Surgical Instrument Tracking for Vitreo-retinal Eye Surgical Procedures Using ARAS-EYE Dataset

F. Lotfi
Faculty of Electrical Engineering, K.N. Toosi University of Technology
Tehran, Iran
f.lotfi@email.kntu.ac.ir

P. Hasani
Faculty of Electrical Engineering, K.N. Toosi University of Technology
Tehran, Iran
parisa.hasani@email.kntu.ac.ir

F. Faraji
Faculty of Electrical Engineering, K.N. Toosi University of Technology
Tehran, Iran
farnooshfaraji@email.kntu.ac.ir

M. Motaharifar
Faculty of Electrical Engineering, K.N. Toosi University of Technology
Tehran, Iran
motaharifar@email.kntu.ac.ir

H. D. Taghirad
Faculty of Electrical Engineering, K.N. Toosi University of Technology
Tehran, Iran
taghirad@kntu.ac.ir

S. F. Mohammadi
Tehran University of Medical Sciences
Tehran, Iran
sfmohammadi@tums.ac.ir

***Abstract*—Real-time instrument tracking is an essential element of minimally invasive surgery and has several applications in computer-assisted analysis and interventions. However, the instrument tracking is very challenging in the vitreo-retinal eye surgical procedures owing to the limited workspace of surgery, illumination variation, flexibility of the instruments, etc. In this article, as a powerful technique to detect and track surgical instruments, it is suggested to employ a convolutional neural network (CNN) alongside a newly produced ARAS-EYE dataset and OpenCV trackers. To clarify, firstly you only look once (YOLOv3) CNN is employed to detect the instruments. Thereafter, the Median-flow OpenCV tracker is utilized to track the determined objects. To modify the tracker, every "n" frames, the CNN runs over the image and the tracker is updated. Moreover, the dataset consists of 594 images in which four "shaft", "center", "laser", and "gripper" labels are considered. Utilizing the trained CNN, experiments are conducted to verify the applicability of the proposed approach. Finally, the outcomes are discussed and a conclusion is presented. Results indicate the effectiveness of the proposed approach in detection and tracking of surgical instruments which may be used for several applications.**

***Index Terms*—Vitreo-retinal eye surgery, convolutional neural networks, OpenCV trackers, surgical instrument tracking**

I. Introduction

Vitreo-retinal eye surgery is a type of surgical operations performing on the vitreous humour and the retina. In this type of surgery, the surgeons are expected to have extremely precise handling maneuverability with adequate control on the interaction force between the surgical tool and delicate organs inside of the eye. Generally, vitreo-retinal eye surgery is considered as one of the most challenging surgical operations owing to human limitations, for which limited assistive equipment have been developed [1]–[3]. Specially, learning of the necessary skills for vitreo-retinal eye surgery by a novice surgeon is a complicated task. A number of researchers have been developed methodologies to facilitate surgery training to the novice surgeons [4]–[6]. Any mistake in this surgical procedure might lead to disastrous complications for the patients including damage to the retina and even vision loss.

Notably, detecting and tracking of medical instruments in vitreo-retinal eye surgery is an important task. Having an accurate time-based position of surgical tool provide a systematic way to asses the skill level of novice surgeons during and after the operation. In fact, the new computer-based technologies have enabled the detecting and recording of instrument position, which is a paramount step towards skill assessment [7]. Other applications also require tracking of instruments such as automatic positioning, surgical motion analysis, and visual servoing [8].

Convolutional neural networks (CNNs) are proved to be powerful in detecting objects through a single image. Thus, to extract feature points automatically, one can employ CNNs to detect and recognize the objects [9], [10]. There are several architectures which provide the ability to recognize visual patterns directly from pixel images with minimal preprocessing. For instance, AlexNet, ZFNet, ResNet, VGGNet and YOLO are designed with different number of convolutional layers, pooling layers and fully connected layers for various applications [11]. In addition, OpenCV trackers [12] such as tracking-learning-detection (TLD) [13] and Median-flow [14] can be implemented beside CNNs to reduce the computational cost [15]. In other words, albeit CNNs are vigorous tools for object detection, there are some limitations for them. Blurred image may bear elimination of object features in the frame which causes diverse objects to be scarcely distinguishable for the CNN. However instead, trackers perform much better and rarely lose the object. This is due to the fact that trackers employ the information through a frame sequence. Consequently, they have a kind of prediction about where the object could be located. In addition, in contrast to CNNs, a tracker utilizes a specified region of image which yield a prompt performance.

