For Review Only

Education of TCM Tongue Diagnosis by Automatic Tongue Diagnosis System

Journal:	*Chinese Journal of Integrative Medicine*
Manuscript ID:	CJIM-2013-0434.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	14-Sep-2013
Complete List of Authors:	Lo, Lun-chien Chen, Yung-Fu Chiang, John Y; National Sun Yat-sen University, Department of Computer Science and Engineering Cheng, Tsung-Lin Damdinsuren, Natsagdorj
Keywords:	Tongue Diagnosis, Automatic tongue diagnosis system (ATDS), Intra-agreement, inter-agreement
Speciality Area:	Tongue Diagnosis, Traditional Chinese Medicine(TCM)

SCHOLARONE™
Manuscripts

Education of TCM Tongue Diagnosis by Automatic Tongue Diagnosis System

Lun-chien Lo[1,2], Yung-Fu Chen[3,4,5], John Y. Chiang[6,*], Tsung-Lin Cheng[2], and Natsagdorj Damdinsuren[7]

[1]Department of Traditional Chinese Medicine, Changhua Christian Hospital, Changhua, Taiwan

[2]Graduate Institute of Statistics and Information Science, National Changhua University of Education, Changhua, Taiwan

[3]Department of Healthcare Administration, Central Taiwan University of Science and Technology, 40601 Taichung, Taiwan

[4]Institute of Biomedical Engineering and Material Science, Central Taiwan University of Science and Technology, 40601 Taichung, Taiwan

[5]Department of Health Services Administration, China Medical University, 40402 Taichung, Taiwan

[6]Department of Computer Science & Engineering, National Sun Yat-sen University, 80424 Kaohsiung, Taiwan

[7]"Manba Datsan,"Clinic and Training Centre for Mongolian Traditional Medicine. 2nd Khoroo, Bayanzurkh district, Ulaanbaatar, Mongolia

For Review Only

*Corresponding author: John Y. Chiang, Ph.D.
Department of Computer Science & Engineering,
National Sun Yat-sen University, Kaohsiung, Taiwan
Tel.: +886-934151515; E-mail addresses: chiang@cse.nsysu.edu.tw

Abstract

Traditionally, tongue diagnosis depends solely on personal knowledge and experience of the practitioner. Lack of scientific verification and manifestation of low agreement in diagnosis are critical issues frequently debated and challenged by healthcare providers. The root for causing this skepticism can be traced back to medical school years when TCM doctors receiving formal training. With no other options available, the teaching of tongue diagnosis was resorted to traditional classroom lecture and textbook with a limited number of illustrations. The exemplary illustrations, usually acquired under poorly controlled lighting environment and printed with low color quality, were misleading at least. This situation is expected to change with the development and recent reliability validation of ATDS [16]. If the knowledge base embedded in ATDS, containing a huge number of high-quality clinical images along with the corresponding tongue features extracted, can be demonstrated as a feasible alternative to the traditional teaching methodology of tongue diagnosis, the major cause for the diagnostic discrepancy can be eliminated. In light of this observation, this paper investigates the efficacy of applying ATDS to the training of novice TCM doctors in the tongue diagnosis. Only through the improvement of diagnostic consistency, can the skepticism surrounding TCM be lifted.

A two-round experiment composed of pre- and after-training tests was performed. Two groups of TCM doctors were formed by experienced doctors (ED) and novice doctors (ND). The intra- and inter-observer agreements for ED and ND groups, and the inter-observer agreements between ATDS and ED, and ATDS and ND were derived first. After a one-month training of the novice doctors, the intra- and inter-observer agreements of the ND group, and the inter-observer agreement between ATDS and ND were obtained. The kappa value of agreement data for ND after training is significantly higher than ND before training with a level of $p<0.05$, indicating an obvious improvement of intra-observer agreement for ND after training with ATDS. The inter-observer agreements for ED group, and ND group after training are significantly higher (Student's *t*-test, $p<0.05$) than the ND group before training. In contrast, no significance ($p>0.05$) is observed between the ED group and the ND group after training, indicating that physicians in the ND group achieve an inter-observer agreement to the same level as that of experienced doctors after one-month training. This finding validates the utilization of ATDS as a mean of education in forging consensus during the learning process of tongue diagnosis.

Keywords: tongue diagnosis, Automatic tongue diagnosis system (ATDS), Intra-agreement, inter-agreement.

For Review Only

1. Introduction

Complementary and alternative medicine (CAM) is a medical practice which is not considered as a formal branch of western medicine. With regard to CAM, complementary medicine is conducted in conjunction with the practice of western medicine, while alternative medicine is taken as its substitute [1]. CAM is becoming popular in high-income countries, such as US and some European countries. It was estimated that approximate 38% and 12 % of adults and children, respectively, have been using CAM in the US, and 80% of the people in low-income countries have been using herbal medicine [2]. Hence, the medical practitioners and students express their great favors and supports in providing CAM curricula for nursing and medical students [2-4]. Only when the medical practitioners have a sound knowledge of CAM, the side effect and adverse drug event can be prevented in the clinical setting that integrates CAM practice with modern medicine.

1-1 Traditional Chinese Medicine

Among the CAM practices, traditional Chinese medicine (TCM) is the most popular medical practice. Nicolao et al. (2010) reported that the TCM practices, such as acupuncture and phytotherapy, are considered as the most popular disciplines requested by both medical experts and students in Switzerland [4]. It has received wider acceptance from western medicine in recent years. However, lack of scientific verification and manifestation of low agreement in diagnosis and treatment are critical issues frequently debated and challenged by healthcare providers. Its diagnosis is generally based on four standard but not validated approaches, i.e., observation, smelling/listening, inquiry, and palpation. Among them, inquiry is the most frequently used method, which is followed by palpation of radial pulse and observation of tongue [5]. The low agreement or inconsistency of diagnosis in TCM practices is believed to be caused by subjective assessment of symptoms appeared in tongues, characterized by pulses, and learned from patients' complaints.

Previous studies have conducted on the issue of consistency of TCM diagnosis [6-16] as well as herbal prescription or treatment [10, 11, 13], indicating that inter- and intra-observer agreements are low. The inconsistency of subjective diagnosis and treatment can be improved by the development of validated instruments, such as standard questionnaire for inquiries [5] and manual for guiding treatments [17]. Recently, intra- and inter-observer agreements of the automatic tongue diagnosis system (ATDS) and TCM practitioners have been conducted in our laboratory [16]. The results demonstrate that the ATDS is very consistent even in the face of variations of environmental lighting and extruding tongue with an intra-observer

For Review Only

agreement significantly higher than that of the TCM doctors, while the inter-observer agreements between the ATDS and a group of TCM doctors and among the TCM doctors are both moderate [16]. ATDS serves not only as clinical equipment in providing doctors with consistent tongue features of patients, but also as a feasible teaching and evaluation means for students learning tongue diagnosis. In this study, the ATDS developed by our team was utilized to train novice TCM doctors, with assistance and guidance of a senior physician, to improve skills and consistency in tongue diagnosis. With the vast amount of tongue images collected, accompanied by the corresponding features extracted automatically, learning through the knowledge base of ATDS is expected to be more effective than that guided by senior physicians only.

