Can Masses of Non-Experts Train Highly Accurate Image Classifiers?

A Crowdsourcing Approach to Instrument Segmentation in Laparoscopic Images

Lena Maier-Hein[1,⋆,⋆⋆], Sven Mersmann[1], Daniel Kondermann[2], Sebastian Bodenstedt[3], Alexandro Sanchez[2], Christian Stock[4], Hannes Gotz Kenngott[5], Mathias Eisenmann[3], and Stefanie Speidel[3]

[1] Computer-assisted Interventions, German Cancer Research Center (DKFZ), Germany
[2] Heidelberg Collaboratory for Image Processing, University of Heidelberg, Germany
[3] Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany
[4] Institute of Medical Biometry and Informatics, University of Heidelberg, Germany
[5] Department of General, Visceral and Transplantation Surgery, University of Heidelberg, Germany
l.maier-hein@dkfz-heidelberg.de

Abstract. Machine learning algorithms are gaining increasing interest in the context of computer-assisted interventions. One of the bottlenecks so far, however, has been the availability of training data, typically generated by medical experts with very limited resources. *Crowdsourcing* is a new trend that is based on outsourcing cognitive tasks to many anonymous untrained individuals from an online community. In this work, we investigate the potential of crowdsourcing for segmenting medical instruments in endoscopic image data. Our study suggests that (1) segmentations computed from annotations of multiple anonymous non-experts are comparable to those made by medical experts and (2) training data generated by the crowd is of the same quality as that annotated by medical experts. Given the speed of annotation, scalability and low costs, this implies that the scientific community might no longer need to rely on experts to generate reference or training data for certain applications. To trigger further research in endoscopic image processing, the data used in this study will be made publicly available.

1 Introduction

Computer-assisted minimally-invasive surgery (MIS) as well as computer-assisted surgical training is gaining increasing interest in the past years. One of the main

⋆ Correspondence author.

⋆⋆ This work was conducted within the setting of the *SFB TRR 125: Cognition-guided surgery* funded by the German Research Foundation (DFG). It was further sponsored by the European Social Fund of the State Baden-Wuerttemberg and the Klaus Tschira Foundation.

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 438–445, 2014.

challenges in this context is the image-based tracking of medical instruments in the endoscopic images, which is a prerequisite for surgical navigation [1], skill assessment [2] and workflow analysis [3], for example. State-of-the-art methods apply machine learning techniques to learn the shape and appearance of different objects from labeled training data [4]. However, the performance of the classifiers depends crucially on the availability of reference annotations, which are extremely expensive to obtain because they are typically made by medical experts with very limited resources. As a consequence, the data sets used to train or validate a new method are typically small and thus not able to capture the wide range of anatomical/scene variance.

Crowdsourcing is the process of outsourcing cognitive tasks to many anonymous untrained individuals from an online community. In contrast to *outsourcing*, the work comes from an undefined public rather than being commissioned from a specific, named group. Its advantages include speed of annotation, scalability and low cost. While the concept has already been applied to a variety of different applications, its usage in the context of medical image processing is extremely limited. According to a very recent review article, the few medical applications can be classified into four main areas [5]: *Problem solving*, *surveying*, *surveillance* and *data processing*. Tasks related to the last category include shape-based classification of polyps in computed tomography(CT) [6], skill assessment [7], and medical image classification [8].

The purpose of our work is to investigate whether crowdsourcing is an appropriate tool for training instrument tracking algorithms in the context of laparoscopic surgery. Using a set of endoscopic video images with reference instrument segmentations we (1) quantify segmentation performance of the anonymous crowd using the raw annotated data as well as segmentations obtained via majority voting and (2) determine the performance of a basic instrument segmentation algorithm on data sets labeled by experts, the crowd or both groups.