In this article, a hybrid approach based on CNNs and trackers

978-1-7281-7296-5/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on December 21,2020 at 13:09:22 UTC from IEEE Xplore. Restrictions apply.

is proposed to aptly detect and track surgical instruments. Furthermore, to train the YOLOv3 deep CNN [16], a new dataset is produced which is associated with the eye surgical instruments and is called ARAS-EYE dataset. To verify the applicability of the method, experimental results are conducted which indicate the applicability of the presented approach. The rest of the paper is organized as follows. Section II concentrates on the methodology of the suggested approach. The produced dataset is presented in section III. Experimental results are reported in section IV. Section V discusses the outcomes and finally, section VI concludes the paper.

II. Methodology

The objective of this paper is to track the medical instruments in vitreo-retinal eye surgery which is an important step towards systematic skill assessment of the novice surgeons. Hence, an image-based tracking methodology is presented utilizing the camera mounted on the surgical microscope. Our method is elaborated in below.

Common problems encountered in tracking surgical instruments are the cluttered backgrounds, motion blur, shadows cast by the tool, changes in light, specular reflections on the tool surfaces, and deformable shape [17], [18]. Recently, the YOLO deep CNN is employed in a variety of applications. Furthermore, the YOLOv3 as the third version of this vibrant CNN, yields a promising performance in detecting delicate objects. Its structure consists of 130 layers, which enables a robust detection in the presence of different illumination conditions. Basically, training a deep CNN is all about an optimization problem in which the initial conditions are one of the most important things in obtaining proper results. In this regard, there are two approaches to train such a deep CNN. First, training the deep CNN utilizing initial random variables, which a sufficient dataset is indispensable in this case. Second, using the transfer learning methods in which pretrained weights of the deep CNN are going to be used as the initial values for the training on a new dataset. The latter one is prefered when there is a small dataset. Since in this research, surgical instruments are possibly discriminative with respect to the background, a big dataset is not vital for the training. Thus, the second approach is considered. A comprehensive description about the dataset is given in the next section.

To precisely detect fragile objects, a deep structured CNN is an advantage. In this regard YOLOv3 is employed. Furthermore, to moderate the computational cost, an OpenCV tracker is utilized, as well. This way specified objects are tracked with an appropriate performance. Utilizing the outcome of the suggested hybrid method, it is possible to not only obtain the instrument trajectory, but also to determine its orientation at each time.

III. The produced dataset

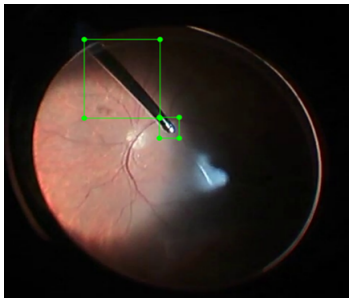
There are several methods presented for detection and tracking of surgical instruments in Robot-Assisted minimally invasive surgery (RAMIS). Recently, deep learning is becoming more alluring and attracts increasing attention. As it is mention earlier, in this research, a method is presented based on deep learning approaches which can detect the surgical instrument in the eye area. One important thing in using deep learning methods is to have an appropriate dataset. In this regard, there are two ways to train such a deep CNN. First, training the deep CNN from scratch utilizing initial random variables which needs sufficient dataset. Second, using the transfer learning methods in which pretrained weights of the deep CNN are going to be used as the initial values for the training on a new dataset. Deciding what type of transfer learning should be performed on a new dataset, is a function of the size of the new dataset (small or big), and its similarity to the original dataset. There are commonly 4 rules to choose type of transfer learning: 1. New dataset is small and similar to original dataset. The best idea might be to train a linear classifier on the CNN codes. 2. New dataset is large and similar to the original dataset. Since there are more data, we can have more confidence that we won't overfit if we were to try to fine-tune through the full network. 3. New dataset is small but very different from the original dataset. it might work better to train the SVM classifier from activations somewhere earlier in the network. 4. New dataset is large and very different from the original dataset. In this case, we would have enough data and confidence to fine-tune through the entire network. Since in this research, the surgery is done in a specified area and objects are determinate there is no need for a big dataset. As the following dataset is small, classifiers are retrained and weights are replacing by optimizing the lost function. In order to fine tuning the CNN, last weights of the trained YOLOv3 on COCO [19] dataset are used as initial values. A comprehensive description about the dataset is given in what fallows.