1-2 Tongue Diagnosis

Tongue diagnosis plays an important role in TCM [18-24]. It is widely believed that the tongue is connected to the internal organs through meridians; thus the conditions of organs, qi, blood, and body fluids, as well as the degree and progression of disease are all reflected on the tongue [25, 26]. Hence, organ conditions, properties and variations of pathogens can be revealed through observation of tongue. For example, changes in the tongue property primarily reflect organ status and the flow of qi and blood; variations in tongue fur can be employed to determine the impact of exogenous pathogenic factors and the flow of stomach qi. In clinical practice of TCM, practitioners observe the characteristics of tongue, such as the color, shape and the amount of saliva before deducing the primary ailment of a patient. However, observation diagnosis is often biased by subjective judgment, originating from personal knowledge, experience, thinking patterns, diagnostic skills, and color perception or interpretation. There are no precise or quantifiable standards existing. Different practitioners may pass varying judgments on the same tongue, while a practitioner may even reach different diagnoses on the identical tongue if examined at different time. For example, it was reported that the mean intra-observer agreement reached only 61% and the inter-observer agreement was as low as 18.2% for 30 practitioners observing 10 tongue images [15]. Such inconsistency inevitably leads many people to be skeptical of TCM, which raised our motivation to develop the ATDS [22]. It was demonstrated that ATDS can establish reliable diagnoses with higher intra-observer agreement than the TCM practitioners [16]. To gain the trust of practitioners and patients and be employed in clinical inspections, the reliability of ATDS in the face of environmental lighting variations and differences, e.g., orientation, forces applied in extruding tongue, etc., for tongue images taken at different time has been established [16].

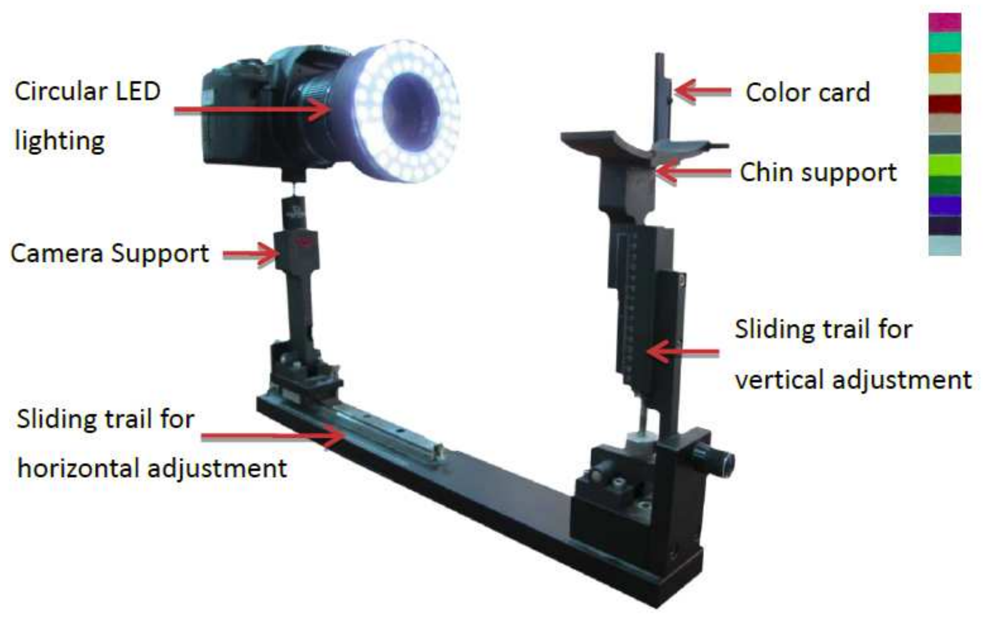
For Review Only

Zhang et al. reported that training through case discussion significantly increased consensus on diagnosis and prescription of rheumatoid arthritis [28]. Schnyer and Allen promoted manualization of protocols used in clinical trials to facilitate the systematic delivery of replicable and standardized, yet individually-tailored treatments of acupuncture [17]. In light of these reports, it is promising that ATDS will serve as an effective tool in training the novice TCM doctors by providing them with standardized procedures as well as objective, reliable and quantified data. The goal of this study is to investigate the efficacy of ATDS in training the novice TCM doctors to improve their skills and consistency in tongue diagnosis by comparing the intra-observer agreements regarding tongue features observed before and after training.

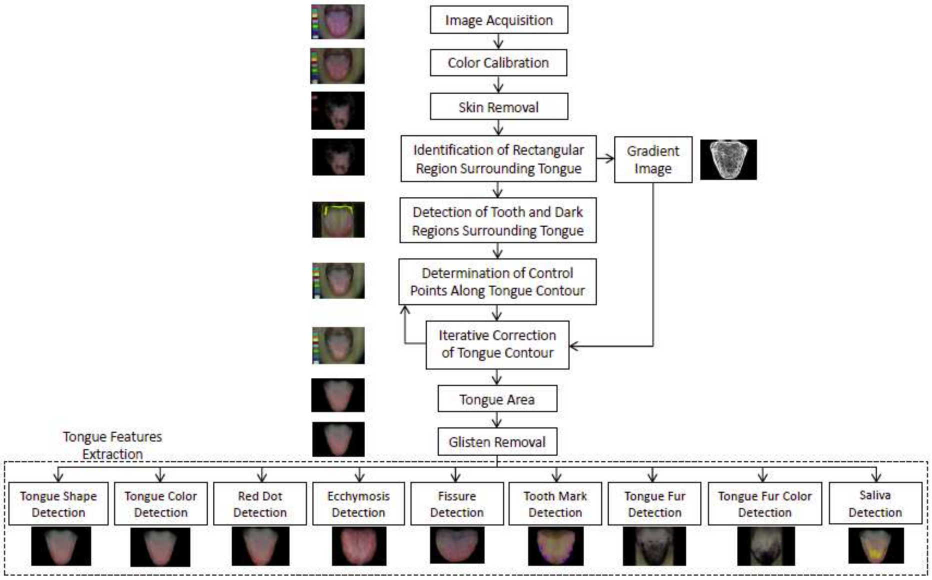
2. Automatic Tongue Diagnostic System

As shown in Fig. 1(a) (figure illustrating components of the ATDS), the ATDS was developed to capture tongue images and extract features reliably to assist the diagnosis of TCM practitioners [16]. Fig. 1(b) demonstrates the steps in the three major functions, i.e., image capturing and color calibration, tongues area segmentation, and tongue feature extraction, included in the ATDS [16, 27].

