Crowd-Algorithm Collaboration for Large-Scale Endoscopic Image Annotation with Confidence

L. Maier-Hein[1(✉)], T. Ross[1], J. Gröhl[1], B. Glocker[2], S. Bodenstedt[3], C. Stock[4], E. Heim[1], M. Götz[5], S. Wirkert[1], H. Kenngott[6], S. Speidel[3], and K. Maier-Hein[5]

[1] Computer-assisted Interventions Group, German Cancer Research Center (DKFZ), Heidelberg, Germany
l.maier-hein@dkfz-heidelberg.de
[2] Biomedical Image Analysis Group, Imperial College London, London, UK
[3] Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[4] Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany
[5] Medical Image Computing Group, DKFZ, Heidelberg, Germany
[6] Department of General, Visceral and Transplant Surgery, University of Heidelberg, Heidelberg, Germany

Abstract. With the recent breakthrough success of machine learning based solutions for automatic image annotation, the availability of reference image annotations for algorithm training is one of the major bottlenecks in medical image segmentation and many other fields. Crowdsourcing has evolved as a valuable option for annotating large amounts of data while sparing the resources of experts, yet, segmentation of objects from scratch is relatively time-consuming and typically requires an initialization of the contour. The purpose of this paper is to investigate whether the concept of crowd-algorithm collaboration can be used to simultaneously (1) speed up crowd annotation and (2) improve algorithm performance based on the feedback of the crowd. Our contribution in this context is two-fold: Using benchmarking data from the MICCAI 2015 endoscopic vision challenge we show that atlas forests extended by a novel superpixel-based confidence measure are well-suited for medical instrument segmentation in laparoscopic video data. We further demonstrate that the new algorithm and the crowd can mutually benefit from each other in a collaborative annotation process. Our method can be adapted to various applications and thus holds high potential to be used for large-scale low-cost data annotation.

L. Maier-Hein and T. Ross—Contributed equally to this paper.

K. Maier-Hein—Many thanks to Carolin Feldmann for designing Fig. 1 and Pallas Ludens for providing the annotation platform. This work was conducted within the setting of the *SFB TRR 125: Cognition-guided surgery (A02, I04, A01)* funded by the German Research Foundation (DFG). It was further sponsored by the Klaus Tschira Foundation.

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 616–623, 2016.
DOI: 10.1007/978-3-319-46723-8_71

1 Introduction

With the paradigm shift from open surgical procedures towards minimally invasive procedures, endoscopic image processing for surgical navigation, context-aware assistance, skill assessment and various other applications has been gaining increasing interest over the past years. The recent international endoscopic vision challenge, organized at MICCAI 2015, revealed that state-of-the-art methods in endoscopic image processing are almost exclusively based on machine learning based techniques. However, the limited availability of training data with reference annotations capturing the wide range of anatomical/scene variance is evolving as a major bottleneck in the field because of the limited resources of medical experts.

Recently, the concept of *crowdsourcing* has been introduced as a valuable alternative for large-scale annotation of endoscopic images [6]. It has been shown that anonymous untrained individuals from an online community are able to generate training data of expert quality. Similar achievements were made in other biomedical imaging fields such as histopathological image analysis [1]. A remaining problem, however, is that object segmentation from scratch is relatively time-consuming and thus expensive compared to other tasks outsourced to the crowd. On the other hand, state-of-the-art annotation algorithms are already mature enough to annotate at least parts of the input image with high confidence.

To address this issue, we propose a collaborative approach to large-scale endoscopic image annotation. The concept is based on a novel atlas-forest-based segmentation algorithm that uses atlas-individual uncertainty maps to weigh training images according to their relevance for each superpixel (Spx) of a new image. As the new algorithm can estimate its own uncertainty with high accuracy, crowd feedback only needs to be acquired for regions with low confidence. Using international benchmarking data, we show that the new approach requires only a minimum of crowd input to enlarge the training data base.

2 Methods

The following sections introduce our new approach to confidence-guided instrument segmentation (Sect. 2.1), our concept for collaborative image annotation (Sect. 2.2) as well as our validation experiments (Sect. 2.3).