In this article, an image dataset is produced for vitreo-retinal eye surgery.There are different videos recorded for vitreoretinal eye surgeries data bases. In order to have a sufficient dataset, 594 images have been extracted from several videos in [6]. Images are captured from video frames to consider all important tools during the surgery. Specified instruments which are employed in this surgery are vitrectomy probe, forceps, and laser. Images are labeled manually to provide reliable dataset. Bounding boxes consist of "shaft","center","laser", and "gripper". As a result, a .xml file is made as an annotation for each image that is required for training and validation process in supervised learnings. As the surgery is done in a closed area and tools are specified, it seems that 594 images are enough. Since images are extracted from videos, it is possible to have more images if necessary.

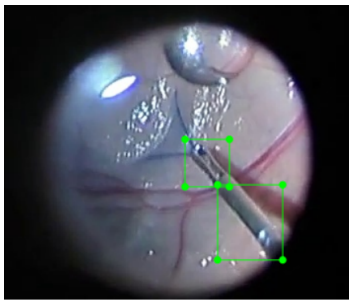
To clarify the procedure, three labeled images are shown as samples of the produced dataset. In Fig. 1a both "center" and "shaft", and in Fig. 1b "shaft" and "gripper" are depicted as two instances. Furthermore, Fig. 1c contains "Laser" label. It is noteworthy to mention that, label and location of each bounding box is mentioned in an annotation file associated with each particular image.

This dataset is called ARAS-EYE dataset and it is published in Github. The aim is to utilize the ARAS-EYE dataset in

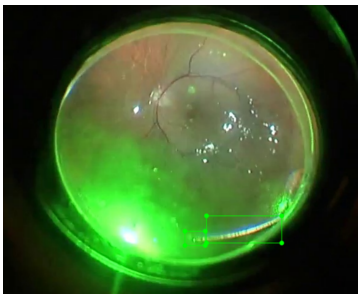
Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on December 21,2020 at 13:09:22 UTC from IEEE Xplore. Restrictions apply.



(a) center and shaft



(b) gripper and shaft



(c) laser, center and shaft

Fig. 1: Sample images of the produced dataset

our current eye surgery training system [6], [20] and eye teleoperated surgery system [21] and fuse the position signals from the visual system with the kinematic data to obtain more powerful results.

IV. EXPERIMENTAL RESULTS

This section concentrates on the performed experiments. The aim is to detect and track surgical instruments employing a hybrid approach based on CNN and conventional OpenCV trackers. The procedure is summarized in three steps which are presented and expanded in what follows. Since a proper performance speed is always an advantage, the proposed method is implemented on an Nvidia GeForce GTX 1080 Ti GPU. It is noteworthy to mention that, the speed of utilizing the YOLOv3 CNN individually on this GPU is about 30 frames per second (FPS) while the hybrid approach may bear at least 30 FPS. This is directly related to how much the tracker is got involved. According to what remarked before, the YOLOv3 CNN is applied to detect the demanded objects. To train this CNN a dataset is produced which its detail is clarified in section III. Furthermore, the transfer learning method is utilized in the training process.

