2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Machine Learning based Classification of Local Robotic Surgical Skills in a Training Tasks Set

L. Juarez-Villalobos[1], N. Hevia-Montiel[2] and J. Perez-Gonzalez[2]

Abstract— **During surgical training, it is important for the surgeon develops good motor skills throughout his training. For this reason, various surgical training systems have been developed to enhance these skills. However, one of the great challenges in these training systems is being able to objectively measure the ability and performance of the main surgical tasks, where currently only a global measurement is obtained once the task is completed. In this work, a temporal evaluation scheme is proposed, that is, an evaluation of local surgical performance at different time intervals during the training of typical tasks (knot-tying, needle-passing and suturing). The goal is to automatically classify expert (experience >100 hrs) and non-expert (experience <10 hrs) surgeons according to their performance during training, based on three classifiers: K-Nearest Neighborhood, Random Forest, and Support Vector Machine Unlike other previously reported methods, this work proposes a new evaluation scheme based on segments or time intervals, which can be an indicator of the surgeon's local performance during a robotic surgical task, without the need for direct labeling of the data at the segment level. The classification performance from obtained results was in accuracy 83% to 100%, 88% to 100% of AUC-ROC, and 88% to 100% of F1-Score in the final test between experts and non-experts surgeons, where the Support Vector Machine classifier presented the best performance. These results suggest that this proposed method by time intervals could be used in various surgical trainers to evaluate the local performance of a surgeon during trainingand thus be able to provide a tool for the quantitative visualization of opportunities to improve surgical skills.**

Clinical relevance— **We consider that the proposed method to carry out a local performance evaluation during surgical training can provide useful information in the learning and improvement of surgical skills.**

I. INTRODUCTION

The surgical training, whether apprentice or expert, presents obstacles in its training process, for example, by overcoming the start of learning, economic pressures, time, pressure on the least amount of errors to avoid putting patients at risk in actual surgery. Therefore, it is necessary to develop training systems with artificial reproduction that consider the visualization, manipulation of the instrument, and spatial orientation. In some cases, simulation of complete surgical procedures [1].

While the development of simulation and task planning systems is a work of great interest today, objective methods for the measurement of surgery skills are still researched [2]. That is why the main interest of the proposed work is developing a system to measure performance in set of robotic surgery tasks.

In the reported paper by Evans et al. [3], wireless inertia sensors are used to evaluate surgical skills in a laparoscopic surgery simulator. Their findings suggest that the proposed metrics can be used to generate a score for a given laparoscopic simulated task. Pérez et al. [4], performed three tasks in the EndoViS training system. Motion data from the instruments were captured with a video tracking system integrated into the EndoViS simulator, using 13 Motion Analysis Parameters (MAP). Three classifiers were trained: Radial-Based Function Networks (RBFNets), K-star (K*), and Random Forest (RF) to classify participating surgeons. They found that K* method showed the best performance in the expert and non-expert surgeons' classification.

In the work proposed by Siyar et al. [5], the K-Nearest Neighbors (KNN) classifiers, Parzen window, Support Vector Machine (SVM), and Fuzzy K-NN were trained. The obtained results show similar performances among the classifiers; however, the SVM method presented the best performance. So, this proposed approach could be implemented to improve the surgeons skills during a surgery simulator training.

These three described studies showed a high automatic classification performance according to the level of experience between expert and non-expert surgeons; however, all these methods focus on the global surgical task performance. In this context, a global performance is a binary classification metric that indicates whether person is an expert or a non-expert at the end of a surgery simulator training.

A gesture-based analysis was proposed by Vedula et al. [6]. They fit logistic generalized estimation equations models to classifier the skill level (expert vs. novice). For this analyses, they studied the close incision task divided in the following maneuvers: suture throw, 2-loop knot, 1-loop knot, and several gestures required to complete the task. However, they only classify between expert and novice surgeons at the level of task, maneuvers and gestures.

Unlike these reported works, the main objective of this research is to give a local performance metric during the robotic surgery tasks. In this way, we would have a local performance metric by time intervals (segments) as the task is carried out, that is not only at the end of a training task; allowing us to know about error behavior in each time interval.

