Applied Soft Computing 11 (2011) 3697–3708

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Knowledge elicitation for performance assessment in a computerized surgical training system

Mario Riojas[a,*], Chuan Feng[a], Allan Hamilton[a,b], Jerzy Rozenblit[a,b]

[a] Electrical and Computer Engineering, University of Arizona, 1230 E. Speedway Blvd., Tucson, AZ 85721, United States
[b] Arizona Health Sciences Center, Department of Surgery, 1501 N. Campbell Avenue, United States

ARTICLE INFO

Article history:
Received 3 December 2009
Received in revised form
18 September 2010
Accepted 30 January 2011
Available online 24 February 2011

Keywords:
Fuzzy logic
Knowledge elicitation
Membership functions
Minimally invasive surgery (MIS)
Objective assessment
Surgical training systems

ABSTRACT

Effective training is the key to minimizing the dangers of minimally invasive surgery (MIS). At present, the assessment of laparoscopic skills relies on the expertise of senior surgeons. The judgment is typically based on and expressed in ordinal variables that can take values such as low, medium, high or other comparable terms. This limited assessment, along with the lack of expert surgeons' metacognitive awareness of how the judgment process takes place, results in imprecise rules for the evaluation of laparoscopic surgical skills. In this work, we present the knowledge elicitation process to model the performance metrics and the rules involved in the assessment of minimally invasive surgical skills. We have implemented a scoring system for the evaluation of laparoscopic skills based on five performance metrics capable of distinguishing between four proficiency levels while providing a quantitative score. Our assessment model is based on fuzzy logic, so that it is easier to mimic the judgment that is already performed by experienced surgeons. The presented framework was empirically validated using the performance data of 38 subjects belonging to five groups: non-medical students, medical students with no previous laparoscopic training, medical students with some training, residents, and expert surgeons.

1. Introduction

Minimally invasive surgery (MIS) is a modern surgical technique requiring small incisions or no incisions. It is performed with an endoscope and several long, thin instruments. The drawbacks associated with large incisions, operative blood loss and post-operative pain are limited, and recovery time is shorter compared to traditional open surgery. Unfortunately, from a surgeon's perspective, laparoscopic surgery is more challenging than conventional surgery because of the restricted vision, hand–eye coordination problems, limited working space and lack of tactile sensation. These issues make MIS a difficult skill for medical students and residents to master.

To minimize the potential risks inherent in MIS, special training processes must be performed to help students adapt to the new surgical technique. When it comes time to evaluate students in minimally invasive surgical skills, the apprenticeship model prevails, as it is the most used method in medical schools around the world. In this model, the trainer serves as both an observer and an evaluator while the trainee or student performs a surgical exercise or procedure. In the apprenticeship model, the metrics used in the evaluation are typically recorded by expert surgeons as linguistic variables that can take values such as low, medium, high or other comparable terms. Then, the judgment process takes place by following mental rules or guidelines to compare the values of the metrics with the expert's standard criteria. The judgment process is expressed by experts in a qualitative manner. This qualitative judgment can take the form of natural language statements, if–then rules, textual descriptions of their assumptions in reaching an answer, reasons for selecting or eliminating certain data and/or information considered in the evaluation process [1].

This limited assessment along with the lack of expert surgeons metacognitive awareness of how the judgment process takes place results in imprecise rules for the evaluation of laparoscopic surgical skills. The assessment by observation does not meet the validity and reliability criteria necessary for any objective evaluation [2].

Moreover, the apprenticeship model is becoming increasingly difficult to sustain. Because the requirement of basic skill increases rapidly, traditional surgical education methods are not suitable for MIS training. Using the operating room for teaching surgical skills is impractical and raises cost-effectiveness and patient safety concerns. In a similar way, using animals and cadavers have limitations due to ethical issues, animal rights, high cost and low efficiency [3].

Therefore, a key aspect is the development of methods for training both residents and practicing surgeons as technology and

* Corresponding author. Tel.: +1 520 626 2117; fax: +1 520 626 2117.
E-mail address: mriojas@email.arizona.edu (M. Riojas).

1568-4946/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2011.01.041

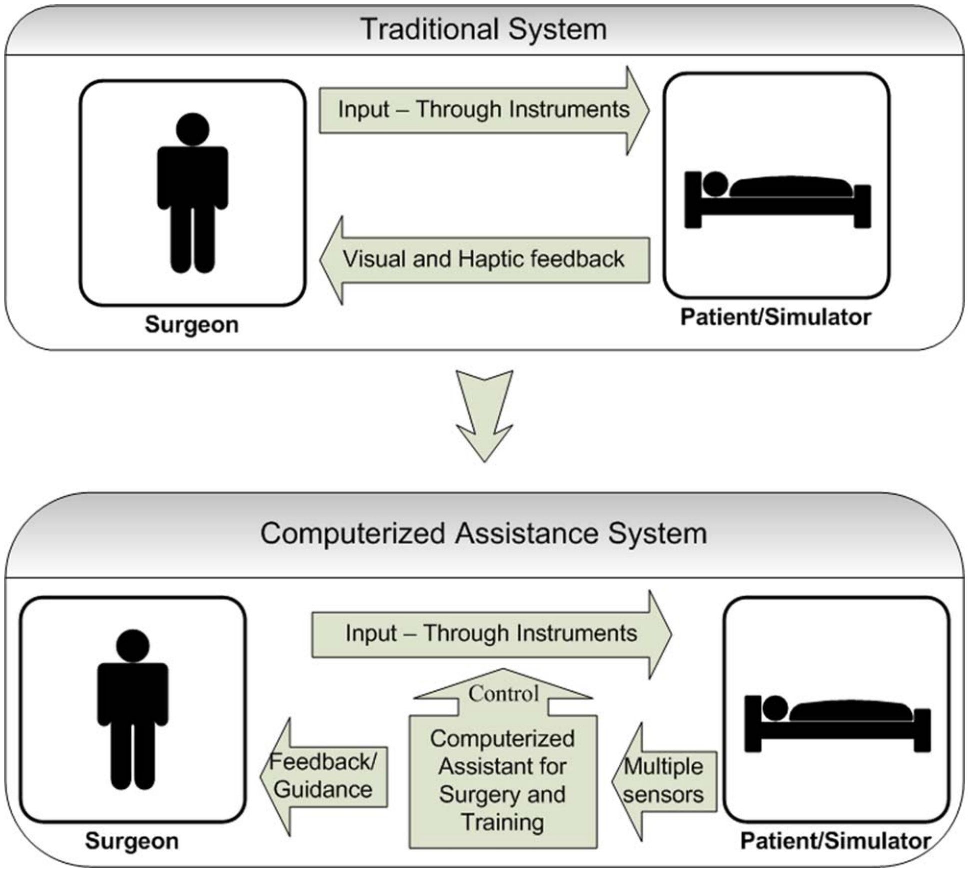

Fig. 1. Traditional VS. CAST assessment approach. CAST generates performance assessment by processing motion tracking data while providing visual and haptic feedback to the trainee.

procedures continue to evolve. The objective of this research is to design and realize a novel prototype that advances the state of the art in surgical training, assessment, and guidance. The system should provide multiple training scenarios, a high fidelity training environment, repeatable, structured exercises, and objective performance assessment capabilities.

In answer to the field's need, The Computer Assisted Surgical Training System (CAST) [4] is being developed at The University of Arizona through collaboration between the *Electrical and Computer Engineering Department* and *The Arizona Simulation Technology and Education Center (ASTEC)* at the College of Medicine. The objective of the CAST system is to implement a simple and effective training method to bridge the gap between existing training approaches and combining their advantages. We propose a knowledge-based sensing system to provide training prior to surgery and possible assistance in the operating room. Our design features the embedding of micro-sensors into the instruments employed for simulation training. The detection and recording of the users' operation permit our system not only to measure a trainee's progress in acquiring psychomotor skills and compare these data to normative databases, but also to evaluate instrument effectiveness in reducing errors. Fig. 1 contrasts the CAST System with the traditional approach. In the CAST system, the surgeon acts upon the patient or simulator through instruments and receives visual and force feedback from the CAST both in the operating room and in training settings. Our approach implements a "hybrid" system in which joint optimization of actuation, sensing, and computing is performed within a closed loop.