2.1 Confidence-Weighted Atlas Forests

The instrument segmentation methods presented in the scope of the *MICCAI Endoscopic Vision Challenge 2015* were typically based on random forests (cf. e.g. [2,3]). One issue with commonly used random forests [4] is that the same classifier is applied to all new images although it is well known that endoscopic images vary highly according to a number of parameters (e.g. hardware applied, medical application) and hence, the relevance of a training image can be expected to vary crucially with the test image. Furthermore, non-approximative

addition of new training data requires complete retraining of standard forests. A potentially practical way to give more weight to the most relevant training images without having to retrain the classifier is the application of so-called *atlas forests* [9]. Atlas forests are based on multiple random forests (*atlases*), each trained on a single image (or a subset of training images). A previously unseen image can then be annotated by combining the results of the individual forests [8,9]. To our knowledge, atlas forests have never been investigated in the context of medical instrument segmentation in particular and endoscopic image processing in general. The hypothesis of our work with respect to automatic instrument segmentation is:

Hypothesis I: *Superpixel-specific atlas weighting using local uncertainty estimation improves atlas forest based medical instrument segmentation in laparosopic video data.*

Hence, we assume that the optimal atlases vary not only from image to image but also from (super)pixel to super(pixel). Our approach is based on a set of

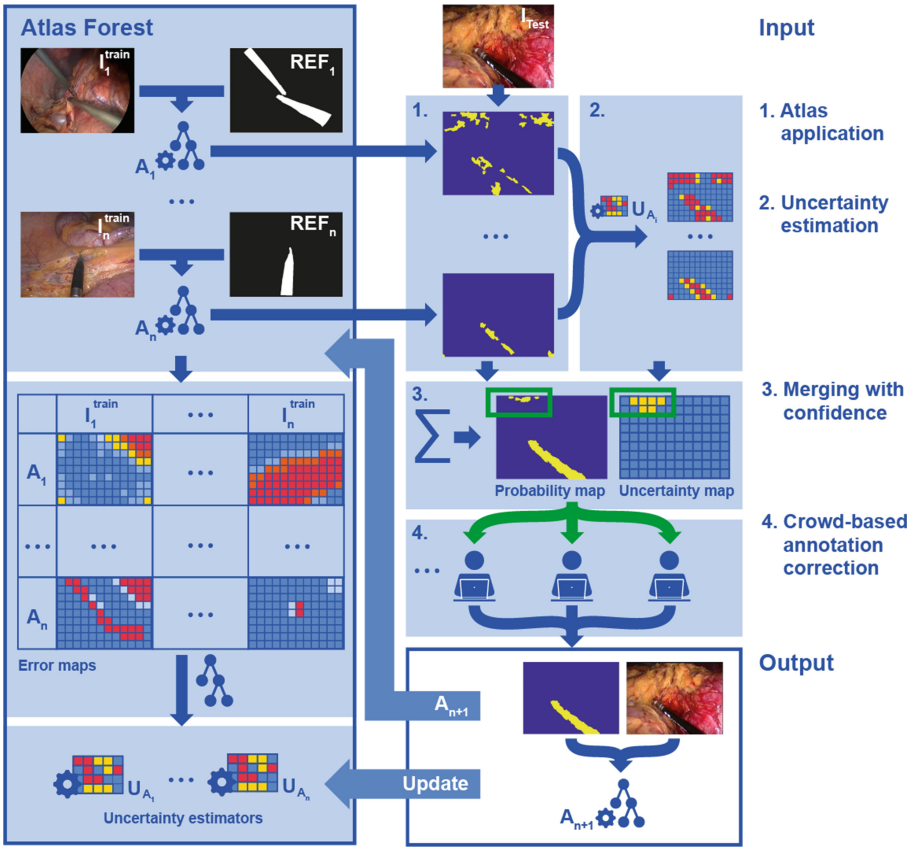


Fig. 1. Concept of collaborative large-scale data annotation as described in Sect. 2.2.

training images $\left\{I_1^{train}, \ldots I_{N_{img}}^{train}\right\}$ with corresponding reference segmentations $\left\{REF_1, \ldots REF_{N_{img}}\right\}$ and involves the following steps (cf. Fig. 1):

Atlas forest generation: As opposed to training a single classifier with the available training data, we train one random forest A_i per training image I_i^{train}. To be able to learn the relevance of a training image for a Spx (see *confidence learning*), we apply each atlas A_i to each image I_j^{train}. Based on the generated instrument probability map $P_{ij} = A_i(I_j^{train})$, we compute a Spx-based error map $E_{ij}(x) = |REF_j(x) - P_{ij}(x)|$ that represents the difference between the probability assigned to a Spx x and the true class label $REF_j(x)$, defined as the number of pixels corresponding to an instrument (according to the reference annotation) divided by the total number of pixels in the Spx.

