Automatic detection of surgical instruments' state in laparoscopic video images using neural networks

C. Martín Vicario[1], I. Oropesa[1], J.A. Sánchez Margallo[2], F.M. Sánchez Margallo[2], E.J. Gómez[1,3], P. Sánchez-González[1,3]

[1] Grupo de Bioingeniería y Telemedicina, ETSI Telecomunicación, Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, España celia.martinvic@alumnos.upm.es, {ioropesa, egomez, psanchez}@gbt.tfo.upm.es

[2] Centro de Cirugía de Mínima Invasión Jesus Usón, Cáceres, Spain; {jasanchez, msanchez}@ccmijesususon.com

[3] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, Spain

Abstract

Software-based solutions such as virtual reality simulators and serious games can be useful assets for training minimally invasive surgery technical skills. However, their high cost and lack of realism/fidelity can sometimes be a drawback for their incorporation in training facilities. In this sense, the hardware interface plays an important role as the physical connection between the learner and the virtual world. The EVA Tracking System, provides computer vision-based information about the position and the orientation of the instruments in an expensive and unobtrusive manner, but lacks information about the aperture state of the clamps, which limits the system´s functionalities. This article presents a new solution for instrument´s aperture state detection using artificial vision and machine learning techniques. To achieve this goal, videos in a laparoscopic training box are recorded to obtain a data set. In each frame, the instrument clamp is segmented in a region of interest by means of color markers. The classifier is modeled using an Artificial Neural Network. The trained prediction model obtains accuracy results of 94% in the validation dataset and an error of 6% in independent evaluation video sequences. Results show that the model provides a competent solution to clamp´s aperture state detection. Future works will address the integration of the model into the EVA and a virtual environment, the KTS serious game.

1. Introduction

Training of technical skills required for minimally invasive surgery (MIS) has advanced from mentor-apprentice model towards structured learning programs priming patient safety. The execution of real surgeries is delayed until surgeons have acquired the necessary skills in a patient-free laboratory. These laboratories allow surgeons to train MIS skills in an efficient, effective and safe environment using mannequins, box trainers or virtual reality (VR) systems [1].

The use of software-based solutions has become a major trend in MIS technical skills learning. VR simulators offer a standardized and safe environment to simulate tasks and to practice skills necessary in the operating room (OR), and have the advantage of measuring the surgeon's performance, which enables a real-time feedback. Another variant of software-based learning are serious games, interactive computer applications with a challenging goal which provide the user with skills, knowledge and attitudes that are useful in real life [2].

One such example of a virtual training environment is the Kheiron Training System (KTS) serious game [3]. KTS combines a box trainer and a video game (Fig. 1A) in order to provide residents and medical students a tool which allows them to train basic laparoscopic technical skills. The player plays the role of a young wizard searching for the Philosopher's Stone and is requested to complete a set of tasks which mimic the movements required in MIS procedures.

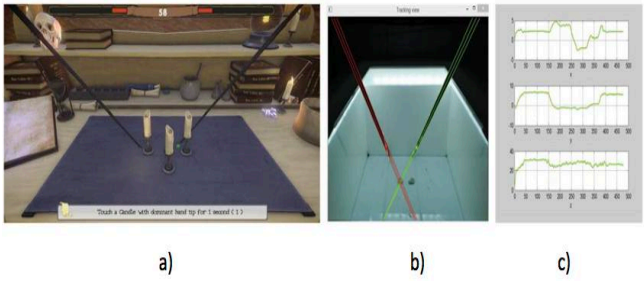

Figure 1 *KTS Serious game [3] and EVA tracking system [4] a) Graphic interface of KTS videogame b)Tracking of instruments inside the training box using EVA Tracking System c) Final path followed by instruments performing a task.*

Virtual environments require a hardware interface to connect the learner and the virtual world with the greatest possible fidelity to the OR environment. The physical connection is achieved using active sensors; affixing magnetic, mechanical or optical devices to real laparoscopic instruments. However, since these disturb the instruments´ ergonomics, new techniques are being applied. This problem is solved using video-based tracking of laparoscopic instruments [2], which applies computer vision techniques to track instruments and perform motion analysis by means of laparoscopic video images.

The KTS serious game is designed to be controlled using real laparoscopic instruments (Fig. 1B). Instrument tracking is carried out using the EVA Tracking System [4]. EVA Tracking System is employed to provide real time information on the position of the laparoscopic instruments and send the coordinates to the game in order to control movements of the wands. EVA is a computer vision-based tracking system that obtains the instrument tip position (Fig. 1C) based solely on the intrinsic camera parameters and the geometrical properties of the instruments without the need of external sensors [4].

