

# EVA: Endoscopic Video Analysis of the Surgical Scene for the Assessment of MIS Psychomotor Skills

I. Oropesa P. Sánchez-González J.A. Sánchez-Margallo J. García-Novoa  
F.M. Sánchez-Margallo E.J. Gómez

*Abstract*—The present work covers the first validation efforts of the EVA Tracking System for the assessment of minimally invasive surgery (MIS) psychomotor skills. Instrument movements were recorded for 42 surgeons (4 expert, 22 residents, 16 novice medical students) and analyzed for a box trainer peg transfer task. Construct validation was established for 7/9 motion analysis parameters (MAPs). Concurrent validation was determined for 8/9 MAPs against the TrEndo Tracking System. Finally, automatic determination of surgical proficiency based on the MAPs was sought by 3 different approaches to supervised classification (LDA, SVM, ANFIS), with accuracy results of 61.9%, 83.3% and 80.9% respectively. Results not only reflect on the validation of EVA for skills' assessment, but also on the relevance of motion analysis of instruments in the determination of surgical competence.

*Keywords*—MIS, competence, assessment, EVA, TrEndo.

## I. INTRODUCTION

Skills' acquisition in minimally invasive surgery (MIS) is gradually adapting from the mentor-apprentice-based Halstedian paradigm towards structured, objective training and assessment programs, where direct involvement of residents in real surgeries is delayed until becoming proficient in the required skills. Several motivators can be identified behind this: media and public awareness towards medical errors, the need to reduce costs in hospitals, or the overloaded schedules of surgeons [1].

In this context, the first stages of basic psychomotor training take place in controlled laboratory settings by means of box trainers and virtual reality simulators [2]. The incorporation of tracking technologies allows them to capture data on the laparoscopic instruments' movements when performing an exercise. This data, when handled properly, may yield a series of motion analysis parameters (MAPs) providing useful, objective information on performance [3].

Tracking technologies for box trainers and VR simulators have traditionally relied on sensor-based systems, based on optical, electromagnetic or mechanic technologies [4]. However, their use may modify the ergonomics and constrain movements of the instruments, altering the users' experience and performance. Moreover, transfer of these

technologies for training and assessment of skills in the OR is compromised as they are often bulky, are not easily sterilized, may require a clear line of sight (optical devices) or be affected by ferromagnetic materials (electromagnetic devices), and thus present errors in position tracking [5].

The present work covers the first validation efforts of the EVA Tracking System for the assessment of MIS psychomotor skills. In its current implementation, the tracking algorithm allows offline determination of the 3D position of the laparoscopic instruments with respect to the endoscope, based solely on their physical and geometrical characteristics. The paper will briefly describe the working principles of the algorithm, as well as some of the findings of this technology when applied to a real training scenario.

## II. MATERIALS AND METHODS

### A. Working Principles

Tracking of the instruments in EVA works under the following assumptions:

- Performing color-based segmentation and applying an edge filter, instruments may be isolated from the video frame. Ideally, a clean image showing the instrument edges yields two neighboring peaks in Hough space. Detecting these maxima provides information regarding borders' position on the screen [6].
- In case of there being two instruments, it is typical in a box trainer setting that they will have opposing entrance angles. In this case, a division of Hough space for positive and negative values of theta can isolate detection of the left and right instruments respectively.
- Detection of the instrument's tip in the screen can be performed by color-based gradient analysis along its bisecting line (obtained from borders' information) [7].
- 3D reconstruction of the tip and orientation information is obtained by means of the geometrical properties of the instrument [8].

### B. Data Acquisition and Preparation

A series of experiments were conducted to validate the use of EVA for the assessment of MIS psychomotor skills.

Forty-two participants (4 expert surgeons, 22 residents, 16 novice medical students) performed a box trainer task at the skillslab of the Leiden University Medical Centre (LUMC, Leiden, The Netherlands), where the goal was to place a number of chickpeas into different target holes using one laparoscopic grasper. Video recordings of performance were taken from the laparoscopic feed for offline analysis with EVA. Simultaneously, instrument movements were recorded with an optical sensor-based tracking system developed at the Delft University of Technology (TUDELFT, Delft, The Netherlands), the TrEndo [9]. Nine MAPs were recorded using both tracking systems (Table 1).

