Crowdsourcing annotation of surgical instruments in videos of cataract surgery

Tae Soo Kim[1], Anand Malpani[2], Austin Reiter[1], Gregory D. Hager[1,2], Shameema Sikder[3] *, and S. Swaroop Vedula[2]

[1] Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA[1]
[2] The Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA[2]
[3] Wilmer Eye Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA[3]

Abstract. Automating objective assessment of surgical technical skill is necessary to support training and professional certification at scale, even in settings with limited access to an expert surgeon. Likewise, automated surgical activity recognition can improve operating room workflow efficiency, teaching and self-review, and aid clinical decision support systems. However, current supervised learning methods to do so, rely on large training datasets. Crowdsourcing has become a standard in curating such large training datasets in a scalable manner. The use of crowdsourcing in surgical data annotation and its effectiveness has been studied only in a few settings. In this study, we evaluated reliability and validity of crowdsourced annotations for information on surgical instruments (name of instruments and pixel location of key points on instruments). For 200 images sampled from videos of two cataract surgery procedures, we collected 9 independent annotations per image. We observed an inter-rater agreement of 0.63 (Fleiss' kappa), and an accuracy of 0.88 for identification of instruments compared against an expert annotation. We obtained a mean pixel error of 5.77 pixels for annotation of instrument tip key points. Our study shows that crowdsourcing is a reliable and accurate alternative to expert annotations to identify instruments and instrument tip key points in videos of cataract surgery.

1 Introduction

Automated comprehension of surgical activities is a necessary step to develop intelligent applications that can improve patient care and provider training [1]. Videos of the surgical field are a rich source of data on several aspects of care that affect patient outcomes [2]. For example, technical skill, which is statistically significantly associated with surgical outcomes [3], may be readily assessed by observing videos of the surgical field. Specifically, movement patterns of surgical instruments encode various types of information such as technical skill [4],

*Correspondence: ssikder1@jhmi.edu. Funding: Wilmer Eye Institute Pooled Professor's Fund and grant to Wilmer Eye Institute from Research to Prevent Blindness.

activity [5], surgical workflow and deviation from canonical structure [4], and amount or dose of intervention. Thus, algorithms to detect surgical instruments in video images, identify the instruments, and to detect or segment instruments are necessary for automated comprehension of surgical activities.

Although algorithms to segment instruments in surgical videos have been previously developed [6], it is by no means a solved problem. Specifically, prior work included segmentation of instruments in surgical videos [7–9, 6], and tracking instruments over time in videos [7, 8]. However, currently available algorithms to identify instruments and segment them in part or whole are not sufficiently accurate to annotate videos of different surgery procedures at scale.

Crowdsourcing has become a popular methodology to rapidly obtain various forms of annotations on surgical videos, including technical skill [10]. Prior work shows that crowdsourcing can yield high quality segmentation of instruments in surgical video images [11]. However, it is unclear how accurately a surgically untrained crowd can identify instruments in surgical video images. Therefore, our objective was to establish the reliability and accuracy of crowdsourced annotations to identify and localize key points in surgical instruments.

2 Methods

Our study was approved by the Institutional Review Board at Johns Hopkins University. In this study, we recruited crowd workers (CWs) through Amazon Mechanical Turk [12]. We evaluated reliability and accuracy of annotations on the identity and outline of the surgical instruments used in cataract surgery. We reproduced our study using an identical design 10 months after the initial survey. We refer to the original survey as Study 1 and the repeat survey as Study 2.

2.1 Surgical Video Dataset

We used images from videos of cataract surgery. Based upon input from an expert surgeon, we defined the following ten tasks in cataract surgery: paracentesis/side incision, main incision, capsulorhexis, hydrodissection, phacoemulsification, removal of cortical material, lens insertion, removal of any ophthalmic viscosurgical devices (OVDs), corneal hydration, and suturing incision (if indicated). We randomly sampled 10 images for each phase from videos of two procedures. Before sampling images, we processed the videos through optical flow based filtering to remove images with high global motion blur. All sampled images had a resolution of 640 by 480 pixels. The images did not contain any identifiers about the surgeon or the patient.

