Crowdsourcing for Reference Correspondence Generation in Endoscopic Images

Lena Maier-Hein[1,⋆,⋆⋆], Sven Mersmann[1], Daniel Kondermann[2], Christian Stock[3], Hannes Gotz Kenngott[4], Alexandro Sanchez[2], Martin Wagner[4], Anas Preukschas[4], Anna-Laura Wekerle[4], Stefanie Helfert[4], Sebastian Bodenstedt[5], and Stefanie Speidel[5]

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

Abstract. Computer-assisted minimally-invasive surgery (MIS) is often based on algorithms that require establishing correspondences between endoscopic images. However, reference annotations frequently required to train or validate a method are extremely difficult to obtain because they are typically made by a medical expert with very limited resources, and publicly available data sets are still far too small to capture the wide range of anatomical/scene variance. *Crowdsourcing* is a new trend that is based on outsourcing cognitive tasks to many anonymous untrained individuals from an online community. To our knowledge, this paper is the first to investigate the concept of crowdsourcing in the context of endoscopic video image annotation for computer-assisted MIS. According to our study on publicly available *in vivo* data with manual reference annotations, anonymous non-experts obtain a median annotation error of 2 px (n = 10,000). By applying cluster analysis to multiple annotations per correspondence, this error can be reduced to about 1 px, which is comparable to that obtained by medical experts (n = 500). We conclude that crowdsourcing is a viable method for generating high quality reference correspondences in endoscopic video images.

1 Introduction

Computer-assisted minimally-invasive surgery (MIS) is often based on algorithms that require establishing image correspondences using at least two different views of the same scene[1]. Example applications are shape recovery, camera calibration, structure and camera-motion estimation or augmented reality (AR). Validation and training of correspondence-based algorithms depend crucially on

[⋆] Corresponding author.

[⋆⋆] This work was funded by the German Research Foundation (DFG) (Collaborative Research Center 125: Cognition-guided surgery).

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 349–356, 2014.

the availability of reference annotations, which, however, are extremely difficult and time/cost-intensive to obtain because they are typically made by medical experts with very limited resources. To address this issue, first efforts have been undertaken to make manually annotated endoscopic data publicly available [2]. However, the available data sets are still far too small to capture the wide range of anatomical/scene variance. *Crowdsourcing* is a new trend that is based on outsourcing cognitive tasks to many anonymous untrained individuals from an online community. Advantages of crowdsourcing include speed of annotation, scalability and low cost. While the concept has already been applied for a variety of different applications, its usage in the context of medical image processing is extremely limited. According to a recent review article, the few medical applications can be classified into four main areas [3]: *Problem solving* (e.g. manipulation of the three-dimensional structures of proteins in order to find the most likely tertiary structure), *surveying* (e.g. to have access to a more diverse population than present in the typical university research subject pool), *surveillance* (e.g. questionnaire on disease symptoms in order to assess that disease's prevalence in a certain country) and *data processing.* Tasks related to the last category include disease detection based on cell analysis [4], shape-based classification of polyps in computed tomography data [5] and medical image classification [6]. To our knowledge, this study is the first to apply the concept of crowdsourcing in the context of endoscopic video image annotation for computer-assisted MIS. The aim is to investigate whether crowdsourcing can be used for generating high quality image correspondences given two different views on the surgical field. Based on a recently published publicly available data set of 100 endoscopic image pairs, each containing up to 27 correspondences [2], the following research questions (RQs) shall be addressed:

RQ1. How is the accuracy of image correspondences established by an anonymous *crowd*?
RQ2. How can multiple annotations be applied to assure highly reliable reference correspondences?
RQ3. How does the accuracy compare to that obtained from medical experts?

2 Methods

2.1 Data Annotation Software

Amazon Mechanical Turk (MTurk) [7] 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 one endoscopic image pair. Given a set of $N_C = 10$ points in one endoscopic image, the task is to find the set of corresponding points in the second endoscopic image. For each HIT, our software recorded (1) the user ID, (2) the coordinates of the points as well as (3) the time needed for the completion of one HIT.