Table 1 MAPs definition.

MAP	Definition
Time	Total time to perform a task (s).
Idle time	% of time where the instrument is considered to be still (speed < 5mm/s).
Path length	Total path covered by the instrument (m).
Depth	Total path covered in the instrument's axis direction (m).
Average speed	Rate of change of the instrument's position (mm/s).
Average acceleration	Rate of change of the instrument's velocity (mm/s <sup>2</sup> ).
Motion smoothness	Jerky movements caused by abrupt changes in acceleration (m/s <sup>3</sup> ).
Economy of area	Ratio between maximum area covered by the instrument on the task surface and path length (-).
Economy of volume	Ratio between maximum volume covered by the instrument on the setting and path length (-).

Since the purpose of the experiment was to validate the task for evaluation purposes, no prior trials were allowed. A brief explanation of the task's objectives was given to participants to let them infer, based on their own experience and skills, the best strategy to perform it.

### C. Construct and Concurrent Validation

Construct validation was performed along the three experience groups (Kruskal-Wallis analysis) and in pairs (Mann-Whitney test) to measure statistical significant differences between MAPs. Differences were considered significant at  $p < 0.05$ .

Concurrent validation was performed between EVA and TrEndo. Pearson's correlation ( $\rho$ ) was employed to measure the degree of concurrence. Values between 0.4-0.7 were considered as medium correlating values, whilst values  $> 0.7$  showed strong correlation between MAPs [10].

### D. Supervised Classification of Performance

Three supervised classification techniques were applied to discern whether surgical experience may be derived from

performance in box trainer tasks. Participants were arranged according to the number of surgeries performed: more than 10 (Experienced, Ex: 4 experts + 14 residents) and less than 10 (Non-experienced, NEx: 16 students + 8 residents). Based on the 9 MAPs, classifiers yielded a binary output according to performance: {S: skilled, NS: not skilled}.

MAPs were normalized and principal component analysis (PCA) applied to reduce the number of input dimensions. Three different classifiers were tested: linear discriminant analysis (LDA), support vector machines (SVM) and artificial neuro fuzzy inference systems (ANFIS).

Leave one out cross validation was performed, in which data from all subjects was employed to train the classifiers except for one, who was used for validation. The process is repeated until each subject has been used both for training and testing. The following parameters were sought:

- *Accuracy*: % of subjects correctly classified according to the input categories.
- *Sensitivity (S)*: % of Ex classified as S.
- *Specificity (E)*: % of NEx classified as NS.
- *RMSE*: Main error measurement between expected (Ex, NEx) and predicted (S, NS) values per classifier.

Additionally, receiver operator curves (ROC) were obtained, plotting specificity vs. sensitivity based on the posterior probability of each classifier.

Significant differences between the three classifiers were sought by means of Cochran's Q test ( $p < 0.05$ ) to determine the degree with which each classifier is coherent with the rest when evaluating a participant. For a more intuitive representation of this idea, classifier plots were made for each technique, in order to show the expected and predicted values, as well as their posterior probability.

Finally, significant differences between classification results obtained by using TrEndo and EVA for data acquisition were measured by means of McNemar's test ( $p < 0.05$ ).

## III. RESULTS

### A. Construct and Concurrent Validation

Figure 1 graphically reflects results for construct and concurrent validation. Construct validation was obtained along the three groups for time, path length, depth, average speed, average acceleration and economy of area/volume.

Paired comparisons showed significant differences for all valid MAPs between students and residents/experts except for average speed, where differences occurred only between novice medical students and residents.

For concurrent validation, all MAPs except motion smoothness obtained values of  $\rho > 0.7$ , reflecting a strong correlation between values obtained with TrEndo and EVA.

