same self-evaluation score: score 1 in concentration-distraction, case 2

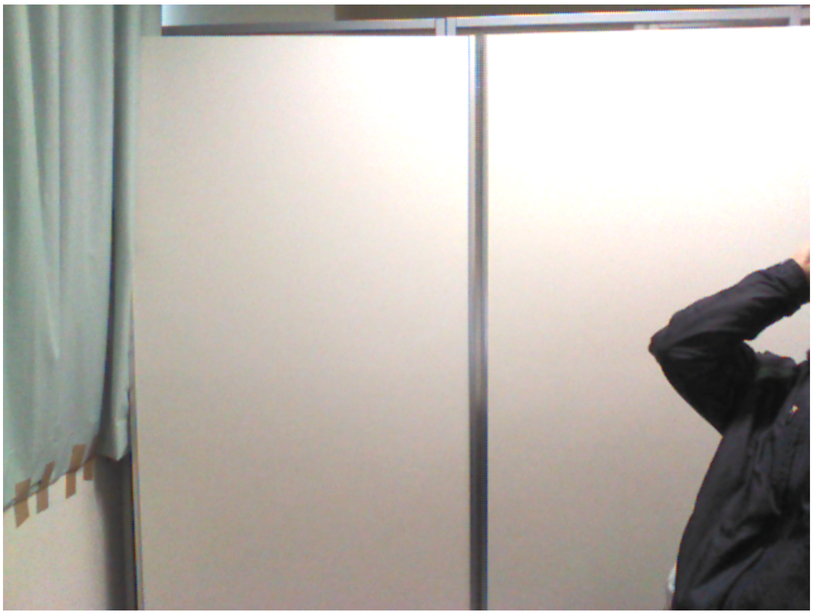


Fig. 2.19 Different situations for same self-evaluation score: score 1 in concentration-distraction, case 3

of internal states.

As the results of E2, we have serious performance degradation for strict matching if we apply classifiers trained for a different person, as shown in Figure 2.12. Figure 2.20, Figure 2.21, and Figure 2.22 show an example of similar appearances for different learners. Figure 2.20 is the reference sample with score five for concentration-distraction, for which L1 distances to the other students' samples are calculated. The two closest samples, whose self-evaluation scores are three and one, respectively, are shown in Figure 2.21, and Figure 2.22. Based on the images, they are not significantly different except in the detailed conditions of the eyes that are not included in feature vectors in our experiments. Detection of those detailed features is left for future work. The performance degradation is, however, relaxed with the lenient matching criterion, as shown in Figure 2.13.

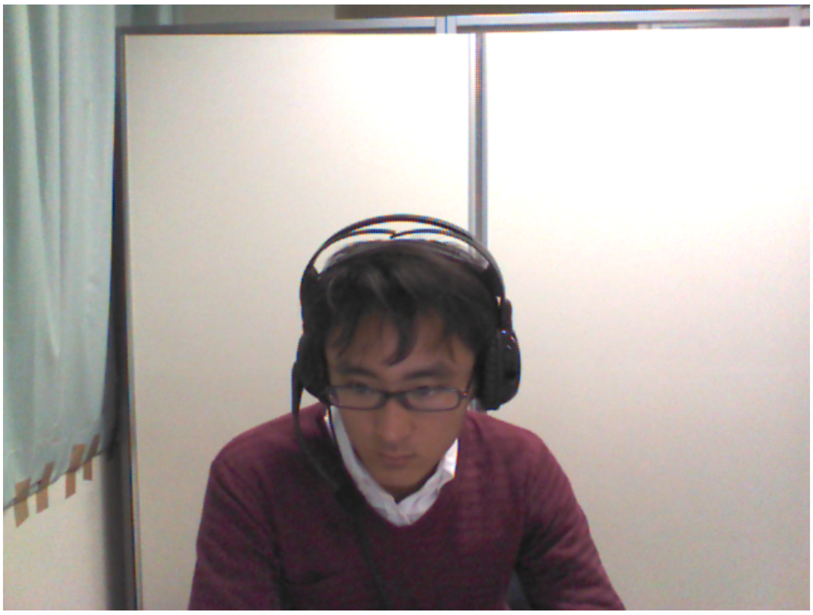


Fig. 2.20 Similar appearances but different self-evaluation scores of different learners. Reference sample, self-evaluation score 5.

The E3 results show the performance expectation of the unified classifier for a new learner. The large gaps between E1 and E3 indicate that simple aggregation of a large number of samples does not necessarily provide a good classifier. On the other hand, the performance difference between E3 and E2 shows that the performance improvement resulting from gathering samples from multiple persons is not negligible. Those results imply that we need a more sophisticated method to utilize a variety of samples from a variety of learners effectively. This problem is discussed in next

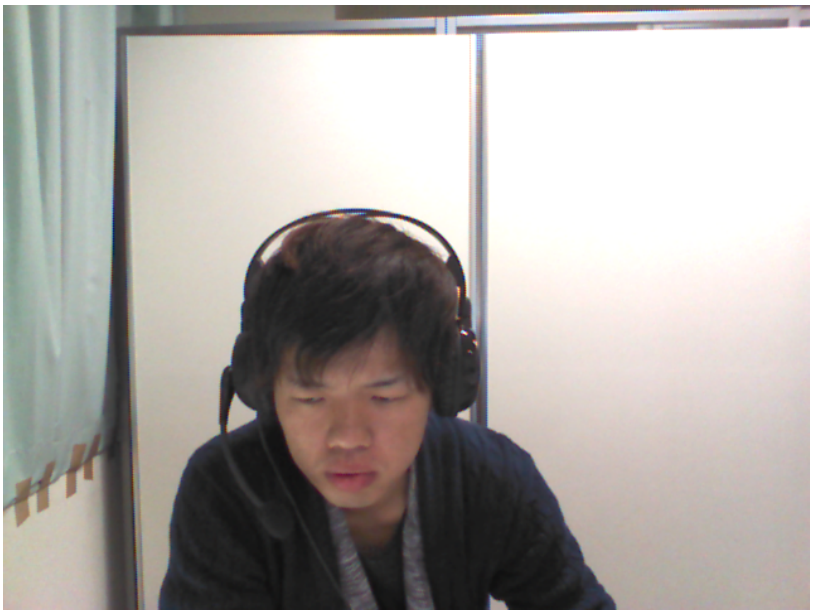


Fig. 2.21 Similar appearances but different self-evaluation scores of different learners. Closest sample 1, self-evaluation score 3.

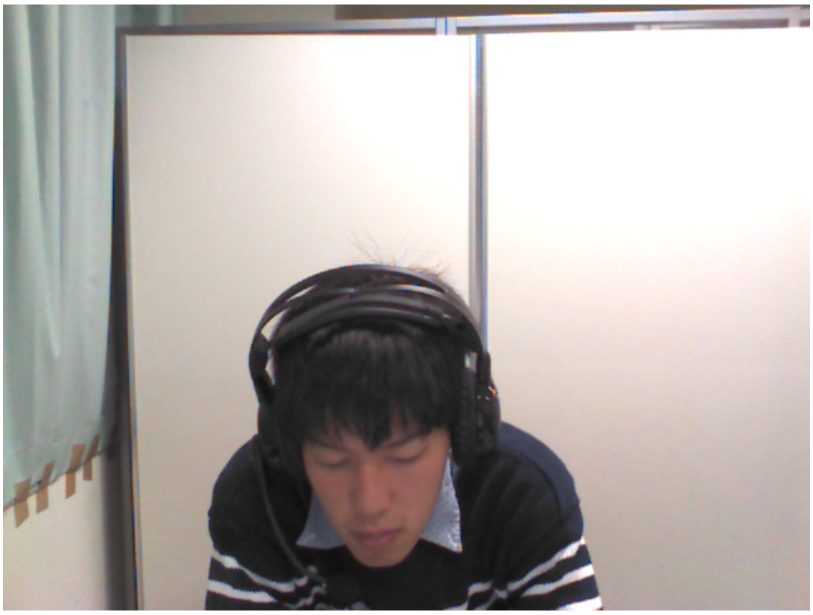


Fig. 2.22 Similar appearances but different self-evaluation scores of different learners. Closest sample 2, self-evaluation score 1.

chapter.

Concerning the differences among learning states, performance is not significantly different. However, we observe the tendency that the accuracy is better in the order of interest-boredom, concentration-distraction, and difficulty-ease. One possible reason might be a negative mood that learners choose to conceal.

The distribution of 3 learning states is plotted in the 3-dimensional space which each of axes corresponds to one learning state, and Principal Component Analysis (PCA) is applied on them to analyze the distribution tendency of all learning states together. Figure 2.23, Figure 2.24, and Figure 2.25 demonstrate the distribution along with the first principal component axis, second principal component axis and third principal component axis, respectively. Samples are denoted by the scattered points, different students are represented in different markers and colors. The explained variance ratio by the principal axes are 0.66, 0.26, and 0.08, respectively. The first and second component axis contains the most information, the third principal component axis is negligible.

The first principal component axis is one diagonal of the cube that goes from low score area of all 3 learning states to high score area of all 3 learning states. The second principal component axis is almost another diagonal that goes from the corner of low score in concentration-distraction and interest-boredom and high score in difficulty-ease, to the corner of high score in concentration-distraction and interest-boredom and low score in difficulty-ease. The distribution tendencies indicate that concentration-distraction and interest-boredom have high positive correlation. And the correlation between concentration-distraction and difficulty-ease, and the correlation between difficulty-ease and interest-boredom are situation-dependent. Those tendencies are consistent with the research and investigation in education. That concentration-distraction and interest-boredom can foster each other, which implies the importance of designing learning with high interestingness. The relation with difficulty-ease reflects the thinking of theory of zone of proximal development. That if the learning content is located in the zone of proximal development of the learners, they would be aroused, and tend to keep concentrated on learning. On the other hand, if the difficulty is beyond the level that students can accomplish with the aid of learning material, they would lose interest and stay focus would also become difficult. It implies that to keep the learning materials in appropriate level is important, too difficult and too easy may be counterproductive. These facts indicate the 3 learning states are different with each other but also correlated.

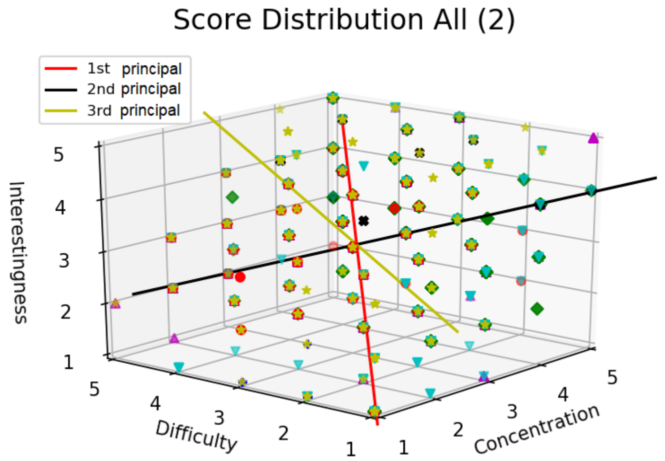


Fig. 2.23 First principal component axis direction

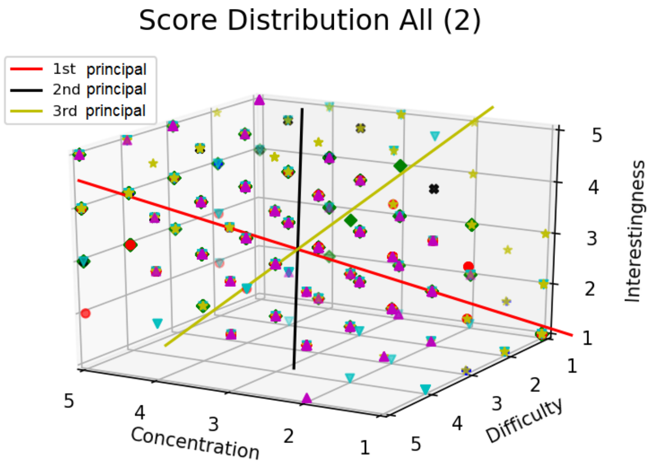


Fig. 2.24 Second principal component axis direction

2.3.4 Discussion

The results of E1 show much room for improvement, especially in strict matching. One reason is the wide variety of external expressions, and another is the ambiguity of self-evaluation. For future improvements, incorporating other measuring modali-

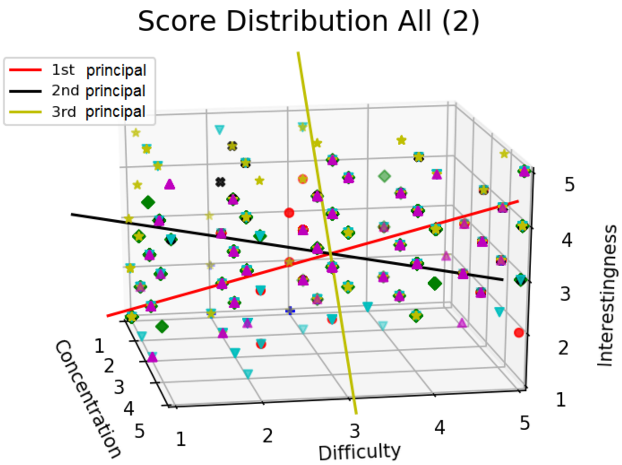


Fig. 2.25 Third principal component axis direction

ties might be a partial solution. Detailed information regarding the eyes is expected to improve the performance. Additionally, we can consider the use of non-intrusive physiological measurements and/or the use of a mouse and keyboard to input information.

E2 results show serious inter-personal differences, and E3 results reveal that simply mixing all samples does not provide sufficient improvement. To clarify this point, we verified how performance changes with the number of prototype learners. Figure 2.26 shows the result. The horizontal axis indicates the number of prototype learners used for training, and the vertical axis indicates the accuracy of estimating concentration-distraction for a new learner. The three lines show maximum accuracy, average, and minimum accuracy.

The figure illustrates the trend that the maximum accuracy is better with one or a few prototype learners, and it worsens as the number of prototype learners increases. This fact suggests that a classifier trained by the samples from one or a few similar prototype learners tends to provide good performance. On the other hand, the minimum accuracy improves gradually as the number of prototype learner increases. A classifier trained for one or a few prototype learners with different characteristics yields poor performance, and this problem will be relaxed with more prototype learn-

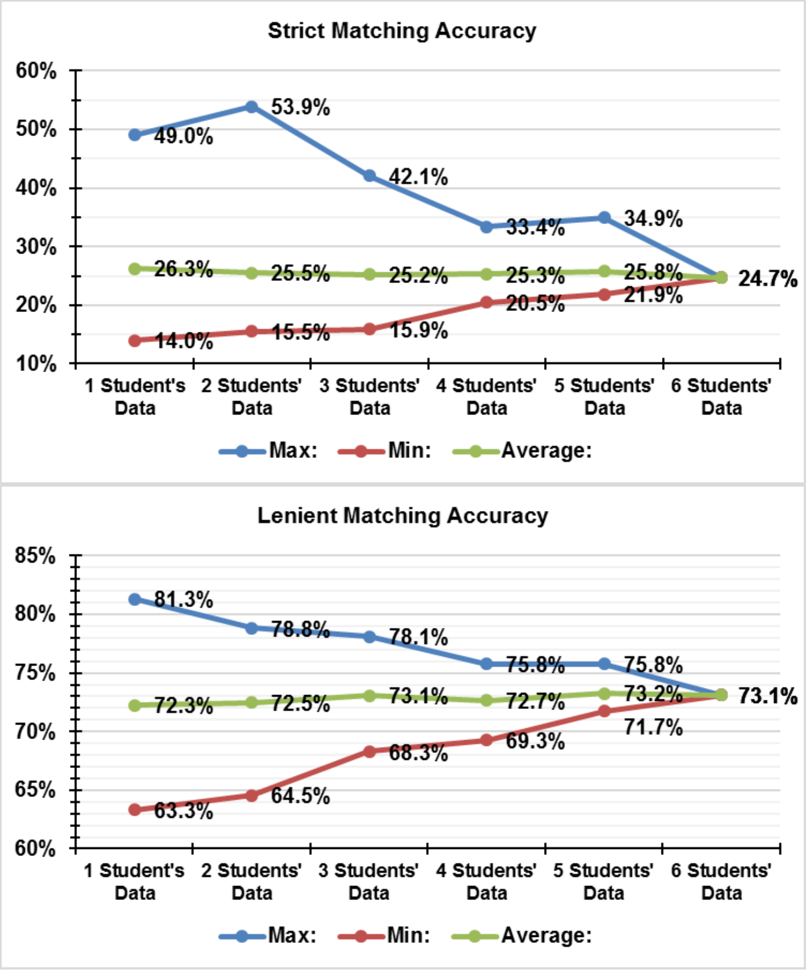


Fig. 2.26 Performance changes by the number of prototype learners

ers. As the number of prototype learners increases, the chance that samples with behavioral characteristics similar to a new learner's are included will increase. This fact suggests that it is useful to consider the unified classifier based on all of the samples as the baseline.