We selected six instruments that CWs had to identify and mark key points for, *viz.* keratome blade (KB), cystotome, Utratas (forceps), irrigation/aspiration (I/A) cannula, anterior chamber (A/C) cannula, and phacoemulsification probe (Phaco) as shown in Figure 1. For each image, we instructed CWs to identify the visible instrument(s) and to mark the corresponding predefined key points.

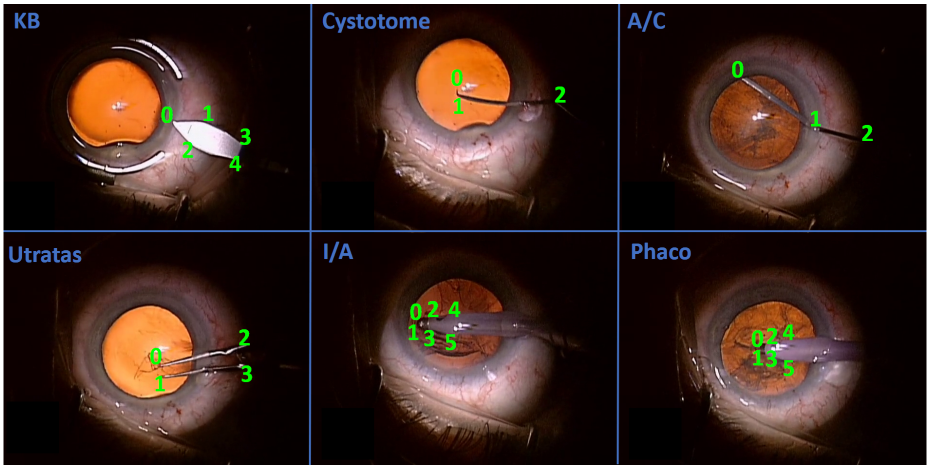


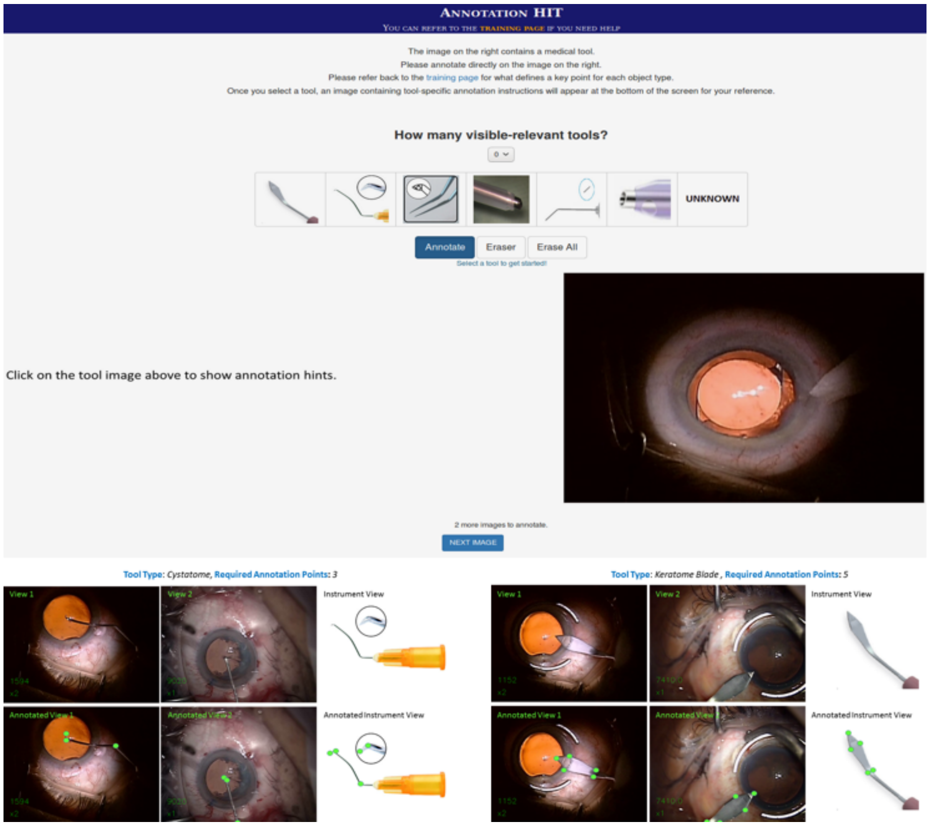
Fig. 1. Six instruments selected for study along with pre-defined key points. Annotators were trained to mark these key points. The qualification HIT contains the six images above with an image without an instrument of interest. KB = keratome blade; A/C = anterior chamber cannula; I/A = irrigation/aspiration cannula; Phaco = phacoemulsification probe. Best viewed in color.