Confidence learning: Based on E_{ij}, we train a regressor (uncertainty estimator) $U_{A_i}(x)$ for each atlas that estimates the error made when applying atlas A_i to Spx x, where the Spx is represented by exactly the same features as used by the atlas forests. The error estimator can be used to generate atlas-specific confidence maps $C_i(I)$ for an image, where the confidence in a Spx x is defined as $|1 - U_{A_i}(x)|$.

Confidence-weighted segmentation merging: To segment a new image I, all atlases A_j are applied to it, and the resulting probability maps P_j are merged considering the corresponding uncertainty maps $U_j(x)$ in each Spx.

In our first prototype implementation, we instantiate our proposed concept as follows. We base our method on the most recently proposed random-forest-based endoscopic instrument segmentation algorithm [3], which extends a method presented at the MICCAI 2015 challenge by classifying Spx rather than pixels and combines state-of-the-art rotation, illumination and scale invariant descriptors from different color spaces. In this paper, we apply this method to individual images rather than sets of images. For error estimation, a regression forest (number of trees: 50) is trained on a Spx basis with the same features as used by the atlas forests.

To classify a Spx x of a new image I, we initially select the most confident atlases:

$$S(x) = \{A_i | i \in \{1 \ldots N_{img}\}, U_{A_i}(x) < e_{max}\} \tag{1}$$

If $S(x)$ is empty, we add the atlases A_j with the lowest $U_{A_j}(x)$ to $S(x)$ until $|S(x)| == N_{min}^A$. For each Spx (type: SEEDS), the mean of the classification results of all confident atlases is used as probability value, and Otsus method is applied to the whole image to generate a segmentation.

2.2 Collaborative Image Annotation

In previous work on crowd-based instrument segmentation [6], the user had to define the instrument contours from scratch. A bounding box placed around the instruments was used to clarify which object to segment. With this approach,

the average time used for annotating one image was almost 2 min. The second hypothesis of this paper is:

Hypothesis II: *Crowd-algorithm collaboration reduces annotation time.*

Our collaborative annotation concept involves the following step (cf. Fig. 1).

Atlas forest initialization. The confidence-weighted atlas forest is initialized according to Sect. 2.1 using all the available training data and yields the initial segmentation algorithm AF^0.

Iterative collaborative annotation. A previously unseen image is segmented by the current atlas forest AF^t. The regions with low accumulated confidence are distributed to the crowd for verification. The crowd refines the segmentation, and the resulting crowd-generated reference annotation is used to generate a new atlas $A^{N_{img}+t}$. The corrections of the crowd may be used to retrain the uncertainty estimators, and the new atlas is added to the new atlas forest AF^{t+1} along with the corresponding uncertainty estimator $U_{A^{N_{img}+t}}$.

2.3 Experiments

The purpose of our validation was to confirm the two hypotheses corresponding to Sects. 2.1 and 2.2. Our experiments were performed on the data of the laparoscopic instrument segmentation challenge that had been part of the MICCAI 2015 endoscopic vision challenge. The data comprises 300 images extracted from six different laparoscopic surgeries (50 each).

Investigation of Hypothesis I. To investigate the benefits of using confidence-weighted atlas forests, we adapted the recently proposed Spx-based instrument classifier [3] already presented in Sect. 2.1. 200 images from four surgeries were used to train (1) an atlas forest AF with simple averaging of the individual probability maps which served as baseline and (2) an atlas forest with confidence weighting AF_w according to Sect. 2.2 ($e_{max} = 0.1; N^A_{min} = 0.1 \cdot N^{train}_{img}$). The remaining 100 images from two surgeries were used for testing. For each classifier and all test images, we determined descriptive statistics for the distance between the true label of a Spx and the corresponding computed probability. In addition, we converted the probability maps to segmentations using Otsu's method and computed precision, recall and accuracy.