2.4 Summary

The mainstream of educational theory manifests that learning is a complicated mental process that learners participate actively, and that knowledge is internalized during this learner-centered process. The learning outcome will be excellent and learning efficiency will be high if the learning is designed to attract and suit learners well in this learner-content interaction. However, to keep students engaged and motivated in self-paced e-learning is a challenge, since little feedback information can be retrieved from the places where students actually perform e-learning. Lack of feedback information hinders students to get timely and appropriate guidance. It is a hazard for learners' motivation and learning efficiency, and also for learning content design and improvement. Learning states of learners during e-learning are important and valuable feedback information for instructors to understand how learners react to learning. Three categories of learning states, concentration-distraction, difficulty-ease, and interest-boredom, are selected according to their direct and close relation to learning, which enable teachers to know how students learn readily. With such information, teachers can provide mentoring, modify materials, or conduct educational analysis etc.

A novel system is designed to estimate learners' internal states during learning through introducing pattern recognition technology which takes image processing results. Visual sensing of learners is employed to sense learners' behaviors during learning, because it is non-intrusive and non-contact, and a small camera can be integrated easily with existing e-learning systems with current technologies, without heavy constraints and cost is preferable. The proposed scheme has three principal components: (a) learning and capturing, a video of a learner's face and upper-body is captured in RGB-D format during learning; (b) processing and recognition, the captured video is processed to get visual features. Subsequently, they are input to SVM to recognize learners' internal states; (c) surveying and mentoring, the detected information is summarized and presented to teachers in a comprehensible manner for them to provide mentoring, modify materials, or conduct educational analysis etc. Actual design of this utilization is left for future work.

Ground truth for developing recognition component is obtained through self-evaluation of learners based on a five-level scale for each learning state index. A feature vector with 33 elements is calculated based on visual features for each interval which covers presence information, head and facial parts information, and probability of gazing at the screen information. In verification experiment, 3,119 valid samples involved seven participants were gathered. The experimental results showed the potential of our classification method by SVM using the abovementioned visual features: approximately 60% average accuracy in strict matching and approximately 90% average accuracy in lenient matching can be achieved.

Chapter 3

Inter-personal Differences in Recognition

One critical problem for practical use, which has been less explored, is how to deal with new students whose characteristics are not well known in advance. E-learning is designed to accommodate a variety of students. We need to consider that learners are considerably different in their behaviors, i.e., inter-personal differences among learners. Even if we achieve sufficiently good performance for a specific learner, we might not obtain good performance for another learner. The analysis based on E2 and E3 in previous chapter has confirmed such phenomenon experimentally. However, it would be difficult to have all types of learner models beforehand. We need to consider a mechanism to adapt the system to new learners. For this purpose, a scheme that can quickly adjust to a new learner with a little effort of the learner was designed and verified.

3.1 Strategy for Adjusting to a New Learner

For learners to record their learning states after actual e-learning requires considerable effort. Scoring, often takes time longer than the time for actual learning, and requires considerable mental effort in video reviewing, introspection, and marking. Provided scores by learners would be unreliable due to the additional load, if learners are simply forced to mark scores after actual e-learning. On the other hand, during the phase of system development, we can expect that a considerable number of samples can be obtained from multiple participants. From this point of view, we need a system that can adjust quickly to a new learner with little effort by the learner. For this purpose, we consider the following scheme:

Step 1: Gather sufficient number of samples with ground truth from multiple learners who collaborate for data collection. Hereafter, we call those learners and samples "prototype learners" and "prototype samples," respectively.

Step 2: Train classifiers by using typical combinations of the prototype samples, details are explained in next section.

Step 3: The system asks a new learner to provide ground truth scores for a small number of intervals in an actual e-learning. Hereafter, we call these samples "representative samples."

Step 4: The system selects the most appropriate classifier from above classifiers by using both representative and prototype samples.

3.2 Classifier Selection Strategy

Assuming that we have sufficiently many prototype samples from multiple prototype learners, we can think of various classifiers as follows:

(C1) Classifiers that are each trained with prototype samples from a single prototype learner.

(C2) Classifiers that are each trained with prototype samples from a combination of two or more prototype learners.

(C3) A classifier that is trained with all prototype samples from all prototype learners. Hereafter, we call this classifier the "unified classifier."

The possible variations comprise the power set of the learners. With those classifiers, our target is to develop a method for choosing the best classifier for a new learner. We consider the following strategies for this problem:

(M1) Accuracy-based method: The system applies every classifier to the representative samples from a new learner and then chooses the classifier that results in the best performance.

(M2) Similarity-based method: The system measures the similarities between representative samples of a new student and prototype samples and then chooses the classifier that has the greatest number of similar samples in its training data within a specified number of resembling samples. In this approach, L1 distance is chosen as the metric for similarity measurement. This strategy has the advantage that it does not require self-evaluation by a new learner.

3.3 Experiment and Discussion for Classifier Selection

Experiments were conducted for checking the possibility of choosing an appropriate classifier by using a small number of representative samples from a new learner. For this purpose, one participant was chosen as a new learner in turn, and the other participants were regarded as prototype learners. Five representative samples were chosen randomly from the new learner's samples. We applied the accuracy-based and similarity-based methods for choosing a classifier from among all classifiers, i.e., C1, C2, and C3 introduced in previous section. Then, the chosen classifier was applied to all the samples of the new learner, and the performance was evaluated. We repeated this process 20 times for every new learner, and the average performance was recorded.

Figure 3.1 shows the results for concentration-distraction. The baseline is the performance by the unified classifier. The average accuracy of seven new learners is displayed with a thick bar. The values of the highest and the lowest accuracy are presented by the thin lines on the bars. We can see that the accuracy-based method provides better performance than the unified classifiers. On the other hand, the similarity-based method does not provide good results. Figure 3.2 shows one example that might explain this. For each of five representative samples of new learners, the ten closest samples are extracted from all prototype samples, and their scores are counted. Although prototype learner A has the most samples similar to representative samples of the new learner, none of them has the same score. This fact clearly shows the difficulty posed by inter-personal differences.

Next, we examined the integrated method, a combination of the above two methods. It has 2 processes when selecting classifier. The first process is the same as the accuracy-based method. If there are multiple classifiers that yield the best performance for given representative samples, the second process chooses the classifier that has the greatest number of similar samples out of the 50 closest samples. As shown in Figure 3.1, the integrated method shows no significant improvement over

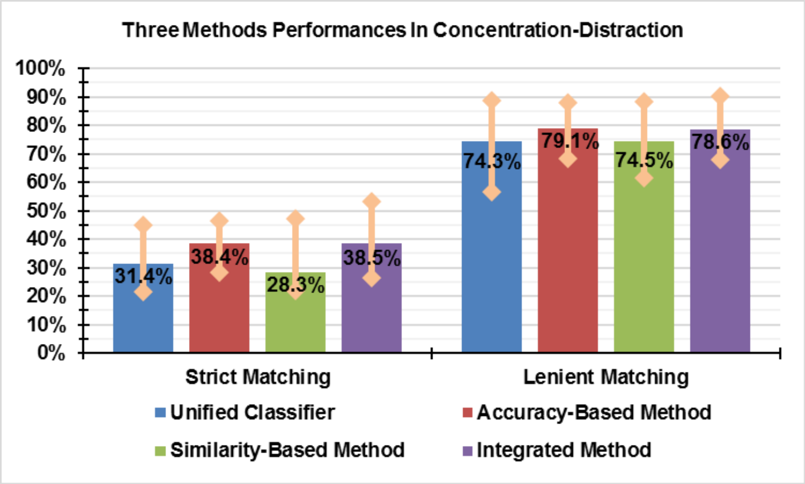


Fig. 3.1 Three methods for performances in concentration-distraction

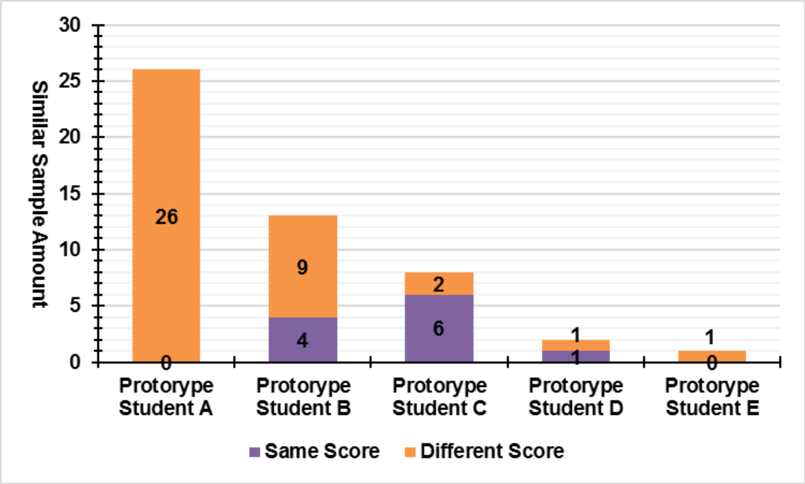


Fig. 3.2 Score distribution of similar samples

the accuracy-based method. This might be a result of the small number of representative samples, in addition to the facts disclosed in Figure 3.2. The results for difficulty-ease and interest-boredom are shown in Figures 3.3 and 3.4, respectively. They also indicate no significant improvements from the integrated method.

As for difficulty-ease, the results in E2 and E3 that have the smallest average accuracies and the smallest range between the highest and the lowest accuracy suggest larger inter-personal differences among learners. This diversity makes improvements difficult. For the interest-boredom case, the performance of the baseline, i.e., E3, is better than that of the other two learning states. The room for improvement is

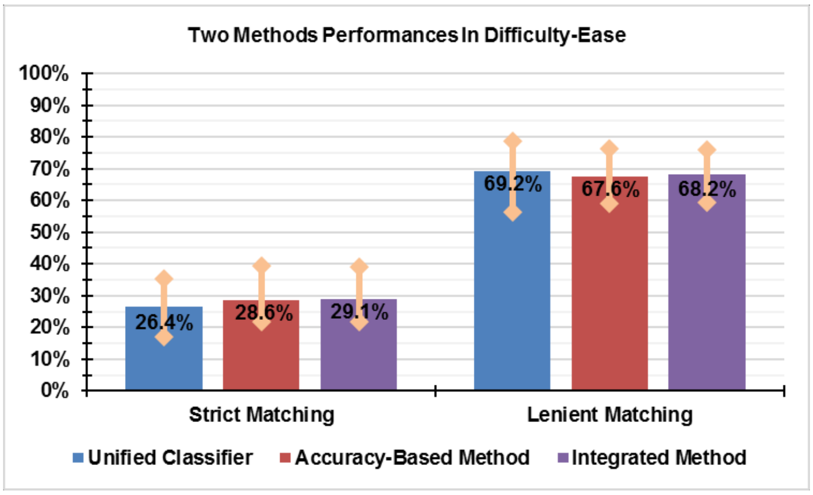


Fig. 3.3 Two methods for performances in difficulty-ease

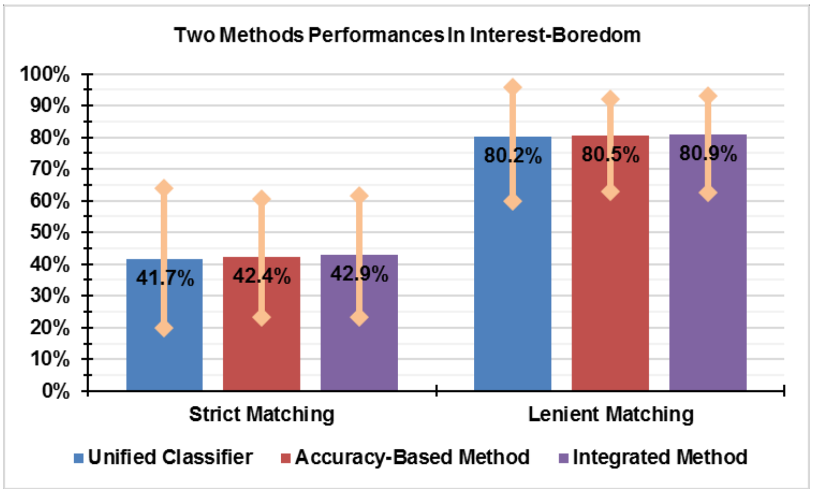


Fig. 3.4 Two methods for performances in interest-boredom

consequently small.

3.4 Investigation for Representative Sample Selection

The accuracy-based and the integrated methods improved the estimation accuracy in some cases, but did not provide significant improvements in other cases. From the results of E1 and Figure 2.26 in Chapter 2, we see that there are many classifiers that provide better performance. We have much room for improvement for further study, especially compare to the classifier that demonstrates the best performance for the learner. Therefore, reducing the gap between the selected classifier of a new learner and the most appropriate classifier of that learner is an important objective.

The classifier selection performance given by accuracy-based method becomes better with the increase of number of representative samples which selected randomly. Figure 3.5 shows the results investigated in concentration-distraction. Twelve numbers started from five until sixty were tested as the number of representative sample for random selection. Two thousand of trials were conducted for each number of random selection to get more statistically stable accuracies for examining the performance tendency. The horizontal axis indicates the different numbers of representative samples selected for classifier selection, and the vertical axis indicates the accuracy of selected classifiers for new learners in concentration-distraction. The three lines show maximum accuracy, average, and minimum accuracy of taking different participants as new learners. The last column in the figure demonstrates the performance given by ideal classifiers. However, the burden of video-reviewing and self-scoring for ground truth by new students also increases. Moreover, the performance improvements are hardly comparable to the increases of effort dedicated to them. We need to find a method that gives good performance with a reasonable number of representative samples. Further research was investigated for choosing better representative samples instead of random selection.

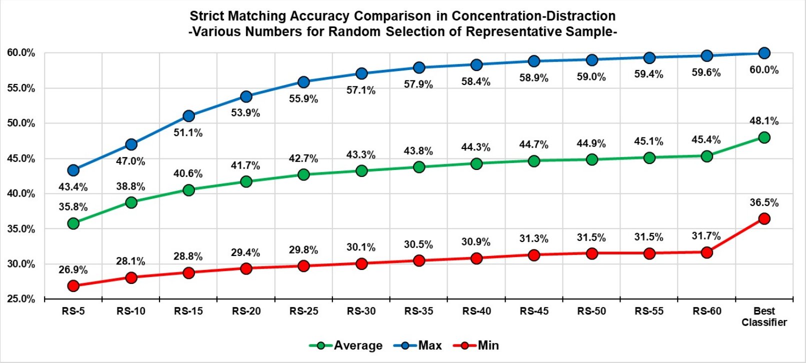


Fig. 3.5 Performance changes by the number of representative samples

Basic Assumptions for Selecting Representative Samples

The proposed method of choosing representative samples is based on the following assumptions.

A1: Frequently appearing samples can be good representative samples. Given a representative sample that is correctly classified, neighboring samples of a representative sample are also correctly classified if the self-scoring is consistent throughout the learning period. A representative sample that lies in a region of higher occurrence probability has more neighboring samples, which are expected to be correctly classified.

A2: A set of representative samples that covers a wider area of the feature space provides better accuracy. Similar to A1, not only samples neighboring to representative samples, but also in-between representative samples have a higher probability of being correctly classified compared to samples far from representative samples. This implies that a set of representative samples are preferably separated from one another.

A3: A set of representative samples with enough variety of classes gives better accuracy. Samples with different scores may correspond to different behaviors. Simple interpolation or extrapolation of certain feature values might not be appropriate. Therefore, it is important to obtain enough variation in the classes of a set of representative samples.

Representative Sample Selection based on Assumptions

According to the above assumptions, a set of samples with high occurrence probability are given priority as representative samples, and the following scheme is proposed:

–Step 1: Estimate the probability distribution of samples obtained from a new learner.