A prototype system is being developed which is capable of high fidelity motion tracking of surgical instruments and objective performance assessment analysis. The system represents the sensing interface and the knowledge-based computer system. It consists of a sensor fusion engine at the front-end and a knowledge based inference system at the backend.

A pelvic-trainer (a covered box with several openings, one for the laparoscope and the others for the surgical instruments) is used to simulate the patient's inner body. Each of the surgical instruments employed during the training tasks has a microsensor mounted on it to detect and record its movements in real time. The data acquired from the sensors is time stamped with an $x\,y\,z$ coordinate that matches the position of the sensor relative to a global orientation at that specific time. This time stamped data can be computed to obtain performance metrics such as time, path length, and continuity of movement [4].

Surgeons maneuver instruments to perform a variety of tasks. During training, they need to gain a clear feel for what constitutes safe and correct movement. Furthermore, during a surgical procedure in the operation room, potentially harmful movements must trigger an alarm if they cannot be prevented altogether. To achieve this, we implement capabilities for objective performance assessment and feedback. An overall score is computed upon completion of the procedure. Context rules are constructed based on empirical expert knowledge about laparoscopic surgical processes; thus, a single objective standard is difficult to define. To formulate a usable standard, and to provide an accurate scoring method, a fuzzy logic method is proposed. Initial experimental results are presented to show the feasibility of the proposed method.

The organization of the paper is as follows. Section 2 briefly surveys related work on performance assessment in MIS training. Section 3 presents underlying assumptions and defines the problem. In Section 4, we describe the details of the performance assessment method including fuzzy logic approach and

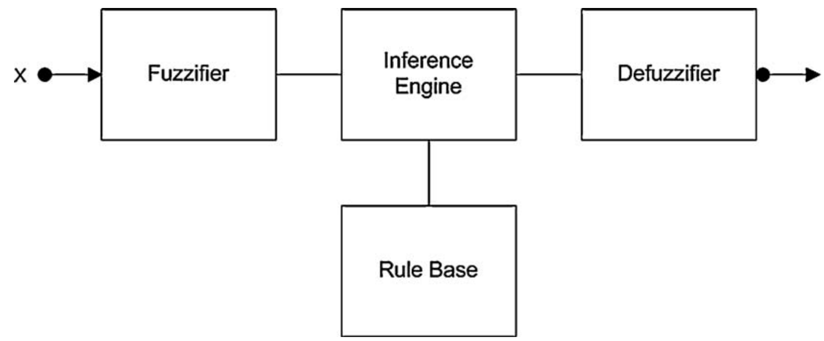

Fig. 2. Basic components of a fuzzy logic system.

the knowledge elicitation process. In Section 5, we present the implementation of the method and the usability experiment result. Section 6 discusses the features that leverage our framework to provide deliberate practice in a valid and reliable training environment. Section 7 concludes the paper and outlines directions for future work.

2. Literature review

Any assessment method should be feasible, valid and reliable. Unfortunately, this is not the case when the assessment of laparoscopic surgical skills is done by observation. "As the assessment is global and not based on specific criteria, it is unreliable. As it is influenced by the subjectivity of the observer it would possess poor test–retest reliability and also be affected by poor interobserver reliability as even experienced senior surgeons have a high degree of disagreement while rating the skills of a trainee" [16].

The development of methods to objectively evaluate surgical skills range from checklists and global scores to complex virtual reality based systems [5–9]. To model expert judgment we have to design a standard method of evaluation capable of correlating with the opinion of experts while overcoming the obstacles of poor test–retest reliability and subjectivity currently present in the apprenticeship model [2,9].

Objective structure assessment of technical skills (OSATS) is a methodology consisting of six stations where residents and trainees perform procedures such as placing sutures in a pad of synthetic skin; joining two cut ends of bowel; and inserting an intercostal catheter. The procedures are performed on live animals or bench models in fixed time periods. Staff surgeons observe the performance of the trainees and evaluate them using a checklist and a global scoring sheet. The checklist is a series of items that should be marked by the staff surgeons in regards to the trainee's performance during a task. Each task has a customized checklist. Examples of items are: select appropriate instruments, correct needle holding technique, suture spacing 3–5 mm, and knots placed to one side of the suture line. The global scoring sheet comprises eight items, each of which is marked from 1 to 5. The items assessed include tissue handling skills, flow of operation, and familiarity with the technique. Examples of poor (score 1), average (score 3) and excellent performance (score 5) are given as guidelines for the observer. OSAT's drawbacks are the resources and time involved in getting several staff surgeons to observe the performance of trainees. However, global and checklist scoring systems have been previously validated [5].

Minimally Invasive Surgical Trainer-Virtual Reality (MIST-VR) is a popular commercial training simulator. Based on virtual reality, the system is able to distinguish between novice, junior and experienced surgeons. MIST-VR is a "low fidelity" system which attempts to replicate the skills of laparoscopic operating but not the appearance (the virtual environment consists of a wire cage in which various geometric objects may be manipulated with two tools, each having 5 degrees of freedom). The final scores on MIST are derived from weighted averages of performance metrics such as: time to task completion, errors and economy of movement. The main drawback of this system is the lack of haptic feedback and the unfamiliarity of the trainees with a virtual environment [6].

Computer Enhanced Laparoscopic Training System (CELTS) is capable of tracking the motion of two laparoscopic instruments while the trainee performs a variety of surgical training tasks. CELTS consists of a box-type video trainer platform that uses conventional laparoscopic imaging equipment coupled with real laparoscopic instruments that are placed through a Virtual Laparoscopic Interface. Using kinematics analysis theory, CELTS generates five quantitative metrics: time to complete the task, depth perception, path length of the instruments, motion smoothness and response orientation. Using standard scores statistics, the performances of trainees are compared to the performances of expert surgeons and assigned a standardized overall score from 0 to 100 [7].

Blue Dragon is a system for acquiring the kinematics and the dynamics of two endoscopic tools along with the visual view of the surgical scene. Hidden Markov models based on haptic information were proposed by Rosen et al. as a method to objectively evaluate laparoscopic surgical skills. In their study, they generate quantitative knowledge of the forces and torques (F/T) applied by the surgeons on their instruments during minimally invasive surgery. Through the use of modified surgical graspers containing embedded sensors that are capable of measuring the F/T, they developed a database of F/T signals. Statistical models of the F/T data allowed them to characterize surgical skills. The methodology of decomposing the surgical task is based on a fully connected finite states (28 states) Markov model where each state corresponds to a fundamental tool/tissue interaction based on the tool kinematics and associated with unique F/T signatures [18].

A fuzzy logic approach to categorize minimally invasive surgical skills has been designed and implemented by Hajshirmohammadi et al. [9]. Using the commercial trainer MIST-VR they collected data from an exercise performed by subjects with different MIS experience. Their work suggests formulating rules for MIS performance assessment and defining proficiency levels from data patterns. Our framework proposes that the assessment of MIS skills should also be a function of the theoretical knowledge of what experienced surgeons conceive as competitive performance.

A limitation shared by many of the above mentioned scoring systems is that they rely on judgment criteria generated at an early stage of the system's design phase. They present no way to offset the subjectivity inherited from relying solely on the performance data of the selection subjects used during the system's design phase. Moreover, the judgment criteria are meant to remain intact over time, making it difficult for a trainee who has reached the "expert" level to keep improving. One of the key characteristics behind the development of an "expert" (in any field) is the engaging of trainees in deliberate practice. Deliberate practice is defined by Ericsson et al. [17] as the practice of tasks beyond the trainee's current level of competence and comfort. Practicing should be oriented toward advancing and not simply maintaining a standardized performance of the task [17]. Therefore, any MIS scoring system should be able to raise the criteria standards used in their assessments to avoid the risk of leaving trainees in a performance plateau. Our work is an effort to overcome these obstacles in the current performance assessment of MIS skills.

3. Problem formulation

Simulations with computerized surgical training systems offer an opportunity to teach and practice skills outside the operating room before attempting them on living patients. Technical skills acquired on low-fidelity simulators can be transferred to improve performance on higher fidelity models such as live animals, cadavers and, eventually live human patients [10].