2 Methods

2.1 Data Annotation Software

Amazon Mechanical Turk (*MTurk*) [9] is an internet-based crowdsourcing platform that allows requesters to distribute small computer-based tasks, referred to as *human intelligence tasks* (HITs), to a large number of untrained workers, referred to as *knowledge workers* (KWs). The KWs can freely choose the HITs they want to perform and receive a small monetary reward for each completed one from the requester (typically a couple of cents for a task of a few minutes). Our annotation user interface was integrated into MTurk by supplying a dynamic webpage (HTML5, JavaScript). In this study, each HIT refers to the segmentation of one medical instrument in a given endoscopic image. Figure 1 shows a screenshot of the annotation process. Based on a bounding box and a very rough contour specifying which of potentially multiple instruments in the image to segment, the KW needs to place a polygon around the object under investigation. For each HIT, our software records (1) the user ID, (2) the coordinates of the points as well as (3) the time needed for the completion of one HIT.

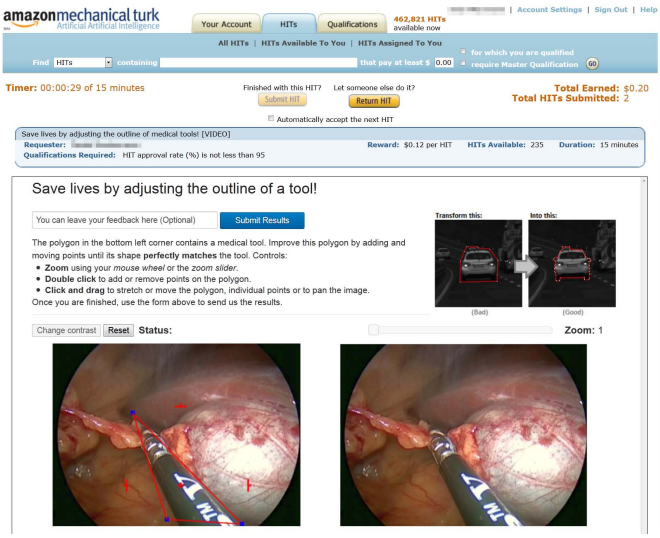

Fig. 1. Screenshot of the *Amazon Mechanical Turk* [9] image annotation software developed for this study. In the header, the micro task to be performed is described to the user. The user generates and moves the blue points from which a red contour is generated.

2.2 Endoscopic Video Data

The training data was generated from a total of 6 surgical procedures, three from laparoscopic adrenalectomies and three from laparoscopic pancreatic resections. From each surgery, 20 images containing one or several medical instruments were extracted, yielding 120 images in total.

Half of the data from each surgical procedure was annotated by a medical expert with experience in laparoscopic surgeries, as shown by means of example in Fig. 2. The $6 \cdot 10$ images each contained 2.1 instruments on average, such that 122 reference instrument segmentations (data set Y^I) and 60 background segmentations (data set Y^B) were obtained for these images.

All images (i.e. twice as many as those annotated by the experts) were further annotated by 10 KWs each, yielding 2350 instrument segmentations in total (data set X^I).

2.3 Quality of Crowd Segmentations

The quality of the crowd segmentation was determined via the dice similarity coefficient (DSC) as follows. Let Y^I represent the set of binary images corresponding to the reference segmentations of the instruments ($|Y^I| = 122$), and let $X^{I^{REF}} \subset X^I$ represent the set of crowd segmentations for which a reference annotation was available ($|X^{I^{REF}}| = |Y^I|$). The elements of Y^I and $X^{I^{REF}}$ are