2.2 Annotation Framework

The Amazon Mechanical Turk framework defines a Human Intelligence Task (HIT) as a self-contained task that a crowd worker (CW) can complete. Our study contained two HITs: a qualification HIT to vet potential CWs, and a main HIT with the actual data collection task.

Training Potential participants in our survey provided consent to be part of the study on the main landing page. We then directed them to a page with detailed instruction about each instrument, including textual description and two images, one with a surgical background and another was a stock catalog image. The images for each instrument included predefined key points that CWs had to annotate (see bottom row in Figure 2).

Qualification HIT We directed CWs who completed training to a qualification HIT for quality assurance of their annotations. In the qualification HIT, CWs were required to annotate seven images in one sitting. To qualify, we required the CW to correctly identify the instrument in each image. In addition, we required the CW to annotate the key points with an error of less than 15 pixels (ground truth was pre-specified). We allowed each CW a total of two attempts to successfully complete the qualification HIT. CWs who didn't qualify could no longer participate in the study. CWs who successfully completed the qualification HIT were eligible to further participate in the study.



ANNOTATION HIT

YOU CAN REFER TO THE TRAINING PAGE IF YOU NEED HELP

The image on the right contains a medical tool.
Please annotate directly on the image on the right.
Please refer back to the training page for what defines a key point for each object type.
Once you select a tool, an image containing tool-specific annotation instructions will appear at the bottom of the screen for your reference.

How many visible-relevant tools?

0

Select a tool to get started!

Click on the tool image above to show annotation hints.

2 more images to annotate.

Tool Type: *Cystatome*, Required Annotation Points: *3*

Tool Type: *Keratome Blade* , Required Annotation Points: *5*

Fig. 2. Top: A screenshot from a live annotation task we hosted on Amazon Mechanical Turk. The crowd worker (CW) initially identifies and selects the instrument(s) visible in the image from the survey. An instructional image, corresponding to the selected instrument is then shown to guide the CW on key points. The CW directly clicks on the survey image to mark key points. Bottom: examples of instructional images for a *cystotome* and a *keratome blade*. Best viewed in color.

Main HIT The main HIT shared the same user interface as the qualification HIT (Figure 2). The interface contained three blocks — target image, instrument choices as visual buttons, and an instructional image based on the instrument chosen by the CW to demonstrate key points to annotate. There were seven visual buttons: one for each of the six instruments of interest and one for annotating presence of any unlisted instrument. We instructed CWs to not select any of the instrument buttons if the image contained no instrument. We did not enforce the order in which CWs annotated key points on instruments seen in the image. We allowed the CWs to annotate a maximum of two instruments per image because we considered it unlikely that more than two instruments may be visible in any image from a video of cataract surgery.