For Rev



(a)

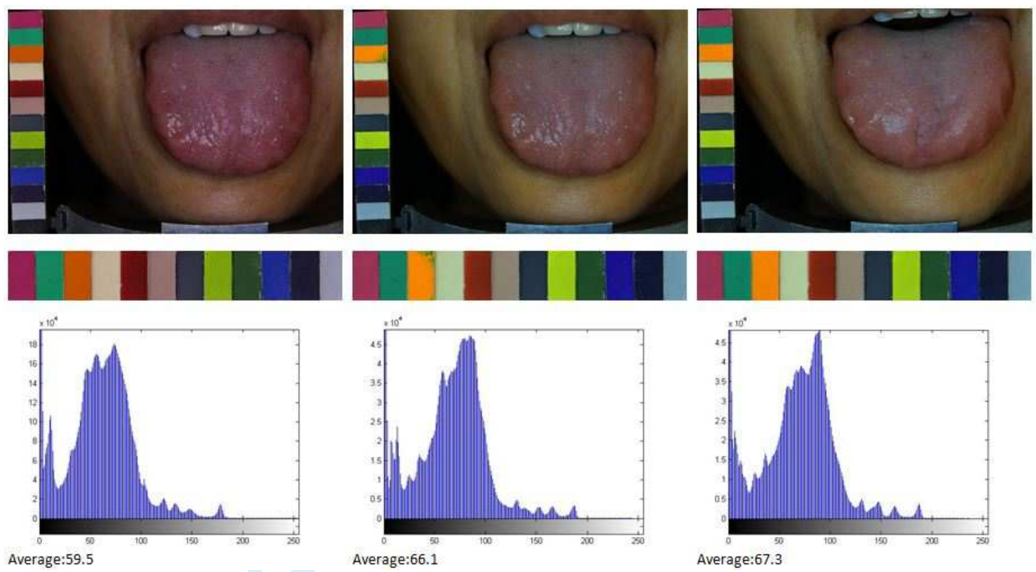


(b)

Figure 1: Illustration of (a) the components of the ATDS and (b) the processing steps of ATDS analysis

Variations in background lighting may change the color and brightness of the acquired images, greatly affecting consistency and stability of the extracted tongue features. The consistency and stability of tongue images captured and features extracted are achieved by calibrating brightness and color to compensate variations in intensity and color temperature of the light source and imaging hardware. The ATDS developed can automatically correct lighting and color deviation caused by the change of background lighting with a color bar attached in the ATDS. The color bar placed beside the patient is used for color calibration to make sure the image quality is consistent even taken at different circumstances. Figures 2 (a) and (b) display the images taken at T_1 before and after color calibration, respectively, whereas (c) demonstrates the image taken at T_2 after calibration. The second and third rows show the color bars clipped from the tongue images and their corresponding histograms. ATDS automatically compensates the color deviation of the original image (Figure 2 (a) with a mean gray level 59.5) to allow colors in images taken at different time intervals consistent with each other (Figures 2 (b) and (c), with a mean gray level 66.1 and 67.3, respectively).

For Review Only



(a) (b) (c)

Figure 2: Calibration of image color using the color bar accompanied with the ATDS to make image quality consistent for images taken at difference circumstances.

Tongue images are analyzed by first isolating the tongue region within an image to eliminate irrelevant lower facial portions and background surrounding the tongue, thereby facilitating feature identification and extraction; and then extracting the tongue features by employing criteria such as the aspect ratio, color composition, location, shape, and color distribution of the tongue, as well as the quantity of neighboring pixels. Features including tongue color, tongue fissure, fur color, fur thickness, ecchymosis, tooth mark, red dot, saliva, and tongue shape are extracted to further generate detailed information regarding length, area, moisture, and number of fissures, marks, and dots to be employed in tongue diagnosis. As depicted in Figure 3, nine primary tongue features, including tongue color (slightly white, slightly red, red, dark red, dark purple), fur color (white, yellow, dye), fur thickness (none, thin, thick), saliva (none, little, normal, excessive), tongue shape (thin and small, moderate, fat and large), tongue fissure, red dot, ecchymosis, and tooth marks (the last four features are divided into categories of none, mild, moderate, and severe), are selected for tongue diagnosis [16].

For Review Only

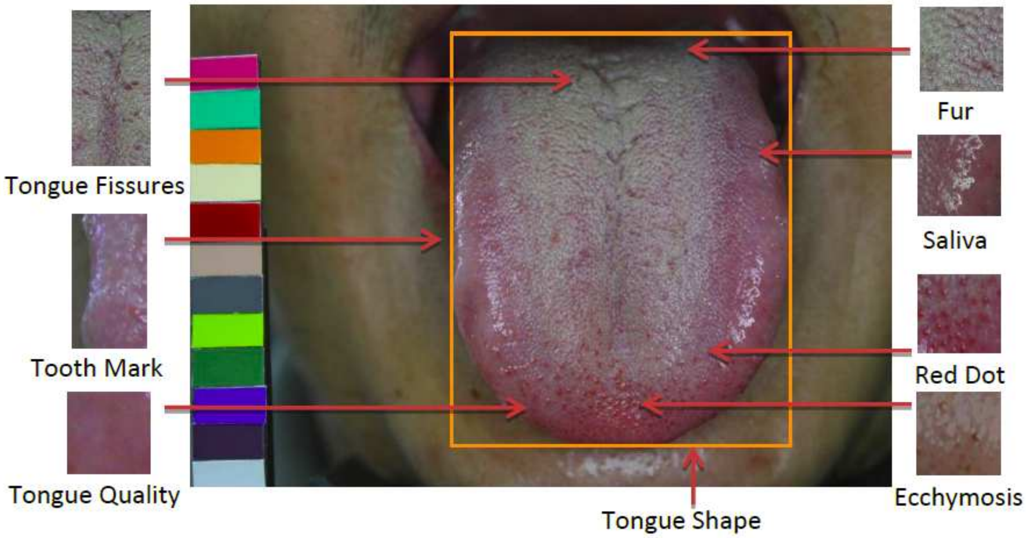


Figure 3. Illustration of nine tongue features extracted by ATDS for tongue diagnosis.

2.1 Subjects

Two groups of TCM doctors, including 12 experienced doctors (ED) and 14 novice doctors (ND), from the Chinese medicine department at Changhua Christian Hospital (CCH) in Taiwan, were invited to participate in this study. The experienced doctors have been in clinical practice for 3 to 15 years with a mean of 5.5 years, while the novice doctors are in their first year of resident training. All of them were educated in Taiwan, each holding a BS degree or higher. The doctors in the ED group were asked to attend regular weekly meetings in the past two years to examine over 1000 tongue images collected through the outpatients of CCH. The images with the consensuses reached by this group of doctors were employed as the training data during the development of the ATDS [16], and for the training of the novice doctors in the current agreement study.

The physicians in the ED and ND groups were later asked to make tongue diagnosis by observing the tongue images acquired from newly recruited patients for the agreement tests. The experiments were conducted in two rounds separated by one month's training. The physicians in both the ED and ND groups participated in the first round observing the same set of 20 newly acquired tongue images, while only the physicians in the ND group took part in the next round following one month of training, assessing the second set of tongue images captured from another 20 different patients. The training protocol is described as follows:

(1) Randomly select 10 images from the tongue image database which contains 1000 images.

(2) Extract the corresponding tongue features from the images selected with the

ATDS.

(3) Ask the physicians in the ND group to observe the selected images with the assistance of features extracted by ATDS.

(4) Physicians in the ND group discuss their observation, and reach agreements in diagnosis.

(5) Randomly select 10 alternative images from the database and repeat the experimental procedure from step (2) to (4) for 5 times.

2.2 Assessment of tongue features

As shown in Figure 4, nine primary features (Figure 3) contained in the questionnaire were administrated automatically by a computer program. The same set of tongue images shown to each of the participating physicians were displayed in a random order.

