Crimson Publishers
Wings to the Research

Research Article

Hand and Gesture Module for Enabling Contactless Surgery

Martin Žagar[1]*, Alan Mutka[1], Ivica Klapan[2,3,4] and Zlatko Majhen[5]

[1]RIT Croatia, Croatia

[2]Klapan Medical Group Polyclinic, Croatia

[3]School of Medicine, Croatia

[4]School of Dental Medicine and Health, Croatia

[5]Bitmedix, Croatia

ISSN: 2689-2707

***Corresponding author:** Martin Žagar, RIT Croatia, Croatia

Submission: August 27, 2021

Published: September 24, 2021

Volume 3 - Issue 1

How to cite this article: Martin Žagar. Hand and Gesture Module for Enabling Contactless Surgery. Trends Telemed E-Health 3(1). TTEH. 000553. 2021. DOI: 10.31031/TTEH.2021.03.000553

Abstract

Nowadays, there are many new approaches and techniques in telemedicine and surgery with different kinds of innovations and a growing need for contactless control of surgery parameters. Our proposal is aiming to resolve the problem of standard surgical parameters with gesture-controlled surgical interventions. We designed a contactless interface as a plug-in application for the DICOM viewer platform using a hardware sensor device controller that supports hand/finger motions as input, with no hand contact, touching, or voice navigation. Our proposed approach enables surgeons to get complete and aware orientation in the operative field (which currently isn't the case and where the problem lies), where 'overlapping' of the real and virtual anatomic models is inevitable. Human mind and understanding of this new surgery work by creating entirely new models of human behavior and understanding spatial relationships, along with devising assessment that will provide an insight into our human nature. That's why the essential part of our solution is to build the adequate hand and gesture module for motion which we will describe in this paper.

Keywords: Contactless surgery; Hand and gesture module; Motion tracking; Spatial relationships

Introduction

Motion tracking enables more precise virtual movement, rotation, cutting, spatial locking, and measuring as well as slicing through datasets. To provide the most immersive experience, we use a camera for depth and motion tracking that has active stereo depth resolution with precise shutter sensors for depth streaming with a range up to 2-3 meters which is essential in the OR and which gives a sense of freedom to the surgeon during the surgery. That's why we approached the design of this part of the solution with special care to build an accurate, but still an intuitive and straightforward way of activating positioning control which is based on waiting for the users' hands to enter the central trigger area to activate the interaction with the interface, with touch-free surgeon's commands. We intend to offer an alternative to closed SW systems for visual tracking and develop the SW framework that will interface with depth cameras and provide a set of standardized methods for medical applications such as hand gestures and tracking, face recognition, navigation, etc. We found it possible to significantly simplify movement gestures in the virtual space of virtual endoscopy. Our clinical and technology research is already at the high maturity level of accomplishing the proof-of-concept phase. Our clinical tests and technological achievements where we already tested our previous solution with Leap Motion in OR are demonstrated in the results of several research papers [1-3]. After updating the needs in clinical workflow based on the inputs from several different medical specialists, we now identified two primary tasks for the hand and gesture module for motion tracking. For the part of hand tracking, it is important to provide hand coordinates in two dimensions and surrounding in 3D. For the gesture recognition and tracking part of the module, we designed gesture states based on the hand tracking.

The rest of the paper presents an overview of the possible solution we have tested with advantages and disadvantages. In the end, the two best solutions for our application are selected and implemented in the final software explication, where we will show the results in the Results section.

Materials and Methods

In this section, we will provide an overview of the existing hand and gesture algorithms.

OpenPose

Open Pose [4,5], has represented the first real-time multi-person system to jointly detect human body, hand, facial, and foot key points (in total 135 key points) on single images. It is primarily used to track human pose but can also be used to track hand and face pose. The main advantage is easy usage and that it can also be implemented in OpenCV. Disadvantages are:

a) For commercial use license should be acquired (the latest price was 25000 dollars).

b) Less accurate than other tested algorithms,

c) Slow if not used with GPU.

Open CV + DLIB (HoG)

This is a widely used approach for image classification tasks based on Histogram of Oriented Gradients (HoG) feature descriptors and Support Vector Machines (SVM) [6]. This approach's main advantage is the high processing of FPS on a regular CPU and a high and stable detection rate. On the other hand, the main disadvantage is low robustness on the scene's light/background condition changes. We have used this solution in our implementation, and we will describe more about this in the following chapters.