–Step 2: Apply clustering to samples with a large probability density.

–Step 3: Collect representative samples by choosing one sample from each cluster.

–Step 4: Select the classifier that demonstrates the best performance for the set of selected samples.

For Step 1, kernel density estimation with Gaussian kernel [39] is applied to the samples of a new learner. In this process, the samples in which no face was detected are excluded because they are less important in estimating the learning state.

$$\rho f(x) = \sum_{i=1}^{N} K(\frac{x - x_i}{h})$$

$$K(x; h) \propto exp(-\frac{x^2}{2h^2})$$

Where ρ is a coefficient, K represents the Gaussian kernel function, x_i is the observation $(x_1, x_2, \cdots, x_N)$, and h is the smoothing parameter referred to as the bandwidth.

For Step 2, hierarchical agglomerative clustering [39] is applied to samples with a high probability density. Clustering is not applied to all the samples at once because it is difficult to know the appropriate number of clusters for a large number of samples. Clustering is also avoided if the results would be poor with respect to the growth of the sample population. Linkage type of "ward" is used, which minimizes the sum of the squared differences within all clusters, and the units of the Euclidean units are metric for the purposes of computing the linkage.

In Step 3, the sample corresponding to the highest probability density is chosen as the representative sample from each cluster. After the selection, each new learner is asked to provide ground truth scores of learning states for the selected representative samples.

Then, in Step 4, the accuracy-based method is applied to these samples to select the suitable classifier.

Experimental Results and Discussion

Experiments were conducted to verify assumptions for representative sample selection. Similar experiment setting was applied. Each participant was considered to be a new learner, and the others were considered as prototype learners. Steps 1 through 4 were applied to the data of each new learner to obtain five representative samples, and a suitable classifier was selected for each of the three learning states. The selected classifier was applied to all the samples of the new learner, and the accuracy was calculated for both strict and lenient matching conditions.

An example of kernel density estimation for all the samples of one learner is shown in Figure 3.6. The samples were sorted according to the estimated probability density at each sample point in descending order. The points in upper-left corner inside the black dashed circle are the samples for which no face was detected.

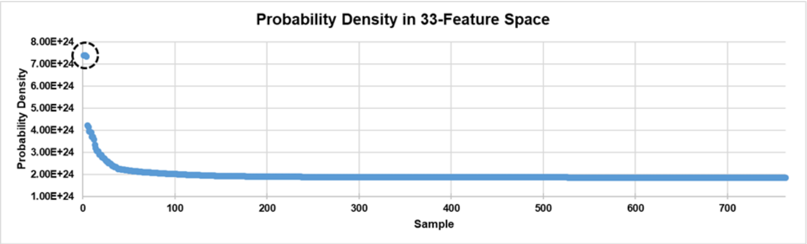


Fig. 3.6 Example of kernel density estimation

Samples with the top 10-40% of high occurrence probability were chosen as clustering targets. An example of dendrogram in hierarchical clustering for a new learner is shown in Figure 3.7, which illustrates how the samples are hierarchically clustered from the bottom to the top. The horizontal axis indicates the sample index, and the vertical axis represents the distance; the height of the legs of the inverse U-shape links are indicative of the distance between the two children. The top of the link denotes the cluster agglomerating, and five clusters were identified within this type of hierarchical structure at the height of the dotted line. The obtained clusters (3, 4), (156, 157), (124, 8, 12, 13, 10, 9, 14), (93, 94), and (172, 170, 171, 168, 169) as shown by the dashed closures.

The average accuracy of the selected classifiers with five representative samples are shown in Figures 3.8 and 3.9 in terms of strict matching and lenient matching accuracy, respectively. The average accuracy of the selected classifiers with ten representative samples are shown in Figures 3.10 and 3.11. The performance obtained by randomly choosing representative samples is presented as the baseline for comparison, which is the average of 2000 random selection trials. Each thin bar represents the range between max and min value of new learners. The best overall performance among the 63 classifiers is also presented, which is indicative of the upper bound of performance.

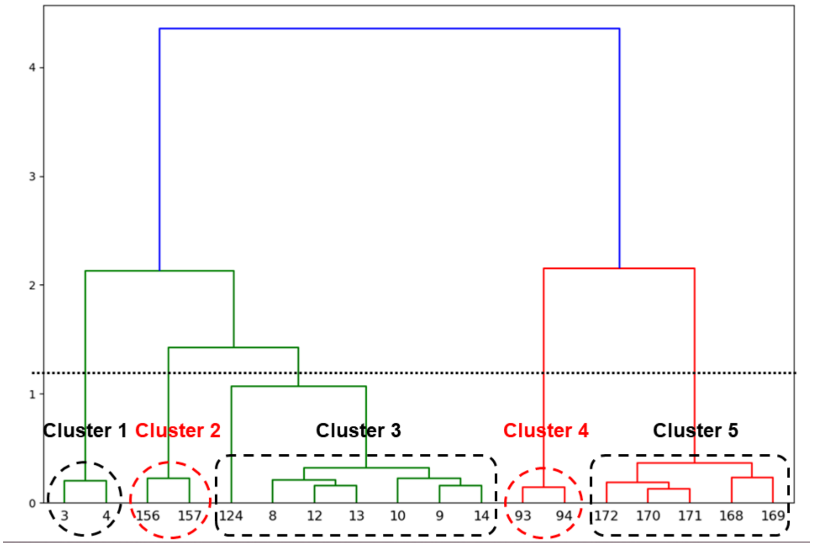


Fig. 3.7 Example of dendrogram in hierarchical clustering

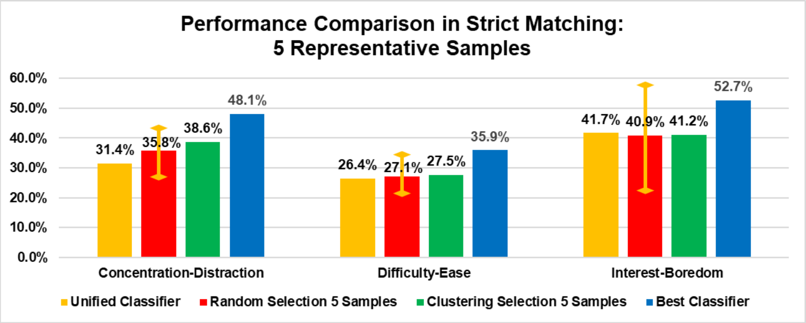


Fig. 3.8 Classifier selection performance comparison in strict matching accuracy: 5 representative samples

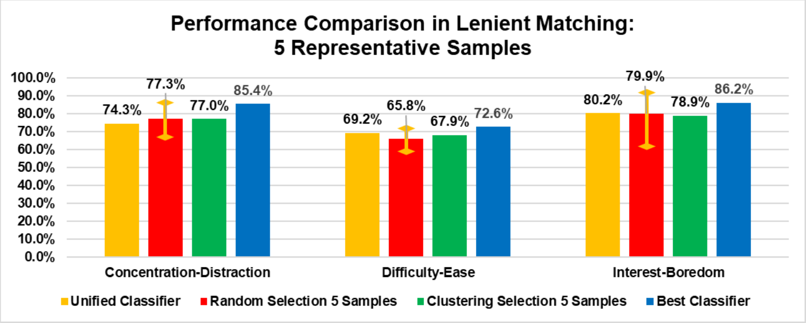


Fig. 3.9 Classifier selection performance comparison in lenient matching accuracy: 5 representative samples

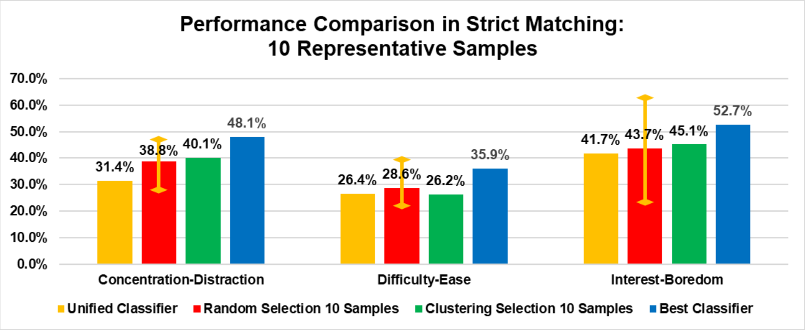


Fig. 3.10 Classifier selection performance comparison in strict matching accuracy: 10 representative samples

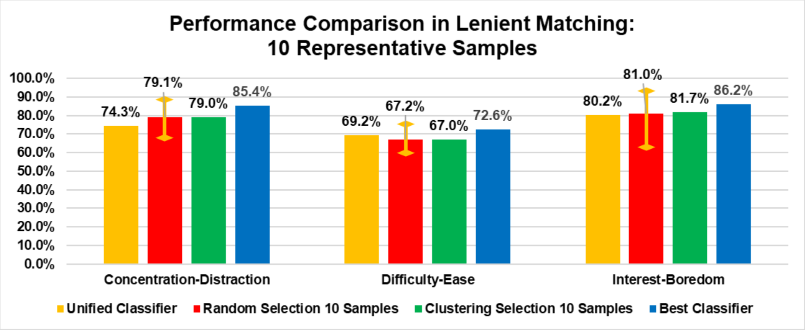


Fig. 3.11 Classifier selection performance comparison in lenient matching accuracy: 10 representative samples

The proposed scheme indicated slightly better performance than random selection, although the improvements were not significant. However, the proposed method did exhibit one advantage. The accuracy of random selection represents the average, and approximately half of the cases are lower this value. By contrast, the proposed method deterministically obtains the presented accuracy, which makes it possible to clearly avoid the cases that are worse than the average associated with random selection. This is potentially valuable because it is typically not possible to know the actual accuracy without systematically scoring a new learners' samples, which is a heavy task for the learner.

Here, discussion about the result in detail based on the assumptions in previous section is presented. First, concerning assumption A1, we examined the performance of the selected classifiers. Figure 3.12 shows the accuracy in a 5-cluster case with strict matching conditions for the representative samples, for the samples within the cluster of each representative sample, and for the overall samples. It is surprising that the accuracy for the representative samples is only around 50% with respect to the difficulty–ease state, which is much lower than the other two learning states. This implies that expressions and behaviors are diverse with respect to the difficulty–ease state of the learner, and a more sophisticated method is needed to deal with them in a more accurate and effective manner. In addition to this issue, the results indicated that a higher degree of accuracy for the samples within the clusters compared to the accuracy for the overall samples is possible, as stated in assumption A1. However, it also implies that the amount of improvement is not enough to achieve significant improvements in the total accuracy.

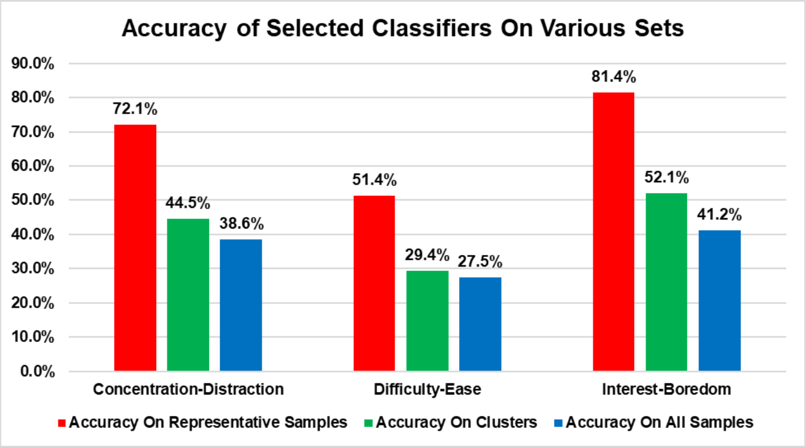


Fig. 3.12 Classification accuracy of different sets of samples. The red bar represents the selected classifiers' performance of representative samples; the green bar represents the performance of the selected classifiers associated with the samples in the clusters; the blue bar represents the accuracy of the selected classifiers associated with all samples.

Concerning assumptions A1, A2, and A3, an improvement in 10-cluster cases compared to 5-cluster cases was observed for concentration–distraction and interest–boredom. With the increased number of representative samples, more samples of neighboring or in-between representative samples were obtained. Variations in the representative samples were also obtained. This effect can be seen based on the improved performance, with the exception of the difficulty–ease state. A possible reason for the worse difficulty–ease performance may be partially due to the diverse behaviors in these cases, as mentioned above.

Concerning assumption A2 and A3, a high degree of accuracy is expected if the characteristics of the representative samples are similar to the characteristics of all samples associated with a new learner. To verify this, we examined the relationship between the accuracy of selected classifiers and score distribution of representative samples and all samples. The distribution of the scores for each learning state can be obtained from ground truth for both the representative samples and all samples, and Pearson correlation was observed between the score distributions of representative samples and all samples. The results are shown in Figure 3.13 and 3.14. Each dot represents a selected classifier. The horizontal axis represents the correlation, and the vertical axis indicates the accuracy of the classifier (i.e., between the worst (0.0) and the best (1.0)). Two or more classifiers are often selected in Step 4, and they are represented by a group of vertically aligned dots, as shown in Figure 3.13.

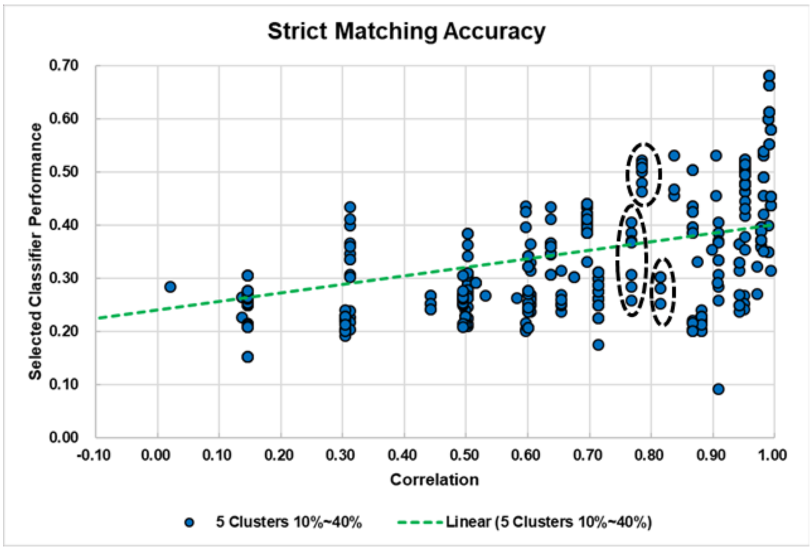


Fig. 3.13 Tendency of correlation and classifier performance in strict matching. Some of the cases in which multiple classifiers were selected are notified by dashed closures.

Based on these results, a rough relationship between correlation and classifier accuracy was observed, and selected classifiers tended to have a higher degree of accuracy

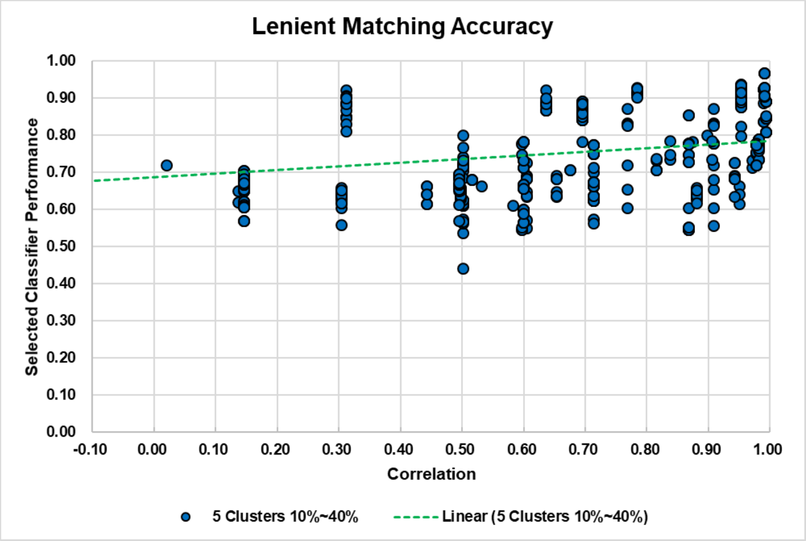


Fig. 3.14 Tendency of correlation and classifier performance in lenient matching

if they had a larger correlation. The linear regression value was 0.16 ($p < 0.01$) for strict matching criterion and 0.1 ($p < 0.01$) for lenient matching. It is important to note that the variations among multiple classifiers for each selection were not negligible. With the same correlation value, the accuracy varied among classifiers. If the best one can be chosen for a set of representative samples, the accuracy would be much higher. The best accuracy increases as the correlation increases. However, an effective method for this purpose was not clearly identified based on the scope and results of this study. Further investigation regarding this issue should be part of future research.