To allow medical experts participating in our study to perform a controlled set of HITs using the same software, we made use of the MTurk *sandbox*, which is typically used for testing the user interfaces supplied by the requester and does not involve payment of the user.

2.2 RQ1: Quality of Crowd Annotation

The first experiment was designed to address RQ1, i.e., to assess the accuracy of correspondences established by the crowd. For this purpose, the data annotation tool, introduced in sec. 2.1, was used to generate 100 different HITs using the first $N_C = 10$ correspondences of each image pair in the publicly available data set [2] of $N_I^{KW} = 100$ images. For each of the $N_I^{KW} \cdot N_C = 1,000$ different correspondences, $N_U^{KW} = 10$ annotations were requested (i.e., HITs repeated by ten users), leading to $n = 10,000$ annotations in total. For each annotation p_{ijk}^{KW} ($i = 1, \ldots, N_I$: image ID; $j = 1, \ldots, N_C$: feature ID; $k = 1, \ldots, N_U$: annotation counter) we then determined the Euclidean distance d_{ijk}^{KW} in pixels (px) to the publicly available reference correspondence p_{ij}^{REF}.

For d_{ijk}^{KW}, descriptive statistics, including mean, median, interquartile range (IQR), minimum, and maximum were determined. As the distances have a lower bound of zero and their distribution can be expected to be strongly right-skewed (rendering the standard deviation an invalid measure), we applied log transformation to d_{ijk} to obtain an approximately normal distribution of the observed distances for further analysis. In order to assess the variability among images, correspondences and users, we applied a random-effects (variance-components) model for the log-transformed distances which had the following form:

$$log\left(d_{ijk}^{KW}\right) = \beta_0 + b_i + b_{ij} + b_k + \epsilon_{ijk} \tag{1}$$

where β_0 denotes the overall population mean, and the remaining variables are random effects denoting the deviation from the overall mean for the ith image (b_i), the deviation from the image-specific mean for the jth correspondence within the ith image (b_{ij}), the deviation from the overall mean for the kth user (b_k) and random within-group errors ϵ_{ijk}. Random effects and errors were assumed to be normally distributed with zero mean, i.e. $b_i \sim \mathcal{N}(0, \sigma_1^2)$, $b_{ij} \sim \mathcal{N}(0, \sigma_2^2)$, $b_k \sim \mathcal{N}(0, \sigma_3^2)$, and $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$. Hence, σ_1^2, σ_2^2, and σ_3^2 represent the variance in annotation accuracy introduced by images, correspondences and users, respectively. The models were estimated by maximum likelihood methods in R version 3.0.2 (2013-09-25) [8] using the 'nlme'-package [9]. We tested for possible model simplification by likelihood ratio tests. The estimate of β_0 and its 95% confidence interval (CI) were back-transformed on the original scale to obtain an estimate of annotation accuracy which is adjusted for all statistical dependencies in the data. For these statistical measurements we only used the first annotation in case a user annotated the same correspondence more than once.

2.3 RQ2: Improving Quality by Redundancy

To investigate RQ2, i.e., whether multiple annotations can lead to higher quality references, we applied a clustering algorithm based on Gaussian finite mixture models fitted by an expectation-maximization (EM) algorithm [10,11]. Given the observed bivariate data points p_{ijk}^{KW} the likelihood of a mixture model with G components is

$$\mathcal{L}(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_G \,|\, p_{ijk}^{KW}) = \prod_{l=1}^{n} \sum_{m=1}^{G} \tau_m f_m(y_l | \theta_k),$$