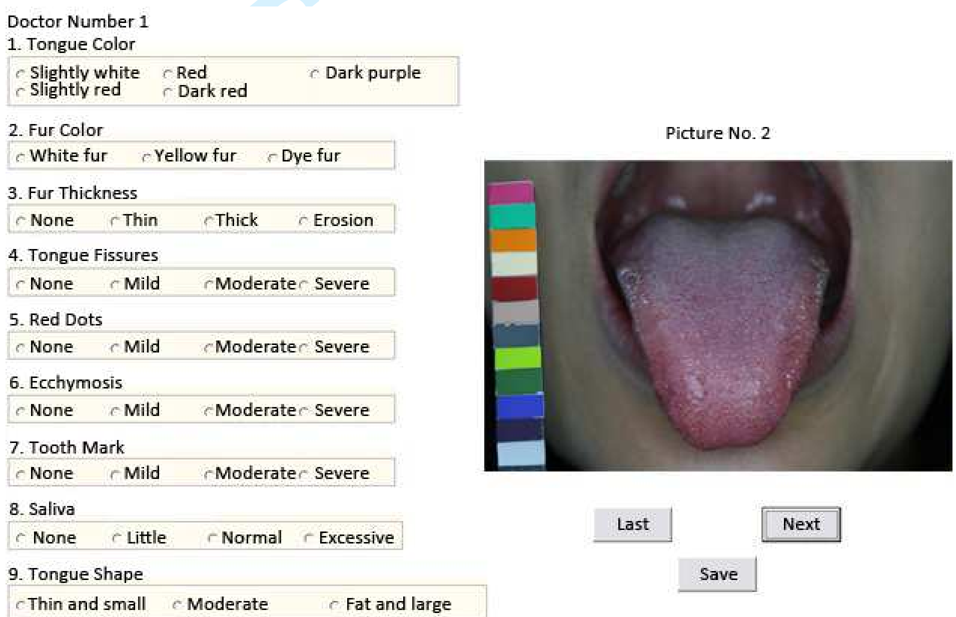


Figure 4: A snapshot of the computerized tongue diagnosis questionnaire

2.3 Experiment Procedure

2.3.1. Image Samples

The tongue images of 40 patients were captured by ATDS first. These 40 tongue images collected were separated into two image sets, P_1 and P_2, each containing 20 images. The image set, P_1 is provided in the first round of testing, while P_2 the second, as shown in Figure 5. Each round contains two stages served with the same set of test images, yet in different display order. In the first round, the image set P_1 was used to derive the intra- and inter-observer agreements for ED and ND groups, and the inter-observer agreements between ATDS and ED, and ATDS and ND, respectively. After a one-month training of the novice doctors, the second round initiated. Another set of tongue images, P_2, was employed to obtain the intra- and inter-observer agreements of the ND group, and the inter-observer agreement between ATDS and ND. The agreement data of novice doctors obtained before and after training were compared to assess the efficacy of ATDS as a mean of education in learning tongue diagnosis.

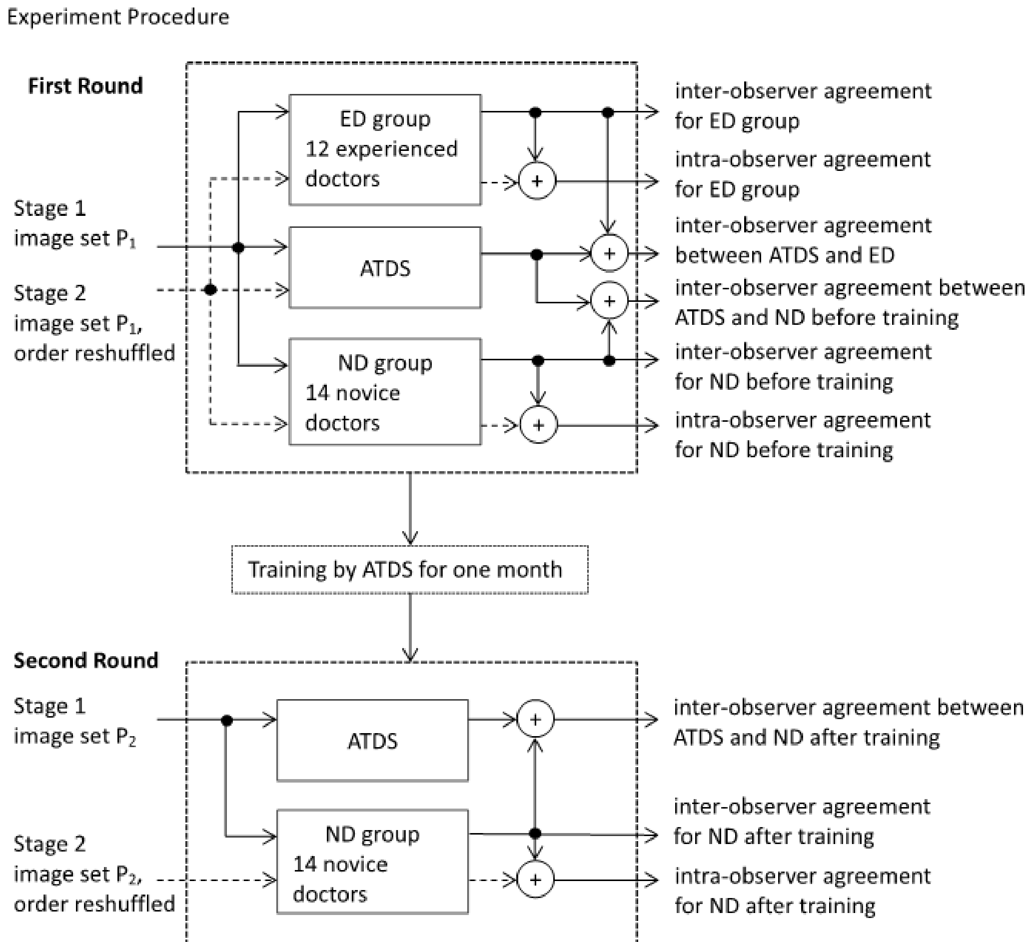


Figure 5: The experimental flow for deriving relevant intra- and inter-observer agreements.

2.3.2. Training

The 1000 tongue images captured with features extracted by ATDS were used to train the novice TCM doctors during the week days, two 2-hour sessions every week, for one month. In addition to inspecting the features extracted by the ATDS for diagnosis, the novice doctors were also encouraged to make discussion with each other to reinforce consensus regarding the extracted features of individual tongue images. A senior TCM physician, Dr. L. C. Lo, was also invited to attend the meeting to resolving cases which an agreement could not be reached after exhaustive discussions.

2.3.3. Agreement Test

First Round: Before Training Session

a. Intra-observer agreements for the ED and ND groups

The intra-observer agreements were conducted for both the ED and ND groups. In the first stage, each of the participating TCM doctors observed the tongue image set P_1 and filled out the questionnaire. After a week, the TCM doctors were asked to review the same set of images P_1 presented with a different order, and fill out the questionnaire again during the second stage. By comparing the answers collected in the two-stage questionnaires, the intra-observer agreements of the ED and ND groups were evaluated. The intra-observer agreement reflects the consistency of observations performed by a group of doctors in two different stages separated by a timespan of one week.

b. Inter-observer agreements for the ED and ND groups

According to the diagnostic results gathered during the first stage of tongue diagnosis through the assessment of image set P_1, the inter-observer agreement among the members in the ED group, and similarly among those in the ND group can be derived. The inter-observe agreement among doctors belonging to a certain group indicates the degree of consensus reached within that specific group during a session of observation.

c. Inter-observer agreements between ATDS and ED, and between ATDS and ND

According to the diagnostic results of the image set P_1 determined by ATDS, the ED and ND groups during the first stage, the strength of inter-observer agreement between ATDS and ED, and between ATDS and ND can be obtained, respectively. Since ATDS has been validated to possess high degree of agreement [16], the inter-observer agreement obtained serves as an indicator for the degree of coincidence between the diagnoses reached by ATDS and doctors.