3.5 Summary

One critical problem is how to deal with a variety of students, i.e., inter-personal differences among learners. If target samples are taken from different participants, the performance of the system is much lower than the intra-personal results, i.e., the test data were applied to the classifier trained by the data from the same participant. For this purpose, a scheme that can quickly adjust to a new learner with a little effort of the learner was designed. First, it is assumed that the system has sufficient number of samples with ground truth from multiple learners during system development stage. Next, the system requests a new learner to give a small number of samples with self-evaluation, which are called “representative samples.” Then, the system automatically selects the classifier that fits the new learner’s representative samples. For this selection, several methods are considered and evaluated. Five representative samples were chosen randomly from a new learner’s learning records, and their performance was evaluated. The results demonstrate the accuracy-based method provides better performance than the unified classifiers, while the similarity-based method does not provide good results.

In further study, we investigated about method of better representative sample selection than random selection. Specifically, based on our basic assumptions, the selection of representative samples that commonly appear during e-learning of a new learner was examined. In our experiments, a slight improvement in the average accuracy was observed, although it was not significant. However, the proposed method did demonstrate the advantage of avoiding bad cases. Through experiments, certain characteristics of the data and classifications were confirmed. For example, neighboring samples around representative samples were not classified well, especially with respect to the difficulty–ease state. In addition, the distribution of the representative samples displayed a significant correlation with respect to accuracy, which make them a potentially valuable indicator for future investigations of new methods.

We also can expect better accuracy if we can obtain more representative samples with less effort by a new learner. Therefore, a method for reducing effort in and distress over self-scoring can be expected to lead to accuracy improvements too.

Chapter 4

Pronunciation Support for Language Learning through 3D Sensing of a Face

4.1 Characteristics of Language Learning

In the field of second language (L2) education, instructors and researchers try to improve the four language skills of students, i.e. speaking, listening, reading and writing [40], [41], [42] in many different ways. Pronunciation is a fundamental factor in speaking and listening. Intelligible pronunciation not only enables students to understand and be understood but also capacitate them to monitor their speech on the basis of input from the environment. Moreover, it has a great impact on their confidence to communicate in an engaging manner [43].

Correct pronunciation requires the accurate position of various articulators, such as the tongue, lips and jaw. To make clear and distinct word sounds, different speech organs work together to create diverse obstructions and/or oral cavities so as to shape the air in a particular fashion as it goes from the inside to the outside of the body.

Some evidence supports the importance of visual information for correct pronunciation. Even with regard to native speakers, sighted people can produce vowels more distinguishably and precisely than congenitally blind people can, according to research in [44]. They also found that due to visual deprivation, congenitally blind adults use more tongue variations in the implementation of vowel contrasts; however, this does not completely compensate for the reduced magnitude of upper lip protrusion in those native speakers. Another piece of evidence is the effect of visual information in speech perception. Our perceived pronunciation is the result of the interaction between hearing and vision. The McGurk effect [45] is a typical phenomenon that demonstrates such a symbiotic outcome in speech perception. Inconsistency between the auditory component and the visual component in terms of pronunciation can lead to illusion, i.e. the perception of a third sound.

Different languages do not have the same phonetic systems. This diversity requires learners to use new movements of articulators for new pronunciations and also necessitates a lot of practice to achieve proficiency. As a typical example, Japanese learners often encounter difficulties in 'rounded pronunciations' that require lip protrusion lacking in Japanese pronunciation, since the Japanese do not have clear distinction between rounded vowels and unrounded vowels in daily conversation. In fact, Duan et al suggested that one of the dominant Chinese mispronunciation patterns by Japanese learners is "Lip rounded and spreading" [46] which indicates a need to learn those pronunciations. Other languages, such as German and French, also have clear distinction between rounded and unrounded vowels, and its learning is important for Japanese learners. For instance, '*yu*', that is, '*fish*' with a second tone in Chinese or '*über*', that is, '*over*' in German. The Chinese example has an additional difficulty because the Japanese language has a different pronunciation for the same Romanized symbol. Japanese learners often try to produce a 'rounded' pronunciation with the mouth shape for 'unrounded' pronunciation for the same Romanized symbol, and it makes an incorrect sound.

4.2 Computer Assisted Language Learning system

In language education, CALL, the widespread e-learning system which has been designed, developed, and utilized for decades, has advanced from early one-way instruction to being able to provide feedback to students regarding their pronunciation [42]. According to [41], the simplest application just works as a digital recorder; learners can record their own pronunciations for comparison with a native speaker's by hearing them. A step further than a digital recorder is the use of speech visualization, for instance, waveforms, spectrograms, formant frequencies, pitches, contours and so on, as seen in [47]. A more sophisticated application can be achieved by employing automatic speech recognition (ASR), it can assess how similar the students' speech is to the native speakers'. Tsubota et al [48] developed an English pronunciation learning system for Japanese learners. The system is able to estimates the intelligibility of Japanese students' speech and rank their errors. Practice exercises for improvement, as well as instructions for correcting errors, are provided afterwards.

Most researchers and practitioners using CALL systems have paid considerable attention to audio, whereas some important aspects of articulation have not been well explored. We have seen that Japanese learners, particularly beginners, experience difficulty in learning some Chinese pronunciations even with the support of a CALL system. One typical difficulty is 'rounded pronunciation' which introduced previously. We focused on this problem and investigated a method that would provide a more comprehensive visual explanation of articulation for students.

For training in such pronunciations, explanations given by written texts are far from comprehensible, and even conventional multimedia approaches of showing pictures or videos do not give sufficient explanations about the articulations. We need a more comprehensible presentation of articulations because roundedness of vowel has a three-dimensional shape deformation, and lip protrusion cannot be easily recognized through ordinary pictures or movies. For this reason, we focused on a visual aid that makes Japanese learners aware of the differences of mouth shapes of rounded and unrounded vowels.

Since current CALL systems do not provide enough functions for this purpose, we need to devise new functions. In this study, we consider the following new functions of CALL systems:

(a) Demonstrate how native speakers pronounce words by showing articulation clearly and distinctly with 3D information.

(b) Show, in the same way, how learners pronounce the words, and enable them to check their correctness and/or weakness.

To achieve these functions, we obviously need to carry out a 3D sensing of the face, particularly around the mouth. For this purpose, we use an RGB-D camera. This type of camera has become precise, inexpensive, and can easily be connected to an ordinary computer.

4.3 Framework and System Design

The framework is designed as follows. Nowadays, the camera and image processing functionality can be easily installed on the computers of both teachers and learners. Videos with 3D information are obtained using an RGB-D camera, and then image enhancement, that is pseudo-coloring, is applied to the captured images or videos, particularly around the mouth. On teachers' computers, pronunciations of native speakers are recorded as they are used as educational materials for learners to watch and learn. On the learners' computers, images and data are obtained in the same way, and they are used to check the correctness of their articulation.

To verify the potential of above framework, a prototype system for function (a) was implemented, and verification experiment was conducted. The method of visualization in implementation of prototype system for function (b) on learners' side can be realized in the same way as (a), however, the learning effects are different. Therefore, we concentrated on the former and evaluated how learners could improve their pronunciation by watching the enhanced videos obtained by teachers.

4.3.1 Raw Data Acquisition

We employed Kinect v2 as an RGB-D sensor, which is ordinarily available at a low cost. The accuracy of Kinect v2 was investigated in [49]. The average depth error is less than 2 mm in a certain area; however, it often has worse measurement. It is reported that the measured depth value ranges from 1996 mm to 2004 mm for the true depth of 2000 mm. On the other hand, precision is much better as reported in [50] and [51]; average standard deviation is lower than 1.5 mm and remains stable during recording. For our purposes, precision is essential because protrusion can be measured as the relative depth change from the normal position. Lip protrusion for Chinese rounded pronunciation approximately ranges from 5 mm to 10 mm. In most cases, the precision is enough to detect it.

The quality of audio recording in Kinect v2, in contrast, is not satisfactory for language learning via examination. For this purpose, we used another audio recording device, which is commonly used in audio recording. The process of obtaining pronunciation clips are as follows (Figure 4.1). First, audio and video data are recorded with timestamps provided by the SDK of Kinect v2. They are segmented based on the values of audio data in which a silent period has significantly smaller values than pronunciation periods. Audio is recorded at the same time using the additional audio device. Second, audio data recorded by the additional audio device and the video data recorded by Kinect are synchronized by aligning audio data recorded by Kinect and audio data recorded by the audio device.

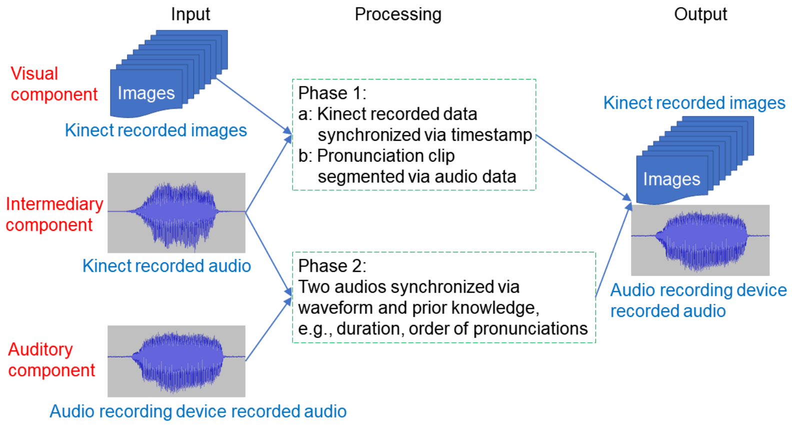

Input
Processing
Output
Visual component
Images
Kinect recorded images
Intermediary component
Kinect recorded audio
Auditory component
Audio recording device recorded audio
Phase 1:
a: Kinect recorded data synchronized via timestamp
b: Pronunciation clip segmented via audio data
Phase 2:
Two audios synchronized via waveform and prior knowledge, e.g., duration, order of pronunciations
Kinect recorded images
Images
Audio recording device recorded audio

Fig. 4.1 Synchronization of video data taken by Kinect and audio data taken by audio recording device

4.3.2 Face and Feature Detection

More than one thousand facial points can be detected through deploying Kinect SDK. Figure 4.2 gives examples of detection results. We can use the results to locate the face and the mouth area in particular. However, they do not cover the mouth region seamlessly and points around the lips are influenced by mouth movements as shown on the right-hand side of Figure 4.2. Therefore, another type of processing that emphasizes lip articulation is necessary, details of which are explained in the next section.

In the SDK of Kinect v2, the face-tracking algorithm requires that the torso is also

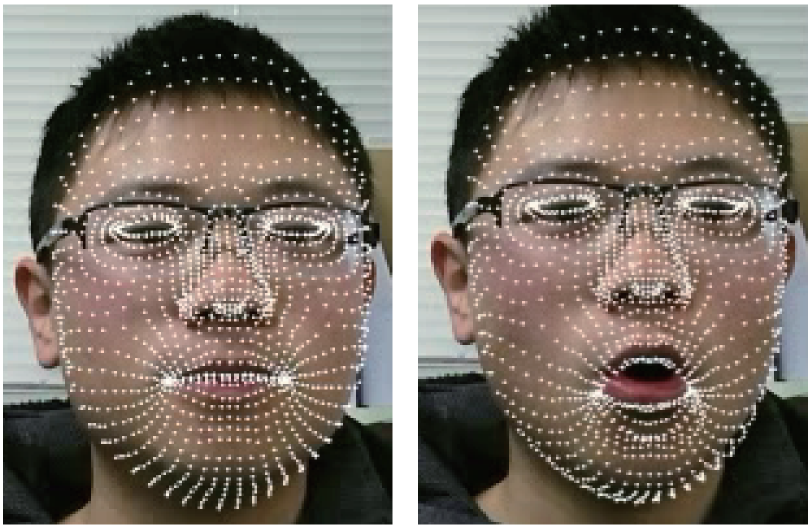


Fig. 4.2 Face and facial points detection result

detected simultaneously. A speaker needs to sit at a certain distance from Kinect to ensure that the upper body is visible.

4.3.3 Reference Point and Protrusion Measurement

The depth data measured by Kinect v2 camera need to be transformed to lip protrusion data. Lip protrusion is the relative depth change of the lip to the other parts of the face. To measure the amount of protrusion, we need a reference point that satisfies two conditions that are as follows:

(1) It should be steadily observed regardless of ordinary body movements. It means the depth of reference point should not be set to a constant value, it needs to be adaptive to the ordinary movements of learners. Meanwhile, it is still able to provide reliable information for lip protrusion measurement.

(2) Its location should not be affected by protrusion or any other kind of mouth movement. This requirement is the result of the observation shown in Figure 4.2, as the consequence, the points or pixels which are very close to lips are excluded from the candidates. However, the reference should not be far away from the mouth region. The closer to lips, the more reliable measurement can be obtained.

After various attempts, the tip of the nose satisfies these requirements well and is able to provide relatively reliable measurement as a reference point. Therefore, the relative distance to the tip of the nose is calculated for each pixel around the lips and is used to judge protrusion.

First, the amount of depth difference at each point p_i is calculated as follows:

$$\Delta d = d_i - d_0$$

where d_i is the depth of the i-th pixel, p_0 is the reference point, the depth of which is d_0.

Then, for each point on or around the lips, if its depth difference is smaller than the predetermined threshold, we regard the point as being protruded.

4.4 Visual Enhancement of Lip Protrusion by Pseudo-coloring

Face deformation parallel to the image plane can easily be perceived using ordinary videos. In contrast, lip protrusion is perpendicular to the image plane and its perception is often difficult. To solve this problem, we use pseudo-coloring on the basis of the measured depth of the lips.

4.4.1 Colorizing Area Selection

We designed the system so that only a fixed area around the mouth is pseudo-colorized because color changes over a wide area of the face would make viewers feel uncomfortable. We, therefore, consider two different subareas, the lip subarea and the non-lip subarea, as illustrated in Figure 4.3. For the lip subarea, it is colored vividly with an attention-grabbing color if the lips are protruded. For the non-lip subarea, it is colored to provide a contrast if the lip subarea is protruded.

The lip subarea is further segmented into the upper lip subarea and the lower lip subarea as illustrated in Figure 4.4. The physiological characteristic that means that the upper lip bulges slightly more in comparison to the lower lip sometimes results in unbalanced colorizing.