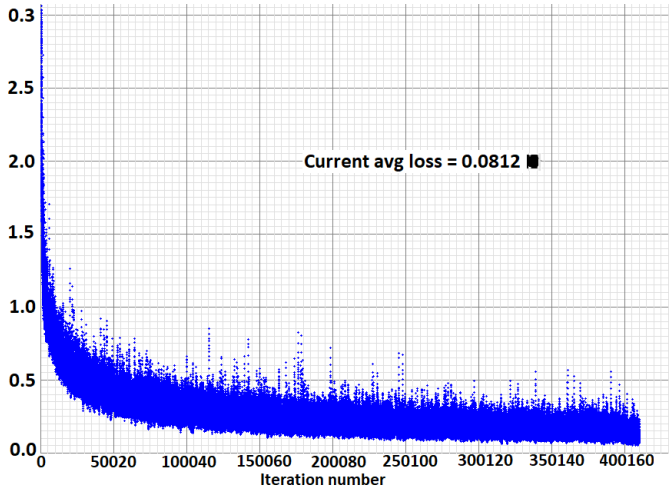


Fig. 2: YOLOv3 training error on the produced dataset

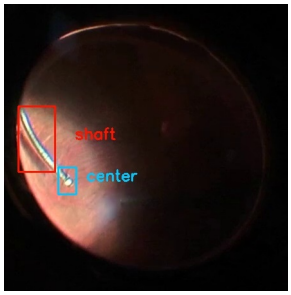
Fig. 2 illustrates the training error. As it is seen, the CNN is trained for more than 400160 batches. Since the error has converged to a local minimum point with 0.0812 value, it is concluded that the CNN is stable and it has been trained appropriately. A point to ponder is, the convergence point value represents a level which the CNN may distinguish diverse objects. Test results of the trained CNN is reported in Fig. 3. As it is indicated, YOLOv3 has been capable of detecting the determined objects. Moreover, this figure shows how the CNN may discriminate between a gripper and a center as two different surgical instruments. Another important fact is evident in Fig. 3c in which even the laser usage is detected.

The main objective of this paper is to suggest an approach for detection and tracking of surgical instruments. Hence, another experiment is performed on a video this time. The video consists of 38 frames in which an operation is being done through a surgery. To clarify, a sequence of 8 images is shown in Fig. 4. As it is evident, surgical instruments are tracked via the CNN and the tracker. The position and the orientation of the surgical instrument are two important variables to obtain. Each bounding box center is taken as a representative point of an object. Note that for an instrument both shaft and center are crucial objects to track. Thus, to obtain the trajectories, center points of the bounding boxes are stored in a matrix. Since there is a single surgical instrument in this experiment, the matrix size is 2×38. Although this may be utilized to get each object trajectory, orientation is still remained to be calculated. A simple way to deal with this problem, is to employ the following equation in each frame:

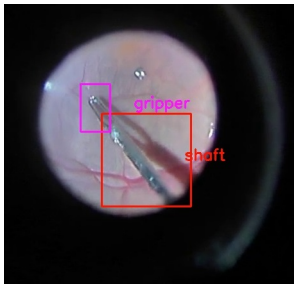
$$\theta = \arctan \frac{y_2 - y_1}{x_2 - x_1} \tag{1}$$

where (y_2, x_2) and (y_1, x_1) represent each center position of the bounding boxes, respectively.

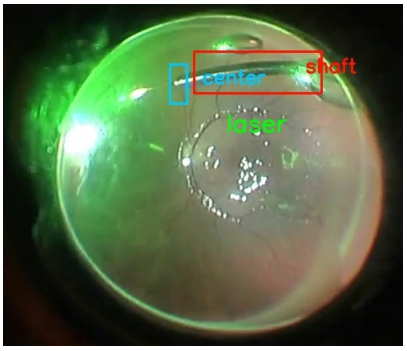
Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on December 21,2020 at 13:09:22 UTC from IEEE Xplore. Restrictions apply.



(a) center and shaft



(b) gripper and shaft



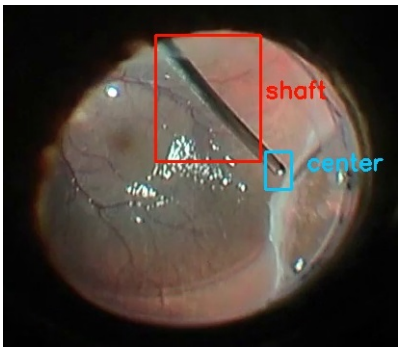
(c) laser, center and shaft

Fig. 3: YOLOv3 bounding boxes outputs after the training. In each image, the corresponding labels are set by the CNN.