Each main HIT in our study contained nine assignments to ensure that as many independent CWs annotated each video image. Each assignment included 30 video images that a CW annotated in one sitting. To control quality of annotations, i.e., to avoid uninformative annotations, we included a reference video image at a randomly chosen location in the image sequence within each assignment. We revoked qualification for CWs who failed to accurately annotate the reference video image, and terminated their participation in the study. We paid CWs $1 for each successfully completed assignment. In addition, we paid a bonus of $0.10 for the initial assignment a CW successfully completed. We did not enforce an upper bound on how long CWs took to complete a given assignment.

2.3 Data Analysis

To determine reliability in identification of instruments, we computed the macro-average of the percent of CWs comprising the majority annotation (percent CW agreement or macro-averaged percent agreement [MPA]). We also computed the Fleiss' kappa, κ_F, [13] as a measure of inter-annotator agreement accounting for agreement expected by chance.

To compute the MPA, we construct a response matrix, $R \in \mathbb{R}^{N \times K}$, where N is the number of samples and K is the number of categories. An element R_{ij} of the response matrix corresponds to the count of the i-th sample being annotated as the j-th instrument class. Then, the MPA is defined as:

$$f_i = \frac{max(R_i)}{\sum_{j=1}^{K} R_{ij}} \tag{1}$$

$$MPA = \frac{\sum_{i=1}^{N} f_i}{N} \tag{2}$$

where $i \in [1, N]$ and $j \in [1, K]$.

We compute Fleiss' kappa as follows:

$$p_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} R_{ij}(R_{ij} - 1) \tag{3}$$

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} R_{ij} \tag{4}$$

$$\overline{P} = \frac{1}{N} \sum_{i=1}^{N} p_i, \quad \overline{P}_e = \sum_{j=1}^{K} p_j^2 \tag{5}$$

$$\kappa_F = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \tag{6}$$

where n is the number of annotations per sample ($n = \{3, 5, 7, 9\}$ in this study). The instrument label for a given sample image is the mode of n annotations.

In addition, we computed accuracy in identification of instruments against ground truth specified in consultation with an expert surgeon. For agreement and accuracy statistics, we first analyzed the original annotations from nine CWs. We then analyzed annotations with $n = 3$, 5, and 7 CWs. For this secondary analysis, we randomly sampled with replacement annotations of the desired n from those captured for each image. We computed measures of reliability and accuracy averaged across 100 iterations for each n. Finally, we evaluated annotations of key points, i.e., instrument localization, using the mean pixel error across annotators averaged across all images.

3 Results

In Study 1, we recruited 19 CWs, of whom 11 CWs qualified to participate in the survey. We captured a total of 1919 annotations within 48 hours from the qualified CWs. In Study 2, 26 CWs qualified to participate in the survey from whom we captured 1916 annotations within 24 hours.

Table 1. Reliability and validity of crowdsourced annotations to identify surgical instruments. Study 2 is an independent replication of Study 1. (MPA = Macro-averaged percent agreement)

	Study 1			Study 2		
n	**MPA**	**Fleiss' κ**	**Accuracy**	**MPA**	**Fleiss' κ**	**Accuracy**
3	0.85 ± 0.01	0.63 ± 0.02	0.86 ± 0.04	0.85 ± 0.01	0.65 ± 0.02	0.77 ± 0.01
5	0.83 ± 0.01	0.63 ± 0.01	0.87 ± 0.03	0.82 ± 0.01	0.66 ± 0.01	0.78 ± 0.01
7	0.83 ± 0.05	0.63 ± 0.09	0.88 ± 0.02	0.81 ± 0.01	0.65 ± 0.01	0.80 ± 0.01
9	0.82	0.63	0.89	0.81	0.65	0.80

Table 1 shows our findings on reliability and accuracy for identifying surgical instruments in the video images. Neither reliability nor accuracy appeared to be affected by the number of CWs rating each image. These findings are consistent across independent samples of CWs.