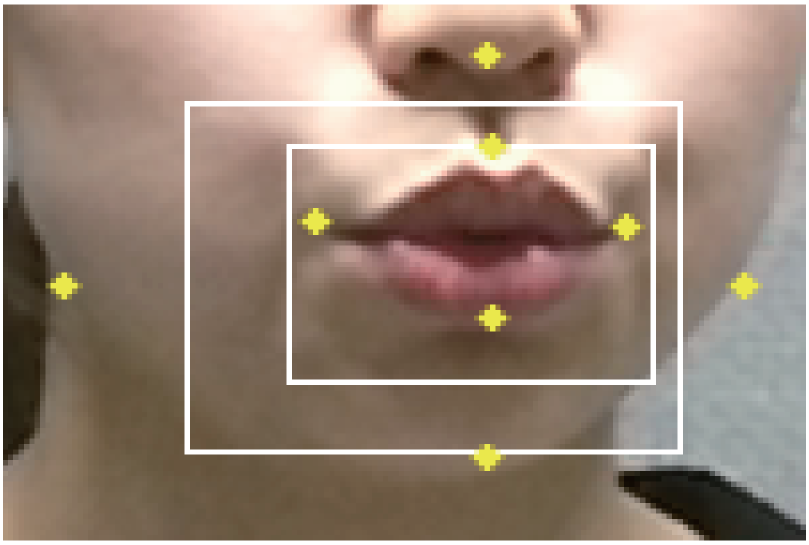


Fig. 4.3 Mouth area: lip subarea and non-lip subarea; yellow points are facial landmarks used for area determination, white rectangles are determined results

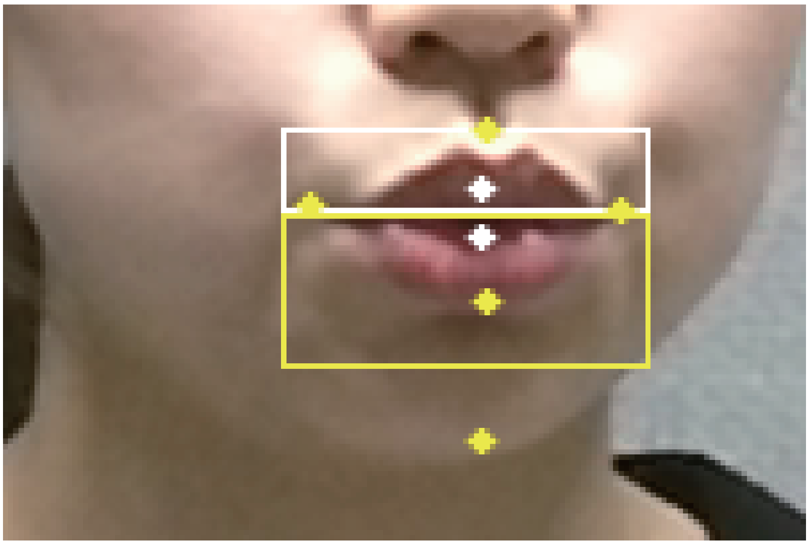


Fig. 4.4 Upper lip and lower lip subarea: yellow points are used for lip subarea segmentation as in Figure 4.3; white points are used for further segmentation

4.4.2 Pseudo-coloring Method

We designed the pseudo-coloring on the basis of the Hue-Saturation-Intensity (HSI) color space. Hue is used to represent the amount of depth change. Saturation is used to emphasize the important area of lip protrusion. Intensity is maintained as much as possible to give the original two-dimensional information as it is.

First, we convert RGB values to HSI values according to [52]. Details are given in the Appendix. Then, the HSI value is modified using the following method.

As we mentioned above, the intensity is kept the same as the input, i.e. $\tilde{I} = I$. Hue is determined according to the relative depth difference to the reference point for each pixel. One important phenomenon that we noticed in our experiments is that if the hue changes continuously, it does not draw the viewers' attention very much. The examples in Figure 4.5, Figure 4.6, Figure 4.7, Figure 4.8, Figure 4.9, and Figure 4.10 demonstrate the comparison of different segmentations of depth difference ranges and hue value ranges with using the original saturation values in both unrounded pronunciation and rounded pronunciation case. Each case has 2 small images which are the original image, pseudo-colorized image under border effect elimination but without lip-subarea division, respectively. The mouth region rectangle is drawn by white line on image to emphasize the processing region.

Figure 4.5 and Figure 4.6 give the examples of hue value changes continuously within mouth region. The lip protrusion is not emphasized so much, because the colors on lips are not sufficiently noticeable. It is much clear if we compare Figure 4.5 and Figure 4.6 to Figure 4.7, Figure 4.8, Figure 4.9, and Figure 4.10. The lip-protrusion in these 2 examples is much easier to be noticed than Figure 4.5 and Figure 4.6. Figure 4.7, Figure 4.8, Figure 4.9, and Figure 4.10 are the examples of pseudo-colorized images with discontinuous hue value ranges, the difference between them is the depth difference ranges segmentation. In Figure 4.7 and Figure 4.8, the background color, i.e., blue, is too much around lips in unrounded pronunciation case which is unnatural for viewing. Figure 4.9 and Figure 4.10 show the examples with enlarged depth difference ranges compare to the above 2 examples, which looks best among these 3 parameter settings.

Through the comparison, significant discontinuity, i.e. coarse quantization, is necessary for the changes to be noticed, i.e. protrusion. For this purpose, we quantize hue into three levels that correspond to three ranges of relative depth (see Table 4.1). Hue value is then modified via the formula shown in the Appendix. The upper lip and the lower lip usually have a slightly different amount of protrusion; hence,

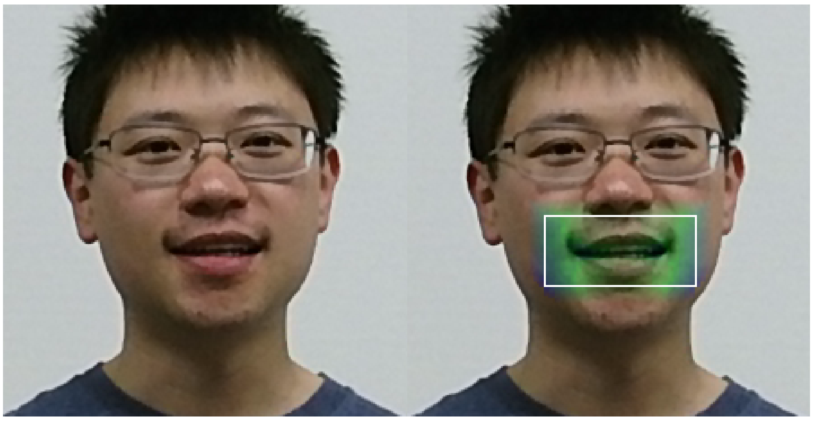


Fig. 4.5 Examples of hue value changes continuously (unrounded pronunciation case)

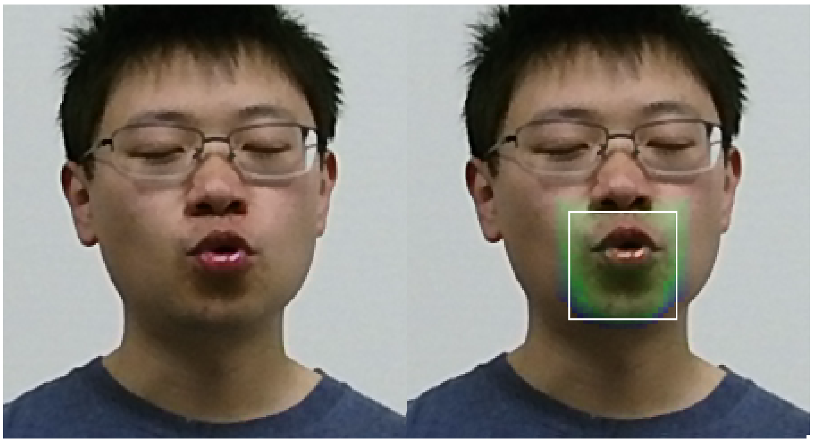


Fig. 4.6 Examples of hue value changes continuously (rounded pronunciation case)

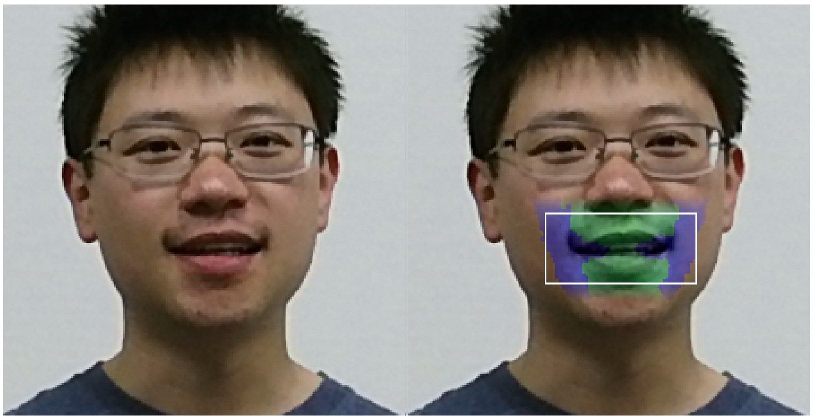


Fig. 4.7 Examples of hue value changes discontinuously case 1 (unrounded pronunciation case)

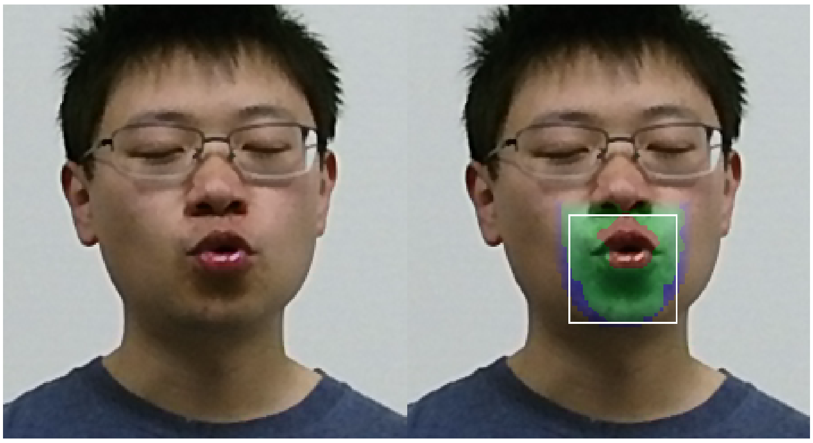


Fig. 4.8 Examples of hue value changes discontinuously case 1 (rounded pronunciation case)

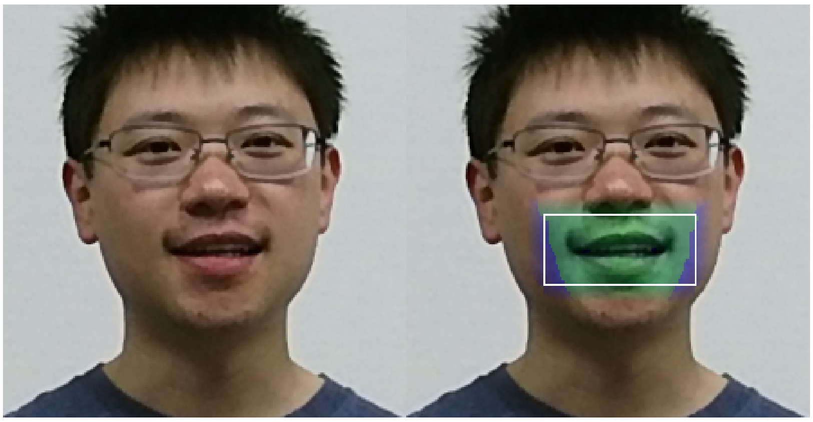


Fig. 4.9 Examples of hue value changes discontinuously case 2 (unrounded pronunciation case)

we designed slightly different corresponding ranges of distance between them. Note that the parameters concerning depth ranges are adjusted to the native speaker who provided pronunciation samples for our video-based materials.

The lip-protrusion is not sufficiently salient with only using coarse quantization of hue values. Saturation value is vital in emphasizing the important area of lip protrusion. During the investigation for saturation parameter, we found that saturation is the most difficult part because not all the value combinations of hue, saturation and intensity can be converted into a valid range of RGB values. It means even the saturation shows different values after construction based on depth difference values,

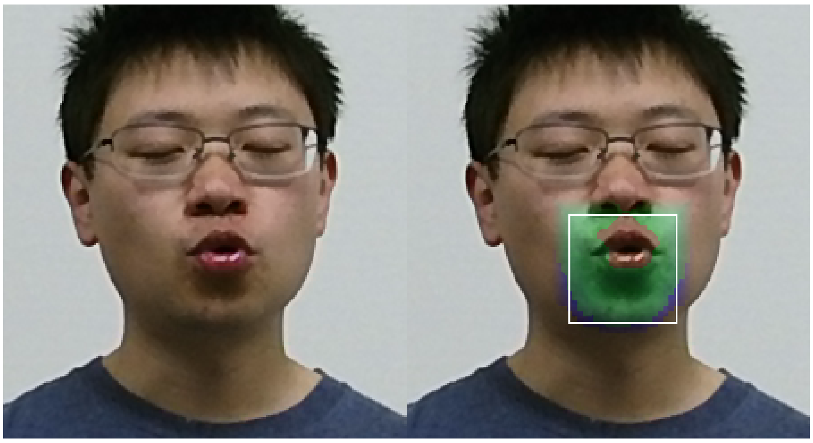


Fig. 4.10 Examples of hue value changes discontinuously case 2 (rounded pronunciation case)

the converted RGB value would demonstrate the same value sometimes. To cope with this problem, the maximum saturation value allowed with the values of hue and intensity that are modified in the above ways is calculated first. (The Appendix gives the actual method). The saturation value is determined on the basis of the maximum value and depth difference Δd for each hue range (shown in Table 4.1).

$$\tilde{S}_{lip} = \begin{cases} S_{max} & \text{hue is around red} \\ S_{max}\left(0.8 - \left(\Delta d - 20\right)/100\right) & \text{hue is around green:} \\ & \text{upper lip} \\ S_{max}\left(0.8 - \left(\Delta d - 27\right)/100\right) & \text{hue is around green:} \\ & \text{lower lip} \\ S_{max}\left(0.8 - \left(\Delta d - 60\right)/100\right) & \text{hue is around blue} \end{cases}$$

$$\tilde{S}_{non-lip} = \begin{cases} S_{max}\left(1.0 - \left(\Delta d/2\right)/100\right) & \text{hue is around} \\ & \text{green} \\ S_{max}\left(0.6 - \left(\Delta d - 60\right)/100\right) & \text{hue is around} \\ & \text{blue} \end{cases}$$

The maximum allowed saturation is assigned to a red-based hue of significant protrusion. The saturation value assigned to a green-based range of minor or no protrusion and that assigned to a blue-based range of skin background are determined

Table. 4.1 Hue calculation

Mouth area level	Depth difference range Upper lip / lower lip (mm)	Hue value range (degree)
Significant protrusion	Upper lip subarea: $\Delta d \leq 20$ Lower lip subarea: $\Delta d \leq 27$	Red-based: [0, 20] [0, 27]
Low-protrusion or non-protrusion	Upper lip subarea: $20 < \Delta d \leq 60$ Lower lip subarea: $27 < \Delta d \leq 60$ Non-lip subarea: $\Delta d \leq 60$	Green-based: (120, 160] (120, 153] [120, 180]
Skin background	$60 < \Delta d \leq 100$	Blue-based: (240, 280]
Non-target area	$100 < \Delta d$	Original hue value

as being proportional to the maximum saturation. After the HSI values are determined using the above method, they are converted into RGB values as shown in the Appendix.

The following figures give the examples of unsatisfactory results in saturation construction with different parameters under the same quantization of hue values and depth difference ranges with the setting in Figure 4.9 and Figure 4.10.