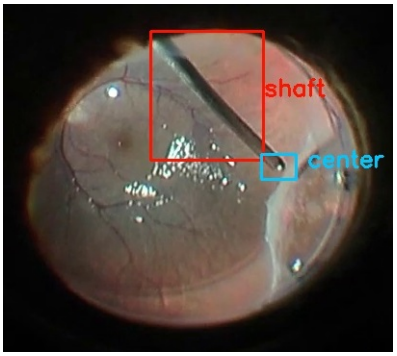
The trajectory of each object is depicted in Fig. 6. Values are normalized with respect to the frame size. It is seen that, while the center position in Fig. 4 and Fig. 5 is lower in comparison with the shaft, its trajectory is higher. This is due to the fact that a frame origin is always located at left top corner of the image. Thus, the lower the object, the higher its trajectory. One may use the middle point of the center and shaft positions as the instrument position representer. Fig. 7 indicates the orientation of the instrument. The region in which the surgeon is performing the surgery may be determined in this figure. It is evident that the surgeon is trying to be remained in a specified region. A rich information about the surgery detail, is given in Fig. 6 and 7. There are a variety of applications for these results which next section is presented to cover them.

V. Discussion

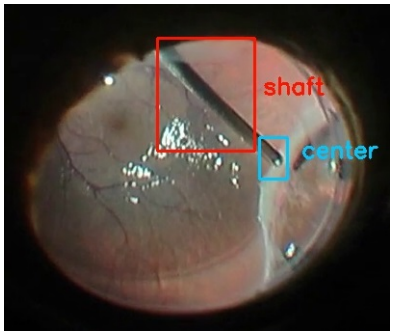
As illustrated in the the previous sections, a methodology is presented for surgical instrument tracking in vitreo-retinal eye surgical procedures using a new produced dataset. The proposed method is based on state of the art deep learning approaches to further ascertain a proper performance in diverse conditions. Although there are other methods, structures and algorithms to solve the problem of detection and tracking of the surgical instrument, it is discernible from experimental results that the capability of the presented solution is convincing. Since a celebrated CNN is employed, the suggested approach may be implemented simply. The objective is to utilize the produced dataset in the following ways



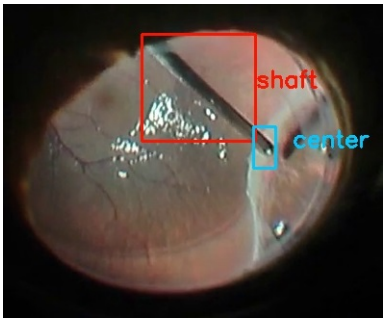
(a) 1st frame



(b) 2nd frame



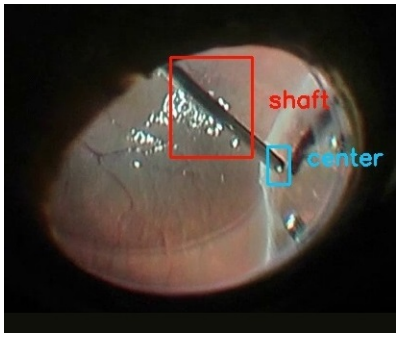
(c) 3rd frame



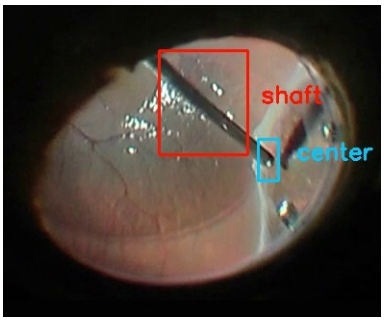
(d) 4th frame

Fig. 4: The results of the proposed hybrid approach on a video. Part I

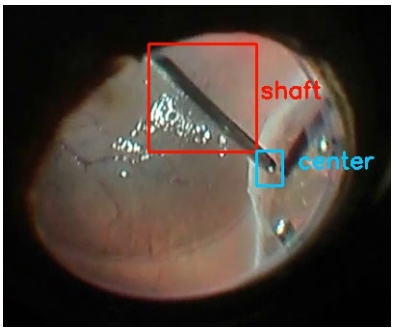
Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on December 21,2020 at 13:09:22 UTC from IEEE Xplore. Restrictions apply.