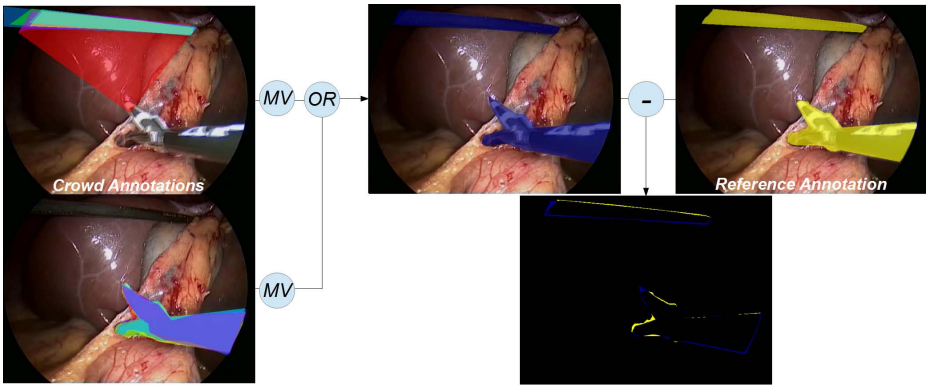

Fig. 2. Concept of endoscopic video annotation using crowdsourcing: For each instrument in a given image, 10 annotations are acquired. Majority voting (MV) is applied to remove outliers. Multiple instruments are then combined using the OR operator to yield the final result (blue) which is very similar to the reference segmentation (yellow). In this example a Dice Similarity Coefficient (DSC) of 0.95 was achieved

again sets (of pixels) representing a particular instrument. Let $X_{ik} \in X^{I^{REF}}$ (i: instrument ID; k: KW id) further correspond to the same instrument as Y_i^I. Then the DSC

$$DSC_{ik} = \frac{2|X_{ik}^{I^{REF}} \cap Y_i^I|}{|X_{ik}^{I^{REF}}| + |Y_i^I|} \tag{1}$$

quantifies the similarity of crowd segmentation k for instrument i with the corresponding reference segmentation.

To investigate whether multiple segmentations for one object can be applied to improve the segmentations of the crowd, the crowd segmentations for one particular instrument $X_i^{I^{REF}}$ was obtained by majority voting, i.e. a pixel was classified as instrument, if and only if at least 5 KWs had marked it as instrument. The resulting DSC with the reference annotations was determined as follows:

$$DSC_i = \frac{2|X_i^{I^{REF}} \cap Y_i^I|}{|X_i^{I^{REF}}| + |Y_i^I|} \tag{2}$$

The corresponding background annotations were defined as the complement of all (merged) instrument segmentations. To quantify segmentation performance by the crowd, a boxplot of DSC_{ik} and DSC_i was generated.

2.4 Classifier Performance–Experts vs Crowd

For this experiment, we developed a basic instrument classification algorithm based on random forests [10] that classifies each pixel into instrument or back-

ground. Based on a preliminary evaluation on different data sets, we used the following features: the B channel from RGB color space, the S channel from HSV color space and o1 from opponent color space. The forests were trained with the bagging approach [10] and consisted of five trees, each with a maximum depth of five. Based on this algorithm and the data described in section 2.2 we trained three classifier types:

- $C^{EXP}(n)$: The classifier was trained exclusively on expert data ($n = 10, 20, \ldots, 50$ images)
- $C^{KW}(n)$: The classifier was trained on images resulting from merging 10 crowd annotations for each instrument via majority voting (cf. sec. 2.4) ($n = 10, 20, \ldots, 100$ images)
- $C^{EXP-KW}(n)$: A combination of 1. and 2.: The classifier was trained on images from the experts and the crowd. For this purpose, half of the data was taken from the experts, half was taken from the crowd. In this process, each image was included at most once for each training (i.e., it was annotated *either* by an expert *or* by a KW) ($n = 10, 20, \ldots, 100$ images)

For a given n and a given classifier type (1.-3.), each of the six video sequences described in sec. 2.2 was used for testing in a leave-1-out approach. To ensure comparability of results, the n training images and 10 testing images used were identical for all methods for a given n. For each training set, 10 random forests were trained leading to a total of $6 \cdot 10 \cdot 10 = 600$ automatically annotated images for testing.