Transfer of learning is the application of skills and knowledge gained in one context being applied in another context. To effectively apply transfer of learning in surgical procedures it is necessary to abstract their essential constructs in developed models that allow the trainees to perform abstract tasks. Hamstra et al. [10] presented the theoretical basis for interpreting the effectiveness of low-fidelity models using the constructivism theory which assumes that "knowledge is constructed from collections of structures, which are essentially mental representations of information." When the right abstract tasks are selected for training, the trainee learns to ignore the physical limitations of the training platform and focus on the procedure, allowing for the suspension of disbelief and the learning of transferable knowledge. We can obtain performance metrics from these abstract tasks to provide feedback and for evaluation purposes. If the performance metrics are properly chosen and validated, technical ability can be then determined to demonstrate fitness to practice independently [10].

It is possible to design a series of tasks in low-fidelity simulators such as CAST that allow an effective transfer of knowledge. A separate issue is determining the proficiency level of a trainee using these abstract tasks. The key questions are: how we can measure proficiency levels of minimally invasive surgery skills? If computerized surgical training systems can compute performance metrics, how do we correlate these metrics' values with proficiency levels?

The goal of this study was to perform the knowledge elicitation process to formulate expert judgment for the assessment of laparoscopic surgical skills. We designed a scoring system based on fuzzy logic capable of distinguishing between four proficiency levels while providing students with a quantitative score. This goal was composed of the following objectives: (1) defining a set of relevant performance metrics in the assessment of laparoscopic surgical skills; (2) eliciting and generating membership functions to model performance metrics; (3) eliciting a set of production rules to model experts' judgment; and (4) defining a set of proficiency levels to categorize subjects. The scoring system can be used to objectively quantify competency in MIS skills.

4. Performance assessment system design

4.1. Theoretical fit between the fuzzy logic model and the assessment of laparoscopic skills model

Fuzzy logic is a multivalued logic initiated in 1965 by Zadeh [15] to form part of what is currently known as soft computing. In contrast to traditional computing which strives for exactness and full truth, soft computing techniques exploit the given tolerance of imprecision, partial truth, and uncertainty for a particular problem. Fuzzy logic provides the opportunity for modeling concepts and dependencies that are inherently imprecisely defined.

In fuzzy sets, the membership value of an element x to a set A can take values in the interval [0,1]. Fuzzy sets represent common sense linguistic variables. The membership of the elements in fuzzy sets is not mutually exclusive (i.e., element x may belong to fuzzy set A at a certain degree but also to fuzzy set B at the same or a different degree). The representation of a fuzzy set A in the universe U is given by

$$A = \{x, \mu_A(x) | x \in A\} \tag{1}$$

where x is called a support value of A if $\mu_A(x) > 0$ and $\mu_A(x)$ is the membership function that defines how each point in the input space is mapped to a membership value between 0 and 1.

Fuzzy sets are used in inference systems to map in a non-linear way crisp inputs to crisp outputs by the application of production rules. A basic fuzzy inference system has four elements: a fuzzifier, an inference engine, a rule base and a defuzzifier, Fig. 2.

The core of a fuzzy inference system is the set of if–then rules contained in the rule base also known as fuzzy rules that specify a relationship between input and output fuzzy sets. Fuzzy rules take the form of:

If A then B

where A is a proposition or collection of propositions that represent the antecedent and B is generally a proposition representing the consequent. In order to interpret a fuzzy rule it is necessary to evaluate its antecedent which involves fuzzifying its crisp input and applying any necessary fuzzy operators (AND, OR, NOT). This is known as the antecedent's evaluation and is done by the system's fuzzifier. After the antecedent is evaluated the inference engine applies the fuzzified input to the fuzzy rule through the implication method to generate a fuzzy output or consequent. Each consequent is multiplied by its rule weight. All consequents are aggregated into one final output fuzzy set. Finally, the defuzzifier generates a crisp number from the aggregated output applying a defuzzification method [12].

Fig. 3 shows the inference process of two fuzzy rules that fired to determine a proficiency *level* in MIS according to a given value 0.869 of the metric *continuity of movement*:

If *continuity of movement* is Moderate Positive.
Then *proficiency level* is Proficient.

If *continuity of movement* is Strong Positive.
Then *proficiency level* is Expert.

After the fuzzification of the value (0.869) corresponding to *continuity of movement*, its membership degree is determined to belong to two fuzzy sets: Moderate Positive and Strong Positive. Therefore, the two rules in Fig. 3 are triggered. The inference engine fires them in parallel, generating two consequents in the form of fuzzy outputs with some membership degree to the Proficient *proficiency level* and a greater membership degree to the Expert *proficiency level*. These two fuzzy outputs were aggregated using a pointwise summatory. Then, the defuzzification method was applied to obtain the assessment score of 87.2.

There is a natural relationship between the laparoscopic skills performance assessment model and the fuzzy logic model. Our motivation to use fuzzy logic is summarized in the following points.

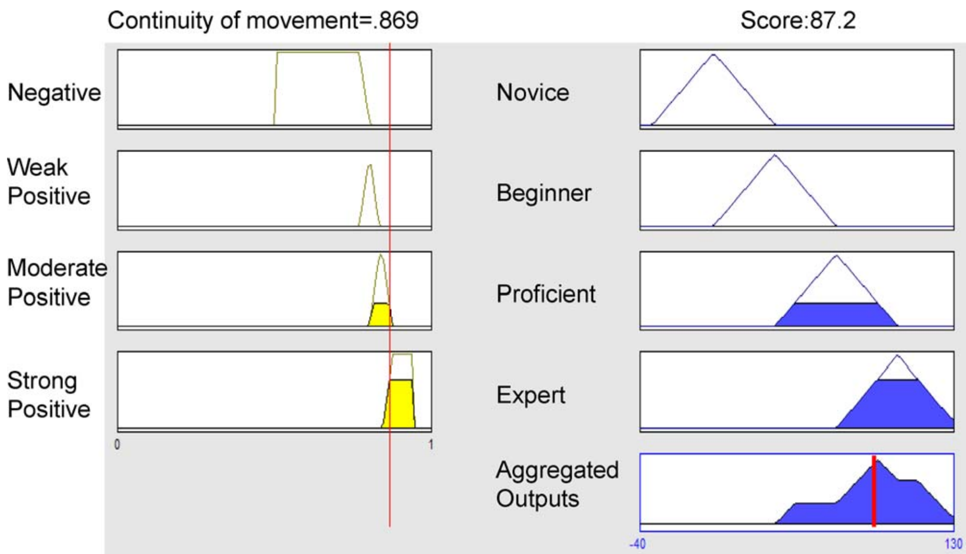

Fig. 3. In this example, two rules have been fired by the input value 0.869 of the metric *continuity of movement* which is used to derive a score for the performance assessment (87.2).

There exists a lack of personal and interpersonal agreement of expert surgeons on defining proficiency levels in laparoscopy surgery. To explain this ambiguity in the proficiency levels definition, let us assume the range of values corresponding to a valid performance metric. This range is sorted from less desirable value to most desirable value and divided into four crisp intervals: Negative, Weak Positive, Moderate Positive and Strong Positive (Fig. 4). If we match proficiency levels such as *Novice*, *Beginner*, *Proficient* and *Expert* with the intervals that characterize them, *Negative*, *Weak Positive*, *Moderate Positive* and *Strong Positive* respectively, then there exist values where it is not possible to define unambiguously if they are characteristic of one proficiency level or another.

We use fuzzy sets to categorize those values that fall within the boundary areas between two neighboring intervals characteristic of different proficiency levels. Fuzzy logic works with non-crisp sets where partial membership is possible. In other words, we can say that there exist subjects with performances containing the metrics' values characteristic of two neighboring proficiency levels at a certain degree, e.g., a membership degree of 0.5 to the *Beginner* proficiency level and a membership degree of 0.5 to the *Proficient* proficiency level.

Fuzzy logic models imprecise dependencies based on natural language. This simplifies the judgment knowledge elicitation process since it is possible to interview expert surgeons in their own terms, i.e., we can take the rules that they already use in the judgment process and model them as fuzzy rules to work within an inference system. We are building a scoring system using the experience of expert surgeons. We are basically taking what expert surgeons know about the judgment process and designing an objective scoring system based on their knowledge.

4.2. Knowledge elicitation process

To learn more about the judgment criteria used in the assessment of laparoscopic surgical skills, we started our elicitation process through an open interview with an expert surgeon from the University of Arizona Medical Center. Our expert used four levels (*Novice*, *Beginner*, *Proficient* and *Expert*) to describe proficiency in minimally invasive surgical skills.