4.4.3 Smoothing Around Borders

The discontinuity of color around borders can make images very unnatural and draws unnecessary attention. To avoid this effect, pseudo-colors and original colors are blended proportionally on the basis of the distance from the border. Figure 4.17 shows an example of the final result of pseudo-coloring. Figure 4.18 shows an example

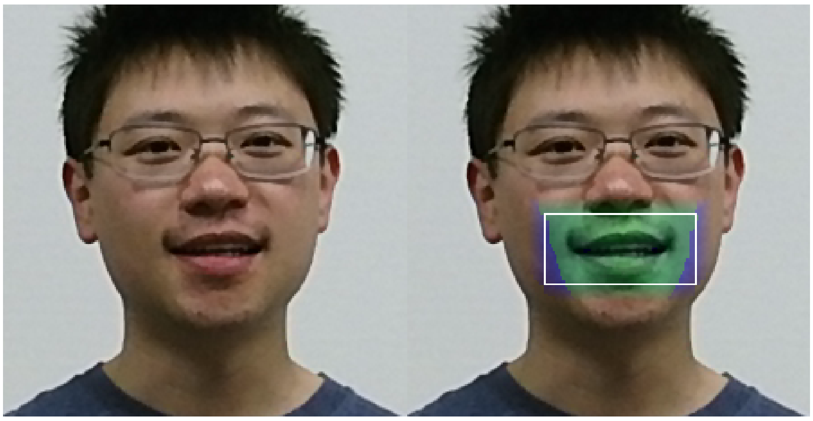


Fig. 4.11 Examples of saturation parameter case 1 (unrounded pronunciation case)

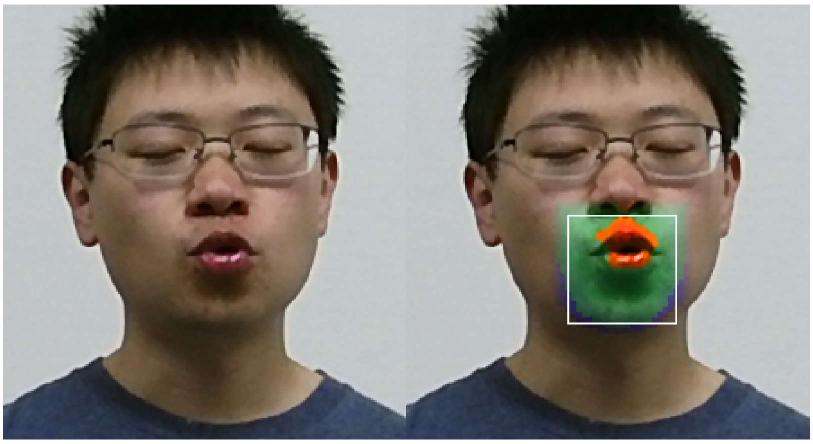


Fig. 4.12 Examples of saturation parameter case 1 (rounded pronunciation case)

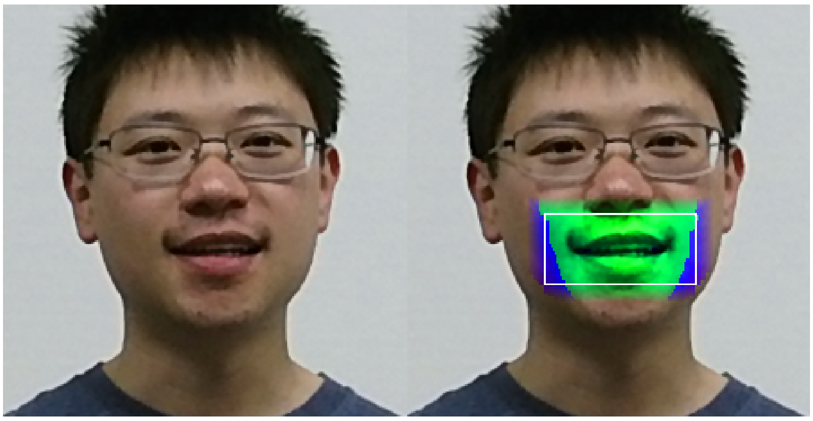


Fig. 4.13 Examples of saturation parameter case 2 (unrounded pronunciation case)

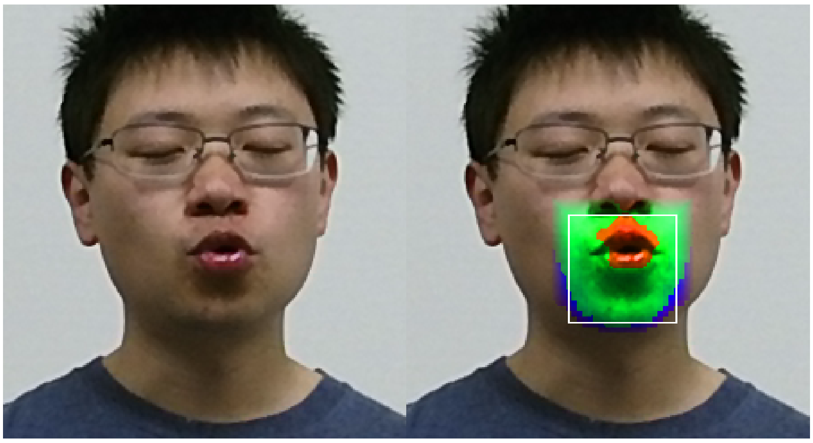


Fig. 4.14 Examples of saturation parameter case 2 (rounded pronunciation case)

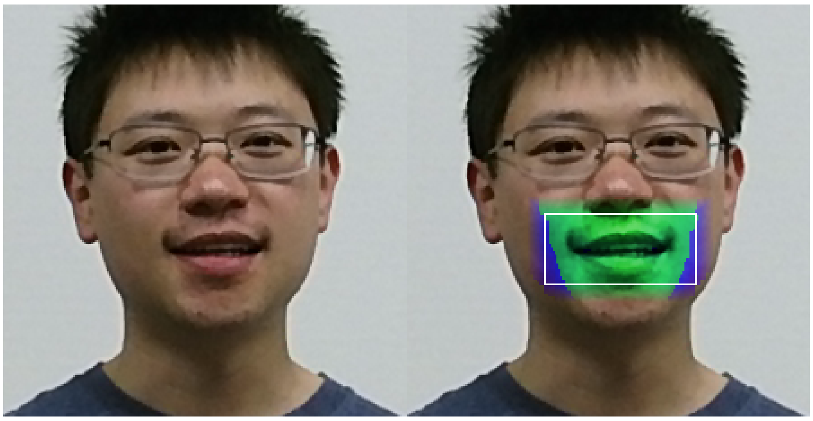


Fig. 4.15 Examples of saturation parameter case 3 (unrounded pronunciation case)

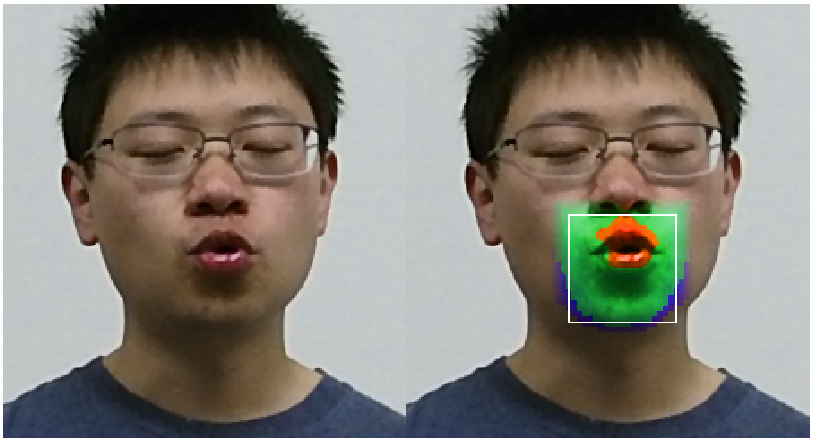


Fig. 4.16 Examples of saturation parameter case 3 (rounded pronunciation case)

with non-negligible noise, and it is still acceptable.

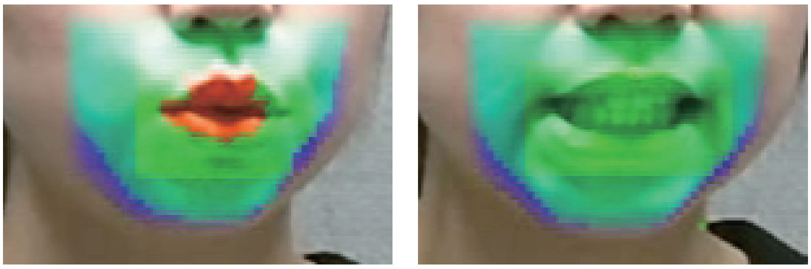


Fig. 4.17 Example of final result of pseudocolouring: rounded and unrounded

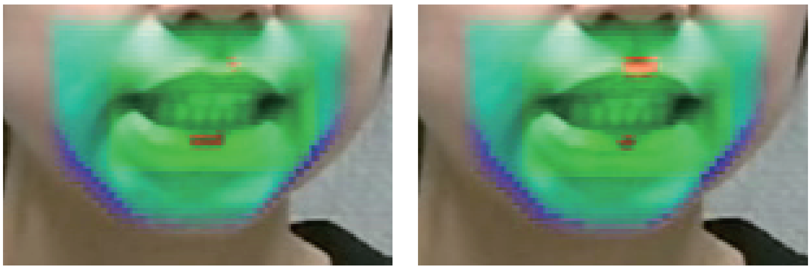


Fig. 4.18 Example of final result of pseudocolouring: noise included

4.5 Experiment and Result

4.5.1 Objective

The experiments were conducted to confirm the effect of our proposed method on actual learners. We chose Chinese pronunciation as a target. We gathered forty-three students from Kyoto University as participants: ten were beginners who had not learned Chinese, twenty-three had learned Chinese for nearly one semester and ten were students with experience of learning Chinese for approximately a year and a half.

To compare the effect on learning that our method has with the effect that conventional videos have, participants were randomly separated into two groups: Group A used conventional videos of a native speaker, whereas Group B used the native speaker's videos prepared using our method. We did not separate learners by learning periods because the statistics can be more reliable by the number of samples without separation. Conversely, it is not intended to evaluate the effects of learning periods.

Given that the two groups might have different levels of knowledge or ability as starting point, for better reliability of statistics, we used an indirect method in which advantages to the baseline methods described below are parameterized for both groups, and then the efficacy is compared on the basis of the parameters. This idea is based on common methods in education, medicine, and others. For statistics in those fields, applying two or more different treatments to the same person is often avoided, e.g., different teaching methods, different medicine, etc. Otherwise, learning effects or curative effects would inevitably affect the statistics. Moreover, pre-test often affects the internal states of learners because they could be aware of what is focused upon and they could recall them if the pre-test is closely related to the content of statistics. This problem has been intensively examined, and not a few ideas and methods were proposed [53]. Rubin defined three categories of missing data [54] [55], based on which our experiments correspond to missing completely at random (MCAR) condition. If MCAR condition is satisfied, mean, regression, or some of other statistics can safely be compared between two groups.

4.5.2 Content

We chose a pair of unrounded and rounded vowels, /i/ and /y/ in the International Phonetic Alphabet (IPA) format. Both vowels are situated in the same position on the IPA chart; however, they have different degrees of lip protrusion. The corresponding formats in Chinese Pinyin notation are '*i*' and '*ü*'-symbolised as '*u*' when associates with '*j*', '*q*', '*x*' and '*y*', respectively.

They can be associated with six consonants-'*j*', '*l*', '*n*', '*q*', '*x*' and '*y*'-and four tones in Chinese. Consequently, we have forty-eight words in total. Each word is used twelve times for each participant to obtain sufficient samples for statistical analysis.

Four types of learning materials for those words were prepared: (a) pinyin symbol, (b) pinyin symbol with audio, (c) combination of pinyin symbol and common video (audio included) and (d) combination of pinyin symbol and pseudo-colorized video (audio included). The forty-eight words are randomized into three groups, each group contains eight rounded words and eight unrounded words. One group is presented in pinyin material (a); another group is in audio material (b); the last group is in two types of material, both (c) and (d). Figure 4.19 gives the examples of (a) and (b). Figure 4.20 provides examples for (c) and (d). The left half demonstrates the examples of the mouth area used in the experiment; upper one is for (c), lower one is for (d). On the right is a full-face image of another native that gives an overall impression of video-based learning materials. The sequence of words is randomized every time to avoid an order effect. For (c) and (d), the pinyin symbol is placed near the mouth. This maneuver can keep participants' attention focused around the mouth. Participants in Group A perform the pronunciation for (a), (b) and (c); participants in Group B perform for (a), (b) and (d).

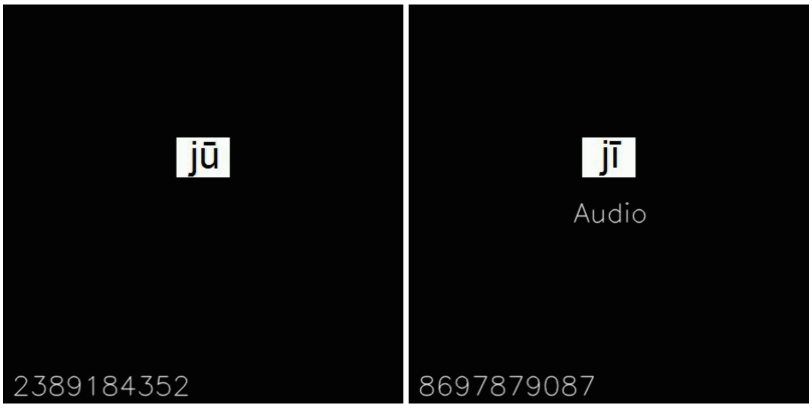


Fig. 4.19 Examples of pinyin symbol (a) and audio (b)

4.5.3 Scoring of Pronunciation

Every pronunciation made by the participants was evaluated by three Chinese native speakers who had experience in teaching the Chinese language to Japanese students. They judged whether the participants had made correct pronunciations by listening to audio files, with informed which pronunciations the participants were requested to pronounce. In particular, each pronunciation was evaluated as the correct pronunciation or another incorrect pronunciation. The final evaluation results were

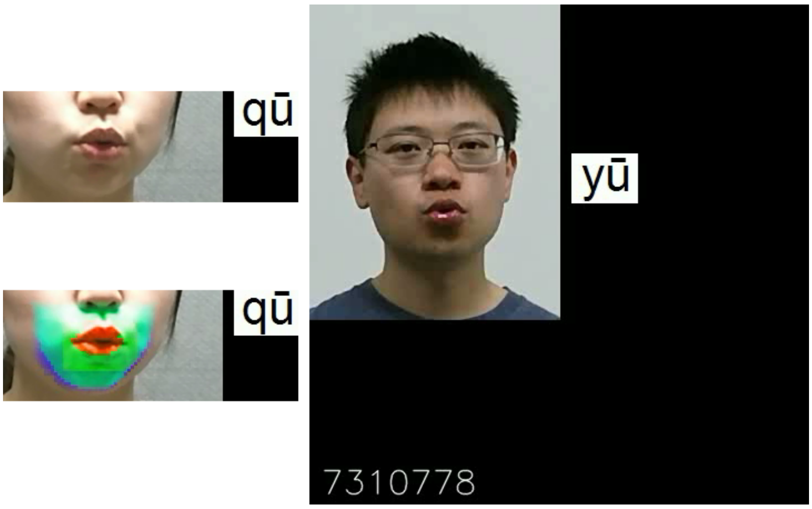


Fig. 4.20 Examples of common video (c) and pseudocoloured video (d)

obtained on the basis of a majority vote, i.e. we accepted the results that two or more evaluators supported.

Arithmetic mean of Cohen's kappa [56], [57] of every pairwise native evaluators on rounded vowel, unrounded vowel, consonant and tone were calculated to evaluate inter-rater reliability (Table 4.2). According to the guidelines for interpreting kappa values [58], consonant and tone have almost 'perfect' agreement, rounded vowel has 'fair' agreement and unrounded vowel has 'moderate' agreement. This result suggests that learners' pronunciation of rounded and unrounded vowel is sometimes ambiguous for which evaluators do not have a good match. However, the value of kappa indicates 'fair' or 'moderate' based on the guideline, which is still acceptable. Therefore, we conclude that each evaluation can be trusted if majority of evaluators support them.

Table. 4.2 Inter-rater reliability values

Pronunciation element:	Arithmetic mean of Cohen's kappa:
Consonant	0.89
Rounded Vowel	0.39
Unrounded Vowel	0.49
Tone	0.96

4.5.4 Result and Discussion

Since rounded vowels are our focus concerning articulation that requires lip protrusion, we first present the detailed results on rounded vowels, and next, we briefly show the results regarding unrounded vowels, consonants and tones for comparison.

Figure 4.21 shows the overall result for rounded vowels. The graph shows the number of average and standard deviations of incorrect pronunciations. The left and middle columns show the results of the cases in which pinyin symbol material (a) and audio material (b) were presented to participants, respectively. The third column shows the results of presenting a conventional video (c) and a pseudo-colored video (d) to Group A and Group B, respectively. From the average values, we can see that pronunciations made good improvements in the case of both the conventional video and the pseudo-colored video. They show the superiority of video-based materials over the pinyin symbol and audio-based materials.