To investigate the influence of different variables (group: crowd, expert, mixed; surgery: OP1,...,OP6; number of training samples: n) on classification performance, we applied multiple linear regression modelling. Models of the following form were fitted for the three outcomes true positive (TP) rate, true negative (TN) rate and precision:

$$Y_i = \mu + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \beta_1 X_{i1} + \ldots + \beta_4 X_{i4} + \gamma_1 X_{i1} + \ldots + \gamma_5 X_{i5} + e_i,$$

where Y_i represents the ith observation of the respective outcome and μ denotes the common mean. Further, α_1 and α_2 are regression coefficients corresponding to the dummy variables X_1 and X_2, where $X_1 = 1$ represents a crowd segmentation and $X_2 = 1$ represents a mixed segmentation ($X_1 \cdot X_2 = 0$). Thus, they quantify the difference in the outcome to expert segmentations that serve as the reference. In an analogous manner, β_1 to β_4 are regression coefficients pertaining to the number of training images (20, 30, 40, 50) with 10 training images being the reference, and γ_1 to γ_5 similarly adjust for differences between surgeries. The e_i denote normally distributed random errors with mean zero. We considered p values <0.05 of the regression coefficients to indicate statistical significance.

For $n > 50$, we only had data from the crowd and the mixed group. In this case, we used an analogous model with X_2 omitted.

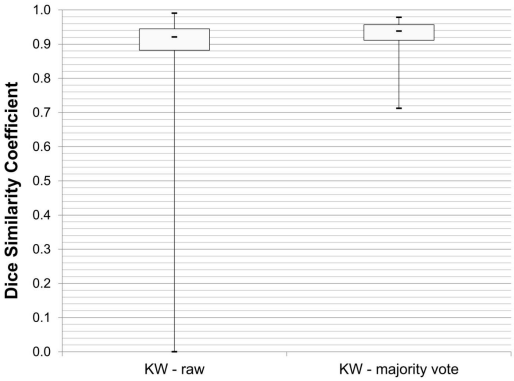

Fig. 3. Box plot (median, first and third quartiles, minimum and maximum) of the Dice Similarity Coefficient (DSC) for all crowd segmentations for which reference data was available (60 images; n = 1200) as well as for the annotations obtained with majority voting (n = 120).

3 Results

The mean time required for obtaining one segmentation for each tool in 20 images (i.e. for one surgery) was 39 ± 11 min averaged over 10 requests (i.e. uploads of HITs). Hence, all 2350 annotations were available in less than 24 hours. The mean DSC for the crowd was 0.89 ± 0.13 (n = 1200 [1]) averaged over all instruments in all images. This could be increased to 0.93 ± 0.05 (n = 120) using the concept of majority voting. Figure 3 shows boxplots of the DSC for the individual KWs as well as for the majority voting.

Table 1. Selection of model coefficients from multiple linear regression models of true positive (TP) rate, true negative (TN) rate and precision for the experiment with $n \leq 50$.

Variable	**TP rate**	**TN rate**	**Precision**
(Intercept)	0.519	0.970	0.769
Experts	(ref.)	(ref.)	(ref.)
Crowd	0.012	−0.002*	−0.007
Mixed	0.005	−0.001	−0.004

*$p < 0.05$

With respect to the type of segmentation (by crowd, experts or mixed) the only statistically significant impact on any of the three outcomes (TP rate, TN rate, precision) was observed for the TN rate in the case of the experiment

[1] $2 \cdot 10$ data sets had to be excluded due to a misplaced bounding box in two images.

with ≤ 50 training images ($p = 0.04$). However, in this case, the difference in performance was only -0.2%. According to the model applied, the surgical data set used for testing explained by far the most variation in segmentation performance and was highly statistically significantly associated with TP rate, TN rate and precision (all $p < 0.001$). A selection of regression coefficients for the experiment with $n \leq 50$ are shown in Tab. 1.

4 Discussion

To our knowledge, we are the first to apply the concept of crowdsourcing for training classification algorithms in the context of computer-assisted MIS. In this study, we showed that the segmentations of medical tools generated by anonymous untrained workers are comparable to those made by medical experts. Outliers can be removed with high reliability when using the concept of majority voting. The number of annotations required to reliably remove outliers, however, remains to be investigated. An important result of our study is that the performance of an endoscopic object classifier was not statistically different when trained with crowd data compared to expert data. Given the speed of annotation, scalability and low costs, this implies that the community might no longer need to rely on experts to generate reference or training data.