*This work was supported by UNAM-PAPIIT IT100220 and IA102920.

[1]L. Juarez-Villalobos is with the Postgraduate Program in Computer Science and Engineering. Yucatán Campus. Universidad Nacional Autónoma de México. México. luis.juarez.villalobos@comunidad.unam.mx

[2]N. Hevia-Montiel and J. Perez-Gonzalez are with the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Km 5.5 Carretera Sierra Papacal - Chuburna, Mérida Yucatán 97302, México. nidiyare.hevia@iimas.unam.mx, jorge.perez@iimas.unam.mx

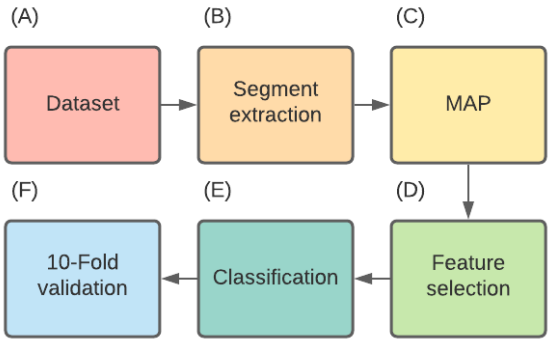

Fig. 1. Proposed methodological diagram.

The proposed method is based on three supervised classifiers that use a set of motion characteristics; these features were previously selected through a statistical analysis. These classification methods will assign a binary performance label (expert or non-expert) both for the classifier training and during the local segment analysis. The proposed methodology, as well as the results and discussion are presented in the following sections.

II. MATERIALS AND METHODS

The pipeline proposed for data processing, classification and validation is presented in figure 1, which is consistent with the following subsections.

A. Dataset

The dataset used in this work was JIWSAWS [7], which consists of kinematic and video information from eight surgeons of different skill levels. They performed five repetitions of three elementary surgical tasks on a table model using a da Vinci robotic surgical system.

The tasks included in the JIWSAWS database are: suturing, knot-tying, and needle-passing, which are standard components of most surgical skills training curricula. In addition, the dataset includes manual annotations of surgical gestures for each task and surgical skills using global scores.

1) Suturing: The subject takes the needle, enters the incision (designated as a vertical line in the desktop model), passes the needle through the "tissue", enters the marked point on one side of the incision, and exits the side of the marked incision at the corresponding point on the other side. After the first puncture, the subject removed the needle from the tissue, passed it to his right hand, and then repeated it three times.

2) Knot-tying: The subject tied one end of the suture to a flexible tube, and one end of the flexible tube was connected to the surface of the desktop plaster. Make a simple knot.

3) Needle-passing: The subject raised the needle (not captured in the video in some cases) and passed four small metal rings from right to left. These rings are connected to a short height above the surface of the desktop model.

B. Segment extraction

The dataset, contains the recorded kinematic data of the surgery instruments for each task. From the kinematic data, the x, y and z axes displacements are acquired during training. To obtain a local performance metric, it is proposed to divide each performed surgery task (knot-tying, needle-passing and suturing) into overlapping temporal segments, using temporal shifts of samples between the current and the next segment.

Figure 2 shows a representative example of segments extraction of a kinematic variable.

C. Motion analysis parameters (MAP)

The kinematic data of each segments were analyzed using ten MAP described in table I. These parameters were calculated from the displacements $x[n]$, $y[n]$ and $z[n]$ of the instruments. The parameters were calculated using software developed in Python 3.7. In this work, only right-hand MAP values were considered as input vectors of classifiers. Mathematical explanation of motion analysis parameters are presented in Perez et al. [4].

D. Feature selection

In order to determine which MAP (table I) can be used to train and improve the performance of three classifiers, the Mann-Whitney U test was performed to obtain MAP that show statistically significant differences between surgeons with more than 100 hours of experience (experts) and surgeons with less than 10 hours of experience (non-expert). A probability of ($p <= 0.05$) was considered statistically significant.