Second Round: After Training Session

a. Intra- and inter-observer agreements for the ND group after training

Following the training protocol for a month, the ND group was presented with the second tongue image set P_2 in a similar two-stage test, each separated by a week. Based on the assessments gathered from the first stage, the inter-observer agreements of ND group for nine tongue features can be calculated. According to the observation data compiled in two stages, the intra-observer agreement of novice doctors after

For Review Only

training can be derived. The intra- and inter-observer agreements of ND group before and after the training are further compared to evaluate the efficacy of ATDS as an educational tool in facilitating the learning of tongue diagnosis.

b. Inter-observer agreement between ATDS and ND after training

According to the diagnostic results of the image set P_2 determined by the ATDS, and novice doctors during the first stage, the strength of inter-agreement between the ATDS and ND group after training was assessed. Also, the inter-observer agreements of ND group before and after the training were compared to justify the feasibility of applying ATDS to facilitate the learning of tongue diagnosis.

2.4 Data Analysis

For tongue and fur color, the intra-agreement and inter-agreement of ATDS, ED and ND were represented by Cohen's kappa coefficient [29]. Cohen's kappa treats all disagreements equally. The weighted kappa incorporates the magnitude of each disagreement and provides partial credits for disagreements when complete agreement is not reached. In fur thickness, tongue fissure, red dot, ecchymosis, tooth mark, saliva, and tongue shape, agreement is expressed by a quadratic-weighted kappa coefficient [30]. The inter-agreement among TCM doctors is indicated with Fleiss' kappa coefficient [31], which is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. While Cohen's kappa is only applicable to assessing the agreement between two raters, the measure of agreement proposed by Landis and Koch [32], on the other hand, is divided into six levels according to the kappa value: poor (less than 0.00), slight (0.00 to 0.20), fair (0.21 to 0.40), moderate (0.41 to 0.60), substantial (0.61 to 0.80), and almost perfect (0.81 to 1.00).

In this study, the levels of intra-observer agreements for the ED and ND groups were measured using Cohen's kappa coefficients. The intra-observer agreements of the TCM physicians in ED and ND groups were calculated, and then the mean values and standard deviations (STDs) of individual tongue features were compared. Inter-observer agreements between ATDS and ED, and between ATDS and ND were measured using Fleiss' kappa coefficients. Means and STDs of the 9 tongue features and the mean of 9 tongue features were compared with two-sample *t*-test. The significant level is set as $p<0.05$.

3 Results

Table 1 shows an example of tongue feature, i.e. tongue color, evaluated by 12 TCM doctors in ED group for 20 patients. As shown in Figure 4, the nominal values

For Review Only

corresponds to different characteristics of tongue color with 1, 2, 3, 4, and 5 representing slightly white, red, dark purple, slightly red, and dark red, respectively. As listed in this table, the observations of tongue color for patients 5, 8, 10, 11, 12, and 16 by ED group are highly consistent with 100% consensus, while patients 2, 9, and 19 demonstrates greater difference in opinion. The root for causing this discrepancy can be traced back to medical school years when TCM doctors receiving formal training. Traditional teaching methodology, namely, classroom lecture and textbook with a limited number of illustrations, were usually the major means of acquiring knowledge relevant to tongue diagnosis. The exemplary illustrations were usually captured under poorly controlled lighting environment and printed with low color quality. No other options existed until the development and recent reliability validation of ATDS [16]. As a clinical device, ATDS can provide tongue diagnosis results with high degree of consistency. If the knowledge base embedded in ATDS, containing an ever-increasing collection of high-quality clinical images along with the corresponding tongue features extracted, can be successfully applied as a means of education in tongue diagnosis. The major cause for the inconsistency due to subjective judgment can be eliminated to reach consensus. In light of this observation, this paper investigates the efficacy of applying ATDS to the training of TCM doctors in the tongue diagnosis. Only through the improvement of diagnostic consistency, can the skepticism surrounding TCM be lifted.

Table 1. An example of tongue features, i.e. tongue color, evaluated by 12 TCM doctors in ED group for 20 patients

Doc. / Pat.	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
P1	4	3	3	3	4	3	4	3	3	4	3	3
P2	5	1	1	5	5	2	5	5	5	2	1	1
P3	2	2	2	2	2	2	2	2	2	2	4	3
P4	5	5	3	5	5	5	5	5	3	5	5	5
P5	2	2	2	2	2	2	2	2	2	2	2	2
P6	4	4	4	4	4	4	5	5	4	4	4	4
P7	3	3	2	3	3	3	3	3	3	3	2	2
P8	2	2	2	2	2	2	2	2	2	2	2	2
P9	5	3	3	4	4	3	5	3	5	4	3	3
P10	2	2	2	2	2	2	2	2	2	2	2	2
P11	2	2	2	2	2	2	2	2	2	2	2	2
P12	2	2	2	2	2	2	2	2	2	2	2	2
P13	2	2	1	2	2	2	2	1	2	2	2	2
P14	2	2	2	3	3	2	2	2	2	2	3	3
P15	3	2	2	2	2	2	2	2	2	2	2	3
P16	2	2	2	2	2	2	2	2	2	2	2	2
P17	3	2	3	3	2	3	2	3	3	3	2	3
P18	2	3	2	2	2	2	2	2	3	2	2	2
P19	3	2	1	1	2	2	1	1	3	2	2	2
P20	2	3	2	2	2	2	2	2	2	2	3	2

For Review Only

3.1 Intra-observer agreement

The results of intra-observer agreement analysis for the ED group, and the ND group before and after training are listed in Table 2. For the ED group, the kappa value (0.64±0.13) is significantly higher (Student's *t*-test, $p<0.01$) than that of ND counterpart before training (0.47±0.11) by considering the mean of 9 tongue features. By comparing the agreement data of ND before and after training, it can be observed that the value of ND after training is significantly higher than that of ND before training with a level of $p<0.05$, indicating an obvious improvement of intra-observer agreement for ND after training with ATDS. With respect to individual tongue features, all the features show improvement after training for ND. Among them, 5 features, i.e., tongue color, fur color, tongue fissure, red dot and tooth mark, are improved significantly (Student's *t*-test, $p<0.05$). These features correspond to major subjects of focus in tongue diagnosis starting from formal school education of TCM practitioners. Criteria to differentiate these tongue features are already established and deep-rooted, even though they may vary from one to another. On the other hand, there are no precise or quantifiable standards existing for features such as fur thickness, ecchymosis, saliva and tongue shape. The concepts of thick v.s. thin, the size threshold for determining the existence of an ecchymosis, the amount of saliva and the classification of shape, all require more complicated and relative reasoning. Taking the tongue shape as an example, whether the classification of a tongue as thin and small, moderate, or fat and large, as listed in Figure 4, shall be relative to the width-height ratio or the ratio between the width of the mouth and that of a tongue, remains arguable and not yet finalized. A short-term intensive training lasting for one month might be sufficient to forge consensus on features with in-depth understanding long existing. However, features with delicate details, less-intuitive reasoning and yet-to-be determined differentiating criteria demand a longer period of education.

For Review Only

Table 2: Comparisons of intra-observer agreements (kappa values) between ED group and ND group before training with a significant level of $^{*}p<0.05$, $^{**}p<0.01$, and $^{***}p<0.001$, as well as between ND group before and after training with a significant level of $^{\dagger}p<0.05$ and $^{\ddagger}p<0.01$ (Student's *t*-test).