ISBN: 978-84-9082-797-0, pags. 293- 296

KTS provides an economical, portable and innovative approximation to laparoscopic training. Integration of EVA adds an inexpensive interaction mechanism, allowing the use of real laparoscopic instruments without significant ergonomic modifications. The main disadvantage of this system is that it lacks information about the clamp´s aperture or rotation state which limits the system functionality.

To our knowledge, only one study reports a solution to the detection of the clamps' aperture. Sahu et al. [5] developed a technique that uses the aperture state of the clamp to move the laparoscopic camera. In this work, a classification model is built using decision trees and boosting. Reported accuracy values are above 95% in ex vivo scenarios. However, the study only uses clamp states that are completely open or closed and with its aperture state perfectly distinguishable.

The goal of this work is to apply computer vision techniques in order to track the laparoscopic instruments and to develop machine learning algorithms in order to detect the clamp´s aperture state in any position of aperture and rotation. In this way, our ultimate goal is to develop an expanded version of the EVA Tracking System not only able to transmit the instruments' spatial coordinates to KTS, but also the current state of its clamps.

2. Material and Methods

2.1. Material

This work is carried out with a personal computer with an intel core i5 2GHz processor and 8GB RAM Memory which allows to program in C++ coding language. Moreover, Boost and OpenCV libraries are used to handle all image and video processing functionalities. To obtain the image set that will serve to train the model, some videos are recorded using the KTS physical environment. This includes a training box, a white box, a webcam in a fixed position and laparoscopic instruments.

Laparoscopic instruments are labeled with color markers in order to track them by means of video processing techniques. The left instrument is labeled with a yellow marker and the right instrument with a green marker.

2.2. Methods

In this work, a neural network model is trained with the goal of classifying the clamps´ aperture state. This is achieved following the steps that are represented in Fig. 2.

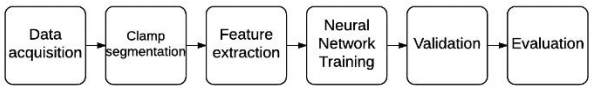

Figure 2. *Steps followed in the implementation, validation and evaluation of the neural network.*

Data acquisition. Images that form the dataset are acquired from videos recorded in the KTS physical environment. In the process of video recording, a rigorous protocol is applied to maximize the diversity of clamps´ positions. Due to the nature of the classifier, every image has to be labeled as *close* or *open*.

Clamp segmentation. Every frame is segmented in a region of interest using the color markers that are placed on the instrument. The region of interest is a window (140x150 pixels) defined around the marker center. The use of this region is justified on two reasons; 1) the information about the clamp state is in this region and 2) processing time is reduced when we minimize the image size.

In order to carry out the segmentation, first it is necessary to carry out a calibration of segmentation parameters [4]. The parameters are saved in a .xml file that will be loaded to segment the image, obtain the region of interest and save it to generate the image dataset (Fig. 3).

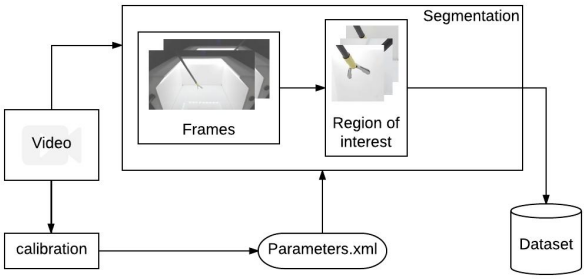

Figure 3 *Process of obtaining the region of interest as a result of image segmentation.*

Feature extraction. In this work the KAZE method [6] is used to extract and describe characteristics from the image. In order to obtain a single vector of characteristics with a fixed dimension, Bag of Words (BOW) technique is applied to KAZE method results [6].

Local characteristics which are obtained with KAZE cannot serve as input for neural networks since the number of descriptor vectors varies. KAZE method is applied to detect and describe characteristics and BOW technique turns a changing number of descriptor vectors into a single vector with constant dimensions. The obtained vector will serve as input for the classifier model.

Neural network training. The classifier model employed is a neural network [7]. In particular, the model is a multilayer perceptron, a supervised learning algorithm. It consists on an input layer, one or more hidden layers and an output layer. Different neural networks are modeled by varying the number of layers and the layer size in order to find the neural network with the best performance.

In addition to a single model, the aggregation of models or bagging technique is applied; three neural networks are aggregated with the objective to improve accuracy. Each neural network is trained separately and has the same weight [8].

Neural network validation. To validate the classifier model, 10-fold cross validation method is applied.

Evaluation. With the goal of evaluate the final trained neural network, an evaluation script is created. This script allows to open a previously recorded video or a live video obtained from the simulator's webcam. Once the video is open, the whole image is segmented in the region of interest, the characteristics are obtained from the images

using KAZE method and BOW technique, and the clamps' state is predicted when the characteristics are introduced in the trained neural network.