where f_m is the bivariate normal density of the mth component in the mixture with parameters θ_m (mean μ_m and covariance matrix Σ_m), and τ_m is the probability that an observation p_{ijk} belongs to the mth component. Model-based clustering was applied separately for each correspondence with different cluster sizes (1 to 10) and different parameterizations of the covariance matrix Σ_m (see [10] for details). The best correspondence-specific model was then chosen based on the Bayesian information criterion (BIC) and the mean μ_m of the largest cluster was chosen as the cluster-based crowd annotation $\check{p}_{ij}^{KW}$. In case of multiple equally sized 'largest' clusters, the mean (in case of 2 clusters) or median (in case of $\geq$3 clusters) of the cluster means was chosen as the crowd annotation. The distance to the reference annotation was determined for $\check{p}_{ij}^{KW}$ and also for the median value of the crowd annotations $\bar{p}_{ij}^{KW}$, and respective descriptive statistics were calculated. We tested for a difference between the cluster-based and the median crowd annotation by a simple random effects model of the form $\bar{p}_{ij}^{KW} - \check{p}_{ij}^{KW} = \beta_0 + b_i + \epsilon_{ij}$, where β_0 reflects the difference, b_i is a random image effect and ϵ_{ij} are within-group errors. This model accounts for the multiple observations per image (which prohibit the use of standard tests such as a t-test). The same assumptions, estimation method, and software as described in the previous section were used.

2.4 RQ3: Comparison to Medical Experts

To address RQ3, the comparison to experts, we repeated Experiment 1 with a group of $N_U^{EXP} = 5$ medical doctors (with experience in laparoscopic procedures) and a reduced number of $N_I^{EXP} = 10$ images. For a fair comparison, we ordered the images from Experiment 1 according to the median annotation performance and then picked 10 images including the first and the last one (i.e., about every 11th image). We then applied the methods described in the previous paragraphs as follows: Accuracy was compared using a model of the logarithm of the distance to the reference annotation similar as in sec. 2.2

$$log\left(d_{ijk}\right) = \beta_0 + \beta_1 x_1 + b_i + b_{ij} + b_k + \epsilon_{ijk}, \tag{2}$$

where x_1 is an additional grouping variable taking the value 0 if the annotation belongs to an expert and 1 if the annotation belongs to a crowd user, and β_1 is the corresponding regression coefficient. Further, the variance of the errors ϵ_{ijk} was allowed to differ between crowd and expert annotations to take into account and test for possible heteroscedasticity. We also determined crowd annotations optimized by cluster analysis as described in sec. 2.3 and compared them to the annotations made by the five experts using descriptive statistics.

3 Results

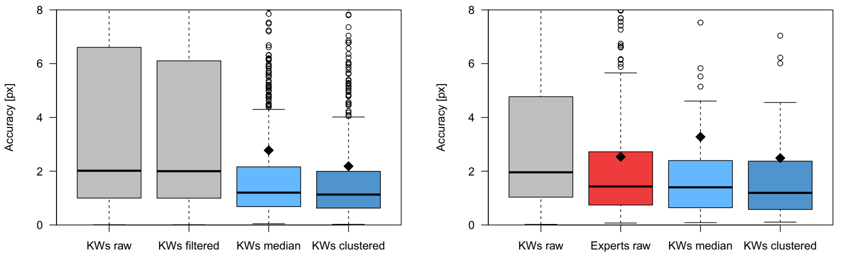
The mean time required for obtaining 100 HITs (one per image) from MTurk was 77 ± 16 min, averaged over 10 requests (i.e., uploads of HITs). Hence, 10,000 annotations could be generated in less than 24 hours. Descriptive statistics for the distance to the reference annotation can be found in Tab. 1 for all three experiments.

Table 1. Distance to reference annotation for research questions RQ1-RQ3. KWs: Knowledge workers. IQR: Interquartile range.

Experiment	Method	Mean	Median	IQR	Min	Max
RQ1	KWs raw	23.3	2.0	(1.0, 6.1)	0.0	430.8
RQ2	KWs median	2.8	1.2	(0.7, 2.2)	0.1	111.0
RQ2	KWs clustered	2.2	1.1	(0.6, 2.0)	0.0	125.0
RQ3	KWs raw	13.7	2.0	(1.0, 4.8)	0.0	291.4
RQ3	KWs median	3.3	1.4	(0.7, 2.4)	0.1	111.0
RQ3	KWs clustered	2.5	1.2	(0.6, 2.4)	0.1	51.4
RQ3	Experts	2.5	1.4	(0.7, 2.7)	0.1	94.1