We followed a four step procedure for the generation of membership functions:

1- Definition of linguistic variables.
2- Semantic decomposition of each variable.
3- Selection and application of the membership function elicitation method.
4- Membership function generation.

4.2.1. Definition of linguistic variables

In this step we define the variables that will serve as input in our fuzzy system. This can be achieved by open question interviews with selected experts or based on previous research on the subject. If the latter is chosen, it is recommended to have an expert's opinion about the relevance of the selected variables before proceeding with the following steps. The definition of each variable should include its range of values, i.e., the set of values that each variable can possibly take.

Application: We defined five relevant metrics for a hand–eye coordination task which were validated by an experienced surgeon: time, movement economy ratio, movement direction profile, peak speed width and continuity of movement.

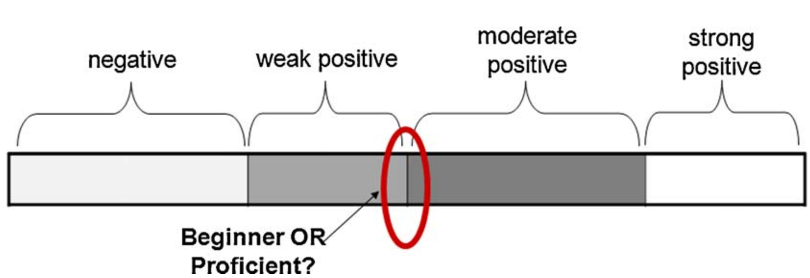

Fig. 4. There are metric values for which the relationship to one and only one proficiency level cannot be established.

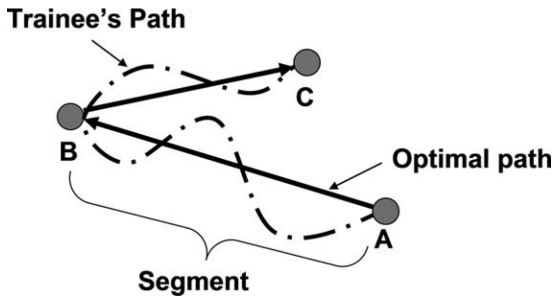

Fig. 5. Movement economy ratio is calculated by dividing the optimal path length between the targets A, B and C (shown in a solid line) by the path length drawn by the trainee (shown as dashed line).

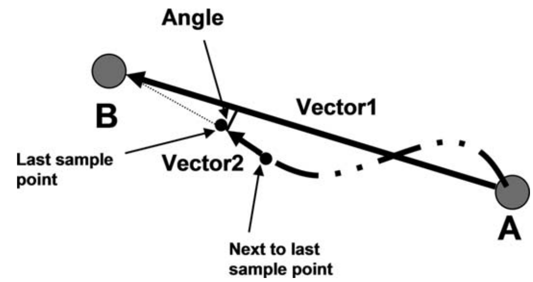

Fig. 6. Movement direction profile quantifies the extent that the laparoscopic instrument (dashed line) deviates from an optimal path (solid line) in moving from target A to target B.

Time: The total time taken by the trainee to perform the task.

Desirable value: →0.
Less desirable value: →∞.

Movement economy ratio: This metric scales the movement track length. A task is divided into segments. Each segment is defined by two targets that have to be reached in sequential order. The movement economy ratio is obtained by dividing the optimal path for performing the complete task by the addition of the path drawn by the instrument's tip while passing through the entire task's targets (i.e., all the segments that comprised the task). Fig. 5 exemplifies the *movement economy ratio* showing s trainee's path, optimal path and segment concepts.

Desirable value: → 1.
Less desirable value: → 0.

$$Re = \frac{\sum_{i=1}^{n} L_{I_i}}{\sum_{i=1}^{n} L_{R_i}} \tag{2}$$

where n is the total task's segmentation number, i is the serial number of each movement segmentation, L_{I_i} is the optimal path length of segment i, L_{R_i} is the trainee's path length of segment i.

Movement direction profile: This metric quantifies the extent that the instrument deviates in moving from target A to target B. It is equivalent to the cosine of the angle formed by two vectors. The first vector has its origin in the starting target A with a magnitude and direction equal to the optimal path between the starting target A and the ending target B. The second vector has its origin at the next to last sample point taken by the position sensor mounted on the instrument's tip with a magnitude and direction equivalent to the shortest path between that point and the last sample point registered by the sensors (Fig. 6).

The movement direction profile (cosine of the angle) will be equal to 1 when the instrument's tip follows the direction of the shortest path between starting target A and ending target B, −1 when the instrument's tip follows a path in the opposite direction of the shortest path, and 0 when the instrument's tip follows a path perpendicular to the shortest path. The closer the value is to 1, the better the movement is rated.

The solid line in Fig. 7 represents the optimal path between point A and point B. Last sample points are represented by the inner circle. The outer circle represents next to last sample points [13].

Desirable value: 1.
Less desirable value: −1.

Peak speed width: The movement speed described by a laparoscopic instrument when moving between two targets by an experienced surgeon goes rapidly from rest to a maximum speed value, maintains that maximum (or a close enough) value until the instrument is near to the target, and then returns to rest at the target. An example of such speed graph is provided in Fig. 8

Peak speed width is obtained by dividing the speed wave's peak amplitude by two and calculating the ratio of the resulting areas (Fig. 9). The peak speed width parameter depends on the wave's horizontal symmetry; waves closer to a trapezoidal shape like the one in Fig. 8 reflect better movement control over the instrument than jitter shapes such the one in Fig. 9. Therefore, their *Peak speed width* value approaches one [13].

Desirable value: → 1.
Less desirable value: → 0.

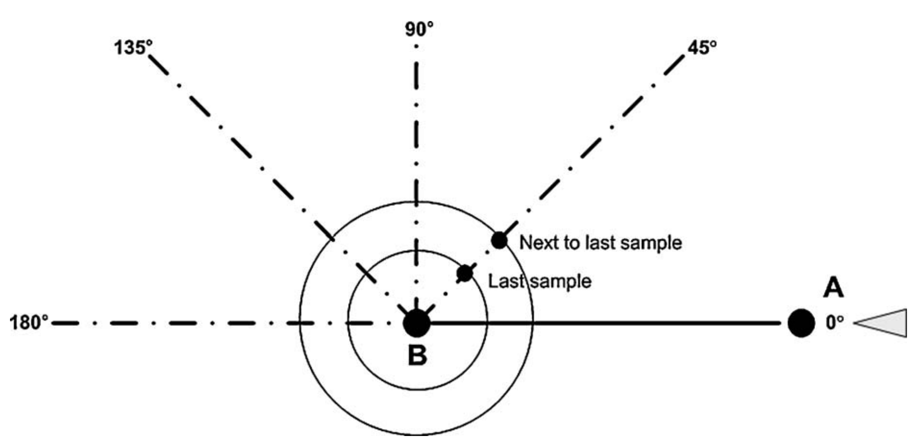

Fig. 7. The terminal angle is measured from the vector described by the optimal path between target A and target B (solid line), and the vector described by the last and next to last sample point captured by the sensor mounted on the instrument.

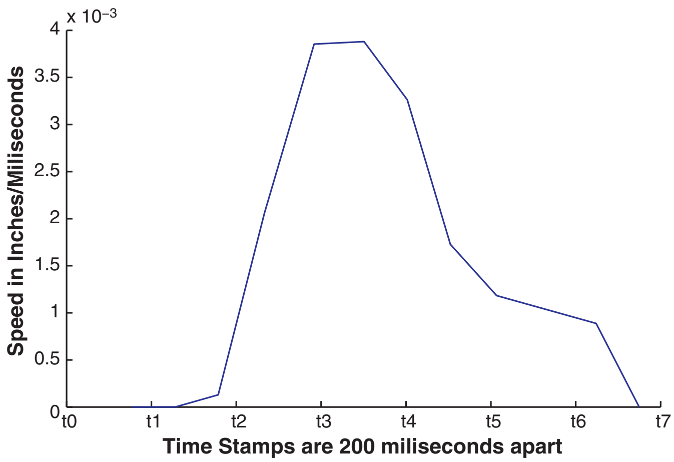

Fig. 8. Speed wave of an experienced surgeon when moving a laparoscopic instrument between two targets.

$$PW = \frac{A}{B} \tag{3}$$

A is the speed curve's upper area and B is the speed curve's lower area.