Once the predicted class is obtained, a condition of transition between states is applied. The condition imposes that at least 4 consecutive frames have to be in a different state from the actual state. The actual state is shown in a label in the region of interest (Fig.4).

Figure 4. *Example of regions of interest that are labeled with its predicted state.*

3. Results

3.1. Dataset

After processing the videos, a repository with 8949 labeled images are obtained. The number of right/left and closed/open clamps are balanced, as can be observed in Table 1.

	Open	Close	TOTAL
Right	2230	2086	4316
Left	2449	2184	4633
TOTAL	4679	4270	**8949**

Table 1. *Image distribution in the dataset*

3.2. Neural network complexity

The complexity of the neural network is defined by the number of layers and the size of its layers. The neural network has an input layer, an output layer and a single hidden layer, following the results of preliminary tests carried out during the training phase.

The output layer has only two neurons due to the fact that prediction model has only two possibilities: open or closed. However, the network input size has to be defined. For this purpose, several cases were tested. The result is represented in the graph in Fig. 5.

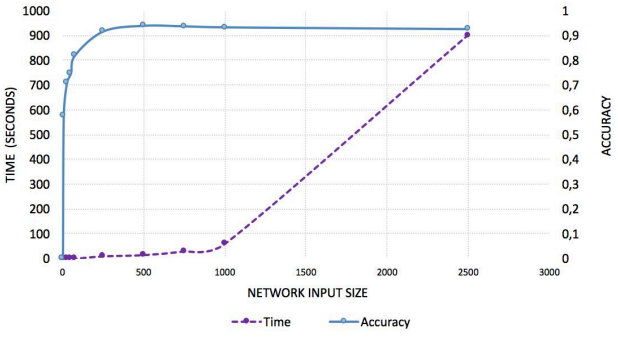

Figure 5. *Accuracy and time vs. Network input size*

Initially, the accuracy grows substantially when the number of neurons rises and reaches a maximum in 500 neurons with 94% accuracy. After that, the accuracy does not improve when the network input size increases.

Another aspect that has been considered is training time (Fig. 5): neural networks with less than 1000 neurons as network input size have a training time under 1 minute. Taking into account every factor, we choose a neural network with 500 neurons in the input layer.

3.3. Learning curve

Figure 6 represents the training error and the cross-validation error when the model is trained with different number of images. The bias and the variance are low and the curve has an asymptotic behavior, so the number of images that comprise our dataset is enough and increasing the dataset will not decrease the error significantly.

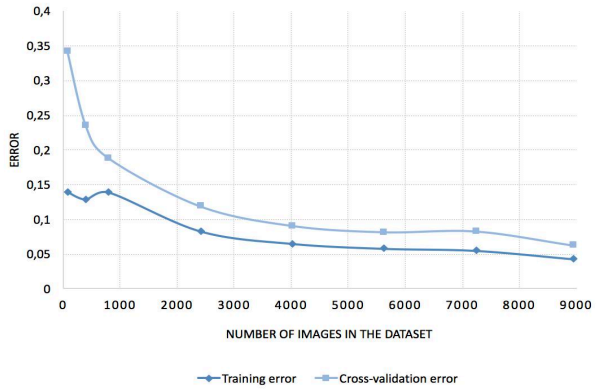

Figure 6. *Learning curve. Training error and cross validation error vs. the number of images that are used to train the neural network..*

3.4. Bagging

Aggregation of models improve performance compared to a simple model when the neural network has low complexity (Fig. 7). However, for a network input size of 500 neurons, the aggregated model does not improve the accuracy of the simple model.

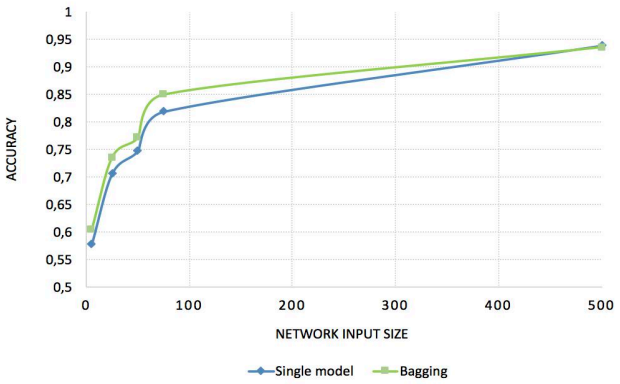

Figure 7. *Accuracy of aggregation of models vs. single model*

3.5. Evaluation

After we have defined all the parameters from the neural network, several tests have been carried out with our evaluation script. New videos were recorded in the KTS box trainer setting. The clamp state was predicted for every frame and displayed in the region of interest (Fig. 8). The

script works with a single instrument or with both of them simultaneously.