Open CV + Haar cascade

Another algorithm for image classification is based on Haar Cascade Classifiers [7]. The implementation is similar to the DLIB (HoG), but there are several disadvantages. DLIB (HoG) is ahead of the Haar cascade classifier over speed, implementation, and accuracy. First, the training is done using a sliding sub-window on the image, so no subsampling and parameter manipulation is required as in the Haar classifier. The main advantage is that this algorithm is fast and easy to train, simple to use, and run on a regular CPU. Disadvantages are:

a) Less accuracy than DLIB (HoG).

b) Higher False Positive prediction than DLIB (HoG).

c) The hand's landmarks are not detected, only the position.

Open vino

Open VINO toolkit [8] (Open Visual Inference and Neural network Optimization) is a free toolkit facilitating a deep learning model's optimization from a framework and deployment using an inference engine onto Intel hardware. The toolkit is highly optimized for Intel platforms, but unfortunately, we had problems transferring the existing TensorFlow models in OpenVino. The examples of trained models are available on this page [9].

Advantages are that the solution is very well documented and has lots of source code and that there is the existing model for ASL (American Sign Language) example. The main disadvantage is difficulties while converting the existing TensorFlow models in Open Vino.

Media pipe hands

Media Pipe Hands [10] is a high-fidelity hand and finger tracking solution. It employs machine learning (ML) to infer 21 3D landmarks of a hand from just a single frame. Current state-of-the-art approaches rely primarily on powerful desktop environments for inference. This method achieves real-time performance on a mobile phone and even scales to multiple hands. Media Pipe Hands utilizes an ML pipeline consisting of various models working together: A palm detection model that operates on the entire image and returns an oriented hand bounding box. A hand landmark model performs on the cropped image region defined by the palm detector and returns high-fidelity 3D hand key points. We have used this solution in our implementation, and we will describe more about this in the following chapters.

Hand and Gesture Module Implementation and Results

All algorithms from the previous section were implemented and tested, and we have selected two of them in the final implementation:

a) DLIB - very fast on a regular CPU computer, easy to train. Since we have constant lighting conditions in OR, the model can be trained in several minutes, and it is ready for usage.

b) MediaPipe Hands - pre-trained R-CNN models for hand detection - there is no need for additional training.

DLIB implementation

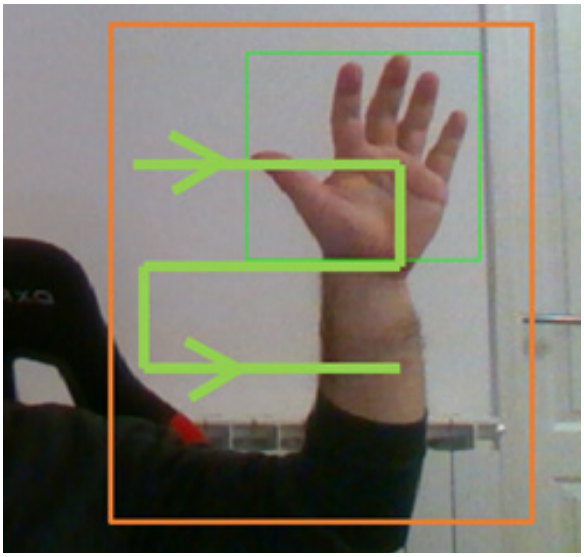

Figure 1: Sliding window image acquisition.

DLIB approach consists of three phases: Acquisition, learning, and a detection phase. The learning phase requires gathering up to 100 images of the hand, showing the specific gesture. For example, Figure 1. shows the sliding window image acquisition principle. First, the region of interest needs to be selected by the user (the orange rectangle). Second, the approximated hand region is also manually selected (the green rectangle). Once the regions are chosen, the algorithm dynamically repositions the rectangle, and the user needs to follow it by placing their hand exactly inside the movable green rectangle. The result is up to 100 images of hands inside the region of interest.

After the acquisition is finished (for example, the open hand or FIVE gesture), the learning algorithms generates an SVM model saved in a dedicated file - FIVE.svm. The whole procedure needs to be repeated for all other gestures. Our project controls the DICOM viewer application only with two gestures (FIVE and TWO). As a result, the training and learning phase generate two files: FIVE.svm and TWO.svm, which are then used in the third detection phase.

The final detection phase is shown in Figure 2 & 3. The input image is processed in parallel with both detectors (FIVE and TWO models). The detected hand center (2D pixel coordinates) as output is provided if one detectors' confidence is more significant than the predefined threshold Figure 2 shows the result when the input hand gesture is TWO, while Figure 3 shows the result of gesture FIVE.