Table 2 shows the MPA for the individual instruments. The MPA was both high and consistent between the two studies for the keratome blade and Utratas

Table 2. Percent agreement of CWs per instrument. Study 2 shows lower inter-rater agreement given a larger number of CWs.

n	KB	Cystotome	Utratas	I/A	A/C	Phaco
Study 1						
3	1.00	0.82 ± 0.05	0.97 ± 0.03	0.81 ± 0.03	0.94 ± 0.02	0.85 ± 0.01
5	1.00	0.80 ± 0.03	0.97 ± 0.02	0.77 ± 0.02	0.93 ± 0.02	0.84 ± 0.01
7	1.00	0.79 ± 0.02	0.97 ± 0.01	0.76 ± 0.01	0.93 ± 0.01	0.84 ± 0.01
9	1.00	0.79	0.97	0.75	0.94	0.83
Study 2						
3	1.00	0.62 ± 0.02	1.00	0.85 ± 0.02	0.86 ± 0.02	0.77 ± 0.02
5	1.00	0.64 ± 0.02	0.99 ± 0.01	0.81 ± 0.02	0.87 ± 0.02	0.71 ± 0.02
7	1.00	0.60 ± 0.01	0.98 ± 0.01	0.78 ± 0.01	0.87 ± 0.02	0.68 ± 0.02
9	1.00	0.55	0.98	0.77	0.86	0.72

but not for other instruments. Table 3 shows accuracy of CWs to identify the individual instruments. The accuracy was distinctly low for the cystotome, particularly in Study 2 where we also observe low reliability in annotations. We were unable to replicate (in Study 2) the accuracy we observed in Study 1 for the cystotome, I/A cannula, A/C cannula, and phacoemulsification probe. While increasing the number of CWs appeared to improve accuracy in identifying the phacoemulsification probe, it seemed to reduce accuracy in identifying the cystotome.

Table 3. Instrument identification accuracy of CWs (instrument label is the mode of *n* CW annotations.

n	KB	Cystotome	Utratas	I/A	A/C	Phaco
Study 1						
3	1.00	0.44 ± 0.08	1.00	0.83 ± 0.05	0.97 ± 0.02	0.93 ± 0.02
5	1.00	0.39 ± 0.05	1.00	0.88 ± 0.04	0.96 ± 0.01	0.97 ± 0.02
7	1.00	0.36 ± 0.05	1.00	0.90 ± 0.03	0.96 ± 0.00	0.99 ± 0.01
9	1.00	0.38	1.00	0.91	0.96	1.00
Study 2						
3	1.00	0.36 ± 0.10	1.00	0.82 ± 0.03	1.00	0.59 ± 0.07
5	1.00	0.27 ± 0.05	1.00	0.89 ± 0.02	1.00	0.71 ± 0.03
7	1.00	0.27 ± 0.04	1.00	0.91 ± 0.01	1.00	0.71 ± 0.01
9	1.00	0.27	1.00	0.93	1.00	0.71

3.1 Instrument Localization

Table 4 summarizes the mean pixel error for key points specific to each instrument. For all instruments we studied, the mean pixel error was lower for key points corresponding to the tips than for key points elsewhere on the instrument. The mean pixel errors for several key points in Study 2 were similar, but not identical, to those in Study 1.

Table 4. Mean pixel error for instrument key points annotated by crowd workers (CWs). "x" indicates the CW was not required to mark the key point for instrument.