Continuity of movement: As before, assume that the instrument movement speed of an experienced surgeon goes rapidly from rest to a maximum speed value, maintains that value until the instrument is close to the target then returns to rest at the target (Fig. 8). Such a speed curve should not present any troughs.

Continuity of movement is calculated by eliminating recursively the speed's graph troughs (Fig. 11) to obtain a modified graph and then calculating the ratio of both areas under the curves original speed graph (Fig. 10) over modified speed graph (Fig. 12).

Desirable value: 1.
Less desirable value: → 0.

$$CM = \frac{A}{B} \tag{4}$$

A is the area under original speed curve and B is the area under modified speed curve.

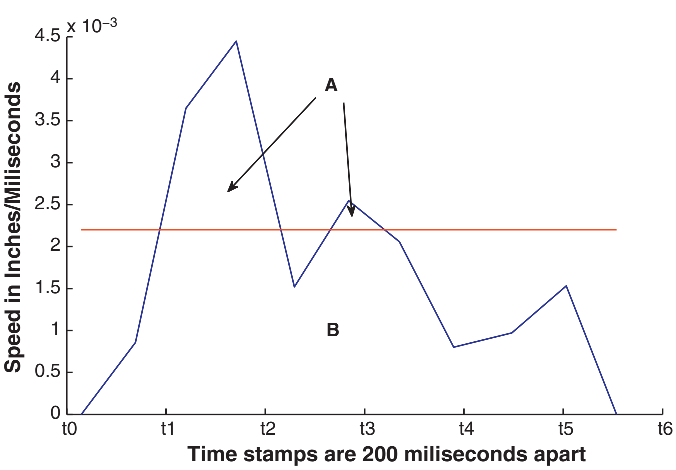

Fig. 9. Speed wave described by moving a laparoscopic instrument between two targets. *Peak speed width* is obtained by calculating the ratio between the two areas (labeled as A and B) that result from dividing the speed wave's peak amplitude by two.

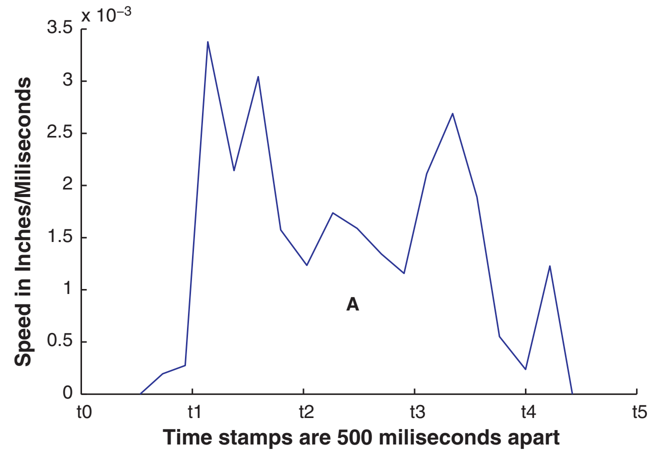

Fig. 10. The area under the speed wave (A) described by moving a laparoscopic instrument between two targets is used to calculate the *continuity of movement* metric.

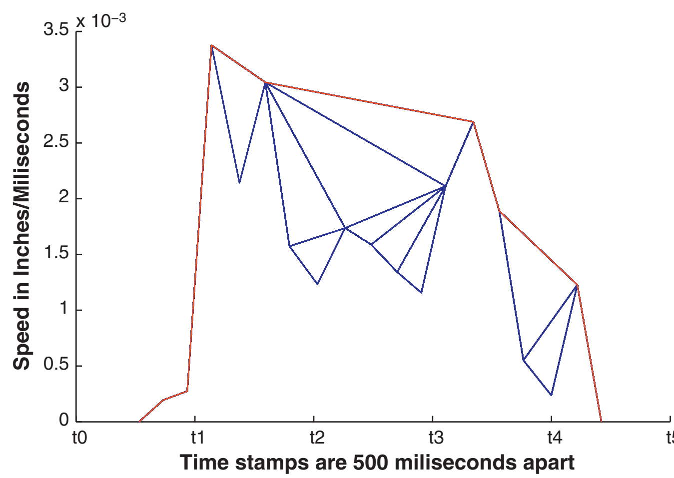

Fig. 11. Troughs were eliminated recursively from the original speed wave to obtain a more stable, smoother version.

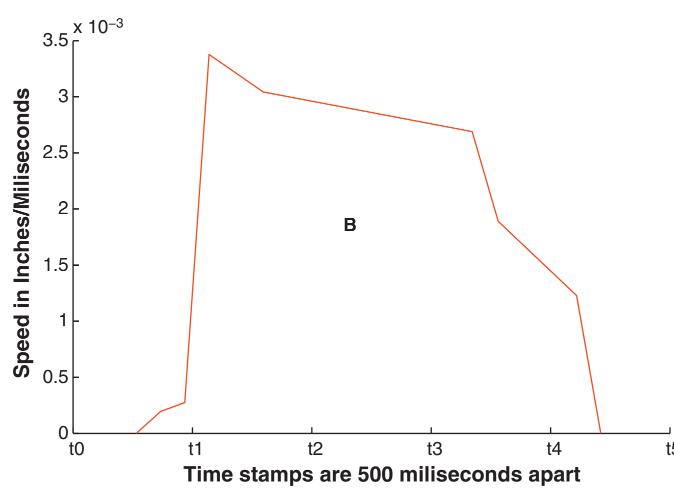

Fig. 12. A smoother modified speed graph exhibits a more desirable movement derived from the trainee's original performance. The area under this modified wave (B) is calculated for the *continuity of movement* metric.

Table 1
Fuzzy sets can be semantically decomposed by using different granularity levels.

Variable	Level 1	Level 2	Level 3
X	Strong Positive Medium Positive Weak Positive Weak Negative Medium Negative Strong Negative	Strong Positive Positive Medium Negative Strong Negative	Positive Medium Negative

4.2.2. Semantic decomposition of each variable

A variable is decomposed into fuzzy terms. Each fuzzy term corresponds to an interval of values that the variable may take. Values belonging to the same fuzzy term share a linguistic meaning. Different degrees of abstraction may be applied to a variable during the semantic decomposition; this depends on the nature of the problem. The abstraction degree is determined by the experts and is related to the granularity level they use when referring to the variable's value in qualitative form. Table 1 shows three different granularity levels a variable may take.

Application: Each metric was decomposed into four fuzzy terms that characterize the performance of the given proficiency level as follows:

Strong Positive → Expert.
Moderate Positive → Proficient.
Weak Positive → Beginner.
Negative → Novice.

These fuzzy terms represent the input sets to our inference system.

4.2.3. Selection and application of the membership function elicitation method

The goal of the elicitation method is to collect the necessary information to generate membership functions. There exist several elicitation methods to choose from in the current literature e.g., polling, direct rating, interval estimation and transition interval estimation [7,8]. Some elicitation methods are more appropriate than others depending on the problem. Factors to consider when choosing an elicitation method are the number of experts involved in the elicitation process and the interpretation given to the membership functions. However, it is our experience that often it is necessary to modify, combine or look for new ways of elicitation depending on the available resources and the restrictions of the problem.

Application: Our elicitation method focuses on the transitional point x for which expert p can make no crisp distinction whether property Ai does or does not apply. This is a modified version of the elicitation method introduced in [8] where an expert is asked for a transition interval.

A simple hypothetical example was used to define the proficiency levels, the expert surgeon assumed an array of sorted samples with 100 values collected from the performance of a selection of subjects with different laparoscopic surgical skills. The expert was asked what percentages belong to each proficiency level, to which he answered the following.

Expert: top 20%.
Proficient: 50–80%.
Beginner: 30–50%.
Novice: bottom 30%.

Given the above definition, our transitional points are located in the positions 30, 50, and 80 in the sorted array. We cannot define unambiguously if the values in these positions are characteristic of one proficiency level or the other. Hence we chose them as the points where the membership functions of two neighboring fuzzy sets intersect at the degree of 0.5.