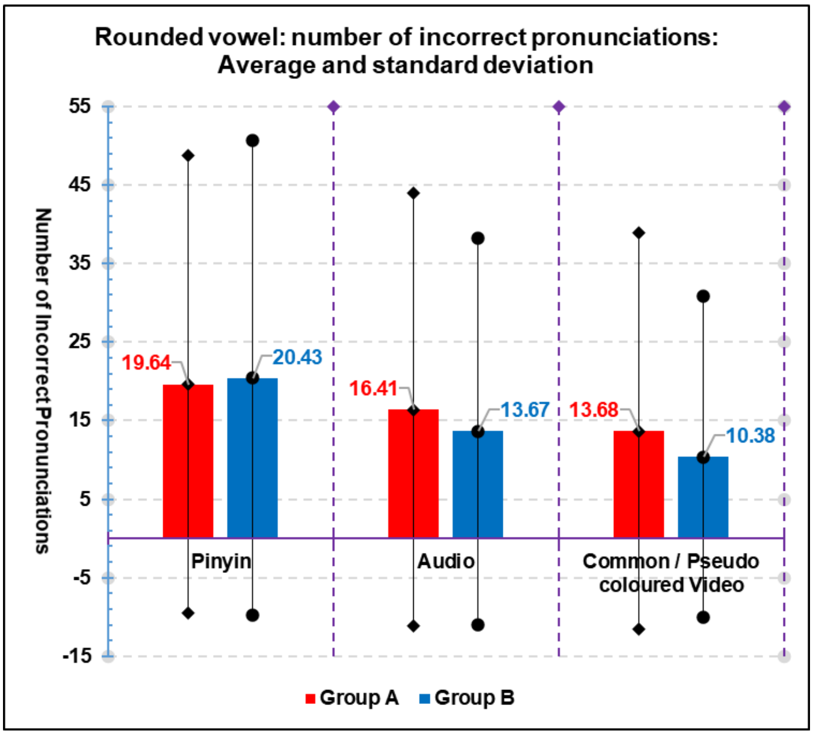


Fig. 4.21 Overall performance concerning the number of incorrect pronunciations of rounded vowels

Next, we compare the effects of conventional videos and pseudo-colored videos. A direct comparison is not possible because the performances of the participants are slightly different between the two groups.

For this purpose, we use regression analysis as shown in Figure 4.22. The graph shows the estimation of how many potentially incorrect pronunciations using pinyin symbol materials also occur with video-based materials. The horizontal axis represents the number of incorrect pronunciations with pinyin symbol materials, and the vertical axis represents the number of incorrect pronunciations with video-based materials. Each dot represents the result of one participant. The regression indicates how many incorrect pronunciations are expected to occur by using each video-based material. In other words, a smaller inclination means a larger amount of potential improvements from the baseline method.

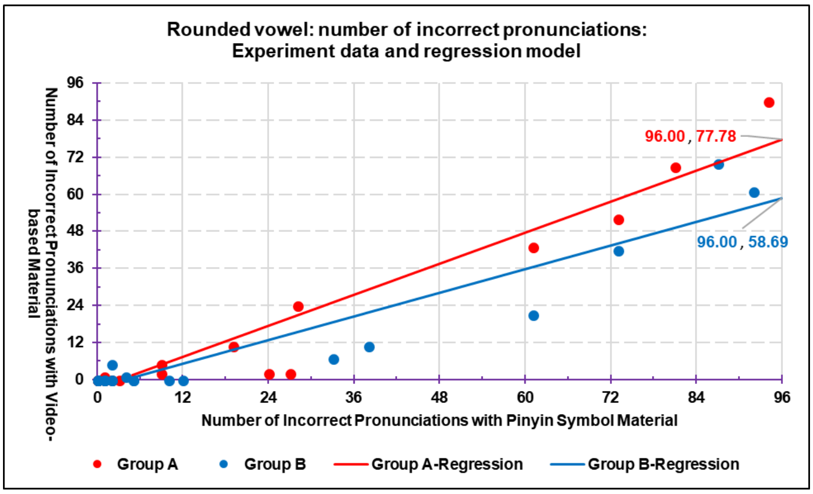


Fig. 4.22 Experiment data and regression models regarding rounded vowels

The red line ($Y = 0.84 \times X - 2.80$) and the blue line ($Y = 0.64 \times X - 2.68$) in Figure 4.22 are the linear regressions for Group A and Group B, respectively, where Y indicates the number of incorrect pronunciation for rounded vowels through the video-based learning materials-material type (c) or (d); X represents the number of incorrect pronunciation for rounded vowels through the pinyin learning material-material type (a), that is the baseline method in this comparison. The inclination for Group A is greater than it for Group B, ($p < 0.01$ for both cases). The R-squared values of Group A and Group B are 0.93 and 0.89, respectively, both of which show the reliability of the regression.

Similar effects were observed when we consider audio learning material-material type (b) as the baseline method. Linear regressions are $Y = 0.91 \times X - 1.18$ and $Y = 0.82 \times X - 0.82$ for Group A and Group B, respectively. Group B has smaller inclination value than Group A ($p < 0.01$ for both cases) too. The R-squared values are 0.97 and 0.98 for Group A and Group B, respectively. Both the results demonstrated that more improvement can be expected through pseudo-colored video material than common video material.

A t-test for rounded vowels using conventional video materials and pseudo-colored video materials was also conducted. In this calculation, we excluded the data of participants who made correct pronunciations for almost all samples with pinyin symbol materials. Those participants obtained almost perfect results for both video-based methods, and improvements are not different between them, i.e. their values are either 0 or very small. To avoid this effect, we gathered the data of participants who delivered incorrect pronunciation for more than five samples, i.e. who have enough room for improvement. For those data, the t-test result shows a significant difference between the two groups (one-tail P-value is 0.0083). Based on these facts, we conclude that pseudo-colored video materials have superiority over conventional video materials.

Figure 4.23 shows the experimental data distribution for unrounded vowels, consonants and tones. Since no reliable regression models were obtained, only experimental data were drawn. In contrast to rounded vowels, the results do not show any superiority of pseudo-colored video materials. For unrounded vowels, Group A and Group B were drawn separately for better viewing. All participants demonstrated good performance, and we cannot see any differences. For consonants and tones, results are more scattered; however, there are no clear performance differences between the two video-based methods. It is reasonable because a pseudo-colored video does not provide better information than a conventional video in showing consonant and tone. This implies that the differences between the videos only work for rounded vowel improvement as we had designed. The result also implies that pseudo-coloring does not have a negative effect on learning consonant and tone pronunciation.

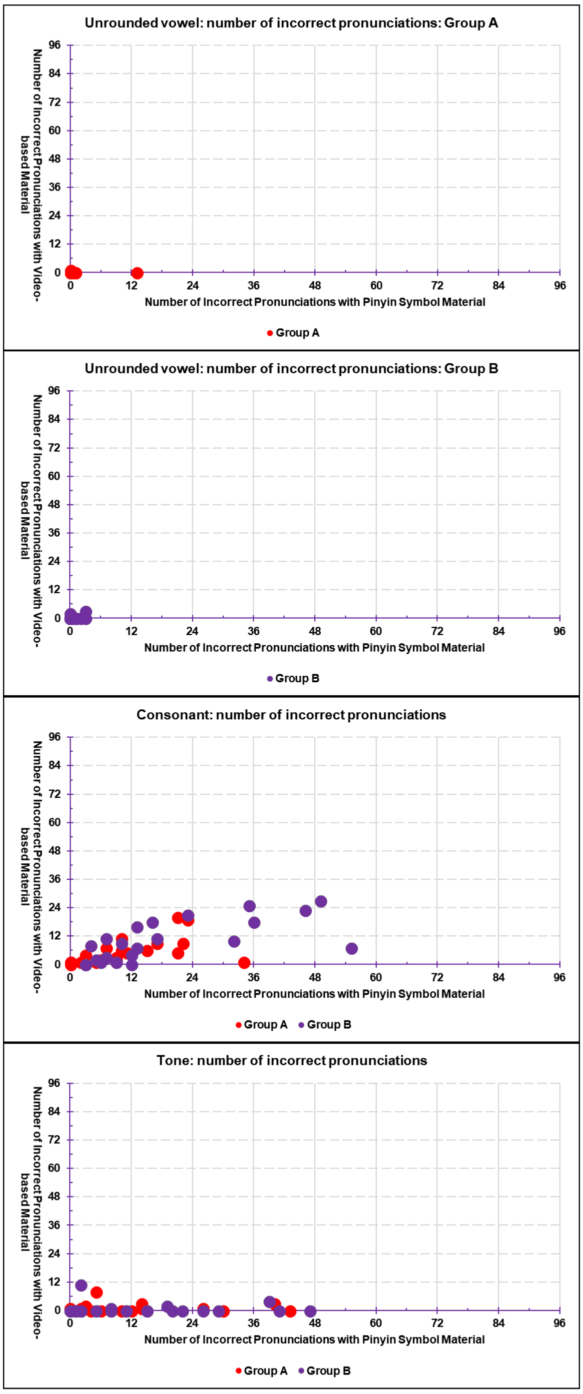


Fig. 4.23 Experiment data regarding unrounded vowels: Group A, unrounded vowels: Group B, consonants and tones

4.6 Summary

In language learning, pronunciation is a fundamental factor in speaking and listening. Intelligible pronunciation not only enables students to understand and be understood but also helps them to monitor their speech on the basis of input from the environment. Correct pronunciation requires the accurate position of various articulators, such as the tongue, lips and jaw. Traditional explanations given by written texts and conventional multimedia approaches do not give sufficient explanations of the articulations.

For providing more advanced supports for pronunciation learning in e-learning, the following visualization functions are considered: (a) demonstrating learners how native speakers pronounce words by showing articulation with 3D information, (b) showing learners how they pronounce the words, and enable them to check their correctness and/or weakness. A prototype system for (a) was implemented. Realization of function (b) on learners' side is left for future work; however, the framework can be the same way as (a). 'Rounded pronunciation' is chosen as the learning target, because Japanese learners often have difficulties in such pronunciation with lip protrusion. It cannot be easily recognized through ordinary pictures or movies, and explanation with 3D information around the mouth provides good supports for Japanese learners.

The framework is designed as follows. Videos with 3D information are obtained by an RGB-D camera, and then pseudo-coloring, is applied to the captured images. Lip protrusion is measured as the relative depth of the lip to the other parts of the face. The tip of the nose is used as the reference point because it is steadily observed regardless of ordinary body movements, and its location is not affected by mouth movements. Pseudo-coloring is applied to the mouth region according to the measured depth. Lip subarea and non-lip subarea are differently colored. For lip subarea, it is colored vividly with an attention-grabbing color if the lips are protruded. For non-lip subarea, it is colored to provide a contrast. The lip subarea is further segmented into the upper lip and the lower lip subarea, and they are colored separately because of the physiological characteristic that upper lip bulges slightly more than lower lip. The color is chosen based on the Hue-Saturation-Intensity color space. Intensity is maintained to give the original two-dimensional information as it is. Hue is modified according to the relative depth. Significant discontinuity, i.e. coarse quantization, is necessary for protrusion to be noticed. Saturation is used to emphasize lip protrusion. The maximum saturation allowed with hue and intensity is calculated. Then, saturation is determined based on the maximum value and relative depth. The color around the border of mouth area is smoothed to attenuate the effect of unnatural and drawing unnecessary attention.

Evaluation was conducted to verify how learners could improve their pronunciation by watching the enhanced videos that captured teachers' pronunciations. Forty-three students in Kyoto University with various learning experiences were gathered. They were randomly separated into two groups for comparing the effects of common video and pseudo-colored video, respectively. Improvements from baseline methods such as the symbol-based method were evaluated by taking linear regression and t-test. As the results, decreased number of incorrect pronunciations for rounded vowel showed the superiority of the pseudo-colored videos.

Chapter 5

Conclusion and Future Work

Learning is a complicated mental activity requires learners to involve themselves in this learner-centered process. The learning outcome will be excellent and learning efficiency will be high if the learning is designed to attract and suit learners well in this learner-content interaction. Self-paced e-learning is supposed to accommodate large number of learners with various types and characteristics, which implies much more effort and labor is required from instructors to design and develop a course in e-learning than in conventional educational form. To keep students engaged and motivated in self-paced e-learning has great importance, and is a challenge. A good instructional design is indispensable. Instructors need to ponder various ways to integrate technology and instruction for optimizing learning experience. Thus, support for self-paced e-learning is the target of investigation in this research, which findings can be generalized to a wide variety of learning.

5.1 Conclusion

In the research of learning state recognition through visual sensing, we designed an e-learning support system that can capture learners' behaviors visually and estimate learners' learning states. We chose concentration-distraction, difficulty-ease, and interest-boredom as a learner's learning states, and these were recognized by using the learner's presence information, head and facial parts information, and probability of gazing at the screen. The experimental results showed the potential of our classification method by SVM using the abovementioned visual features: approximately 60% average accuracy in strict matching and approximately 90% average accuracy in lenient matching can be achieved. We also examined practical methods for adjusting to a new learner who can provide only a few samples as ground truth. Accuracy-based selection of classifiers and our integrated method showed better performance than the unified classifier for which all of the samples were used for classifier training.

Further investigation was conducted to reduce the gap between the selected classifier of a new learner and the most appropriate classifier of that learner via selection of better representative samples than random selection. This problem concerns interpersonal and intra-personal variation in the learner's behavior and internal states. Essentially, the challenge is dealing with those variations using a small number of samples. The proposed method of choosing representative sample is based on three assumptions: Frequently appearing samples can be good representative samples; A set of representative samples that covers a wider area of the feature space provides better accuracy; A set of representative samples with enough variety of classes gives better accuracy. Kernel density estimation and hierarchical clustering were involved in the scheme. The experiments showed a slight improvement in the average accuracy was observed, although it was not significant. However, the proposed method did demonstrate the advantage of avoiding bad cases. Through experiments, certain characteristics of the data and classifications were confirmed. For example, neighboring samples around representative samples were not classified well, especially with respect to the difficulty–ease state. In addition, the distribution of the representative samples displayed a significant correlation with respect to accuracy, which make them a potentially valuable indicator for future investigations of new methods.

In the research of pronunciation support for language learning through 3D sensing of a face, we focused on learning support for rounded pronunciation, specifically, visual enhancement of lip protrusion. To achieve this, we proposed the pseudo-coloring of face images through sensing with an RGB-D camera. This visualization provides learners with an intuitive sense and comprehensible information regarding lip protrusion. This method can be used in preparation of teaching materials and can also be used to check learners' pronunciations. We conducted experiments to verify the former usage. The results of the experiment indicate that our proposed method provides better performance than does the conventional video method in the reduction of incorrect rounded vowel pronunciations. Moreover, the results show that beginners who often make mistakes demonstrate significant improvements with our method. It suggests that the method works for beginners from the early days of learning.

5.2 Future Work

About the future work regarding learning state recognition, we need a variety of investigations to improve recognition accuracy, including improvements in the sensing system, feature selection, and classifier selection. For example, incorporating other measuring modalities, such as non-intrusive physiological measurement, input information vis mouse and keyboard, to improve current measuring system can partly solve the problems. By means of data fusion, we could take full advantage of multiple modalities. The missing information of eyes status increases the difficulty of problems significantly. An economical detection system under current technology is worth studying. Furthermore, design of a more complicated recognition and selection system could be another possible way. From more intelligent aspect, we could consider the utilization of the connections among learning states, currently they are handled individually. For instance, recognition information regarding one state could be used for another state estimation. Developing a user interface for providing educational information for teachers' browsing is also important.

For future studies about pronunciation learning support, we need to verify how the pseudo-coloring of learners' videos helps their learning. This function will provide a new kind of visual feedback to learners. Moreover, the data can be stored in an e-portfolio and can be used for formative assessment of both teachers and students. For this purpose, we need the automatic adjustment of pseudo-coloring parameters for each user. In particular, in our experiments, we determined the coloring parameters for a specific person; however, this adjustment needs to be automated for each learner. As another extension of our work, the automatic evaluation of pronunciation is a good target, whereby feedback such as pronunciation instruction can be provided. This could be a good idea for the future design of language learning environments.

The experiments of learning state estimation were conducted on language e-learning materials. For future work, we need to proceed to other types of e-learning materials such as natural science, humanities, etc. Age and learning place of learners are also critical factors for learning effects. We need to further investigate how our framework contributes for diverse conditions. The similar problems are in pronunciation learning supports. Various pronunciation in various language may be difficult for learners. Much room for applications is left for future works.

Concerning the system design, the proposed system is a prototype that is a combination of a camera, screen capture device, audio recording device, and software for image processing and pattern recognition. They should be well integrated into an ordinary computer or mobile device for the convenience of learners. For this purpose, there are already some mobile phones and tablet devices that have RGB-D camera, and we can expect that those devices will be good platforms.

Acknowledgement

The journey is the reward. When I decided to start my PhD journey in a foreign country, when I decided to dive in a completely different major, I thought I was prepared to face what would come to me, with my enthusiasm and will. However, the reality was tougher and more painful than I imagined. I was majored in Educational Technology for B.S. and M.S., it was a great challenge to pursue PhD in Engineering, but I believed if I could understand technology better, I could do more and better in education. I still believe it now.