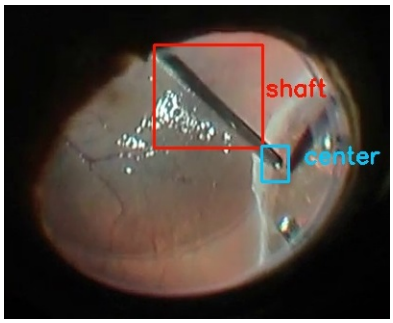
(a) 5th frame



(b) 6th frame



(c) 7th frame



(d) 8th frame

Fig. 5: The results of the proposed hybrid approach on a video. Part II

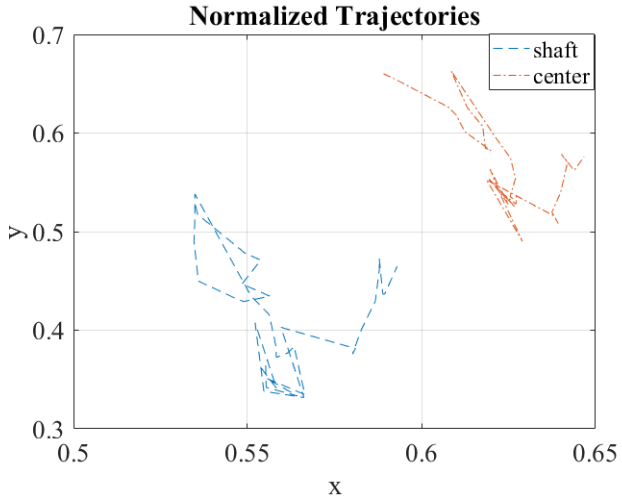


Fig. 6: Shaft and center trajectories

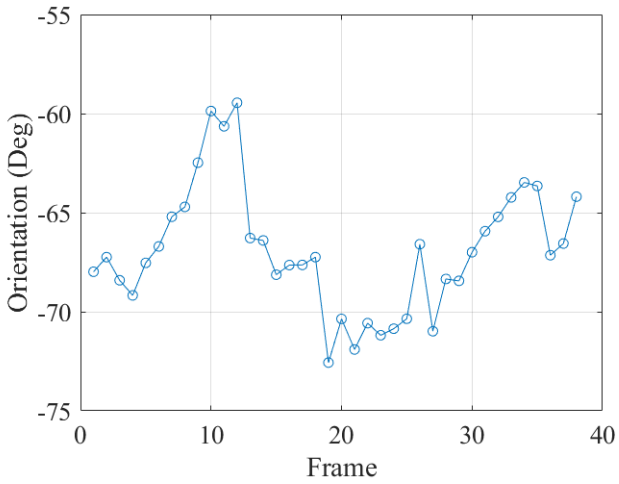


Fig. 7: Surgical instrument orientation

- The skill assessment of novice surgeons in a real-time manner based on the instantaneous task performance is an important topic of research. Our aim is to utilize the performance metrics proposed in the previous researches such as error-based performance index [22] and motion smoothness [7] in order to score the novice surgeon during the operations. In this way, the expert surgeon is notified of he probable mistakes made by the novice surgeon and is able to interfere to the procedure to avoid unfavorable complications for the patient. Notably, real-time skill assessment should be generally based on the time history of the trainee's behavior; thus, it is difficult task to accomplish.
- The majority of previously proposed haptic systems for surgery training require the expert surgeon to perform every detail of surgical operation, while the novice surgeon receives the guidance signals from the expert surgeon [4]–[6]. However, the expert surgeons usually prefer to not be

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on December 21,2020 at 13:09:22 UTC from IEEE Xplore. Restrictions apply.

involved in every detail of operation specially the ones that are easy to accomplish by the novice surgeons. On the other hand, the possibility of incidence of undesired complication highly increases when the novice surgeons perform the operation without any supervision. To resolve the problem, the existing control architectures [4]–[6] may be extended by automatizing some levels of error detection and even guidance. To that effect, a dataset including the trajectory records of the expert surgeons is utilized and the machine learning methods may be applied to those records. The proposed tracking method may be later utilized to enrich the above dataset using the available movies of surgical operations performed by the expert surgeons.

- We have developed a haptic system for eye surgery training in which the position of surgical tool is obtained by the encoders of the haptic devices [20]. An important work is to fuse the results obtained from image-based tracking with the encoder signals. Generally, combining of the sensory data derived from several sources modifies the deficiencies of each sensor to obtain the information with less uncertainty [23].

VI. Conclusions

In this article, a method is proposed which precisely yields the position and orientation of a surgical instrument. In this regard, a newly produced dataset is introduced along with a hybrid approach. The method is based on utilizing the YOLOv3 CNN beside the Medianflow a traditional OpenCV tracker. To train the CNN on the dataset, transfer learning methods are employed. Experimental results are conducted on a video to show the applicability of the suggested solution in facing with the surgical instrument tracking problem. Finally, a discussion on importance of developing such a visual tracker for surgical instruments is given.