4.2.4. Membership function generation

The elicitation process provides the key points needed to generate membership functions. Most of the time the information collected from different experts is averaged. It is also possible to give some weight to the expert's opinions according to their experience (e.g., number of performed procedures or working hours) or proven competency in the field. Membership function generation is straightforward and based on the information collected in the elicitation phase.

Application: We generated membership functions consisting of straight segments i.e., triangular and trapezoidal, also known as polygonal membership functions. Polygonal membership functions have several advantages as mentioned by Piegat [14]:

1- A small amount of data is needed to define the membership function.
2- It is easy to modify the parameters (modal values) of membership functions on the basis of measured values of the input → output of a system.
3- It is possible to obtain an input → output mapping which is a hyper surface consisting of linear segments.
4- Polygonal membership functions meet the condition of a partition of unity. The condition of a partition of unity restricts the sum of memberships of each element x from the universe of discourse to be equal to 1,

$$\sum_{h} \mu_{A_h}(x) \equiv 1, \quad \forall x \in X \tag{5}$$

where h is the number of the fuzzy set [14].

Our elicitation method provided us with the transitional points where two membership functions intersect at the height of 0.5. We use these points to calculate the rest of the critical points needed to construct polygonal membership functions according to the following criteria:

1- We chose a triangular function over a trapezoidal when possible.
2- We encouraged vertical symmetry on non-outer membership functions.
3- We satisfied the condition of a partition of unity.

Fig. 13 shows the membership functions defined for the metric *continuity of movement.* In this example, the identified transitional points were Beginner–Novice: 0.788045, Proficient–Beginner: 0.819006 and Expert–Proficient: 0.862335. We used these transitional points to generate the four membership functions that describe the *Negative*, *Weak Positive*, *Moderate Positive* and *Strong Positive* fuzzy sets.

4.3. Triangular functions

We chose the pair of transitional points ($tp1$, $tp2$) with the smallest difference to generate the first membership function. In our example for the metric *continuity of movement*, these points are Beginner–Novice: 0.788045 and Proficient–Beginner: 0.819006. These two transitional points enclose the range of the most characteristic values of the beginner proficiency level i.e., the *Weak Positive* fuzzy set. Triangular curves depend on three parameters given by:

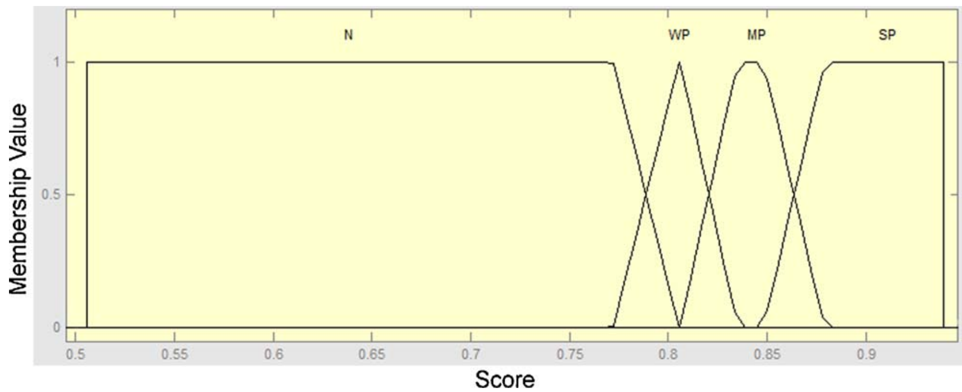

Fig. 13. Membership functions defined for the metric *continuity of movement*. Transitional points between neighboring fuzzy sets intersect at the degree of 0.5 meaning that such values equally belong to both neighboring sets (N = Negative, WP = Weak Positive, MP = Moderate Positive and SP = Strong Positive).

$$f(x;a,b,c)=\begin{cases}0 & \text{for } x<a\\ \dfrac{x-a}{b-a} & \text{for } a\le x<b\\ \dfrac{c-x}{c-b} & \text{for } b\le x\le c\\ 0 & \text{for } x>c\end{cases} \tag{6}$$

Based on Eq. (2) we defined the three parameters to generate a triangular membership functions these are the vertices a, b and c. Vertex b, which is the peak of the triangular function, is determined by the middle point mp between the two transitional points:

$$mp=\frac{tp2-tp1}{2}+tp1 \quad \text{where}:\ tp2>tp1 \tag{7}$$

Vertex a is equivalent to the point where the line that passes through the points $(mp, 1)$ and $(tp1, 0.5)$ intersects with the membership degree of 0.

Vertex c is obtained in a similar way to vertex a: by calculating the intersection with the membership degree equal to 0 of the line that passes through the points $(mp, 1)$ and $(tp2, 0.5)$. The resulting triangular function for the *Weak Positive* set is *f*(0.7725, ⇄0.8034, ⇄0.8346).

4.4. Trapezoidal functions

We gave priority to triangular functions over trapezoidal. However, triangular functions impose restrictions over their neighboring membership functions if we want to satisfy the condition of a partition of unity, making it difficult to represent all of the fuzzy sets as triangular shapes.

Continuing with our example using the metric *continuity of movement*, the triangular membership function describing the *Weak Positive* fuzzy set establishes restrictions over its neighboring membership functions. Therefore, trapezoidal functions are required to make the neighboring membership functions corresponding to *Negative* and *Moderate Positive* fuzzy sets to have a membership degree of 0 at the point where the already defined *Weak Positive* fuzzy set has a membership degree of 1 (i.e., at its middle point mp) as well as a membership degree of 0.5 where the transitional points have been identified.

Trapezoidal curves depend on four parameters and are given by:

$$f(x;a,b,c,d)=\begin{cases}0 & \text{for } x<a\\ \dfrac{x-a}{b-a} & \text{for } a\le x<b\\ 1 & \text{for } b\le x<c\\ \dfrac{d-x}{d-c} & \text{for } c\le x<d\\ 0 & \text{for } d\le x\end{cases} \tag{8}$$

At this point, we already know one out of four parameters required to generate the trapezoidal membership function corresponding to the *Negative* fuzzy set. We know that the vertex d of the *Negative* fuzzy set should be equal to vertex b of the *Weak Positive* fuzzy set (i.e., 0.8034) to satisfy the condition of partition of unity, making the maximum membership point for one fuzzy set (*Weak Positive*) represent the minimum membership point of its neighbor (*Negative*). Vertex c of the *Negative* membership function is equivalent to vertex a of the triangular function of the *Weak Positive* fuzzy set (i.e., 0.7725). Finally, we gave the same value to the vertices a and b because the *Negative* fuzzy set corresponds to the left outside membership function of the continuity of movement variable. This means it has only one neighbor to its right side (*Weak Positive*). We gave a and b the minimum value shown in our samples i.e., 0.5761. The trapezoidal function for the *Negative* set is *f*(0.5761, ⇄0.5761, ⇄0.7725, ⇄0.8034).

We encourage vertical symmetry of trapezoidal membership functions when they are not right or left outside membership functions such as the *Moderate Positive*. As in previous examples vertex a and b can be defined from the neighboring set *Weak Positive*. The middle point mp between the two transitional points that enclose the highest membership degrees of the *Moderate Positive* is calculated using Eq. (9). These three points define a half of a symmetrical trapezoid from where the remaining vertex can be easily deduced. The trapezoidal function for the *Moderate Positive* set is *f*(0.8034, ⇄0.8346, ⇄0.8467, 0.8778).

5. Fuzzy inference system implementation and results

5.1. Framework

We design a hand–eye coordination task consisting of a tray with 5 labeled pegs and a starting point (Fig. 14). The trainee has to touch the tip of each peg with the tip of a 5 mm × 33 mm MIS spatula with an unlimited time in the following order: start → 0 → 1 → 0 → 2 → 0 → 3 → 0 → 4.

CAST emits a sound when the corresponding target has been touched by the spatula's tip. This auditory aid is used by the trainee to recognize when a target has been reached and therefore to continue executing the next movement in the sequence.

We used MATLAB Fuzzy Logic Toolbox 2 [11] for the implementation of the scoring system. Fuzzy Logic Tool Box 2 is a collection of functions built in the MATLAB numeric computing environment that provides tools to create and edit fuzzy inference systems [10].