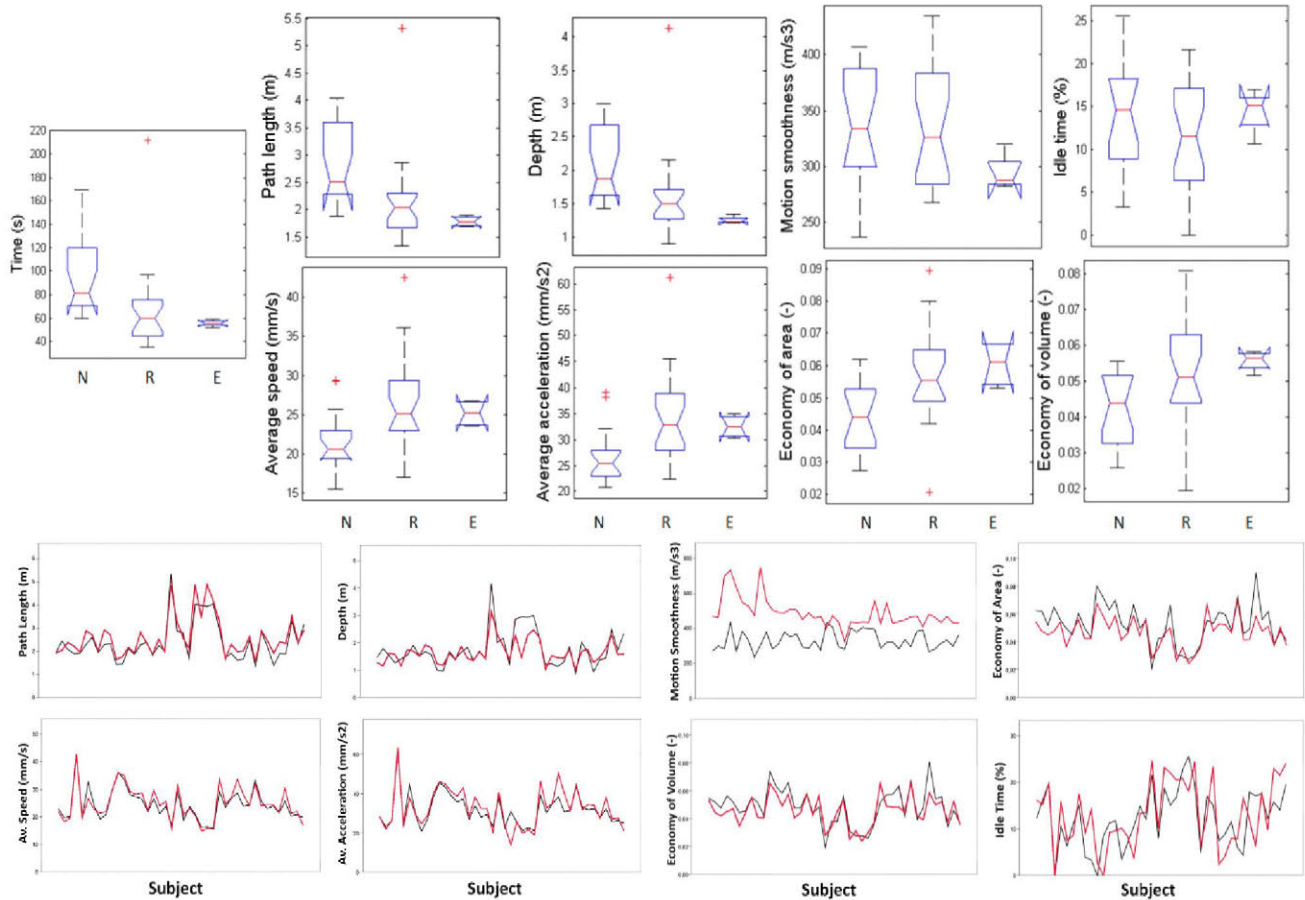


Fig. 1 Top: Construct validation: Results are expressed as notched box diagrams, in which every box distinguishes lower, median and upper quartile value. Significance is shown where the notched sections of the boxes do not overlap each other. N: Novice medical students; R: Resident; E: Experts. Bottom: Concurrent validation scores per subject, in order of participation (x-axis). Black: Scores obtained by EVA. Red: Scores obtained by TrEndo.