E. Classification

The proposed algorithm is based on expert and non-expert automatic classification during three training tasks

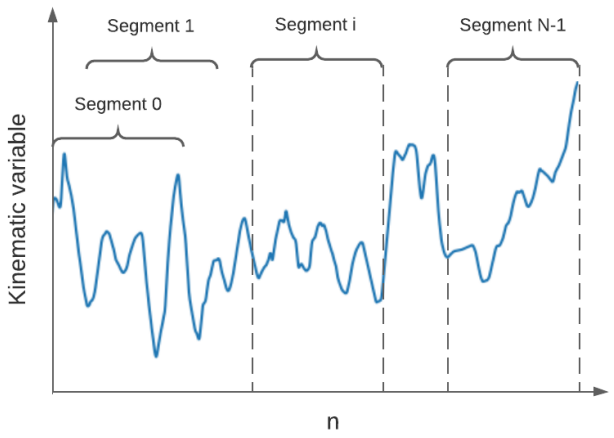

Fig. 2. A representative example of the segment division of a kinematic variable (displacements in $x[n]$, $y[n]$ or $z[n]$), acquired during training. N represents the total number of segments, n is each sample of the vector, and i is an example of an intermediate segment.

TABLE I

DESCRIPTION OF THE USED METRICS (MAP) IN THE CLASSIFICATION.

Metrics	Definition
AX	Average position on X-axis (mm)
AY	Average position on Y-axis (mm)
AZ	Average position on Z-axis (mm)
AVX	Average velocity on X-axis (mm/s)
AVY	Average velocity on Y-axis (mm/s)
AVZ	Average velocity on Z-axis (mm/s)
AAX	Average acceleration on X-axis (mm/s^2)
AAY	Average acceleration on Y-axis (mm/s^2)
AAZ	Average acceleration on Z-axis (mm/s^2)
Path length	Total route followed by the instrument (mm)
Depth perception	Total distance traveled along the axis (mm)
Motion smothness	Abrupt changes of the acceleration (mm/s^3)
Average velocity	Change rate of instrument position (mm/s)
Average acceleration	Instrument velocity change rate (mm/s^2)

using MAP variables. The database provides us with the tasks labeled if performed by an expert surgeon or a non-expert surgeon, which we use to infer performance over time (for each segment analyzed). In this way, a binary label was assigned to each extracted segment, according to the experience level (label equal to -1 for the non-expert class and +1 for the expert class). As classification methods, it was proposed to use three classifiers described below:

1) KNN: Is a nonparametric and supervised regression and classification method. KNN is used as a method to classify items by training short-distance examples in element space. KNN is a type of learning in which the function only approaches locally, and all calculations are postponed for classification [8].

2) RF: It is a combination of predictor trees. Each tree depends on the values of a random vector tested independently and with the same distribution for each of these. It is a substantial modification of bagging that builds a long collection of uncorrelated trees and then averages them [9].

3) SVM: It constructs a hyperplane or set of hyperplanes in a space of very high (or even infinite) dimensionality that can be used in classification or regression problems [10]. In this work a linear kernel was used.

F. Validation

To evaluate the classifiers performance, the mode of all the classified segments were calculated and then compared versus the global labels (expert or non-expert), previously assigned in the database. This was done for each robotic surgical task.

The database was divided into 70% training data and 30% data for the final test. With the training data, a 10-fold validation process was performed for the three tasks analyzed. Accuracy (ACC), F1-Score, and Area Under the Receiver Operating Characteristics (AUC-ROC) metrics were used to measure the classification performance between expert and non-expert surgeons. The obtained results are presented in the following section.