Index	KB	Cystotome	Utratas	I/A	A/C	Phaco
Study 1						
0	**3.21** ± 1.97	**4.35** ± 2.90	**3.64** ± 2.67	**4.38** ± 4.49	**5.18** ± 11.83	**4.86** ± 3.59
1	10.28 ± 13.50	115.5 ± 106.3	**3.47** ± 2.34	**8.73** ± 9.49	18.49 ± 37.93	**7.63** ± 8.65
2	9.70 ± 14.29	38.51 ± 41.84	22.10 ± 23.84	9.15 ± 10.31	19.0 ± 22.16	8.24 ± 9.88
3	25.75 ± 31.26	x	21.89 ± 24.75	7.16 ± 9.79	x	4.96 ± 4.79
4	25.71 ± 30.6	x	x	12.3 ± 13.3	x	10.7 ± 14.3
5	x	x	x	13.4 ± 12.3	x	11.7 + ± 14.3
Study 2						
0	**2.62** ± 1.55	**4.15** ± 2.62	**3.532** ± 2.69	**2.98** ± 3.14	**4.63** ± 8.87	**5.32** ± 3.93
1	13.18 ± 14.55	132.1 ± 102.0	**3.40** ± 1.92	**7.00** ± 8.26	27.30 ± 42.11	**8.62** ± 6.85
2	11.89 ± 16.82	42.62 ± 39.99	25.20 ± 20.43	8.92 ± 8.88	20.13 ± 23.08	7.77 ± 3.57
3	25.32 ± 28.22	x	17.73 ± 25.67	6.89 ± 5.15	x	5.13 ± 7.81
4	22.10 ± 28.01	x	x	11.34 ± 11.03	x	13.42 ± 10.29
5	x	x	x	10.28 ± 12.08	x	12.88 ± 15.09

3.2 Discussion

Our findings show that CWs can reliably identify instruments in videos of cataract surgery, some more accurately than others. Not surprisingly, this accuracy was lower for instruments that closely resembled others. For example, there are few visual cues to distinguish an I/A cannula from a Phaco, and a cystotome from an A/C cannula. Thus, ambiguity in identification is likely a result of similarity between instruments and lack of clarity in images due to pose, occlusion, or illumination. Our inability to replicate accuracy in identifying instruments despite prior instruction and qualification of CWs suggests sampling variability.

The large pixel errors for key points other than those corresponding to tips of instruments appeared to result from erroneous identification of instrument or insufficient instruction provided to CWs. For example, confusion of cystotome for A/C cannula leads to an exceptionally large pixel error for key point 1 of cystotome. Figure 3 illustrates a common failure scenario in which incorrect identification of an instrument leads to erroneous key points.

A potential solution to improve the accuracy of instrument localization is to provide more context to CWs. For example, if an assignment contains a sequentially ordered set of images, then CWs may be able to better discriminate between ambiguous cases. However, regardless of the large pixel errors, the annotated key points provide a sparse outline of the instrument of interest. The structure of the instrument in the scene may be accurately estimated by fitting a

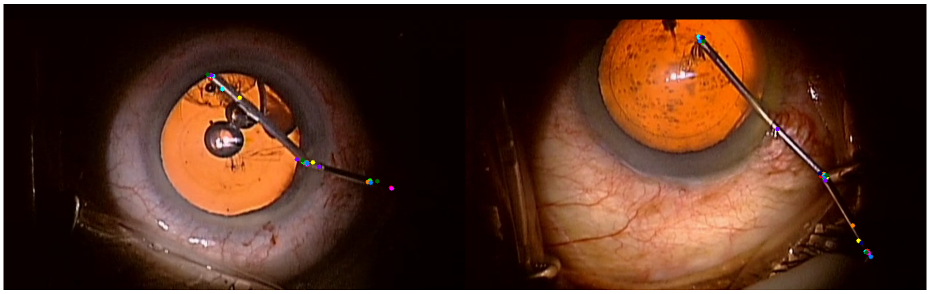


Fig. 3. A common failure mode arising from viewpoint ambiguity where CWs incorrectly annotate anterior chamber cannula (left) for cystotome (right).

prior geometric model to the set of key point annotations provided by CWs. Instead of requesting bounding box annotations from the crowd, key point based inquiry can potentially yield ground truth data for instrument segmentation. Future work should evaluate accuracy and utility of key point annotations for segmentation of instruments in surgical videos.