RQ1: All 1,000 correspondences had a least seven annotations from different users. Removal of repeated annotations by the same user yielded 9,050 annotations. The mean KW annotation time for one HIT comprising 10 correspondences was 225±176 s (min: 28 s, max: 894 s). Figure 1(a) shows the distance of the KW annotation to the reference annotation for all individual annotations (a), all annotations with duplicates removed (for the case that one correspondence was annotated by the same user twice) (b), for the median annotation obtained from all KWs (c), as well as for the annotation obtained from applying the clustering procedure described in sec. 2.4 (d). A typical annotation result is shown in Figure 2. The accuracy of the crowd, represented by the expected difference to the reference annotation, was estimated to be 2.9 px (95%-CI: 2.3, 3.7), with standard error $SE(\beta_1)$=0.12 on the logarithmic scale. While the most variance in the accuracy was introduced by the individual crowd users ($\sigma_3 = 1.25$), a lower amount of variance was due to differences among images ($\sigma_1 = 0.42$) and correspondences ($\sigma_2 = 0.35$). All random effects statistically significantly improved the model fit in likelihood ratio tests with p values <0.0001.

RQ2: By using multiple annotations per correspondence, the median error could be reduced by 88% (median of ten annotations) and 91% (cluster analysis) compared to the individual KWs (cf. Tab. 1). The expected additional



(a) crowd only (1,000 features) (b) crowd and experts (100 features)

Fig. 1. Box plot of the annotation accuracy as defined in sec. 2.2 for all 1,000 correspondences (a) and for the subset of 100 correspondences annotated by five medical experts (b). KWs-raw: all annotations by the crowd (n = 10,000); KW - repetitions removed: all annotations with repetitions removed (n = 9,050); KWs - median: annotations obtained by taking the median of all 10 annotations for a particular feature point as described in sec. 2.3 (n = 1,000); KW - clustered: correspondences obtained by the clustering procedure as described in sec. 2.3 (n = 1,000). The diamonds represent mean values.

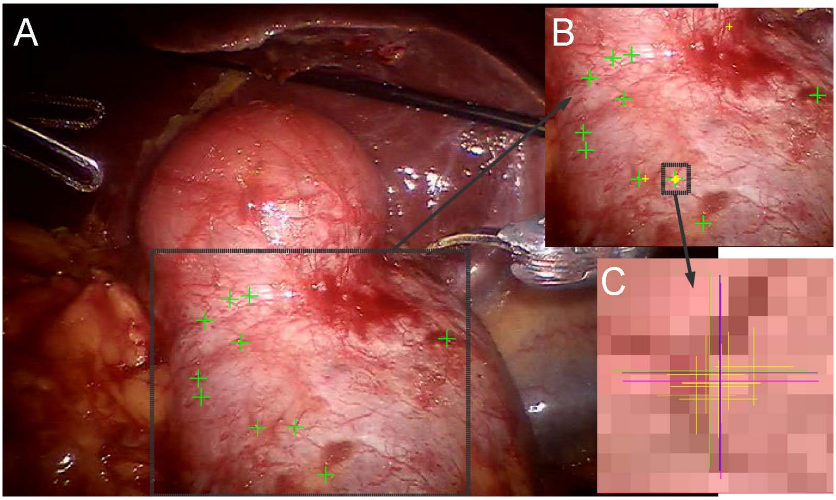


Fig. 2. Typical annotation results showing ten reference correspondences (green crosses) in the full (A) and zoomed-in (B/C) images, as well as the annotations by the knowledge workers (KWs) (yellow crosses; n=10) for one particular correspondence at two different scales (B/C). The median KW annotation is displayed as blue and the clustered KW annotation as pink cross.

improvement in annotation accuracy associated with the cluster-based method compared to the median annotation was estimated to be 0.6 px (95%-CI: 0.2, 0.9) and was statistically significantly different from zero with p <0.02.