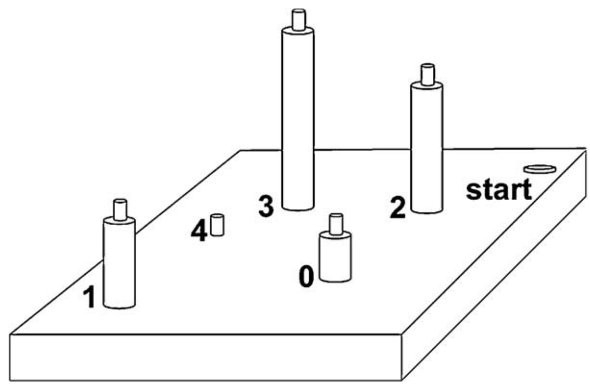

Fig. 14. Hand–eye coordination task composed by five pegs and a starting point. Each peg has a different height. The tip of each peg corresponds to a different target in the exercise.

5.2. Scoring system implementation

The CAST Scoring System is a five input, one output Mamdani fuzzy model with 20 rules developed with the Matlab Fuzzy Logic Toolbox 2. The inference process is performed automatically by MATLAB. However, this had to be programmed with some specifications shown in Table 2.

MAMDANI fuzzy inference system: Considers a fuzzy output. In the assessment model the outputs are fuzzy sets representing proficiency levels: *Novice*, *Beginner*, *Proficient* and *Expert*.

MIN implication method: In this method, the output membership function is truncated at the height corresponding to the rule's antecedent computed degree.

SUM aggregation method: In this method, the aggregated output fuzzy subset is constructed by taking the pointwise sum over all of the fuzzy subsets assigned to the output variable by the inference rules.

CENTROID defuzzification method: In this method, the crisp value of the output variable is computed by finding the center of gravity of a continuous aggregated fuzzy set. The centroid y_i of B is given by:

$$y' = \frac{\int_s y_i \mu_B(y)dy}{\int_s \mu_B(y)dy} \tag{9}$$

A Mamdani model lies at the core of the CAST scoring system. The Mamdani type is one of the most frequently used fuzzy logic structures, characterized by handling rules with fuzzy consequents. As opposed to other fuzzy models, such as Sugeno type, which handles crisp outputs, Mamdani systems demand processing a greater computational load to defuzzify consequents. Nevertheless, working with fuzzy consequents facilitates eliciting expert knowledge in a more transparent and intuitive manner. Experts do not have to provide crisp values or a mathematical equation to describe outputs for Mamdani systems.

CAST fuzzy system features membership function and output sets with triangular and trapezoidal forms with a maximum degree of membership of 1 and a minimum of 0. Rules were designed by matching the fuzzy terms obtained in each variable's semantic decomposition (Negative, Weak Positive, Moderate Positive and Strong Positive) with their corresponding proficiency level (Novice, Beginner, Proficient and Expert). The rules' antecedents have one proposition formed by a variable (performance metric) and one fuzzy term. The rules' consequents have one proposition formed by a variable (proficiency level) and one fuzzy term. A single-input, single-output rule design was preferred because of the difficulty experienced while trying to elicit rules from an expert surgeon that involved more than one proposition in their antecedent and/or consequent. The expert's struggle to articulate rules with more than one antecedent and consequent indicates that the judgment process occurs considering one proposition at the time and this concept is captured in our scoring system. The system's rulebase consists of a total of $v \times m$ rules where v is the number of variables, and m is the number of proficiency levels. For example, for the variable *continuity of movement* we have the following set of rules:

If *continuity of movement* is Strong Positive then Proficiency level is expert.
If *continuity of movement* is Moderate Positive then *Proficiency level* is proficient.
If *continuity of movement* is Weak Positive then Proficiency level is beginner.
If *continuity of movement* is Negative then Proficiency level is novice.

Each rule has an associated weight that can take values from 0 to 1. The weight determines the rule's impact on the final score. For the hand–eye coordination task previously described in Section 5, each rule was provided with the same weight. The rule's weight is related to the variable's degree of relevance. This degree of relevance was determined by an expert surgeon and depends on the task being evaluated since different tasks might require different abilities at different levels.

The fuzzy output is composed by all the possible consequents of the rules in our rule base i.e., the four proficiency levels *Novice*, *Beginner*, *Proficient* and *Expert* (Fig. 15). We distributed the membership functions evenly on a range of values going from −33.33 to 133.32. We followed the same criteria for constructing the membership functions used in the construction of the input fuzzy sets described in Section 4.2.

Although the fuzzy output sets range from −33.33 to 133.32, the maximum score a subject might achieve is 99.99 and the minimum is 0. This is due to the centroid defuzzification method.

5.3. Experimentation and results

A total of 38 subjects participated in this study. Subjects were distributed in five groups according to their expertise in MIS, 17 non-medical students, 11 medical students without previous laparoscopic surgery training, 5 medical students with some laparoscopic surgery training, 4 medical residents and 1 expert surgeon. The hand–eye coordination task presented in Section 5 was performed 8 times by each subject. In total 304 samples were used in this study. Subjects were asked to use only their dominant hand (left or right) to perform each of eight trials. For each subject four trials (odd trials) were used in the system's knowledge base while the other four (even trials) were used for testing purposes.

Subjects' identification numbers were provided according to the five groups of expertise: non-medical students were assigned identification numbers within the range of [1000, 1999], medical students with no laparoscopic training [2000, 2999], medical students with laparoscopic training [3000, 3999], residents [4000, 4999] and surgeons [5000, 5999].

Fig. 16 shows the average score per subject derived from the 4 trials selected for testing purposes. Groups are divided with dashed vertical lines.

Table 2
Specifications underlying the CAST assessment system.

Fuzzy inference system	Mamdani
Implication method	Min
Aggregation method	Sum
Defuzzification method	Centroid
Membership functions	Triangular–trapezoidal [0, 1]

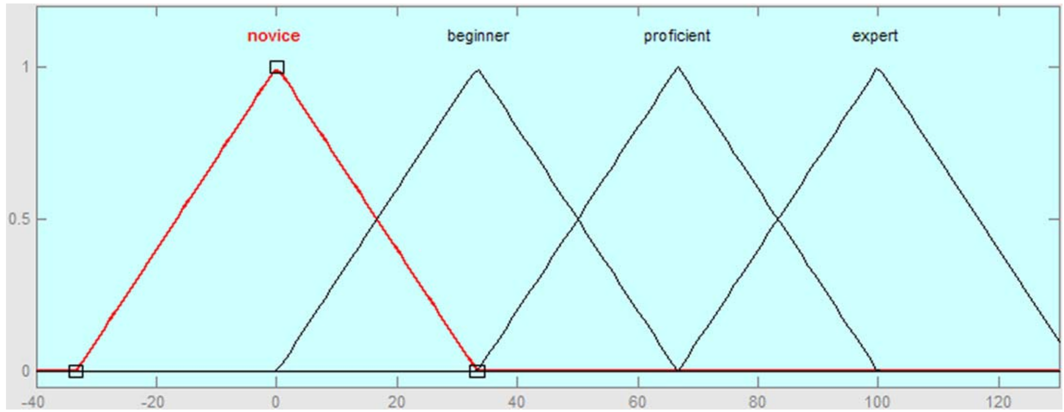

Fig. 15. CAST output fuzzy sets depict four proficiency levels. Outputs are normalized such that holding a membership value of 1 in the novice or expert sets results in a score of 0 and 100 respectively after defuzzification.