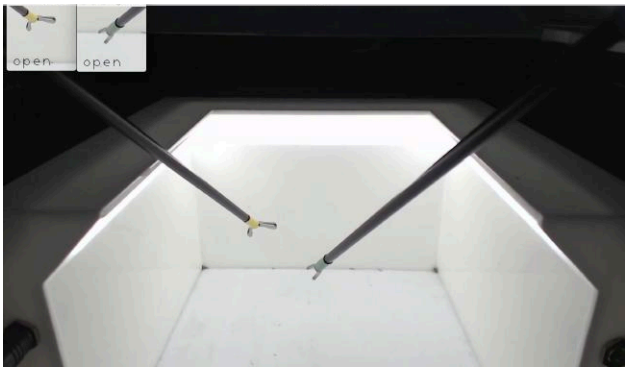

Figure 8. *Screenshot of an evaluation with two laparoscopic instruments, showing the complete image and regions of interest labeled with the current predicted state of clamps' aperture.*

To test how the transition condition affects the performance of the whole model, 3 new videos are employed, counting the number of frames that are incorrectly classified with and without the condition (Table 2). As a result, we obtained that the condition reduces the errors in 37.5 %.

Video		%Errors without condition	%Errors with condition	%Error reduction
Left instrument		7.65	4.74	38.1
Right instrument		14.07	8.95	36.4
Both instruments	Left	8.3	5.62	32.5
	Right	7.2	4.42	39.1
Total		**9.6**	**6.0**	**37.5**

Table 1: *Evaluation results on new video sequences*

The transition condition introduces a delay of 0.1 seconds that is hardly detectable. However, in some cases, the use of only 4 frames is not enough to avoid errors, especially in some positions of clamp rotation.

4. Discussion

The results from validation and evaluation show the feasibility of using an artificial neural network to implement a model that automatically detects the state of a laparoscopic instrument with a high accuracy. In our experiments, the highest accuracy (94%) is reached with a neural network with a single hidden layer and 500 neurons in its input layer. The results that we have obtained are similar to those obtained by Sahu et al. [5], but our work has the advantage of employing images of clamps in any state of aperture or rotation.

The model performs well in new sequences recorded in the box trainer. The transition condition that is imposed reduces the number of erroneously classified frames from 9.6 to 6%; however, it adds a delay that may end up affecting real time performance in a virtual scenario such as KTS. Because of that, a balance between accuracy and delay has to be established. Another possibility is to add a dynamic transition condition.

Training a new model with a larger database would not likely increase the accuracy. The obtained dataset with 8949 images is shown to be sufficient. Moreover, we have proven that aggregation does not improve model' performance. Instead of adding more images to the dataset, a possible improvement would be to maximize the number of positions in which the clamp is recorded.

The next steps in this work include developing new solutions to address the problem of rotation. This is necessary to provide a full, high fidelity experience to trainees and surgeons training in virtual scenarios such as KTS. Parallel to this, we will integrate the trained model in the EVA Tracking System-KTS serious game. This will allow us to validate the performance of the model in an actual training environment with real surgeons.

5. Conclusion

We have presented an approximation to detect the clamp´s aperture state in laparoscopic instruments using computer vision and machine learning techniques. The integration of this information will allow to increase the fidelity of software-based surgical training systems, such as, but not limited to, the KTS serious game.

References

[1] Hiemstra E. Acquiring Minimally Invasive Surgical Skills. Leiden: Department of Minimally Invasive Surgery in Gynaecology, Faculty of Medicine/Leiden University Medical Center (LUMC), Leiden University, 2012

[2] I. Oropesa *et al.*, «Methods and Tools for Objective Assessment of Psychomotor Skills in Laparoscopic Surgery», *J. Surg. Res.*, vol. 171, n.º 1, pp. e81-e95, 2011.

[3] L. F. Sánchez Peralta *et al.*, «Serious game for psychomotor skills training in minimally invasive surgery: Kheiron Training System», en *XXXIII Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2015)*, Madrid, Spain, 2015, pp. 439-442.

[4] I. Oropesa *et al.*, «Controlling virtual scenarios for minimally invasive surgery training using the EVA Tracking System», en *XXXIII Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB 2015)*, Madrid, Spain, 2015, pp. 431-434.

[5] M. Sahu, D. Moerman, P. Mewes, P. Mountney, and G. Rose, «Instrument State Recognition and Tracking for Effective Control of Robotized Laparoscopic Systems», *Int. J. Mech. Eng. Robot. Res.*, 2016.

[6] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, «KAZE Features», en *Computer Vision – ECCV 2012*, 2012, pp. 214-227.

[7] S. S. Haykin, *Neural networks and learning machines*, 3rd ed. New York: Prentice Hall, 2009.

[8] J. Zhang, «Developing robust non-linear models through bootstrap aggregated neural networks», *Neurocomputing*, vol. 25, n.º 1, pp. 93-113, abr. 1999.