III. RESULTS AND DISCUSSION

The trajectory of an expert and a non-expert surgeon for knot-tying task is shown in figure 3. A clear difference can

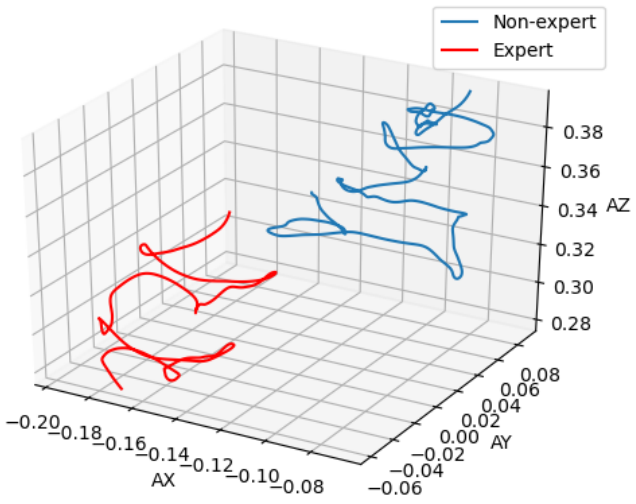

Fig. 3. Knot-tying task instrument trajectory of an expert and a non-expert surgeons.

be observed between the trajectories of expert surgeons (red) and non-Expert (blue) according to displacements on AX, AY and AZ axis. This agrees with the Mann-Whitney U test, in which statistically significant differences were found ($p <= 0.05$) between all MAP analyzed. For this reason, all MAP were used for classifiers training.

The classification results are presented in table II. In first column, three analyzed tasks are presented; the second column shows each implemented classifier. The first section corresponds to cross-validation and the last section of the table is the final test. Regarding knot-tying task, it can be noted that in the three classifiers (KNN, RF and SVM) 100% were obtained in all metrics for cross validation and final test. This agrees with the figure 3(A), where differences are observed between the trajectory of an expert and a non-expert surgeon. Regarding the needle-passing task, it can be observed that all classifiers obtained a good performance. The classifier that showed the best performance for this task was SVM with 100% (for all metrics) for cross-validation and final test, followed by RF with 95% for ACC, 93% for AUC-ROC and 86% for F1-Score in training stage and 100% in final test for all metrics. KNN classifier presented the lowest performance in final test with 83% for ACC, 88% for AUC-ROC and 88% for F1-Score metrics. Finally, for suturing task, KNN and SVM classifiers obtained 100% in all evaluations; RF showed a performance of 97% for ACC, 97% for AUC-ROC and 96% for F1-Score in cross validation and 100% in final test for all metrics. In general, the classification performances obtained show to be consistent with what is shown in figure 3, where visual differences between expert and non-expert surgeons are shown.

The obtained results are comparable with previously reported by Pérez et al. [4], who automatically classified between expert and non-expert surgeons during training in a laparoscopic surgery simulator. They report ACCs of

TABLE II

CLASSIFIERS PERFORMANCE RESULTS (MEAN AND STANDARD DEVIATION).

		10-Fold cross-validation			Final test		
Task	Classifier	ACC	AUC-ROC	F1-Score	ACC	AUC-ROC	F1-Score
Knot-tying	KNN	1.0 +- 0.0	1.0 +- 0.0	1.0 +- 0.0	1	1	1
	RF	1.0 +- 0.0	1.0 +- 0.0	1.0 +- 0.0	1	1	1
	SVM	1.0 +- 0.0	1.0 +- 0.0	1.0 +- 0.0	1	1	1
Needle-passing	KNN	1.0 +- 0.0	1.0 +- 0.0	1.0 +- 0.0	0.83	0.88	0.88
	RF	0.95 +- 0.16	0.93+-0.19	0.86+-0.38	1	1	1
	SVM	1.0 +- 0.0	1.0 +- 0.0	1.0 +- 0.0	1	1	1
Suturing	KNN	1.0 +- 0.0	1.0 +- 0.0	1.0 +- 0.0	1	1	1
	RF	0.97+- 0.11	0.97+- 0.09	0.96+- 0.12	1	1	1
	SVM	1.0 +- 0.0	1.0 +- 0.0	1.0 +- 0.0	1	1	1

93.33% for K-star, 87.58% for RBFNet, and 84.85% for RF classifiers for final test validation. These results are the average performance of three tasks: peg transfer, pattern cutting, and intracorporeal knot suture. As can be seen, our results exceed those reported by Pérez et al. with ACCs between 83% and 100% for final test validation (table II). In contrast, to results reported by these authors, who used MAP and global performance, we proposed local analysis and performance metrics by segments.