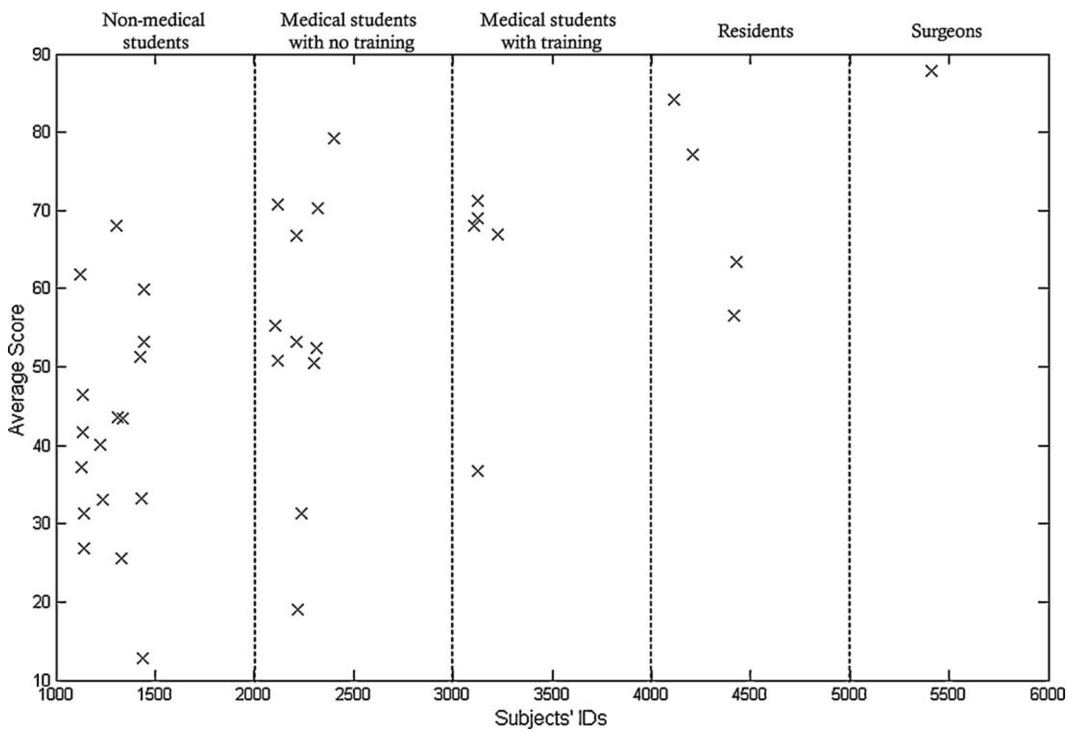

Fig. 16. Average score plot. Identification numbers were assigned to subjects according to their previous MIS training higher id's correspond to greater previous training.

The average scores for each group are shown in Table 3. There is a clear pattern of increasing score according to each group's level of expertise.

6. Discussion

The presented framework holds great potential as it shows scalability features in three key aspects for the constant improvement of an objective scoring system. Those three aspects are as follows:

Integration of new experts' knowledge: Contributions from other expert surgeons can be easily integrated with the system if proficiency levels are defined as before, i.e., by asking, based on performance data, what percentages from a selection of subjects with different surgical skills belong to each proficiency level? For example, if expert 1 defines novices as the bottom 30% and expert 2 defines novices as the bottom 40%, these two definitions can be averaged to reconcile their difference.

Integration of new evaluation metrics: The integration of new performance metrics can also be done in a straightforward manner. The construction of membership functions can be derived from the performance data already recorded on the system. If necessary, for each new metric four new rules need to be added to the system's rule base with consequents that correspond to the four proficiency levels previously defined in this work. This feature is relevant in the cases of task dependent metrics where different tasks demand to compute a different set of metrics.

Table 3
Mean score for each subjects' group participating in our study.

Group	Score
Non-medical students	41.7369
Medical students with NO MIS training	54.5209
Medical students with MIS training	62.3915
Residents	70.32374
Surgeons	87.8574

Integration of new performance data: The constant recording of subjects' performance data and membership function regeneration after deployment gives the system the potential to be adaptable to its trainees' pace of improvement or to new, higher skilled users. We derived fuzzy membership functions through the location of transitional points between proficiency levels defined as percentages. The locations of the transitional points change according to the number of samples used for the generation of membership functions. The greater the number and diversity of the samples, the more accurate the system will be. To offset the subjectivity inherited from relying solely on the data of the subjects used during the system's design phase, the samples of new subjects that exceed a limit of a metric range (i.e., its maximum or minimum value registered in the system) should be considered for its integration with the system knowledge base, and therefore, for the recalculation of new transitional points.

The scalability features of the presented elicitation method are a step forward toward a robust and objective system for the assessment of laparoscopic surgical skills.

7. Conclusions

This work demonstrates how to objectively measure proficiency levels in minimal invasive surgical skills by computing motion tracking data and modeling expertise judgment as an inference system based on fuzzy logic. We conceive the MIS assessment process as a function of two elements the abstract theoretical knowledge of experience surgeons of what constitute competitive performance and the performance data of subjects with a wide diversity of MIS experiences. The presented knowledge elicitation framework can easily accommodate system's enhancements in performance metrics, and expertise knowledge while rising assessment standards according to trainees' acquired skills.

This work also introduced a new performance metric *continuity of movement* to the set of metrics employed in the assessment of laparoscopic surgical skills. This metric can be used not only in simple hand–eye coordination tasks but also in more complex laparoscopic procedures.

References

[1] M.A. Meyer, K.B. Butterfield, W.S. Murray, R.E. Smith, J.M. Booker, Guidelines for eliciting expert judgement as probabilities or fuzzy logic Fuzzy Logic and Probability Applications American Statistical Society, SIAM Series, Los Alamos National Laboratory, 2002.
[2] R.K. Reznick, Teaching and testing technical skills, The American Journal of Surgery 65 (3) (1993) 358–361.
[3] L.W.T. Schuwirth, C.P.M. van der Vleuten, Medical education challenges for educationalists, BMJ 333 (2006).
[4] C. Feng, H. Haniffa, J.W. Rozenblit, J. Peng, A.J. Hamilton, M. Salkini, Surgical training and performance assessment using a motion tracking system, in: Proceedings of the 2nd European Modeling and Simulation Symposium, EMSS, 2006, pp. 647–652.
[5] R.K. Reznick, S. Smee, J.S. Baumber, R. Cohen, A. Rothman, D. Blackmore, M. Bérard, Guidelines for estimating the real cost of and objective structured clinical examination, Academic Medicine 68 (7) (1993) 513–517.
[6] A. Darzi, S. Mackay, Assessment of surgical competence, Quality in Health Care 2 (2001).
[7] N. Stylopoulos, S. Cotin, S.K. Maithel, M. Ottensmeyer, P.G. Jackson, R.S. Bardsley, P.F. Neumann, D.W. Rattner, S.L. Dawson, Computer-enhanced laparoscopic training system: bridging the gap, Surgical Endoscopy 18 (5) (2004).
[8] T. Bilgic, B. Turksen, Measurement of membership functions: theoretical and empirical work, in: D. Duboir, H. Prade (Eds.), Fundamentals of Fuzzy Sets, Kluwer Academic Publishers, Massachusetts, 2000, pp. 195–228.
[9] I. Hajshirmohammadi, S. Payandeh, Fuzzy set theory for performance evaluation in a surgical simulator, Presence 16 (6) (2007) 603–622.
[10] S.J. Hamstra, A. Dubrowski, D. Backstein, Teaching technical skills to surgical residents – a survey of empirical research, Clinical Orthopaedics and Related Research 449 (2006) 108–115.
[11] The Mathworks, Fuzzy Logic Toolbox 2, User's Guide, MATLAB, 2007.
[12] A.D. Kulkarni, University of Texas. Computer Vision and Fuzzy-Neural Systems, first ed., Prentice Hall, New Jersey, 2001.
[13] C. Feng, J. Rozenblit, A. Hamilton, Fuzzy logic-based performance assessment in the virtual, in: Assistive Surgical Trainer (VAST) in Proceedings of the ECBS Conference, 2008, pp. 203–209.
[14] A. Piegat, Fuzzy Modeling and Control. Studies in Fuzziness and Soft Computing, first ed., Physica-Verlag A Springer-Verlag Company, Heidelberg, 2001.
[15] L.A. Zadeh, Fuzzy sets, Information and Control 8 (1965) 338–353.
[16] K. Moorthy, Y. Munz, S.K. Sarker, A. Darzi, Objective assessment of technical skills in surgery, BMJ 327 (2003) 1032–1037.
[17] K.A. Ericcson, M.J. Prietula, E.T. Cokely, The making of an expert, Harvard Business Review (2007).
[18] J. Rosen, J.D. Brown, M. Barreca, L. Chang, B. Hannaford, M. Sinanan, The blue DRAGON – a system for monitoring the kinematics and the dynamics of endoscopic tools in minimally invasive surgery for objective laparoscopic skill assessment, Studies in Health Technology and Informatics – Medicine Meets Virtual Reality 85 (2008) 412–418.