A limitation of our study could be seen in the fact that the segmentation tool only allowed for drawing a polygon, and hence, tools with holes could not be segmented with maximal specificity. As the experts used different software to annotate the images, a dice coefficient of 100% could thus not be achieved. Furthermore, to distinguish multiple instruments in a single endoscopic image from each other, we manually positioned a contour with bounding box in the images. Future work should investigate computing automatic bounding boxes such that manual pre-processing is not necessary at all.

It is worth mentioning that the classifier performance achieved may not appear to be very high. One explanation is that we used endoscopic images with very high variability (6 different surgeries) in order to capture a maximum of scene variance. However, the focus of this study was to investigate the quality of the crowd annotations rather than the performance of a specific algorithm.

Although crowdsourcing platforms are becoming increasingly popular, user performance varies greatly. To improve performance, researchers have explored developing games to motivate workers, using qualification tests to eliminate unqualified workers, incorporating verification tasks to confirm that workers are paying attention, applying the concept of priming and - like us in this study - duplicating effort across many workers. Although majority voting improved the DSC by only 5% in this study, the minimum could be improved from 0 to > 0.7. Based on a parallel study on correspondence establishment [11] and the results of this study, we believe that outlier removal is critical to fully exploit the potential of the crowd in the context of MIS.

We showed that non-experts are able to generate high quality image segmentations, which implies that the physicians' expert knowledge and experience is

not necessary for this particular task. Future studies should explicitly aim to identify further key applications but also limitations of crowdsourcing in the context of medical image computing and computer-assisted interventions.

References

1. Maier-Hein, L., Mountney, P., Bartoli, A., Elhawary, H., Elson, D., Groch, A., Kolb, A., Rodrigues, M., Sorger, J., Speidel, S., Stoyanov, D.: Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. Med. Image Anal. 17, 974–996 (2013)
2. Ahmidi, N., Gao, Y., Béjar, B., Vedula, S.S., Khudanpur, S., Vidal, R., Hager, G.D.: String motif-based description of tool motion for detecting skill and gestures in robotic surgery. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 26–33. Springer, Heidelberg (2013)
3. Katic, D., Wekerle, A.L., Görtler, J., Spengler, P., Bodenstedt, S., Röhl, S., Suwelack, S., Kenngott, H.G., Wagner, M., Müller-Stich, B.P., Dillmann, R., Speidel, S.: Context-aware augmented reality in laparoscopic surgery. Comp. Med. Imag. and Graph. 37(2), 174–182 (2013)
4. Allan, M., Ourselin, S., Thompson, S., Hawkes, D., Kelly, J., Stoyanov, D.: Toward detection and localization of instruments in minimally invasive surgery. IEEE T. Bio-med. Eng. 60(4), 1050–1058 (2013)
5. Ranard, B., Ha, Y., Meisel, Z., Asch, D., Hill, S., Becker, L., Seymour, A., Merchant, R.: Crowdsourcing - harnessing the masses to advance health and medicine, a systematic review. J. Gen. Intern. Med. 29(1), 187–203 (2014)
6. Nguyen, T.B., Wang, S., Anugu, V., Rose, N., McKenna, M., Petrick, N., Burns, J.E., Summers, R.M.: Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. Radiology 262(3), 824–833 (2012)
7. Chen, C., White, L., Kowalewski, T., Aggarwal, R., Lintott, C., Comstock, B., Kuksenok, K., Aragon, C., Holst, D., Lendvay, T.: Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. J. Surg. Res. 187, 65–71 (2014)
8. Foncubierta Rodríguez, A., Müller, H.: Ground truth generation in medical imaging: A crowdsourcing-based iterative approach. In: Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia, CrowdMM 2012, pp. 9–14. ACM, New York (2012)
9. Chen, J.J., Menezes, N.J., Bradley, A.D., North, T.: Opportunities for crowdsourcing research on amazon mechanical turk. Interfaces 5(3) (2011)
10. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
11. Maier-Hein, L., et al.: Crowdsourcing for reference correspondence generation in endoscopic images. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 345–352. Springer, Heidelberg (2014)