Features	ED (N=12)	ND before training (N=14)	ND after Training (N=14)
Tongue Color	0.39±0.18	0.37±0.13	0.51±0.11‡
Fur Color	0.83±0.04**	0.65±0.17	0.77±0.14‡
Fur Thickness	0.68±0.27	0.52±0.14	0.63±0.18
Tongue Fissure	0.72±0.18**	0.53±0.11	0.68±0.14†
Red Dot	0.50±0.15	0.38±0.15	0.50±0.16†

Ecchymosis	0.71±0.57*	0.31±0.14	0.41±0.13
Tooth Mark	0.59±0.24	0.44±0.19	0.60±0.17†
Saliva	0.58±0.34	0.46±0.16	0.51±0.19
Tongue Shape	0.81±0.09***	0.54±0.22	0.59±0.20
Mean†	0.64±0.13**	0.47±0.11	0.57±0.10†

3.2 Inter-observer agreement

As tabulated in Table 3, the results of the inter-observer agreement analyses (kappa value) between the ATDS and ED group, and between the ATDS and ND group before and after training range from 0.25 to 0.79 (0.45±0.17), from 0.17 to 0.68 (0.40±0.16) and from 0.18 to 0.75 (0.44±0.17), respectively, indicating moderate inter-observer agreements. The inter-observer agreements between the ATDS and ND after training increase for all tongue features tested (0.44±0.17). This can be attributed to the fact that ATDS is utilized as the means of education throughout the ND training session. Although higher agreements of several features have been achieved for ND after training, the inter-observer agreements of some other features remain slightly falling behind those of ED. This validates the importance of clinical experience. One-month intensive short-term training cannot compensate the seniority gap between ND and ED groups in terms of clinical practice. Training sessions with a longer period are indispensable in acquiring the skills necessary for consistent tongue features identification.

Table 3: Inter-observer agreements between ATDS and ED, and between ATDS and ND before and after training

Features	**Kappa value** (ATDS *v.s.* ED)	**Kappa value** (ATDS *v.s.* ND Before Training)	**Kappa value** (ATDS *v.s.* ND After Training)
Tongue Color	0.43	0.27	0.40
Fur Color	0.40	0.52	0.55
Fur Thickness	0.55	0.21	0.26
Tongue Fissure	0.79	0.68	0.75
Red Dot	0.34	0.33	0.37
Ecchymosis	0.26	0.17	0.18
Tooth Mark	0.51	0.54	0.57
Saliva	0.25	0.32	0.34
Tongue Shape	0.52	0.52	0.55
Mean	0.45±0.17	0.40±0.16	0.44±0.17

As listed in Table 4, the inter-observer agreements (kappa values) among the TCM doctors in the ED group, as well as in the ND group before and after training are ranging from 0.16 to 0.62 (0.41±0.15), 0.07 to 0.52 (0.29±0.14), and 0.26 to 0.76

For Review Only

(0.43±0.14), representing moderate inter-observer agreements. As indicated in this table, the inter-observer agreements for ED group and ND group after training are significantly higher (Student's *t*-test, $p<0.05$) than the ND group before training. In contrast, no significance ($p>0.05$) is observed between the ED group and the ND group after training, indicating that physicians in the ND group achieve an inter-observer agreement to the same level as that of experienced doctors after one month of training. This finding validates the utilization of ATDS in forging consensus during the learning process of tongue diagnosis.

Table 4: Inter-observer agreements among TCM physicians in ED and ND groups. Note: Comparison with ND group before training with a significance level of $^*p<0.05$.

Features	Kappa value of ED	Kappa value of ND before training	Kappa value of ND after training
Tongue Color	0.50	0.19	0.41
Fur Color	0.62	0.52	0.76
Fur Thickness	0.48	0.32	0.43
Tongue Fissure	0.57	0.46	0.49
Red Dot	0.23	0.21	0.37
Ecchymosis	0.16	0.07	0.26
Tooth Mark	0.35	0.31	0.39
Saliva	0.42	0.26	0.37
Tongue Shape	0.40	0.30	0.40
Mean	$0.41\pm0.15^*$	0.29±0.14	$0.43\pm0.14^*$

4 Discussion and Conclusion

In order to integrate the CAM into current healthcare systems and facilitate the communication between medical doctors and CAM providers, education in CAM is gaining more intention in the developed countries, such as the United States, the United Kingdom, Canada, Australia, Germany, and so on [33]. It has been long argued that the inconsistent diagnosis in TCM practices is mainly caused by subjective assessment of symptoms. With the advancement of information and communication technology (ICT), computer-assisted education, such as e-learning and simulation-based education, has highly elevated the learning efficacy of medical education because of its high fidelity, duplicability, and repeatability [34-36]. The ATDS developed in this study takes the advantage of ICT technology for presenting the tongue images with consistency without being affected by the variations in background lighting, position of chin, and length, shape, and angle of extruding

For Review Only

tongue [16]. Furthermore, accompanied with the 9 extracted features as feedback information, the ATDS can be used as an effective platform for TCM diagnosis and education. It can also be integrated into the e-learning system to facilitate learning effectiveness of TCM practices.

The ATDS developed, with a mean kappa value of 0.93 (almost perfect) for intra-observer agreement, has demonstrated superiority over a team of experienced TCM practitioners, whose intra-observer agreement shows a mean of 0.64 (substantial) [16]. To further prove the feasibility of ATDS in clinical practice, the inter-observer agreement between characteristic information captured and analyzed by the ATDS and through visual observation by the TCM doctors was also performed [16]. In light of the high degree of reliability in reaching the same conclusions under diverse lighting, image-taking angles, and tongue lengths variations, the efficacy of ATDS serving as a mean of education in learning tongue diagnosis has been verified in this study.

To this end, a two-round experiment composed of pre- and after-training tests was performed. Two groups of TCM doctors were formed by experienced doctors and novice doctors. The intra- and inter-observer agreements for ED and ND groups, and the inter-observer agreements between ATDS and ED, and ATDS and ND were derived first. After a one-month training of the novice doctors, the intra- and inter-observer agreements of the ND group, and the inter-observer agreement between ATDS and ND were obtained. For the ED group, the kappa value (0.64±0.13) of intra-observer agreement is significantly higher (Student's *t*-test, $p<0.01$) than that of ND counterpart before training (0.47±0.11) by considering the mean of 9 tongue features. The kappa value of ND after training is significantly higher than that of ND before training with a level of $p<0.05$, indicating an obvious improvement of intra-observer agreement for ND after training with ATDS. With respect to individual tongue features, all the features show improvement after training for ND. The results of the inter-observer agreement analyses (kappa value) between the ATDS and ED group, and between the ATDS and ND group before and after training range from 0.25 to 0.79 (0.45±0.17), from 0.17 to 0.68 (0.40±0.16) and from 0.18 to 0.75 (0.44±0.17), respectively, indicating moderate inter-observer agreements. The inter-observer agreements between the ATDS and ND increase for all tongue features tested (0.44±0.17). This can be attributable to the fact that ATDS is utilized as the means of education throughout the ND training session. The inter-observer agreements (kappa values) among the TCM doctors in the ED group, as well as in the ND group before and after training are ranging from 0.16 to 0.62 (0.41±0.15), 0.07 to 0.52 (0.29±0.14), and 0.26 to 0.76 (0.43±0.14), representing moderate inter-observer agreements. The inter-observer agreements for ED group and ND group after training are significantly

For Review Only

higher (Student's *t*-test, $p<0.05$) than the ND group before training. In contrast, no significance ($p>0.05$) is observed between the ED group and the ND group after training, indicating that physicians in the ND group achieve an inter-observer agreement to the same level as that of experienced doctors after one-month training by ATDS. This finding validates the utilization of ATDS in forging consensus during the learning process and providing a feasible alternative to the traditional teaching methodology of tongue diagnosis.