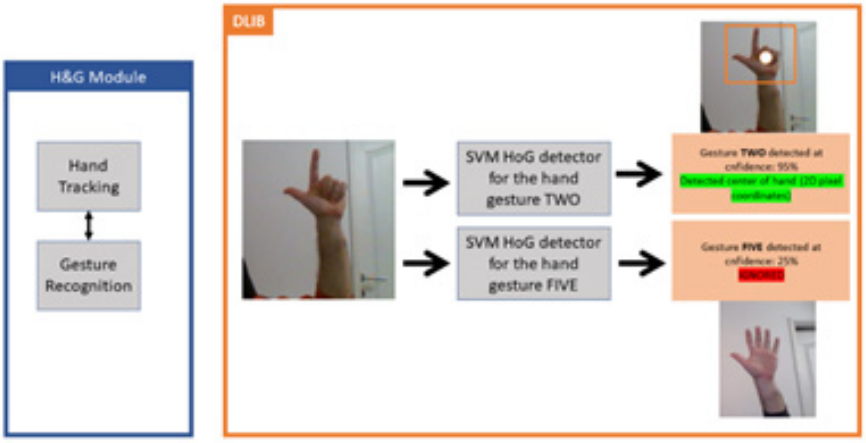

Figure 2: Final detection with gesture TWO.

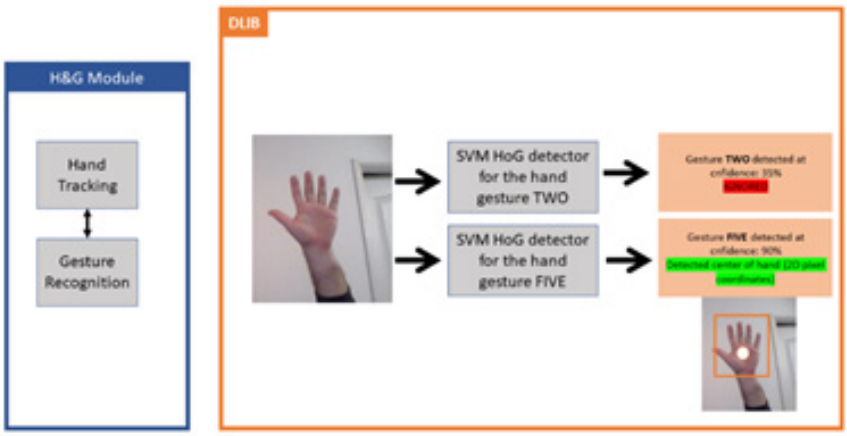

Figure 3: Final detection with gesture FIVE.

The implemented DLIB algorithm provides up to 20 FPS hand/gesture detection, mostly depending only on the CPU power. The algorithm's output are:

a. Center of hand in 2D Image Pixel coordinates

b. Center of hand in 3D world/camera coordinates

c. Hand gesture (the gesture must be learned during the learning phase).

DLIB implementation in the developed software is presented in Figure 4. There exists two tabs: TRACK and TRAIN tab.

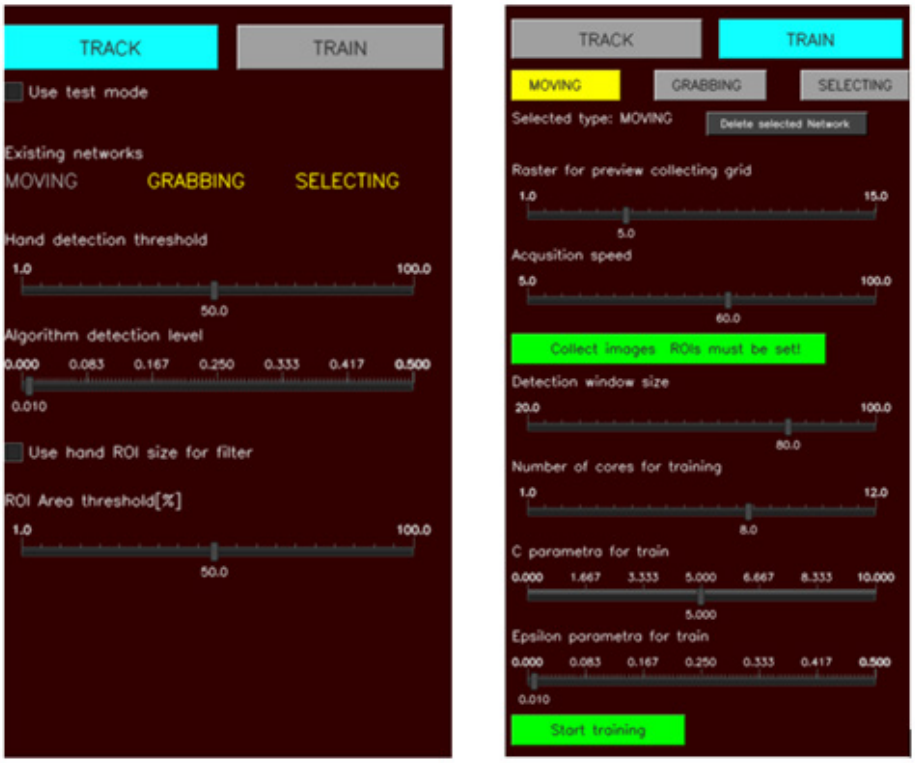

Figure 4: DLIB software implementation - training and track tab.