RQ3: The medical experts achieved an annotation time for one HIT of 243±157 s (min: 102 s, max: 756 s), which is slightly higher compared to the crowd. Figure 1(b) compares the performance of the crowd to the performance of the medical experts for a selected set of 100 correspondences. The expected accuracy was by -0.7 px (95%-CI: -1.7, 0.2) smaller for the experts compared to the crowd with p=0.13. Annotations made by experts exhibited on average less variability. The standard deviation of the random errors was by the factor 1.4 times larger for crowd annotations than for expert annotations with p <0.001. Using crowd annotations optimized by cluster analysis, the mean distance to the reference annotation was smaller for the crowd annotation than for 4 out of 5 experts by 0.02 to 1.31 px, and larger only with respect to one expert by 0.25 px.

4 Discussion

To our knowledge, this is the first study to evaluate the concept of crowdsourcing in the context of endoscopic image annotation. Our statistical analysis took into account all dependencies in the data by hierarchical linear modeling, and enables inferences about different sources of variability in the data. According to the results, KWs can establish image correspondences of high quality. This, however, requires the (automatic) removal of outliers, typically present in the data. We showed that a Gaussian finite mixture models based clustering method applied to multiple feature annotations by the crowd can generate correspondences whose quality is comparable to that achieved by medical experts (median: 1 px). In practice, this means that time-consuming tedious annotations tasks do not necessarily have to be performed by an expert in the future.

It is worth noting that the reference used for this study was also generated by a human observer and thus prone to error. Based on the annotation precision achieved by the five medical experts that participated in this study we can estimate the error of this reference to be in the order of magnitude of more than 1 px. Considering this, the accuracy of about 1 px obtained by the KWs is excellent.

The user task was to find a correspondence for a *given* feature point. In practice, this concept could be applied for finding the corresponding point for a feature point extracted automatically by some algorithm, for example. An alternative approach would be to manually establish (arbitrary) correspondences between a pair of images and then design a *verification* task, in which KWs can rate correspondences.

We showed that non-experts are able to establish image correspondences that are comparable to those generated by medical experts. Apparently, the physicians' expert knowledge and experience is not necessary for this particular task. The method could thus be used to expand the publicly available data set [2] to increase anatomical/scene variance. Future studies should explicitly aim to identify further key applications but also limitations of crowdsourcing. According to a related study, performed by the authors [12], one application with high potential is the segmentation of medical instruments from endoscopic images for training instrument tracking algorithms.

In conclusion, the excellent results of this study illustrate the great potential of crowdsourcing in the context of medical image computing and computer-assisted interventions and should encourage further research in this area.

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. Puerto, G.A., Mariottini, G.-L.: A comparative study of correspondence-search algorithms in MIS images. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part II. LNCS, vol. 7511, pp. 625–633. Springer, Heidelberg (2012)
3. 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, 187–203 (2014)
4. Mavandadi, S., Dimitrov, S., Feng, S., Yu, F., Sikora, U., Yaglidere, O., Padmanabhan, S., Nielsen, K., Ozcan, A.: Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study. PLoS ONE 7, e37245 (2012)
5. 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, 824–833 (2012)
6. 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)
7. Chen, J.J., Menezes, N.J., Bradley, A.D., North, T.: Opportunities for crowdsourcing research on amazon mechanical turk. Interfaces 5 (2011)
8. Team, R.C.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013)
9. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D.: R Core Team: nlme: Linear and Nonlinear Mixed Effects Models (2013); R package version 3.1-113
10. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis and density estimation. J. Am. Stat. Assoc. 97, 611–631 (2002)
11. Fraley, C., Raftery, A.E., Murphy, T.B., Scrucca, L.: mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical report, Technical Report No. 597, Department of Statistics, University of Washington (2012)
12. Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H., Eisenmann, M., Speidel, S.: Can masses of non-experts train highly accurate image classifiers? A crowdsourcing approach to instrument segmentation in laparoscopic images. In Barillot, C., Golland, P., Hornegger, J., Howe, R., eds.: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Volume 17., Springer, LNCS (2014) (in press)