Other than serving as a clinical instrument in providing TCM doctors with consistent tongue features of patients during clinical inspection and as a training system in achieving high degree of agreement during learning process, the ATDS can also play the part as a dynamic evaluation system. Based on the answering pattern of a learner in a test, the ATDS can dynamically supply the relevant information for learning, and subsequently feed the pertinent question targeted on the specific error made by the learner. Learners of the tongue diagnosis can experience learning path tailored to the individual's needs, further supplemented by clinical tongue images and reference materials in the knowledge base to strengthen weak spots uncovered. With the help of ATDS in learning, evaluation and clinical practice, hopefully a new generation of TCM doctors with higher degree of consistency in tongue diagnosis can be educated. Through the improvement of diagnostic agreement, the skepticism surrounding TCM can finally be lifted.

Acknowledgement

This study was supported in part by National Science Council of Taiwan under Grant No. NSC101-2923-B-110-002-MY3. The authors would like to express their appreciation to Mr. Chun-Hung Lu for his assistance in the acquisition of tongue images and data preparation.

For Review Only

References

1. Z. Hussain and E. M. M. Quigley, "Systematic review: complementary and alternative medicine in the irritable bowel syndrome," *Alimentary Pharmacology & Therapeutics*, 23, pp. 465–471, 2006.
2. A. Rao, M. Shakeel, A. Trinidade, G. Rao, A. Pearce and K.-W. Ah-See, "The Importance of Complementary and Alternative Medicine Education in Medical School," *Journal of Evidence-Based Complementary & Alternative Medicine*, 17: 191-198, 2012.

3. Y. Yildirim, S. Parlar, S. Eyigor, O. O. Sertoz, C. Eyigor, C. Fadiloglu and M. Uyar, "An analysis of nursing and medical students' attitudes towards and knowledge of complementary and alternative medicine (CAM)," *Journal of Clinical Nursing*, 19, pp. 1157–1166, 2010
4. M. Nicolao, M. G. Täuber & P. Heusser, "How should complementary and alternative medicine be taught to medical students in Switzerland? A survey of medical experts and students," *Medical Teacher*, 32: 50–55, 2010.
5. G. G. Zhang, W. Lee, B, Bausell, L. Lao, B. Handwerger and B. Berman, "Variability in the Traditional Chinese Medicine (TCM) Diagnoses and Herbal Prescriptions Provided by Three TCM Practitioners for 40 Patients with Rheumatoid Arthritis," *J. Altern. Complement. Med.*, 11: 415-421, 2005.
6. K. A. O'Brien, E. Abbas, J. Zhang, Z. X. Guo, R. Luo, A. Bensoussan, and P. A. Komesaroff, "An investigation into the reliability of Chinese medicine diagnosis according to Eight Guiding Principles and Zang-Fu Theory in Australians with hypercholesterolemia," *J. Altern. Complement. Med.*, vol. 15, no. 5, pp. 259-266, 2009.
7. K. A. O'Brien, E. Abbas, P. Movsessian, M. Hook, P. A. Komesaroff , and S. Birch, "Investigating the Reliability of Japanese Toyohari Meridian Therapy Diagnosis," *J. Altern. Complement. Med.,* vol. 15, no. 10, pp. 1099-1105, 2009.
8. S. Mist, C. Ritenbaugh and M. Aickin, "Effects of Questionnaire-Based Diagnosis and Training on Inter-Rater Reliability among Practitioners of Traditional Chinese Medicine," *J. Altern. Complement. Med.,* vol. 15, no. 7, pp. 703-709, 2009.
9. K. A. O'Brien and S. Birch, "A review of the reliability of Traditional East Asian Medicine diagnoses," *J. Altern. Complement. Med.,* vol. 15, no. 4, pp. 353-366, 2009.
10. R. R. Coeytaux, W. Chen, C. E. Lindemuth, Y. Tan, and A. C. Reilly, "Variability in the diagnosis and point selection for persons with frequent headache by traditional Chinese acupuncture," *J. Altern. Complement. Med.*, vol. 12, no. 9, pp. 863-872, 2006.
11. J. J. Sung, W. K. Leung, J. Y. Ching et al., "Agreements among traditional Chinese medicine practitioners in the diagnosis and treatment of irritable bowel syndrome," *J. Aliment. Pharmacol. Ther.*, vol. 20, no. 10, pp. 1205-1210, 2004.
12. B. Zell, J. Hirata, A. Marcus et al., "Diagnosis of symptomatic postmenopausal women by traditional Chinese medicine practitioners," *J. Menopause*, vol. 7, no. 2, pp. 129-134, 2000.

For Review Only

13. G. G. Zhang, B. Bausell, L. Lao et al., “The variability of TCM pattern diagnosis and herbal prescription on rheumatoid arthritis patients,” *J. Altern. Ther. Health Med.*, vol. 10, no. 1, pp. 58-63, 2004.
14. G. G. Grant, B. Singh, W. Lee, H. Barry, L. Lao, and B. Brian, “Improvement of agreement in TCM diagnosis among TCM practitioners for persons with the conventional diagnosis of Rheumatoid Arthritis: Effect of training,” *J. Altern. Complement. Med.*, vol. 14, no. 4, pp. 381-386, 2008.
15. K. Minah, D. Cobbin, and C. Zaslawski, “Traditional Chinese Medicine tongue inspection: An examination of inter- and intra-practitioner reliability for specific tongue characteristics,” *J. Altern. Complement. Med.*, vol. 14, no. 5, pp. 527-536, 2008.
16. L. C. Lo, Y. F. Chen, W. J. Chen, T. L. Cheng, and J. Y. Chiang, “The Study on the Agreement between Automatic Tongue Diagnosis System and Traditional Chinese Medicine Practitioners,” *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 505063, 9 pages.
17. R. N. Schnyer, J. J. Allen, “Bridging the gap in complementary and alternative medicine research: Manualization as a means of promoting standardization and flexibility of treatment in clinical trials of acupuncture.” *J. Altern. Complement. Med.*, 8:623-634, 2002.
18. T. Kaptchuk, “Chinese Medicine: The Web That Has No Weaver. 2nd ed.,” London: Rider, 2000.
19. Li Gao, Ping Liu, Jue-xian Song, Xiao-guang Wu, Chang-ming Xu, Li-yuan Huang, Ping-ping Wang, Ying Zhang and Yu-min Luo, "A pilot study on the relationship between tongue manifestation and the degree of neurological impairment in patients with acute cerebral infarction," *Chinese Journal of Integrative Medicine*, Vol. 19, Issue 2, pp. 149-152, 2013.
20. Li Gao, Ping Liu, Jue-xian Song, Qian Liu, Chang-min Xu, Li-yuan Huang, Ping-ping Wang, Hai-ping Zhao and Yu-min Luo, "Study on the correlation of tongue manifestation with fibrinogen and neutrophil in acute cerebral infarction patients," *Chinese Journal of Integrative Medicine*, Vol. 18. Issue 12, pp. 942-945, 2013.
21. Tsung O. Cheng, "Tongue as the window to the heart," *Chinese Journal of Integrative Medicine*, Vol. 17, Issue 8, p. 640, 2011.
22. Wen-jie Qi, Miao-miao Zhang, Hong Wang, Yan Wen, Bao-en Wang and Shu-wen Zhang, "Research on the relationship between thick greasy tongue fur formation and vascular endothelial cell permeability with the protein expression of zonula occludens-1," *Chinese Journal of Integrative Medicine*, Vol. 17, Issue 7, pp. 510-516, 2011.