In this work, we evaluated annotations for instruments and not for other aspects of cataract surgery, e.g., phases. Our findings do not shed light on whether crowdsourcing can yield reliable annotations for other aspects of cataract surgery videos, but a couple of study design features are relevant for future work. First, requiring CWs to successfully complete a qualification HIT can assure quality at the outset. Second, our findings suggest the need for adequate instruction and quality control to obtain reliable annotations for features that may be subject to interpretation. The amount of instruction for qualifying CWs and quality control measures should be tailored depending on anticipated subjectivity in interpreting what CWs are asked to annotate in the images.

Finally, we studied annotation for type of instruments because it is informative about the phase of surgery, particularly in cataract surgery. We did not evaluate the effect of annotation accuracy on performance of algorithms for down-stream applications such as phase detection. Our findings suggest that studies developing algorithms for applications using CW annotations should also evaluate their sensitivity to annotation accuracy.

4 Summary

Crowdsourcing can rapidly yield reliable and accurate annotations on identity and location of the tip(s) for a selected set of instruments in video images of cataract surgery procedures.

References

1. Vedula, S., Ishii, M., Hager, G.: Objective assessment of surgical technical skill and competency in the operating room. Annu Rev Biomed Eng 21(19), 301–325 (2017)

2. Puri, S., Kiely, A., Wang, J., Woodfield, A., Ramanathan, S., Sikder, S.: Comparing resident cataract surgery outcomes under novice versus experienced attending supervision. Clinical Opthalmology (9), 1675–1681 (2015)
3. Birkmeyer, J.D., Finks, J.F., O'Reilly, A., Oerline, M., Carlin, A.M., Nunn, A.R., Dimick, J., Banerjee, M., Birkmeyer, N.J.: Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine 369(15), 1434–1442 (2013)
4. Forestier, G., Petitjean, F., Senin, P., Riffaud, L., Hénaux, P., Jannin, P.: Finding discriminative and interpretable patterns in sequences of surgical activities. Artificial Intelligence in Medicine 82, 11–19 (2017), https://doi.org/10.1016/j.artmed.2017.09.002
5. Gao, Y., Vedula, S.S., Reiley, C., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Bejar, B., Yuh, D.D., Chen, C., Vidal, R., Khudanpur, S., Hager, G.D.: The jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: In Modeling and Monitoring of Computer Assisted Interventions (M2CAI), MICCAI (2014)
6. Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, B., Ourselin, S., Pakhomov, D., Sznitman, R., Teichmann, M., Thoma, M., Vercauteren, T., Voros, S., Wagner, M., Wochner, P., Maier-Hein, L., Stoyanov, D., Speidel, S.: Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. ArXiv e-prints (May 2018)
7. Sznitman, R., Becker, C., Fua, P.: Fast part-based classification for instrument detection in minimally invasive surgery. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014. pp. 692–699. Springer International Publishing, Cham (2014)
8. Rieke, N., Tan, D.J., Alsheakhali, M., Tombari, F., di San Filippo, C.A., Belagiannis, V., Eslami, A., Navab, N.: Surgical tool tracking and pose estimation in retinal microsurgery. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 266–273. Springer International Publishing, Cham (2015)
9. Reiter, A., Allen, P.K., Zhao, T.: Appearance learning for 3d tracking of robotic surgical tools. The International Journal of Robotics Research 33(2), 342–356 (2014), https://doi.org/10.1177/0278364913507796
10. Malpani, A., Vedula, S.S., Chen, C.C.G., Hager, G.D.: A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. International Journal of Computer Assisted Radiology and Surgery 10, 1435–1447 (2015)
11. Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H.G., Eisenmann, M., Speidel, S.: Can masses of non-experts train highly accurate image classifiers? A crowdsourcing approach to instrument segmentation in laparoscopic images. Med Image Comput Comput Assist Interv 17(Pt 2), 438–445 (2014)
12. Little, G., Chilton, L.B., Goldman, M., Miller, R.C.: Turkit: Tools for iterative tasks on mechanical turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation. pp. 29–30. HCOMP '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1600150.1600159
13. Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33(3), 613–619 (1973), https://doi.org/10.1177/001316447303300309