### B. Classification Results

SVM and ANFIS showed the highest accuracy rates (Fig.2; Table 2) at 83.3% and 80.9% respectively, and a better identification of NEx subjects (91.7% and 87.5% respectively). On the other hand, the three classifiers performed similarly on classification of Ex surgeons (72.2%). Differences according to Cochran's test were considered significant ( $p = 0.02$ ). Visual inspection of classifiers' plots (Fig. 3) and accuracy results point at the higher number of misclassifications for LDA as possible cause. McNemar test for each classifier revealed no significant differences between using the EVA and TrEndo data.

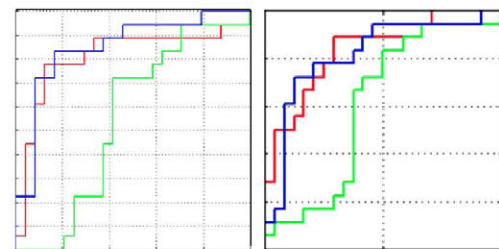


Fig. 2 ROC curves. Left: EVA. Right: TrEndo. Green: LDA; Blue: SVM; Red: ANFIS. X-axis: (1-E), Y-axis: S.

Table 2 EVA classification results for the proposed classifiers (boldface). In parenthesis, homologous results obtained by TrEndo are given.

Classifier	Accuracy	RMSE	S	E
LDA	<b>61.9</b> (61.9)	<b>0.62</b> (0.62)	<b>72.2</b> (66.7)	<b>54.2</b> (58.3)
SVM	<b>83.3</b> (76.2)	<b>0.41</b> (0.48)	<b>72.2</b> (72.2)	<b>91.7</b> (79.2)
ANFIS	<b>80.9</b> (76.2)	<b>0.44</b> (0.48)	<b>72.2</b> (66.7)	<b>87.5</b> (83.3)

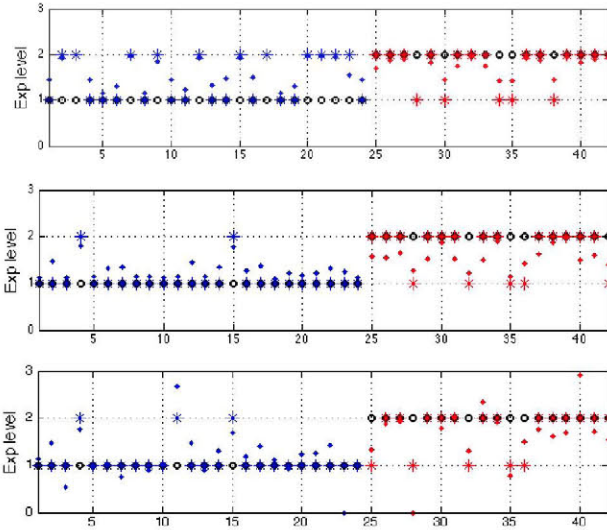


Fig. 3 Classifier plots: Top to bottom: LDA, SVM, ANFIS. Default classes (NEx=1; Ex=2) are shown as circles (o). Output levels (NS=1; S=2) are defined. Probabilistic outputs are shown as dots (\*). A threshold for classification is set at 1.5, results given as asterisks (\*). Blue: NEX; Red: EX

#### IV. DISCUSSION

This study covers the first validation efforts of the EVA Tracking System applied to MIS psychomotor skills' assessment. Results not only give proof of its usefulness for this purpose, but serve as confirmation on the importance of motion analysis when determining surgical competence.