The TRAIN tab consists of the following properties/parameters:

a. MOVING, GRABBING and SELECTING - the general Gesture States naming (more explained later)

b. Raster for preview collection grid - sliding window raster, greater number more images will be acquired

c. Acquisition speed - the acquisition time between two acquired pictures for training - less number, faster acquisition

d. Collect images, ROIs must be set - this button starts the acquisition process

e. Detection window size, number of codes for training, C parameter for training, Epsilon parameter for training - parameters for DLIB learning phase

f. Start training - button to start the training/learning phase.

The training phase needs to be repeated for every general Gesture State (moving, grabbing, and selecting). As a result, three trained models are saved, and the software is ready for the track/detection phase within the TRACK tab. The TRACK tab consists of the following properties/parameters:

a. Hand detection threshold - the confidence threshold - lower value, less restriction, more stable detection

b. Algorithm detection level

c. ROI area threshold.

Conclusion

Whether surgery is planned or urgent, it is always marked by interdisciplinarity with a very strong dependence on the individual expertise of surgeons, anesthesiologists, and other medical staff. Working in team conditions, the surgical procedure is often marked by time pressure and consequently stress. Therefore, operating theaters are also considered high-risk sites prone to errors and surgical complications. In the process of medical data visualization during the surgery, it is important enabling a precise and fast solution for contactless surgery where we built our solution for hand and gesture tracking with the main benefits interactor features in selecting and grabbing:

Selecting

By using a gesture TWO and moving hand UP-DOWN, the user changes the DICOM control modes

a) NONE - none of the modes if selected

b) HU_CENTER - changing the center parameter of the CT's Hounsfield scale

c) HU_WINDOW - changing the window parameter of the CT's Hounsfield scale

d) VOLUMEN_OPACITY - changing the volume opacity

e) PLANE_OPACITY - changing the opacity of the measuring plane

f) CAMERA_YAW_PITCH - changing the scene's camera position in XYZ world

g) SLICES_XYZ - changing the measurement planes in XYZ world

h) SLICES_XYZ_M1 - changing the measurement planes in XYZ world with setting the measurement point M1

i) SLICES_XYZ_M2 - changing the measurement planes in XYZ world with setting the measurement point M2.

Grabbing

By using a gesture FIVE and moving LEFT-RIGHT, for each mode the parameters can be changed. All parameters are changes by moving the hand LEFT-RIGHT, except for the CAMERA and SLICES modes where LEFT-RIGHT, UP-DOWN, and the Z-axis are included (towards the camera) are included.

With these features, a surgeon is able to control the DICOM Viewer scene and to enable fully contactless medical data visualization during the surgery.

Acknowledgment

This work is funded by European Institute for Innovation and Technology, EIT Innostars Health grant.

Conflict of Interest

We declare that we have no conflict of interest.

References

1. Klapan I, Majhen Z, Žagar M, Klapan L, Trampuš Z, et al. (2019) Utilization of 3d medical imaging and touch-free navigation in endoscopic surgery: Does our current technologic advancement represent the future innovative contactless noninvasive surgery in rhinology? What is next? Biomedical Journal of Scientific & Technical Research 22(1): 16336-16344.
2. Klapan I, Duspara A, Majhen Z, Benićl I, Trampuš Z, et al. (2019) Do we really need a new navigation-noninvasive "on the fly" gesture-controlled incisionless surgery? Biomedical Journal of Scientific & Technical Research 20(5): 15394-15404.
3. Klapan I, Duspara A, Majhen Z, Benić I, Kostelac M, et al. (2017) What is the future of minimally invasive surgery in rhinology: Marker-based virtual reality simulation with touch-free surgeon's commands, 3D-surgical navigation with additional remote visualization in the operating room, or ...? Frontiers in Otolaryngology-Head and Neck Surgery 1(1): 1-7.
4. Cao Z, Hidalgo G, Simon T, Wei S, Sheikh Y (2019) OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, USA, pp. 7291-7299.
5. https://github.com/CMU-Perceptual-Computing-Lab/openpose
6. https://www.learnopencv.com/training-a-custom-object-detector-with-dlib-making-gesture-controlled-applications
7. Viola p, Jones MJ (2001) Rapid object detection using a boosted cascade of simple features. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, USA, pp. I-I.
8. https://docs.openvinotoolkit.org/latest/index.html
9. https://software.intel.com/content/www/us/en/develop/tools/openvino-toolkit/pretrained-models.html
10. https://google.github.io/mediapipe/solutions/hands