Investigation of Hypothesis II. For our collaborative annotation concept, we designed two annotation tasks for the crowd, using Amazon Mechanical Turk (*MTurk*) as Internet-based crowdsourcing platform. In the *false positive (FP) task*, the crowd is presented with Spx classified as instrument that had a low accumulated confidence (here: mean of confidence averaged over all atlases that were used for the classification of the Spx) in our weighting-based method. An eraser can be used to delete regions that are not part of medical instruments. In the *false negative (FN) task*, the crowd is presented with Spx classified as background that had a low accumulated confidence in our weighting-based methods.

An eraser can be used to delete regions that are not part of the background. To investigate whether crowd-algorithm collaboration can increase annotation speed by the crowd, we initialized the atlas forest AF_w with the 200 training images according to Sect. 2.2. AF_w^0 was then applied to the testing images, and the resulting segmentations were corrected using the two refinement tasks (majority voting with 10 users). We compared the annotation time required for the collaborative approach with the annotation time needed when segmenting the instruments from scratch.

3 Results

The observed median (interquartile range (IQR)) and maximum of the difference between the true class label (i.e. the number of pixels corresponding to an instrument (according to the reference annotation) divided by the total number of pixels in the Spx) and the corresponding probability value on the test data was 0.07 (0.04, 0.09) and 0.20. Descriptive statistics for the accuracy of non-weighted atlas forests AF and weighted atlas forest AF_w after segmentation using Otsu's method are shown in Fig. 2. There is a trade-off between the percentage of Spx regarded as confident and the quality of classification. When varying the confidence threshold $e_{max} = 0.1$ by up to $\pm$ 75 %, the (median) accuracy decreases monotonously from 0.99 (78 % coverage) to 0.96 (94 % coverage) on the confident regions. This compares to a median accuracy of 0.87 for the baseline method (AF) and to 0.94 for AF_w applied to all Spx ($e_{max} \in \{0.025, 0.05, ..., 0.075\}$). On the confident regions, the (median) precision was 1.00 for all thresholds, compared to 0.38 for non-weighted AFs and 0.84–0.86 for weighted AFs applied to all Spx. These spectacular values come at the cost of a reduced recall (range: 0.49–0.55). Example classification results from the atlas forest and the weighted atlas forest along with the corresponding confidence map are visualized in Fig. 3.

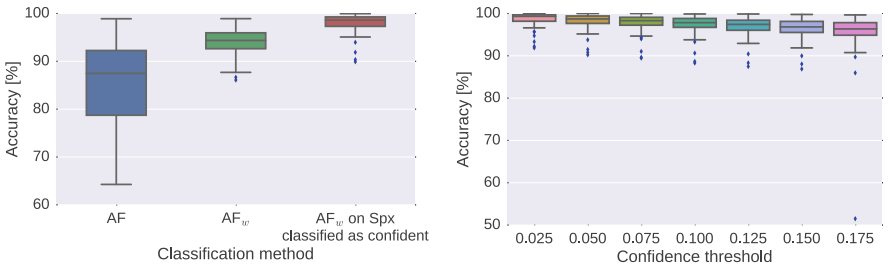
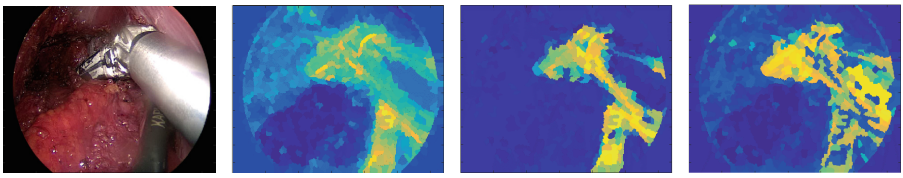


Fig. 2. Accuracy of the standard atlas forest AF and the weighted atlas forest AF_w when using all superpixels (Spx) of 100 test images as well as accuracy of AF_w when evaluated only on the confident Spxs of these images. (b) Accuracy of AF_w on just the confident Spxs for varying confidence threshold e_{max}. The whiskers of the box plot represent the 2.5 % and 97.5 % quantiles.



(a) Test image I (b) $AF(I)$ (c) $AF_w(I)$ (d) Confidence map

Fig. 3. Test image (a) with corresponding AF classification (blue: low probability) (b) AF_w classification (c) and confidence map of AF_w (blue: low confidence) (d). The specular highlight is not recognized as part of the instrument but the associated uncertainty is high.