Construct validation was established for 7 out of 9 possible MAPs. To a greater or lesser extent most of them have been featured in the literature and validated for different tasks and abilities, and thus this study helps corroborate their relevance for assessment purposes [2]. More importantly, concurrent validation was established with data registered by the gold standard system employed, the TrEndo. In this sense, only motion smoothness did not present significant differences between systems. A possible reason for this may reside in the post-processing stage of EVA, which effectively applies low pass filtering of the signal that dampens the influence of movements' jerkiness.

The relevance of motion analysis manifests also in the correlation between experience and expertise detected by

the supervised classifiers. While it is true that SVM and ANFIS performed better than LDA for this specific task, the former fact is more important at this point of our research rather than finding an optimal classifier (if indeed there is one). However, certain trends in the data (e.g.: subjects misclassified for 2 or 3 classifiers) suggest that other factors besides prior experience may be conditioning performance of the task, whether subject- (musical aptitudes, stress, etc.) or setting-related (box trainer, endoscope position, etc.).

#### V. CONCLUSIONS

The EVA Tracking System has been proven valid for assessment of MIS psychomotor skills. The current MATLAB implementation is being migrated to C++/OpenCV to increase processing speed and allow for real time tracking, and research continues in order to increase robustness of the segmentation stages. Our final goal is to achieve a system combining real time tracking and intelligent data analysis to provide immediate, relevant feedback on performance in a stress-free, patient-safe environment.

#### ACKNOWLEDGMENTS

Authors kindly thank the Minimally Invasive Surgery and Interventional Techniques group (TUDELFT) for their support with TrEndo, as well as the LUMC staff for their support during the validation experiments. Authors participate under partial funding of LLP-Leonardo da Vinci project MISTELA (528125-LLP-1-2012-1-UK).

#### REFERENCES

1. Tsuda S, Scott D, Doyle J et al. (2009) Surgical Skills Training and Simulation. *Curr Prob Surg* 46: 271-370.
2. Oropesa I, Sanchez-Gonzalez P, Lamata P et al. (2011) Methods and Tools for Objective Assessment of Psychomotor Skills in Laparoscopic Surgery. *J Surg Res* 171: e81-e95.
3. van Hove PD, Tuijthof GJM, Verdaasdonk EGG et al. (2010) Objective assessment of technical surgical skills. *Brit J Surg* 97: 972-987.
4. Chmarra MK, Grimbergen CA, Dankelman J (2007) Systems for tracking minimally invasive surgical instruments. *Minim Invasive Ther Allied Technol* 16: 328-340
5. Peters TM (2006) Image-guidance for surgical procedures. *Phys Med Biol* 51(14): R505-40
6. Oropesa I, Sánchez-Margallo JA, Sánchez-González P et al. (2012) Comparative study of two video-based laparoscopic instrument tracking algorithms for image guided surgical applications. *Proc. II Joint Workshop New Technologies for Computer/Robot Assisted Surgery*.
7. Allen BF, Kasper F, Nataneli G et al. (2011) Visual tracking of laparoscopic instruments in standard training environments. *Stud Health Technol Inform* 163: 11-17.
8. Cano AM, Lamata P, Gayá et al. (2006) New methods for video-based tracking of laparoscopic tools. In: Harders M, Székely G (eds) *ISBMS 2006, LNCS, vol 4072*. Springer, Heidelberg, 142-149.

9. Chmarra MK, Bakker NH, Grimbergen CA et al. (2006) TrEndo, a device for tracking minimally invasive surgical instruments in training setups. *Sens Actuat: A-Physical* 126: 328-334.
10. Aggarwal R, Grantcharov T, Moorthy K et al. (2007) An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Ann Surg* 245: 992-999.

Corresponding author:

Author: Ignacio Oropesa

Institute: Biomedical Engineering and Telemedicine Centre, ETSI Telecomunicación, Universidad Politécnica de Madrid

Street: Avda. Complutense, 30

City: Madrid

Country: Spain

Email: [ioropesa@gbt.tfo.upm.es](mailto:ioropesa@gbt.tfo.upm.es)