References

[1] A. Üneri, M. Balicki, J. Handa, P. Gehlbach, R. H. Taylor, I. Iordachita, *et al.*, "New steady-hand eye robot with micro-force sensing for vitreo-retinal surgery," in *Biomedical Robotics and Biomechatronics (BioRob), 2010 3rd IEEE RAS and EMBS International Conference on*, pp. 814–819, 2010.

[2] H. Meenink, *Vitreo-retinal eye surgery robot: sustainable precision*. PhD thesis, Technische Universiteit Eindhoven, 2011.

[3] A. Bataleblu, M. Motaharifar, E. Abedlu, and H. D. Taghirad, "Robust h? control of a 2rt parallel robot for eye surgery," in *2016 4th International Conference on Robotics and Mechatronics (ICROM)*, pp. 136–141, IEEE, 2016.

[4] S. S. Nudehi, R. Mukherjee, and M. Ghodoussi, "A shared-control approach to haptic interface design for minimally invasive telesurgical training," *IEEE Transactions on Control Systems Technology*, vol. 13, no. 4, pp. 588–592, 2005.

[5] B. Khademian and K. Hashtrudi-Zaad, "Dual-user teleoperation systems: New multilateral shared control architecture and kinesthetic performance measures," *IEEE/ASME Transactions on Mechatronics*, vol. 17, no. 5, pp. 895–906, 2012.

[6] M. Motaharifar, H. D. Taghirad, K. Hashtrudi-Zaad, and S.-F. Mohammadi, "Control synthesis and iss stability analysis of dual-user haptic training system based on s-shaped function," *IEEE/ASME Transactions on Mechatronics*, 2019.

[7] S. Cotin, N. Stylopoulos, M. Ottensmeyer, P. Neumann, D. Rattner, and S. Dawson, "Metrics for laparoscopic skills trainers: the weakest link!," in *International conference on medical image computing and computer-assisted intervention*, pp. 35–43, Springer, 2002.

[8] N. Rieke, D. J. Tan, M. Alsheakhali, F. Tombari, C. A. di San Filippo, V. Belagiannis, A. Eslami, and N. Navab, "Surgical tool tracking and pose estimation in retinal microsurgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 266–273, Springer, 2015.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 1137–1149, 2017.

[10] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[11] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *IEEE CVPR*, vol. 4, 2017.

[12] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[13] Z. Kalal, K. Mikolajczyk, J. Matas, *et al.*, "Tracking-learning-detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, p. 1409, 2012.

[14] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," in *Pattern recognition (ICPR), 2010 20th international conference on*, pp. 2756–2759, IEEE, 2010.

[15] F. Lotfi, V. Ajallooeian, and H. Taghirad, "Robust object tracking based on recurrent neural networks," in *2018 6th RSI International Conference on Robotics and Mechatronics (IcRoM)*, pp. 507–511, IEEE, 2018.

[16] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[17] D. Wesierski, G. Wojdyga, and A. Jezierska, "Instrument tracking with rigid part mixtures model," in *Computer-Assisted and Robotic Endoscopy*, pp. 22–34, Springer, 2015.

[18] M. Alsheakhali, A. Eslami, and N. Navab, "Detection of articulated instruments in retinal microsurgery," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pp. 107–110, IEEE, 2016.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[20] "Arash-asist." https://aras.kntu.ac.ir/arash-asist/. Accessed: 2019-08-10.

[21] A. Molaei, E. Abedloo, H. D. Taghirad, and Z. Marvi, "Kinematic and workspace analysis of diamond: An innovative eye surgery robot," in *2015 23rd Iranian Conference on Electrical Engineering*, pp. 882–887, IEEE, 2015.

[22] B. Khademian and K. Hashtrudi-Zaad, "Experimental performance evaluation of a haptic training simulation system," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1247–1252, IEEE, 2009.

[23] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Information Fusion*, vol. 35, pp. 68–80, 2017.

Authorized licensed use limited to: UNIVERSITY OF WESTERN ONTARIO. Downloaded on December 21,2020 at 13:09:22 UTC from IEEE Xplore. Restrictions apply.