I would like to extend my gratitude to a number of people, since this dissertation would not have been possible to complete without their assistance and support. My deepest appreciation to my supervisor, Professor Yuichi Nakamura, for his expert instructions, insightful guidance, great patience on my research, and kind help in my daily life for the past years. I have learned much from his attitude towards being an extraordinary teacher and a scholar.

I would like to express my sincere gratitude to Professor Kita and Professor Koyamada for the insightful suggestions and comments on my research and this thesis. Without their warm-hearted supervision, this dissertation would not be possible to be completed.

Special acknowledge to Dr. Kazuaki Kondo, for his constructive suggestions, comments, advices and kindly teaching in expertise during my journey for PhD.

I would like to offer my special thanks to Professor Masatake Dantsuji, Professor Hiroaki Nanjo, and group members in Dantsuji laboratory. Without active and close collaboration with them, my research would be impossible to advance. Their professional suggestions, advices and support extended my knowledge and experiences.

I am indebted to Miss Obata. Without her kindly help in daily life, it would be much tough for me to spend years in pursuing PhD degree in Japan.

I would like to express sincere gratitude to all the members in Nakamura laboratory for their help and support in academic and daily life.

I would like to offer my special thanks to all of my dear friends for their selfless help, support, and care in my life.

A special thank to Professor Li Peng, School of Information Science and Technology, Northeast Normal University, China. His supervision during my M.S. and continuous encouragement opened up my vision towards to wider perspective about Educational Technology.

In the end, I would like to express my deepest gratitude to my family, especially for my parents, Jin Yu and Xiwen Huang, for their warm support, guidance, and education in my life. I would not be able to finish my journey in pursuing PhD without their warm support and encouragement.

Publication

Journal papers

1. Siyang YU, Kazuaki KONDO, Yuichi NAKAMURA, Takayuki NAKAJIMA, and Masatake DANTSUJI, “Learning state recognition in self-paced e-learning,” IEICE TRANSACTIONS on Information and Systems, 2017, 100(2): 340-349
2. Siyang YU, Kazuaki KONDO, Yuichi NAKAMURA, Takayuki NAKAJIMA, Hiroaki NANJO, and Masatake DANTSUJI, “Visual emphasis of lip protrusion for pronunciation learning,” IEICE TRANSACTIONS on Information and Systems, Vol.E102-D,No.1,pp156-164. Jan. 2019.
3. Siyang YU, Kazuaki KONDO, Yuichi NAKAMURA, Takayuki NAKAJIMA, and Masatake DANTSUJI, “Investigation on e-Learning Status Estimation for New Learners—-classifier selection on representative sample selection,” IEICE TRANSACTIONS on Information and Systems, (submitted)

Workshops and Conferences

1. Siyang YU, Kazuaki KONDO, Hiromasa YOSHIMOTO, Yuichi NAKAMURA, Masatake DANTSUJI, “Browsing of concentration by capturing learner’s behaviors in e-learning,” ITE Annual Convention 2014, 2 Sept. 2014
2. Siyang Yu, Kazuaki Kondo, Hiromasa Yoshimoto, Yuichi Nakamura, Takayuki Nakajima, Masatake Dantsuji, “Automatic Learning State Estimation in Actual E-learning,” INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND ENVIRONMENTAL ENGINEERING (CSEE 2015), 1583-1590
3. Siyang YU, Kazuaki KONDO, Yuichi NAKAMURA, Takayuki NAKA-

JIMA, Hiroaki NANJO, Masatake DANTSUJI, “Lip-protrusion Visualization for Language Learning Support,” IEICE Technical Report, MVE2017

Bibliography

[1] S. Yu, K. Kondo, Y. Nakamura, T. Nakajima, and M. Dantsuji, “Learning state recognition in self-paced e-learning,” IEICE TRANSACTIONS on Information and Systems, vol.100, no.2, pp.340–349, 2017.

[2] S. YU, K. KONDO, Y. NAKAMURA, T. NAKAJIMA, H. NANJO, and M. DANTSUJI, “Visual emphasis of lip protrusion for pronunciation learning,” IEICE TRANSACTIONS on Information and Systems, vol.102, no.1, pp.156–164, 2019.

[3] A. Sangrà, D. Vlachopoulos, and N. Cabrera, “Building an inclusive definition of e-learning: An approach to the conceptual framework,” The International Review of Research in Open and Distributed Learning, vol.13, no.2, pp.145–159, 2012.

[4] S. Guri-Rosenblit, “ ` distance education ¨ and ` e-learning ¨ : Not the same thing,” Higher education, vol.49, no.4, pp.467–493, 2005.

[5] A. Koohang and K. Harman, “Open source: A metaphor for e-learning.,” Informing Science, vol.8, 2005.

[6] S. Bermejo, “Cooperative electronic learning in virtual laboratories through forums,” IEEE Transactions on Education, vol.48, no.1, pp.140–149, 2005.

[7] E. Commission, “Communication from the commission to the council and the european parliament: the e-learning action plan: designing tomorrow’s education,” 2001.

[8] L. Amaral and D. Leal, “From classroom teaching to e-learning: the way for a strong definition,” 2006.

[9] R.C. Richey, K. Silber, and D. Ely, “Reflections on the 2008 aect definitions of the field,” TechTrends, vol.52, no.1, pp.24–25, 2008.

[10] C. Central, “A product at every price: A review of mooc stats and trends in

2017," 2018.

[11] M.G. Moore, "Editorial: Three types of interaction," pp.1–7, 1989.

[12] T. Anderson, "Modes of interaction in distance education: Recent developments and research questions," Handbook of distance education, pp.129–144, 2003.

[13] D.H. Schunk, Learning theories an educational perspective sixth edition, Pearson, 2012.

[14] L.S. Vygotsky, Mind in society: The development of higher psychological processes, Harvard university press, 1978.

[15] L.C.P.W.G. of the American Psychological Association's Board of Educational Affairs, "Learner-centered psychological principles: A framework for school reform and redesign," 1997.

[16] R.W. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, and C. Strohecker, "Affective learning ! a manifesto," BT technology journal, vol.22, no.4, pp.253–269, 2004.

[17] R.W. Picard *et al.*, "Affective computing," 1995.

[18] N.J. Butko, G. Theocharous, M. Philipose, and J.R. Movellan, "Automated facial affect analysis for one-on-one tutoring applications," Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp.382–387, IEEE, 2011.

[19] M.B. Ammar, M. Neji, A.M. Alimi, and G. Gouardères, "The affective tutoring system," Expert Systems with Applications, vol.37, no.4, pp.3013–3023, 2010.

[20] J. Whitehill, M. Bartlett, and J. Movellan, "Automatic facial expression recognition for intelligent tutoring systems," Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pp.1–6, IEEE, 2008.

[21] K. Zakharov, A. Mitrovic, and L. Johnston, "Towards emotionally-intelligent pedagogical agents," International Conference on Intelligent Tutoring Systems, pp.19–28, Springer, 2008.

[22] S.K. D ¨ mello, S.D. Craig, A. Witherspoon, B. Mcdaniel, and A. Graesser, "Automatic detection of learner ¨ s affect from conversational cues," User modeling and user-adapted interaction, vol.18, no.1-2, pp.45–80, 2008.

[23] D. Litman and K. Forbes, “Recognizing emotions from student speech in tutoring dialogues,” Automatic Speech Recognition and Understanding, 2003. ASRU’03. 2003 IEEE Workshop on, pp.25–30, IEEE, 2003.

[24] L. Shen, M. Wang, and R. Shen, “Affective e-learning: Using” emotional” data to improve learning in pervasive learning environment.,” Journal of Educational Technology & Society, vol.12, no.2, 2009.

[25] C.H. Yang, “Fuzzy fusion for attending and responding assessment system of affective teaching goals in distance learning,” Expert Systems with Applications, vol.39, no.3, pp.2501–2508, 2012.

[26] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper, and R. Picard, “Affect-aware tutors: recognising and responding to student affect,” International Journal of Learning Technology, vol.4, no.3-4, pp.129–164, 2009.

[27] J.A. Fredricks, P.C. Blumenfeld, and A.H. Paris, “School engagement: Potential of the concept, state of the evidence,” Review of educational research, vol.74, no.1, pp.59–109, 2004.

[28] D.J. Shernoff, “The nature of engagement in schools,” in Optimal Learning Environments to Promote Student Engagement, pp.47–75, Springer, 2013.

[29] M. Csikszentmihalyi, Flow: The Psychology of Optimal Experience, 01 1990.

[30] U. Schiefele, A. Krapp, and A. Winteler, “Interest as a predictor of academic achievement: A meta-analysis of research.,” in The role of interest in learning and development, pp.183–212, Lawrence Erlbaum Associates, Inc, 1992.

[31] L.L. Shirey, “Importance, interest, and selective attention,” in The role of interest in learning and development, pp.281–296, Lawrence Erlbaum Associates Hillsdale, NJ, 1992.

[32] E.L. Deci, “The relation of interest to the motivation of behavior: A self-determination theory perspective.,” in The role of interest in learning and development, pp.43–70, Lawrence Erlbaum Associates, Inc, 1992.

[33] P.J. Silvia, “Self-efficacy and interest: Experimental studies of optimal incompetence,” Journal of Vocational Behavior, vol.62, no.2, pp.237–249, 2003.

[34] S. D’Mello, R.W. Picard, and A. Graesser, “Toward an affect-sensitive autotutor,” IEEE Intelligent Systems, vol.22, no.4, 2007.

[35] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments," Proceedings of the Annual Meeting of the Cognitive Science Society, 2007.

[36] M. Wosnitza and S. Volet, "Origin, direction and impact of emotions in social online learning," Learning and instruction, vol.15, no.5, pp.449–464, 2005.

[37] N. Smolyanskiy, C. Huitema, L. Liang, and S.E. Anderson, "Real-time 3d face tracking based on active appearance model constrained by depth data," Image and Vision Computing, vol.32, no.11, pp.860–869, 2014.

[38] C.C. Chang and C.J. Lin, "Libsvm: a library for support vector machines," ACM transactions on intelligent systems and technology (TIST), vol.2, no.3, p.27, 2011.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," Journal of machine learning research, vol.12, no.Oct, pp.2825–2830, 2011.

[40] R. Blake, "Technology and the four skills," Language Learning and Technology, vol.20, no.2, pp.129–142, 2016.

[41] P. Hubbard, Computer Assisted Language Learning: Critical Concepts in Linguistics, Routledge, 2009.

[42] M. Levy and G. Stockwell, CALL dimensions: Options and issues in computer-assisted language learning, Lawrence Erlbaum Associates, 2006.

[43] M. Celce-Murcia, Teaching English as a second or foreign language 3rd edition, Heinle & Heinle, 2001.

[44] L. Ménard, C. Toupin, S.R. Baum, S. Drouin, J. Aubin, and M. Tiede, "Acoustic and articulatory analysis of french vowels produced by congenitally blind adults and sighted adults," The Journal of the Acoustical Society of America, vol.134, no.4, pp.2975–2987, 2013.

[45] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," Nature, vol.264, no.5588, pp.746–748, 1976.

[46] R. Duan, J. Zhang, W. Cao, and Y. Xie, "A preliminary study on asr-based detection of chinese mispronunciation by japanese learners," Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[47] S.H. Hew and M. Ohki, “Effect of animated graphic annotations and immediate visual feedback in aiding japanese pronunciation learning: A comparative study,” CALICO journal, vol.21, no.2, pp.397–419, 2004.

[48] Y. Tsubota, M. Dantsuji, and T. Kawahara, “An english pronunciation learning system for japanese students based on diagnosis of critical pronunciation errors,” ReCALL, vol.16, no.1, pp.173–188, 2004.

[49] L. Yang, L. Zhang, H. Dong, A. Alelaiwi, and A. El Saddik, “Evaluating and improving the depth accuracy of kinect for windows v2,” IEEE Sensors Journal, vol.15, no.8, pp.4275–4285, 2015.

[50] E. Lachat, H. Macher, M. Mittet, T. Landes, and P. Grussenmeyer, “First experiences with kinect v2 sensor for close range 3d modelling,” The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol.40, no.5, pp.93–100, 2015.

[51] O. Wasenmüller and D. Stricker, “Comparison of kinect v1 and v2 depth images in terms of accuracy and precision,” Asian Conference on Computer Vision, pp.34–45, Springer, 2016.

[52] R.C. Gonzalez and R.E. Woods, Digital Image Processing second edition, Prentice Hall, 2002.

[53] D.T. Campbell and J.C. Stanley, “Experimental and quasi-experimental designs for research,” Handbook of research on teaching, pp.171–246, 1963.

[54] D.B. Rubin, “Inference and missing data,” Biometrika, vol.63, no.3, pp.581–592, 1976.

[55] R. Little and D. Rubin, Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics, Wiley, 2014.

[56] K.A. Hallgren, “Computing inter-rater reliability for observational data: an overview and tutorial,” Tutorials in quantitative methods for psychology, vol.8, no.1, p.23, 2012.

[57] R.J. Light, “Measures of response agreement for qualitative data: some generalizations and alternatives.,” Psychological bulletin, vol.76, no.5, p.365, 1971.

[58] J.R. Landis and G.G. Koch, “The measurement of observer agreement for categorical data,” biometrics, pp.159–174, 1977.

Appendix Calculation in Color Space

Calculation for HSI values:

Converting RGB to HSI [52], we get the following:

$$H = \begin{cases} 0 & B \leq G \\ 360 - 0 & B > G \end{cases}$$

With

$$0 = \cos^{-1} \left\{ \frac{1/2\left[(R-G)+(R-B)\right]}{\left[(R-G)^2+(R-B)(G-B)\right]^{1/2}} \right\}$$

$$S = 1 - \frac{3}{R+G+B}\left[min\left(R,G,B\right)\right]$$

$$I = \frac{1}{3}\left(R+G+B\right)$$

Converting HSI to RGB [52], we get the following:

Red-Green sector $0 \leq H < 2\pi/3$:

$$B = I\left(1-S\right)$$

$$R = I\left[1 + \frac{S\cos H}{\cos\left(\pi/3 - H\right)}\right] \tag{5.1}$$

$$G = 3I - \left(R+B\right)$$

Green-Blue sector $2\pi/3 \leq H < 4\pi/3$:

$$R = I\left(1-S\right)$$

$$G = I\left[1 + \frac{S\cos\left(H - 2\pi/3\right)}{\cos\left[\pi/3 - \left(H - 2\pi/3\right)\right]}\right] \tag{5.2}$$

$$B = 3I - (R + G)$$

Blue-Red sector $4\pi/3 \leq H \leq 2\pi$:

$$G = I\,(1 - S)$$

$$B = I\left[1 + \frac{S\cos\left(H - 4\pi/3\right)}{\cos\left[\pi/3 - \left(H - 4\pi/3\right)\right]}\right] \tag{5.3}$$

$$R = 3I - (G + B)$$

Hue determination according to distance segmentations gives the following:

Lip subarea:

$\Delta d \leq 20$ (upper lip subarea):

$$\tilde{H} = \begin{cases} 0 & \Delta d < 0 \\ \Delta d & 0 \leq \Delta d \leq 20 \end{cases}$$

$\Delta d \leq 27$ (lower lip subarea):

$$\tilde{H} = \begin{cases} 0 & \Delta d < 0 \\ \Delta d & 0 \leq \Delta d \leq 27 \end{cases}$$

$20 < \Delta d \leq 60$ (upper lip subarea):

$$\tilde{H} = (\Delta d - 20) + 120$$

$27 < \Delta d \leq 60$ (lower lip subarea):

$$\tilde{H} = (\Delta d - 27) + 120$$

Non-lip subarea:

$\Delta d \leq 60$:

$$\tilde{H} = \begin{cases} 120 & \Delta d < 0 \\ \Delta d + 120 & 0 \leq \Delta d \leq 60 \end{cases}$$

Lip and non-lip subarea:

$60 < \Delta d \leq 100$:

$$\tilde{H} = (\Delta d - 60) + 240$$

$100 < \Delta d$: Original value that is converted from RGB.

The maximum allowed saturation is calculated on the basis of the variation of formula (5.1), (5.2) and (5.3). For example, when the hue is around red, the maximum allowed saturation calculation is as follows:

$$\tilde{S}_{max-red} = \frac{\left(\tilde{R}_{max}/\tilde{I} - 1\right)\cos\left(\pi/3 - \tilde{H}\right)}{\cos\tilde{H}}$$

where $\tilde{R}_{max}$ is set to the maximum allowed value, i.e. 1.0.