For Review Only

23. Yan Feng, Hao Xu, Dan Qu, Feng Zheng, Da-zhuo Shi and Ke-ji Chen, "Study on the tongue manifestations for the blood-stasis and toxin syndrome in the stable patients of coronary heart disease," *Chinese Journal of Integrative Medicine*, Vol. 17, Issue 5, pp. 333-338, 2011.
24. Wan Su, Zhen-ye Xu, Zhong-qi Wang and Jia-tuo Xu, "Objectified study on tongue images of patients with lung cancer of different syndromes," *Chinese Journal of Integrative Medicine*, Vol. 17, Issue 4, pp. 272-276, 2011.
25. N. Wiseman and A. Ellis, "Fundamentals of Chinese Medicine: Zhong Yi Xue Ji Chu," Brookline, MA: Paradigm Publications, 1996.
26. B. Kirschbaum, "Atlas of Chinese Tongue Diagnosis," Seattle: Eastland Press, 2000.
27. Yu-Cheng Hsu, Ying-Ching Chen, Lun-chien Lo, and John Y. Chiang, "Automatic Tongue Feature Extraction," *Proc. Int'l Comput. Symp.*, Taiwan, pp. 936-941, 2010.
28. G. G. Zhang, B. Singh, W. Lee, B. Handwerger, L. Lao and B. Berman, "Improvement of Agreement in TCM Diagnosis Among TCM Practitioners for Persons with the Conventional Diagnosis of Rheumatoid Arthritis: Effect of Training," *J. Altern. Complement. Med.*, 14: 381-386, 2008.
29. J. Cohen,"A coefficient of agreement for nominal scales, " *Educ. Psychol. Meas.* , pp. 37-46, 1960.
30. J. Cohen, "Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychol. Bull.*, pp. 213-220, 1968.
31. J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, pp. 378-382, 1971.
32. J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *J. Biometrics*., vol. 33, no. 1, pp. 159-174, 1977.
33. N. K. Quartey, P. H. X. Ma, V. C. H. Chung, and S. M. Griffiths, "Complementary and alternative medicine education for medical profession: systematic review," *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 656812, pp. 1-13, 2012.
34. M. Chow, D. K. Herold, T. M. Choo, and K. Chan, "Extending the technology acceptance model to explore the intention to use Second Life for enhancing healthcare education," *Computers & Education*, vol. 59(4), pp. 1136-1144, 2012.
35. F. Consorti, R. Mancuso, M. Nocioni, and A. Piccolo, "Efficacy of virtual patients in medical education: A meta-analysis of randomized studies," *Computers & Education*, vol. 59(3), pp.1001-1008, 2012.
36. C. J. Schroedl, T. C. Corbridge, E. R. Cohen, S. S. Fakhran, D. Schimmel, W. C. McGaghie, and D. B. Wayne, "Use of simulation-based education to improve

For Review Only

resident learning and patient care in the medical intensive care unit: a randomized trial," *J. Crit. Care*, vol. 27(2), pp. 219.e7-13, 2012.

For Review Only

Dear Anonymous Reviewers,

The authors are grateful to your comments and suggestions for improving the quality and presentation of this paper. All advices are followed. It is our sincere hope that this revision will enhance readability to satisfy the requirements of this prestigious journal.

Reviewer 1

Comment 1:

The number of tongue images used in tests is only 20, so it is difficult to prove the tests were effective.

Answer:

The reviewer's comment is highly appreciated. To the best of our knowledge, most of the papers surveyed for intra- and inter-observer agreements are conducted on sample size far less than 20, in terms of the number of images tested, features extracted, and observers participated.

In this paper, for each of the 20 images selected, 9 tongues features are extracted. Even though the number of TCM doctors participated (12 experienced doctors (ED) and 14 novice doctors (ND)), tongue images tested, and features extracted belong to the class of small sample size, yet the conditions of the statistical measure employed. i.e., Cohen's kappa coefficients, Fleiss' kappa coefficients and two-sample *t*-test, are met. The claim that the utilization of ATDS as a mean of education in forging consensus during the learning process of tongue diagnosis does exhibit statistical significance.

For Review Only

Comment 2:

In tongue diagnosis system, the lighting affected the quality of tongue images and the results of automatic analysis to such an extent. Although the color bar and correction were utilized, light proof enclosure is optimal solution. ATDS didn't use light proof enclosure, the tongue images were probably captured without standard environment.

Answer:

The reviewer's comment is highly appreciated. An earlier paper of the authors demonstrates that the ATDS, without light proof enclosure, is highly consistent even in the face of variations of environmental lighting [16].

L. C. Lo, Y. F. Chen, W. J. Chen, T. L. Cheng, and J. Y. Chiang, "The Study on the Agreement between Automatic Tongue Diagnosis System and Traditional Chinese Medicine Practitioners," *Evidence-Based Complementary and Alternative Medicine*, vol. 2012, Article ID 505063, 9 pages.

This desirable characteristic is mainly attributed to the powerful automatic color calibration implemented in the ATDS. Consistent tongue images are acquired even in the presence of lighting variations.

Comment 3:

In the education of TCM, automatic tongue diagnosis system is helpful, but much clinical practices is very necessary and apprentice following master worker is most important part.

Answer:

You are absolutely right. We totally agree with your point of view. Human guidance, especially following the clinical practice of a master, is the single most important factor in education. Even though ATDS proves to be effective, yet it never intends to replace human guidance during the education process. Human guidance and ATDS are not mutually exclusive. ATDS serves as an efficient tool in helping apprentice follow master worker. With the help of ATDS in learning, evaluation and clinical practice, hopefully a new generation of TCM doctors with higher degree of consistency in tongue diagnosis can be educated. Through the improvement of diagnostic agreement, the skepticism surrounding TCM can finally be lifted.

For Review Only

Reviewer 2

Comment 1:

The rationale of this study should be made clearer to readers.

Answer:

In heed of the reviewer's comment, the rationales of this study, summarized below, are added in the "Abstract" and "4. Discussion and Conclusion," respectively.

Abstract:

"this paper investigates the efficacy of applying ATDS to the training of novice TCM doctors in the tongue diagnosis. Only through the improvement of diagnostic

consistency, can the skepticism surrounding TCM be lifted."

4. Discussion and Conclusion:

"The ATDS developed in this study takes the advantage of ICT technology for presenting the tongue images with consistency without being affected by the variations in background lighting, position of chin, and length, shape, and angle of extruding tongue [16]. Furthermore, accompanied with the 9 extracted features as feedback information, the ATDS can be used as an effective platform for TCM diagnosis and education. It can also be integrated into the e-learning system to facilitate learning effectiveness of TCM practices."

For Review Only