The median (IQR) and maximum percentage of atlases that had a confidence above the chosen threshold ranged from 25 % (0.05 %, 50 %) and 87 % (e_{max}: 0.025) per Spx to 89 % (57 %, 96 %) and 100 % (e_{max}: 0.175). The corresponding (median) percentage of Spx classified incorrectly *and* not shown to the crowd ranged from 0.5 % to 3.4 %. With the collaborative annotation approach, the annotation time per image could be reduced from about two minutes to less than one minute (median: 51 s; IQR: (35 s, 70 s); max: 173 s).

4 Discussion

To our knowledge, we are the first to investigate the concept of crowd-algorithm collaboration in the field of large-scale medical image annotation. Our approach involves (1) automatic initialization of crowd annotations with a new confidence-weighted atlas-forest-based algorithm and (2) using the feedback of the crowd to iteratively enlarge the training data base. In analogy to recent work outside the field of medical image processing [5,7], we were able to show that collaborative annotation can speed up the annotation process considerably. Our experiments further demonstrate that the performance of an atlas on previously unseen images can be predicted with high accuracy. Hence, Spx-individual weighting of atlases improves classification performance of atlas forests compared to the non-weighted approach.

It is worth noting that we just presented a first prototype implementation of the collaborative annotation approach. For example, we took a simple threshold-based approach to convert the set of probability maps with corresponding confidence maps into a final segmentation. Furthermore, we did not systematically optimize the parameters of our method. This should be considered when comparing the results of our atlas forest with the results of other methods. According to our experience, the performance of random forests compared to atlas forests is highly dependent on the features used. In fact, when we initially trained all classifiers on point-based features (without local binary patterns), non-weighted atlas forests showed a similar performance to random forests. In the current version, random forests [3] perform similar to the weighted AF_ws when evaluated on all Spxs.

A disadvantage of our approach could be seen in the fact that we currently train one uncertainty estimator for each atlas. Note, however, that there is no need to perform the training on *all* images with reference annotations. Hence, the strong advantages of atlas forests are kept.

A major advantage of our method is the extremely high precision. Given the 100 % precision on the confident regions, we designed an additional fill-up-task, where the crowd was simply asked to complete the segmentation of the algorithm. This way annotation times were further reduced to about 45 s per image.

In conclusion, we have shown that large-scale endoscopic image annotation using crowd-algorithm collaboration is feasible. As our method can be adapted to various applications it could become a valuable tool in the context of big data analysis.

References

1. Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N.: Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans. Med. Imaging **35**, 1313–1321 (2016)
2. Allan, M., Chang, P.-L., Ourselin, S., Hawkes, D.J., Sridhar, A., Kelly, J., Stoyanov, D.: Image based surgical instrument pose estimation with multi-class labelling and optical flow. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 331–338. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24553-9_41
3. Bodenstedt, S., Goertler, J., Wagner, M., Kenngott, H., Mueller-Stich, B.P., Dillmann, R., Speidel, S.: Superpixel-based structure classification for laparoscopic surgery. In: SPIE Medical Imaging, p. 978618 (2016)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). http://dx.doi.org/10.1023/A%3A1010933404324
5. Cheng, J., Bernstein, M.S.: Flock: hybrid crowd-machine learning classifiers. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 600–611. ACM (2015)
6. Maier-Hein, L., et al.: Can masses of non-experts train highly accurate image classifiers? In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 438–445. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10470-6_55
7. Radu, A.-L., Ionescu, B., Menéndez, M., Stöttinger, J., Giunchiglia, F., Angeli, A.: A hybrid machine-crowd approach to photo retrieval result diversification. In: Gurrin, C., Hopfgartner, F., Hurst, W., Johansen, H., Lee, H., O'Connor, N. (eds.) MMM 2014. LNCS, vol. 8325, pp. 25–36. Springer, Heidelberg (2014). doi:10.1007/978-3-319-04114-8_3
8. Zikic, D., Glocker, B., Criminisi, A.: Classifier-based multi-atlas label propagation with test-specific atlas weighting for correspondence-free scenarios. In: Menze, B., Langs, G., Montillo, A., Kelm, M., Müller, H., Zhang, S., Cai, W.T., Metaxas, D. (eds.) MCV 2014. LNCS, vol. 8848, pp. 116–124. Springer, Heidelberg (2014). doi:10.1007/978-3-319-13972-2_11
9. Zikic, D., Glocker, B., Criminisi, A.: Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. Med. Image Anal. **18**(8), 1262–1273 (2014)