As described, an analysis based on several subtasks has been proposed by Vedula et al. [6]. They analyzed three different models (task level), which had similar classification accuracy, with an AUC-ROC of 0.79 for the task level model, 0.78 for the maneuver level model, and 0.7 for the gesture level model. In contrast, we obtained AUC-ROCs from 0.86 to 1 for the three tasks proposed, which overcome that reported by Vedula et al. In addition, the proposed analysis is based on local segments, which can allow continuous feedback throughout the task during robotic surgery training.

IV. CONCLUSIONS

In this work, a new performance evaluation scheme for a surgeon during a given surgical task training was presented. The proposed algorithm is based on a set of motion parameters that feed three classifiers: KNN, RF and SVM. The aim is to automatically classify between expert and non-expert surgeons according to their performance during training. Additionally, in this work a segment-based evaluation scheme is proposed which provides an indicator of the surgeon's local performance. As future work, it is expected to be able to assign a performance rating instead of a label (expert and non-expert surgeons), in addition a manual segment level labeling will be include to evaluate global and local performance. In addition, it is intended to evaluate the classifiers with a laparoscopy training simulator. We consider that this proposal can be used in medical simulators to improve the user experience and strengthen continuous improvement in surgical procedures learning.

ACKNOWLEDGMENT

Acknowledgments to CONACYT for financial support to L. Juarez-Villalobos with scholarship number 1006208.

REFERENCES

[1] G. Chávez-Saavedra, E. Lara-Lona, C. Hidalgo-Valadez, N. Romero-Salinas, and G. J. Méndez-Sashida, "Experiencia en procedimientos laparoscópicos en méxico durante 2015: ¿dónde estamos?" *Cirugía y Cirujanos*, vol. 87, no. 3, pp. 292–298, 2019.

[2] M. K. Chmarra, S. Klein, J. C. De Winter, F. W. Jansen, and J. Dankelman, "Objective classification of residents based on their psychomotor laparoscopic skills," *Surgical Endoscopy*, vol. 24, no. 5, pp. 1031–1039, 2010. [Online]. Available: /pmc/articles/PMC2860557/ /pmc/articles/PMC2860557/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2860557/

[3] R. L. Evans, R. W. Partridge, and D. K. Arvind, "Demonstration paper: A comparative study of surgical skills assessment in a physical laparoscopy simulator using wireless inertial sensors," in *Proceedings - Wireless Health 2014, WH 2014*. New York, New York, USA: Association for Computing Machinery, Inc, oct 2014, pp. 1–8. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2668883.2669588

[4] F. Pérez-Escamirosa, A. Alarcón-Paredes, G. A. Alonso-Silverio, I. Oropesa, O. Camacho-Nieto, D. Lorias-Espinoza, and A. Minor-Martínez, "Objective classification of psychomotor laparoscopic skills of surgeons based on three different approaches," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 1, pp. 27–40, jan 2020.

[5] S. Siyar, H. Azarnoush, S. Rashidi, A. Winkler-Schwartz, V. Bissonnette, N. Ponnudurai, and R. F. Del Maestro, "Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task," *Medical and Biological Engineering and Computing*, vol. 58, no. 6, pp. 1357–1367, jun 2020. [Online]. Available: https://link.springer.com/article/10.1007/s11517-020-02155-3

[6] S. S. Vedula, A. Malpani, N. Ahmidi, S. Khudanpur, G. Hager, and C. Chen, "Task-Level vs. Segment-Level Quantitative Metrics for Surgical Skill Assessment," *Journal of surgical education*, vol. 73, no. 3, pp. 482–489, may 2016. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26896147/

[7] Y. Gao, S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. Yuh, C. C. Chen, R. Vidal, S. Khudanpur, and G. Hager, "Jhu-isi gesture and skill assessment working set (jigsaws) : A surgical activity dataset for human motion modeling," *In Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop*, 2014.

[8] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, p. 175, aug 1992.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, oct 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, sep 1995. [Online]. Available: https://link.springer.com/article/10.1007/BF00994018